Evolution of human gene expression

Anna Maria Linnéa Warnefors

Doctor of Philosophy

University of Sussex

February 2011

Declaration

I hereby declare that this thesis has not been, and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature: ____________________________________

Preface

The research presented in this thesis has been submitted for publication as follows:

Chapter 2

Warnefors, M., V. Pereira and A. Eyre-Walker. Transposable elements: insertion pattern and impact on gene expression evolution in hominids. Mol Biol Evol **27**:1955-1962.

Author contributions: MW, VP and AEW designed the study. MW conducted all analyses and wrote the manuscript under supervision of AEW.

Chapter 3

Warnefors, M. and A. Eyre-Walker. A McDonald-Kreitman-type test for positive selection on gene expression levels. *Submitted.*

Author contributions: MW and AEW designed the study. MW conducted all analyses under supervision of AEW. MW and AEW wrote the manuscript.

Chapter 4

Warnefors, M. and A. Eyre-Walker. The accumulation of gene regulation through time. *In press.*

Author contributions: MW designed the study. MW conducted all analyses and wrote the manuscript under supervision of AEW.

Chapter 5

Warnefors, M., B. Hartmann, S. Thomsen, P. Patraquim and C.R. Alonso. Ultraconserved elements in the *Drosophila Hox* gene *Ultrabithorax. In preparation.*

Author contributions: MW and CRA designed the study. MW conducted all analyses under supervision of CRA. BH, ST and PP performed experiments that have not been included here. MW wrote the chapter under supervision of AEW.

Acknowledgements

This thesis would not have been possible without the support of my wonderful supervisor Adam Eyre-Walker. Thank you for making science so much fun!

I also wish to thank:

Claudio Alonso and Stefan Thomsen for teaching me everything I know about flies.

Past and present members of the Eyre-Walker group and in particular Nina Stoletzki and Gwenaël Piganeau for being the best office mates imaginable.

Miguel Ángel Céspedes and Kate Elliott for being there for me when I needed it the most.

Adina Feldman and Koen Boriau for some great holidays.

Jon Carruthers, John Chesebro, Sylvaine Fouix, Elli Leadbeater, Pierre Nouvellet, Eoin O’Connor, Casandra Villava, David Waxman and many other friends and colleagues who made my time at Sussex such an enjoyable experience.

Och slutligen – tack till mamma, pappa, Emma, Åsa, Thomas och min käraste farmor Valborg.

University of Sussex

Anna Maria Linnéa Warnefors

Doctor of Philosophy

Evolution of human gene expression

Summary

During evolution, biological differences between species can arise not only due to structural differences between genes, but also following changes in how, where and when genes are active. However, we know much less about this second aspect, because large-scale comparative transcriptomics only became feasible relatively recently. In this thesis, I will therefore investigate several aspects of gene expression evolution, with emphasis on our own species.

A first step to understanding regulatory evolution is to determine how variation in gene expression is created. Transposable elements (TEs) are genomic parasites that can affect their host genome in a number of ways, including gene expression. In Chapter 2, I investigate to what extent transposable elements (TEs) have contributed to expression differences between humans and chimpanzees.

Once expression variation has been established, a combination of selection and drift will decide which variants are passed on to future generations. It is of particular interest to identify changes that were established through positive selection, as these are adaptive. In Chapter 3, I describe a new method to detect positive selection acting on gene expression and apply it to data from humans and chimpanzees.

Human gene expression is regulated through several mechanisms associated with transcription and post-transcriptional processing. In Chapter 4, I consider the long-term evolution of the human genome and investigate whether genes have reached their maximum capacity in terms of regulatory complexity. Finally, in Chapter 5, I explore the relationship between gene regulation and sequence conservation by identifying and analysing extremely conserved elements in the genome of the fruit fly *Drosophila melanogaster.*

Table of contents

1.

General introduction

The human genome contains around 20000 protein-coding genes (Flicek et al. 2011), however this structural information is not enough. To build and maintain our bodies, the genome also must detail how, where and when each gene is to be used. It has long been hypothesised that many of the differences between humans and other animals are due to changes in these instructions (Britten and Davidson 1969; King and Wilson 1975; Wray 2007). In this thesis, I will explore the evolution of human gene expression on a genome-wide level. I will begin, in Section 1.1, by providing a brief overview of the many mechanisms that regulate how protein-coding genes are expressed. I will then move on, in Section 1.2, to present some commonly used methods for assessing gene regulation *en masse* and discuss the technical limitations that affect measurements of genome-wide gene expression patterns within and between species. In Section 1.3, I will review our present understanding of differences in gene expression between humans and other species, along with the evolutionary processes that may underlie these differences. Finally, in Section 1.4, I will outline how the work presented as part of this thesis will address four fundamental aspects of gene expression evolution that are not yet fully understood.

1.1. Mechanisms of human gene regulation

The expression of protein-coding genes involves two phases: transcription, in which the gene's DNA serves as a template to produce a messenger RNA (mRNA) molecule and translation, in which the mRNA is decoded and used to assemble a polypeptide (Figure 1.1). In this section, I will describe the steps of this process, with focus on the many regulatory layers that alter the final output both qualitatively and quantitatively.

1.1.1. Transcriptional regulation

Protein-coding genes are transcribed by RNA polymerase II (Pol II) and the first step of transcription is therefore to make sure that the polymerase is recruited to the gene, where it binds, together with several general transcription factors (GTFs), to a region in the immediate vicinity of the transcription start site, known as the promoter (Thomas and Chiang 2006). The details of this recruitment vary, as different promoters contain different combinations of sequence motifs recognized by GTFs (Baumann, Pontiller, and Ernst 2010). In addition, some of the recruited factors have specific functions and are preferentially found in certain tissues (Goodrich and Tjian 2010).

Pol II function is further controlled by transcription factors (TFs) bound to sequence elements that are known as enhancers or silencers depending on whether they activate or repress transcription (Noonan and McCallion 2010). These elements can be situated up to at least 1 Mb away from the promoter (Lettice et al. 2002) and it appears that the DNA forms a loop which brings the relevant sequences together (Tolhuis et al. 2002). In some cases, a third class of sequence elements, referred to as insulators, provide a barrier that prevents

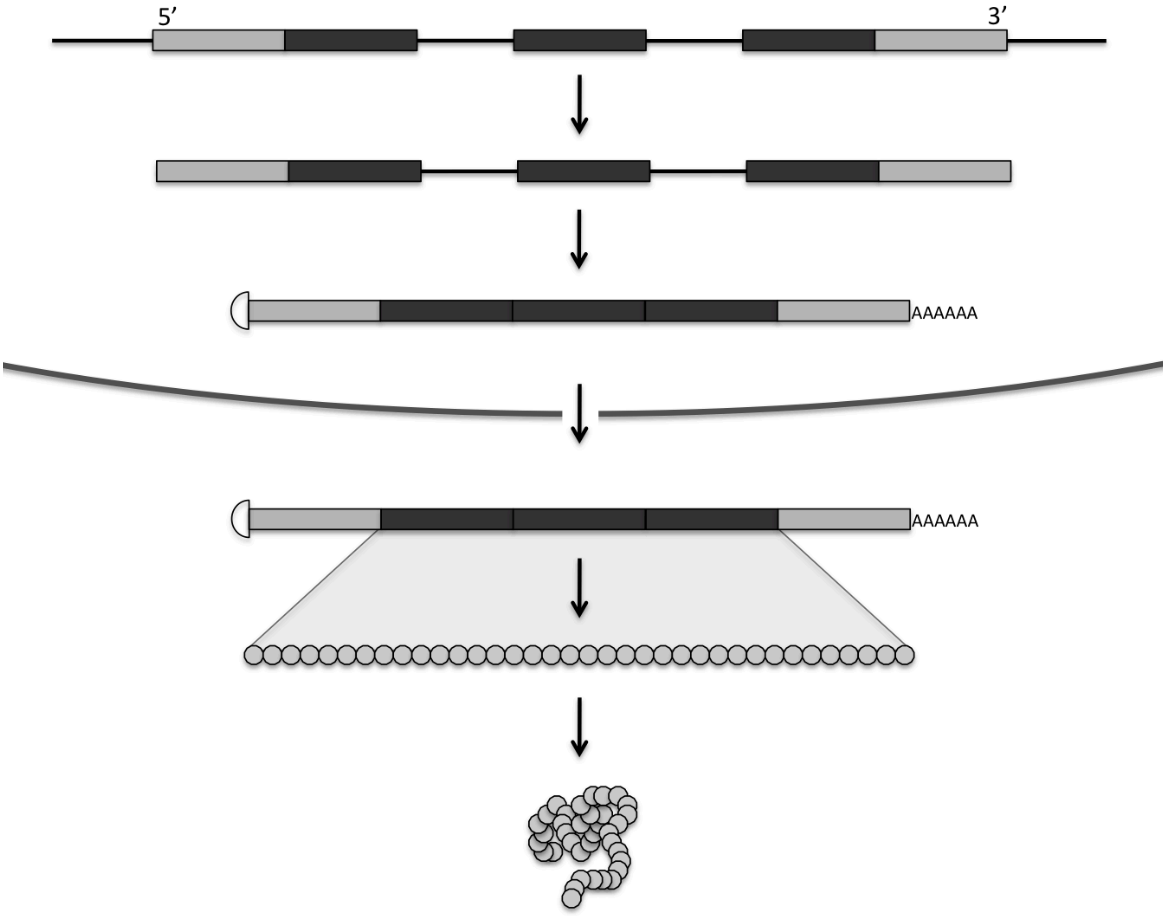


Figure 1.1. Expression of protein-coding genes. The DNA (top) is transcribed from the 5′ to the 3′ end. The resulting unprocessed pre-mRNA contains exons (boxes) separated by introns (lines). The exons, in turn, consist of protein-coding sequences (dark grey) and untranslated regions (UTRs; light grey). The introns are subsequently removed to produce a mature mRNA, which is transported out of the nucleus into the cytoplasm, where it is translated into a protein (bottom).

enhancers and silencers from interacting with the wrong promoter (Noonan and McCallion 2010). Many genes are regulated by arrays of autonomous enhancers, each of which drives expression in a subset of tissues (Visel et al. 2009), and it is therefore thought that enhancers play an important role in evolution, as changes in a given elements are less likely to have widespread side effects (Wray 2007).

Pol II typically begins transcription by producing several short RNA fragments, before it manages to leave the promoter and the GTFs behind and enter the elongation phase (Saunders, Core, and Lis 2006). In some genes, Pol II pauses almost immediately after escaping the promoter, where it remains, ready to quickly resume elongation upon induction (Margaritis and Holstege 2008). This was originally thought to be a relatively rare phenomenon, but it now seems that many genes undergo transcription initiation without subsequent elongation (Guenther et al. 2007).

1.1.2. Chromatin remodelling

Transcription may also be controlled by the accessibility of the DNA. In the cell, DNA is packed into chromatin, meaning that it is wrapped around nucleosomes, which are complexes of histone proteins. It is thought that the presence of nucleosomes prevents transcriptional regulators from binding and consistent with this view, it has been found that active promoters are typically depleted in nucleosomes (Ozsolak et al. 2007). There is also some evidence to suggest that, at least in some cases, this depletion is a requirement, rather than a consequence, of Pol II binding (Bai and Morozov 2010).

Many histone modifications, such as acetylation or methylation of specific residues, are associated with specific gene regions and frequently correlate with transcription rate, however in this case the direction of causality is less clear (Li, Carey, and Workman 2007).

1.1.3. Post-transcriptional modifications

Transcription is not sufficient to produce a fully functional mRNA. Most genes undergo splicing, which is a process where certain regions of the transcript, the introns, are removed and the remaining sequences, the exons, are joined together (Sharp 1987). Also, the ends of the RNA molecule must be modified to avoid degradation. The 5′ end modification is known as a "cap" (Shatkin 1976), whereas the 3′ end receives a tail of adenosines and is therefore said to be polyadenylated (Millevoi and Vagner 2010).

Splicing presents the opportunity to alternatively exclude and include certain exons in the final mRNA. More than 90% of human genes produce such alternative splicing isoforms (Wang et al. 2008). It is also common for genes to have multiple polyadenylation sites (Ozsolak et al. 2010) and transcription start sites (Carninci et al. 2006). The resulting isoforms may encode slightly different proteins or contain different sets of regulatory signals that affect later steps of the gene’s expression.

The mRNA may also be modified through RNA editing, where single bases within the transcript are altered. The most frequent change is from adenosine into inosine, which in most contexts is equivalent to guanosine. Editing can affect both regulatory signals and the encoded protein sequence (Farajollahi and Maas 2010).

1.1.4. mRNA transport

Following transcription, which takes place in the cell’s nucleus, the mRNA is exported to the cytoplasm for translation (Stewart 2010). Some mRNA are further transported to specific subcellular compartments. The sequence elements that direct this localisation are typically located in the 3′ untranslated region (UTR) of the transcript (Martin and Ephrussi 2009) and may differ between alternative isoforms of the same gene (An et al. 2008).

1.1.5. Regulatory RNAs

The cytoplasm also houses many non-coding RNAs (ncRNAs), which play a role in gene regulation. MicroRNAs (miRNAs) are short RNAs that bind to partially complementary target sequences within the mRNA (most commonly located in the 3′ UTR), thereby either causing the mRNA to be degraded or preventing it from being translated (Huntzinger and Izaurralde 2011). It has been estimated that miRNAs affect over 60% of human genes (Friedman et al. 2009). A less well-studied class of RNAs is the endogenous short interfering RNAs (siRNAs), which are similar to miRNAs, but complement their targets perfectly. They too can target mRNAs for degradation (Okamura and Lai 2008).

There are also examples of longer ncRNAs with diverse roles in gene regulation (Ponting, Oliver, and Reik 2009). Considering the large amount of non-coding transcripts (Carninci et al. 2005), it seems likely that many new instances of ncRNA regulation will be identified in the future, although it should be taken into account that some of these transcripts might represent transcriptional noise (Ponting, Oliver, and Reik 2009).

1.1.6. Translation and beyond

The first task of translation is for the small ribosomal subunit to scan the mRNA, starting from the 5′ end, until it reaches the start codon and is joined by the larger subunit. The efficiency of the scanning depends on the secondary structure of the 5′ UTR (in many cases translation initiation can only proceed if a helicase is present), as well as a number of regulatory proteins that bind to the 5′ and 3′ UTRs (Sonenberg and Hinnebusch 2009).

Following ribosome assembly, elongation begins and proceeds until the ribosome encounters a stop codon. If this stop codon occurs in an unexpected position, it can trigger a pathway known as nonsense-mediated decay (NMD), which leads to the degradation of the mRNA. The NMD mechanism is used to detect and remove faulty transcripts, however, it also plays a role in gene regulation, in which case it may be triggered by various other signals, including introns in the 3′ UTR or short open reading frames (ORFs) within the 5′ UTR (Nicholson et al. 2010).

Finally, it should be noted that even if gene expression may be considered over once a protein has been produced, the function, activity and turnover rate of that protein can still be extensively modified by the addition of various post-translational modifications, such as phosphorylation, methylation, acetylation and many others (Walsh, Garneau-Tsodikova, and Gatto 2005).

1.2. Large-scale methods to assess gene expression

As detailed above, human genes can be regulated at virtually every step of their expression. Consequently, no single measurement can capture all facets of gene regulation and conclusions about gene expression evolution may therefore depend on the type of data that is being analysed. Furthermore, all measurement techniques are subject to errors, which, if unaccounted for, may lead to biased results. Before reviewing what is currently known about the evolution of gene expression (see Section 1.3), it is therefore necessary to become familiar with the strengths and limitation of the molecular methods that were used to produce the data. In Section 1.2.1, I will introduce how mRNA levels can be assessed on a genome-wide scale using the microarray technique. The analyses presented in this thesis rely heavily on microarray data and the technique has also played a central role in the general development of the field. I will also discuss the advantages of the more recent method of RNA sequencing. In Section 1.2.2, I will present the chromatin immunoprecipitation (ChIP) method, which can be used to determine which DNA sequences are bound by a given protein and which has been extensively used to reveal the regulatory information encoded in the genomes of humans and other species. In Section 1.2.3, I will discuss some general issues that affect all comparative studies of gene expression, regardless of methodology.

1.2.1. Microarrays

The microarray technique allows simultaneous quantitative assessment of the expression levels of thousands of mRNAs. Briefly, microarrays are chips that contain DNA fragments, “probes”, which are complementary to the mRNAs of given genes. The microarray chip is incubated together with an RNA or cDNA sample, which has been labelled with a fluorescent dye. The RNAs or cDNAs will hybridise with their corresponding probes and by measuring the intensity of fluorescence for each probe, it is possible to estimate the concentrations of the different RNA species in the sample (Schena et al. 1995). Having access to mRNA expression levels makes it possible to study the combined effect of many regulatory steps, without knowledge of the exact regulators or target sequences involved. However, it is nonetheless important to remember that although mRNA level often correlates with protein level (Fu et al. 2007; Gry et al. 2009), it may be the case that differences in mRNA concentrations are buffered or enhanced by subsequent regulation.

While the principle of microarrays is simple, they need to be carefully interpreted to avoid misleading artefacts. Firstly, the technique is sensitive to subtle variations in the experimental procedures, such that there can be substantial variation in observed intensities even between replicates of the same experiment. This technical variation needs to be addressed by statistical normalisation of the data (Reimers 2010). Secondly, the physical location of individual probes on the microarray chip can influence how well they hybridise with their targets and non-random chip designs may therefore skew the results (Verdugo et al. 2009). Thirdly, the hybridisation efficiency is also dependent on the exact sequence of the probes, which means that probes targeting the same gene may give different amounts of signal (Irizarry et al. 2005). If gene expression is measured in two species, using species-specific probes, this probe effect will therefore tend to inflate the observed differences. However, for closely related species, which are not too divergent in terms of mRNA species, identical probe sets can be used: For example, it is possible to compare human and chimpanzee gene expression using human-specific microarrays, provided that all probes with mismatches are removed from the analysis (Khaitovich et al. 2005), as probe-target mismatches influence the hybridisation profile (Gilad et al. 2005).

More recent technologies, such as RNA sequencing, can circumvent some of the issues described above. RNA sequencing provides short reads of the RNAs present in a sample and these sequence fragments can be bioinformatically processed to give information about the full transcripts (Costa et al. 2010). In addition to estimating expression levels, RNA sequencing can be used to detect unknown alternative isoforms or cases of RNA editing. As the technique does not rely on hybridisation with known sequences, it also does not suffer from the same cross-species issues that affect microarrays. However, as this is a relatively new technique, the number of available datasets is limited.

1.2.2. Chromatin immunoprecipitation

For a more in-depth understanding of the different regulatory steps that affect gene expression, it is useful to know which molecular factors are associated with the DNA at different locations and time points. Chromatin immunoprecipitation (ChIP) is a technique where cells are treated with formaldehyde to create covalent bonds between the DNA and its associated proteins. The DNA is then fragmented and those fragments that are associated with a protein of interest can be identified by microarrays (ChIP-chip) or sequencing (ChIP-seq) (Collas 2010). The method has been extensively used to identify TF binding sites and locations of modified histones in the human genome, as part of the Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium et al. 2007). There are also similar projects for two model organisms: the fruit fly *Drosophila melanogaster* (The modENCODE Consortium et al. 2010) and the nematode *Caenorhabditis elegans* (Gerstein et al. 2010).

1.2.3. Sources of variation in gene expression data

The generation and analysis of gene expression datasets is a complex process and it is therefore critical to be aware of the many factors that can contribute to observed variation within and between species. Alongside technical variation, there are biological aspects that need to be taken into account, to ensure that the analysed samples are directly comparable. For example, it has been shown that gene expression can change with the diet (Somel et al. 2008) and age (Lu et al. 2004) of the sampled individuals. Some observed gene expression differences might therefore be explained by changes in environment or by skewed sampling, where the age of the sampled individuals has not been matched across species (Hodgins-Davis and Townsend 2009).

In cross-species studies it is often especially difficult to obtain perfectly matched samples, because of ecological differences between the species (Hodgins-Davis and Townsend 2009). When the different contributing factors are correlated in this way, it can be impossible to tease apart the genetic and environmental effects (Leek et al. 2010). While many confounding factors are impossible to eliminate, care should be taken to keep them to a minimum.

1.3. Gene expression in humans and other species

Already four decades ago, it was suggested that changes in gene expression have played a major role in phenotypic evolution (Britten and Davidson 1969; King and Wilson 1975). This view is still widely held (Carroll, Grenier, and Weatherbee 2004; Wray 2007), but we currently know too little about regulatory evolution to determine its true contribution to phenotype diversity (Hoekstra and Coyne 2007). This Section will introduce the evolutionary patterns observed for paralogous genes within a genome (Section 1.3.1) and between orthologous genes in different species (Section 1.3.2), as well as discuss the contribution of selection to these observations.

1.3.1. Expression diversification of duplicate genes

Although the most likely outcome of a gene duplication event is that one of the copies is subsequently silenced, occasionally both duplicates remain active in the genome (Lynch and Conery 2000). The fate of the two copies may however be very different: it is frequently observed that one gene in the duplicate pair evolves quickly, both in terms of protein-coding sequence (Zhang, Gu, and Li

2003) and gene expression (Gu, Zhang, and Huang 2005). This is consistent with a model where one copy carries out the ancestral function, whereas the other is free to adopt a new role, so-called neo-functionalisation (Zhang 2003). Analysis of the tissue specificity of human and mouse duplicates suggests that in as much as half of all cases, one copy retains the original expression pattern, while the other has lost expression in some tissues: in another 25% of gene pairs the gene expression pattern has been partitioned between the copies, consistent with sub-functionalisation, where each copy performs a subset of the functions originally carried out by the ancestral gene, while the remaining genes show similar expression patterns, suggesting that they act to increase gene dosage (Farre and Alba 2010).

Three main scenarios are possible for a new regulatory mutation: it may be deleterious and subsequently get removed by negative selection, it may be neutral and randomly change in frequency until it is either fixed or lost, or it may be adaptive and get driven to fixation by positive selection. The accelerated evolution of one of the two duplicates could therefore be due either to relaxed negative selection, which would allow neutral accumulation of mutations or to positive selection. It is not known whether the fixation of duplicate genes is primarily a neutral or adaptive process (Innan and Kondrashov 2010) and although some models have been devised to categorise the evolutionary patterns of expression changes in duplicate genes, these do not address the role of positive selection in generation the observed patterns (Gu 2004; Oakley et al. 2005).

The burst of expression change for one of the copies following gene duplication represents a special case and the divergence of gene expression for orthologous genes in different species might not be explained by such periods of rapid evolution. However, it should be noted that lineage-specific gene duplication and expression divergence between lineages are correlated phenomena, such that orthologous genes with a duplication event in one species tend to show more divergent expression patterns between species, possibly indicating a causal relationship (Huminiecki and Wolfe 2004).

1.3.2. Differences in gene expression between humans and other species

It is impossible to give a single measure of the level of conservation between human gene regulation and that of other species. Some aspects are remarkably similar across taxa: for example, TFs encoded by the *Hox* gene family regulate early development in an analogous fashion across all animals (Carroll, Grenier, and Weatherbee 2004). On the other hand, the majority of binding regions for various TFs do not overlap between humans and mice (Kunarso et al. 2010; Schmidt et al. 2010).

Considering the cognitive differences between humans and chimpanzees, one might expect to see large gene expression changes in the human brain. However, expression divergence between the two species is lower in brain samples, than it is in samples from heart, kidney, liver and testis (Khaitovich et al. 2005). Furthermore, Broca's area, which controls human speech, does not show significant expression differences compared to other parts of the human cerebral cortex (Khaitovich et al. 2004a).

Early findings suggested that human gene expression was evolving without constraint, i.e., that the vast majority of all new mutations were neutral (Khaitovich et al. 2004b; Yanai, Graur, and Ophir 2004), but later studies have demonstrated an extensive role for negative selection in comparisons between humans and chimpanzees (Lemos et al. 2005), as well as humans and mice (Liao and Zhang 2006).

The contribution of adaptive mutations to gene expression evolution is still an open question (for a review of how this question has been studied, see Chapter 3). It may well be that selective pressures differ between tissues. In particular, it has been noted that gene expression is unusually divergent between human and chimpanzee testis samples, after correcting for the variation among individuals (Khaitovich et al. 2005). This might indicate that gene expression is positively selected in this tissue, although this has yet to be formally tested and could have other explanations (Khaitovich et al. 2006).

1.4. Objectives of this thesis

The aim of this thesis is to increase our understanding of the principles of gene expression evolution, with emphasis on humans. In this section I will describe how each of the four analytical chapters relates to a fundamental question in the field. Each chapter contains a more thorough review of the relevant literature for the topic.

1.4.1. Generation of regulatory variation

The first step towards a more complete appreciation of gene expression evolution is to identify how new expression variants are created. Mutations that affect gene expression in humans span the range form point mutations that change a single nucleotide to copy number variants (CNVs) of 1 kb or more (Stranger et al. 2007). Transposable element (TE) insertion is a particular type of mutation that has been suggested to play a major role in human evolution (Britten 2010). Chapter 2 of this thesis investigates whether TEs have caused expression differences between humans and chimpanzees.

1.4.2. Selection acting on gene expression

Secondly, we want to know how selection acts on the observed variation. Adaptive mutations are of special interest, as they increase the organism's fitness. However, identifying them remains problematic. In Chapter 3, I present a new method to estimate the proportion of expression variation that is due to adaptive evolution. I apply this method to expression divergence between humans and chimpanzees.

1.4.3. Limits to regulatory diversification

Even in the absence of negative selection on gene function, there will still be limitations to regulatory evolution. For example, expression levels cannot increase beyond the capacity of the transcriptional machinery. Knowledge of such external constraints and the extent to which they curb expression is important for correctly modelling neutral evolution over longer time periods. In Chapter 4, I investigate the accumulation of regulatory mechanisms through time and examine whether regulatory complexity is a limiting factor in humans. I consider many different facets of regulation, including transcriptional regulation, alternative processing, miRNA regulation, NMD and RNA editing.

1.4.4. Sequence signatures of regulatory elements

Sequence conservation is frequently used to identify regulatory elements, as functionally important units are expected to be maintained by negative selection. However, the relationship between function and sequence conservation is far from straightforward. On one hand, many elements with demonstrated roles in gene regulation are not especially well conserved (Blow et al. 2010). On the other hand, some sequences, known as ultraconserved elements (UCEs), are identical across species, even though no known molecular mechanism seems to require that degree of conservation (Bejerano et al. 2004). UCEs have primarily been studied in humans and other vertebrates, but further insights might be possible using model organisms that are more easily manipulated. In Chapter 5, I analyse UCEs that are shared between twelve *Drosophila* species.

2.

Transposable elements: insertion pattern and impact on gene expression evolution in hominids

2.1. Introduction

Almost half of the human genome is made up of transposable elements (Lander et al. 2001). These DNA sequences are able to insert into a new genomic location through the process of transposition. While most such insertions are likely to be subsequently lost due to selection or genetic drift, our lineage has still accumulated more than 7500 TE copies since the split from chimpanzees (Mills et al. 2006), with three families accounting for more than 95% of these transposition events: the Long Interspersed Element 1 (L1), the Alu element, which belongs to the Short Interspersed Elements (SINEs), and the SVA element (SINE-R, VNTR, Alu).

TEs have commonly been viewed as selfish parasites, whose persistence in the genome is best explained by their success as replicating units, rather than any benefit they might bestow on the host (Doolittle and Sapienza 1980; Orgel and Crick 1980). Indeed, the presence of TEs can severely impair genome function, either by direct disruption of functional sequences (Kazazian et al. 1988) or by promoting ectopic homologous recombination, which can lead to potentially harmful duplications, deletions and genome rearrangements (Hedges and Deininger 2007).

On the other hand, some TE-derived sequences are among the most conserved elements of the human genome (Kamal, Xie, and Lander 2006; Lowe, Bejerano, and Haussler 2007), suggesting that some TEs are functional. In particular, some TEs have been found to play a role in transcriptional regulation by providing genes with promoters and enhancers (Jordan et al. 2003; van de Lagemaat et al. 2003; Bejerano et al. 2006; Bourque et al. 2008). Several human genes are transcribed from a promoter situated within the L1 element (Nigumann et al. 2002) and transcripts originating within Alus have also been reported (Faulkner et al. 2009). The evolutionary potential of TE-derived *cis*-regulatory sequences was recently demonstrated in rice, where recent TE insertions have led to upregulation of gene expression and the creation of new regulatory networks (Naito et al. 2009).

Other mechanisms may also contribute to the transcriptional impact of TEs, such as reduced elongation efficiency or premature polyadenylation following intronic L1 insertion (Han, Szak, and Boeke 2004). Furthermore, mammalian TE activity is under epigenetic control, through siRNAs (Yang and Kazazian 2006), histone modifications (Martens et al. 2005) and DNA methylation (Walsh, Chaillet, and Bestor 1998). In *Arabidopsis thaliana*, a side effect of epigenetic silencing has been reduced expression of neighbouring cellular genes (Hollister and Gaut 2009).

With this in mind, it is tempting to ask how the evolution of human gene expression has been affected by TE activity. Expression divergence (ED) is a measure of the difference in gene expression levels between two species. Two previous studies have suggested a relationship between TE insertions and ED.

Firstly, there is a correlation between the number of Alu insertions and ED as measured between human and mouse, although the direction of the correlation depends on the statistic used to measure ED (Urrutia, Ocana, and Hurst 2008). The authors also concluded that Alu elements are enriched around broadly expressed genes, but they do not themselves drive an expansion of gene expression patterns. Secondly, a positive correlation between ED and the number of lineage-specific SINEs and Long Terminal Repeat (LTR) elements has been found in rodents, where, although the amount of variance explained was modest, the average effect of TEs was considerable and appeared to have contributed around 20% of the total ED between mouse and rat (Pereira, Enard, and Eyre-Walker 2009).

Here, we investigate to what extent TE activity has contributed to hominid evolution by analysing quantitative changes in gene expression and transcript diversity between human and chimpanzee.

2.2. Materials and Methods

We used two datasets to study the evolution of gene expression. In the first, microarray expression data for brain, heart, kidney, liver and testis was available from six humans and five chimpanzees (Khaitovich et al. 2005). These experiments were conducted using the Affymetrix U133plus2 array, which was designed for human sequences, but contains a number of probes that match chimpanzee sequences equally well. This array has been shown to perform well in comparison to other arrays, including the newer exon arrays (Robinson and Speed 2007). The raw data was masked using the protocol developed by Toleno et al. (2009), in which probes were removed unless they had a perfect, single match in both the human and the chimpanzee genome. Furthermore, only probe sets that contained at least six such probes were used for further analysis, as probe sets represented by fewer probes tend to give unreliable results (Toleno et al. 2009). Expression values were calculated using the RMA (robust multichip analysis) function in the Bioconductor affy package (Irizarry et al. 2003a; Irizarry et al. 2003b; Gentleman et al. 2004). (Processed data was kindly provided by Joe Hacia of the University of Southern California).

For each gene, we calculated ED between human and chimpanzee as the Euclidean distance between the average log-transformed expression values for each tissue. If a gene was assigned multiple probe sets, a single probe set was chosen at random to represent that gene, in order to avoid bias in the estimation of ED (see Section 2.3). Gene coordinates were downloaded from the UCSC Genome Bioinformatics site (Rhead et al. 2010), using genome build hg18 for human and panTro2 for chimpanzee. For genes with alternative transcripts, a single transcript was chosen at random among those that matched the probe set representing that gene.

To allow lineage-specific analysis of ED, we analysed a second dataset, which included data from rhesus macaque as an outgroup species. Somel et al. (2009) measured gene expression levels in the prefrontal cortex of 39 humans, 14 chimpanzees and 9 rhesus macaques, using the Affymetrix U133plus2 platform as in the first dataset. The raw data was masked using files made available by the authors to include only probes that had a single, perfect hit in the genomes of all three species and to require each gene to be represented by at least eight such probes. Log-transformed expression values were calculated using the RMA function in Bioconductor (Irizarry et al. 2003a; Irizarry et al. 2003b; Gentleman et al. 2004). We calculated ED as the Euclidean distance between the average expression levels for the relevant species and normalised the values by dividing by the mean ED value for that species pair. To determine whether the individuals in the dataset had reached puberty or not, we used life history data from the AnAge database (de Magalhaes and Costa 2009).

Recently inserted TEs in the human and chimpanzee genomes had previously been identified by Mills et al. (2006). We converted the data to current genome coordinates, using the UCSC liftOver tool (Rhead et al. 2010). Due to rearrangements in the updated genome assemblies, conversion failed for 7 human and 440 chimpanzee entries. These were excluded from the set. We then scored each of the genes for which we had expression data, according to the presence or absence of a recent TE insertion within the following seven regions: 0-2 kb, 2-10 kb and 10-20 kb upstream and downstream of the transcript and within the introns.

To identify TEs present in both human and chimpanzees, but not in the rhesus macaque (genome release rheMac2), we used the human-chimpanzee and human-macaque net alignments displayed in the UCSC Genome Browser (Rhead et al. 2010). We identified all gaps in the human-macaque alignment that did not match a human-chimpanzee gap and then compared these to transposable elements in the RepeatMasker track. To allow for slight annotation errors, we isolated all RepeatMasker entries where the coordinates matched a gap in the rhesus macaque sequence, plus/minus 20 bp.

Expression state in the germ line was assigned according to eGenetics/SANBI EST data (Kelso et al. 2003), as incorporated in Ensembl release 56 (Flicek et al. 2010), by considering genes active if they were associated with the Cell Type term “germ cell”.

2.3. Results

We set out to investigate if recent TE insertions in the human or chimpanzee lineage have led to increased ED in nearby genes. Lineage-specific TEs had previously been identified (Mills et al. 2006) by identifying indels in the human-chimpanzee genome alignment and matching these to TEs in RepBase version 10.02 (Jurka 2000). Thus, the set of new TE insertions may also contain a small number of ancient TEs that were precisely deleted in one species. Genes were classified according to the presence of a recently inserted TE within 0-2 kb, 20-10 kb and 10-20 kb either upstream or downstream of the transcribed sequence or within the introns. No exonic TEs were found.

Microarray expression data for both species were available for 8995 genes and five tissues (Khaitovich et al. 2005; Toleno et al. 2009). We calculated ED as the Euclidean distance between the log-transformed tissue-specific expression values for each species. We decided against another commonly used alternative definition of ED, based on the correlation coefficient, as it tends to overestimate ED for genes with conserved uniform expression (Pereira, Waxman, and Eyre-Walker 2009).

Calculations of ED were complicated by the fact that some genes were represented by more than one probe set in the microarray data. Although the platform used to generate the data was not designed to address alternative splicing, some probe sets have still been created to target different transcripts of the same gene. If different numbers of probe sets are used to generate the ED values and if the probability of retaining a TE is related to whether the affected gene undergoes alterative processing, this could introduce a bias into the analysis. Indeed, we found that human genes to which we had mapped at least one recently inserted TE had on average 2.7 annotated Ensembl transcripts, whereas genes without insertions had 2.3 transcripts ($p = 2 \times 10^{-16}$, Mann-Whitney U test). The corresponding values for chimpanzee were an average of

2.0 transcripts for genes with TEs and 1.8 transcripts for genes without ($p = 5 \times 10^{-10}$). To avoid bias in our estimates of ED we therefore decided to let each gene be represented by a single probe set chosen at random.

To evaluate the effect of TE insertions on ED, we compared genes with or without TEs within their upstream, downstream and intronic sequences. An overview of the analysis is shown in Figure 2.1. Although we found a marginally significant increase in median ED for genes with L1 insertions within 0-2 kb upstream and gens with SVA insertions within 0-2 kb downstream ($p = 0.030$ and $p = 0.032$, Mann-Whitney U test), these results are not significant after correcting for multiple tests. We therefore combined the data from each TE family (Figure 2.2). In spite of a general tendency towards an increase in median ED, none of the regions gave significant results when considered separately. However, if we combine these p values, using the Z transformation method (Whitlock 2005), the result is significant ($p = 0.024$), and even more so if we exclude the regions 10-20 kb upstream or downstream ($p = 0.0027$).

It therefore seems that genes with new TEs have higher ED. It is, however, not possible to infer the direction of causality based on these results, as they could be explained either by increased ED as an effect of TE insertion or by a tendency for genes with higher ED to accumulate TEs. To test between these alternatives we identified TE insertions that occurred before the human-chimpanzee split, but after the split from rhesus macaque: we reasoned that these fairly

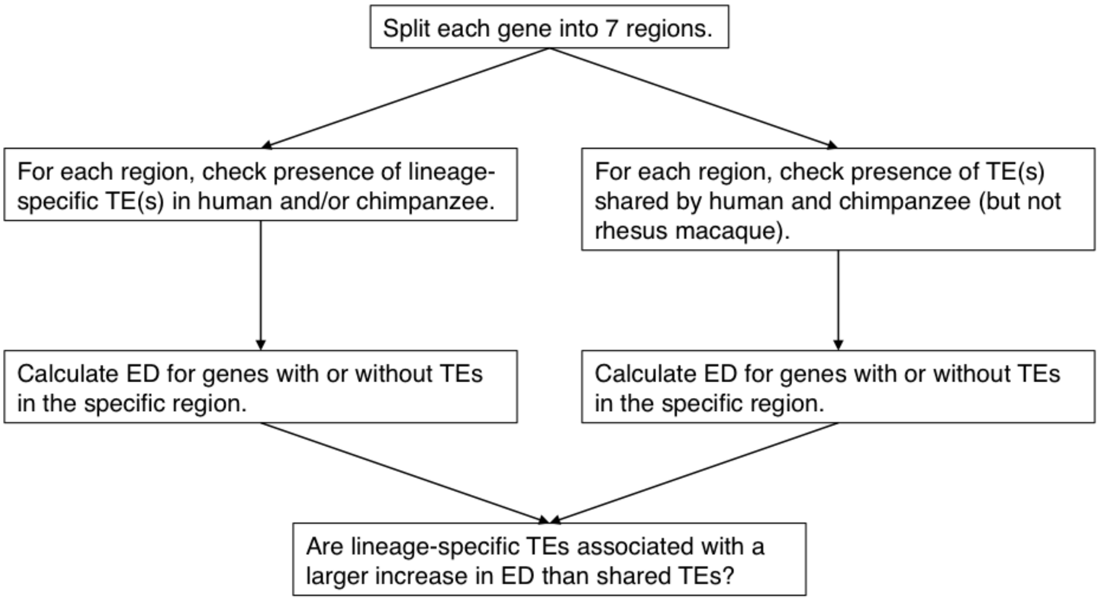
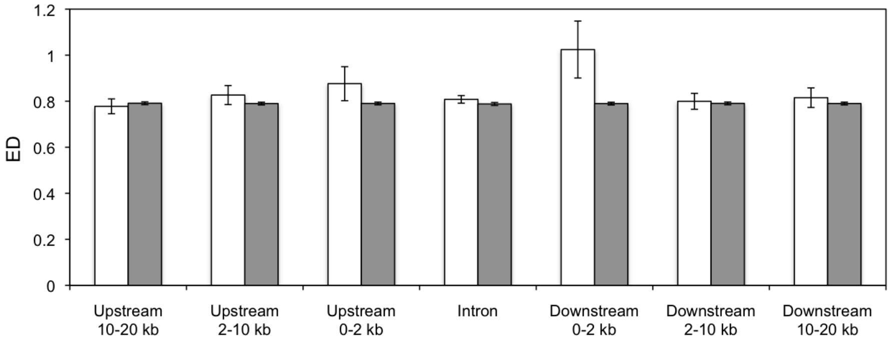

Split each gene into 7 regions.
For each region, check presence of lineage-specific TE(s) in human and/or chimpanzee.
For each region, check presence of TE(s) shared by human and chimpanzee (but not rhesus macaque).
Calculate ED for genes with or without TEs in the specific region.
Calculate ED for genes with or without TEs in the specific region.
Are lineage-specific TEs associated with a larger increase in ED than shared TEs?

Figure 2.1. Flowchart describing the steps of the analysis (see text for details).

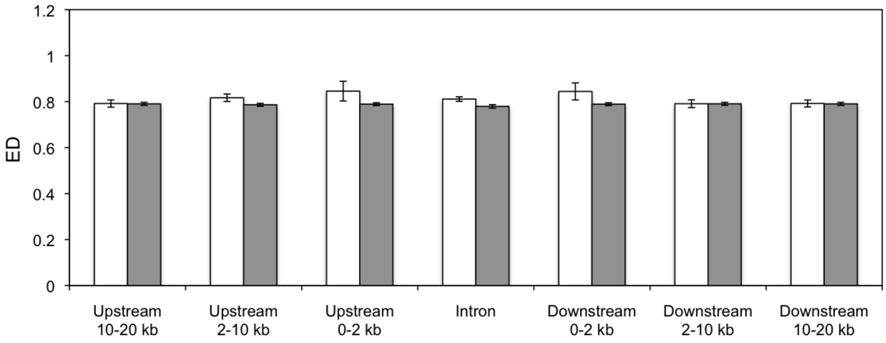


Genes with TEs	
U 10-20	285
U 2-10	237
U 0-2	50
Intron	1180
D 0-2	42
D 2-10	204
D 10-20	264

Figure 2.2. The association between lineage-specific TE insertions and ED. Mean ED for genes with (white) or without (gray) TEs specific to either human or chimpanzee within 0-2 kb, 2-10 kb or 10-20 kb upstream or downstream of the transcribed region or within the introns. The number of genes carrying species-specific TE insertions in a specific region is listed on the right. Standard errors are indicated as bars.

recent insertions should affect humans and chimpanzees equally and therefore not contribute to ED between the two species. We found that genes with shared TE insertions did display a significantly higher level of ED, if we combined TEs within regions and probabilities as above (combined p value = 0.00003), indicating that TEs tend to integrate and/or be retained in genes that for some other reason are more likely to change their expression level (Figure 2.3).

Thus, at least part of the increase in ED for genes with species-specific TE insertions can be explained as a background effect, which also affects genes with shared TEs. Nevertheless, it is possible that TEs induce an additional increase in ED. To investigate this, we calculated the relative effect of TEs on ED as the ratio between the average ED values for genes with species-specific TEs and genes without such TEs, divided by the ratio between the average ED values for genes with shared TEs and genes without such TEs. If the relative effect is above one, it indicates that the presence of species-specific TEs acts to increase ED over and above the general tendency for TEs to integrate into genes with high ED. Howver, we find that the relative effect is not significantly above one for any of the seven regions under consideration (Figure 2.4). The highest relative effect is observed for genes with TEs within 0-2 kb downstream of the transcript, but the 95% confidence interval obtained by bootstrapping is (0.97, 1.48) for this single value and thus the result is not significant. As we cannot detect any increase in ED due to new TE insertions, beyond what can be explained by a general tendency for gene with higher ED to retain TEs, we con



Genes with TEs	
U 10-20	1368
U 2-10	1268
U 0-2	262
Intron	3219
D 0-2	269
D 2-10	1054
D 10-20	1409

.

Figure 2.3. The association between shared TEs and ED. Mean ED is given for genes with (white) and without (grey) a TE shared between humans and chimpanzees, within 0-2 kb, 2-10 kb or 10-20 kb upstream or downstream of the transcribed region or within the introns. The number of genes carrying shared TE insertions in a specific region is listed on the right. Standard errors are indicated as bars.

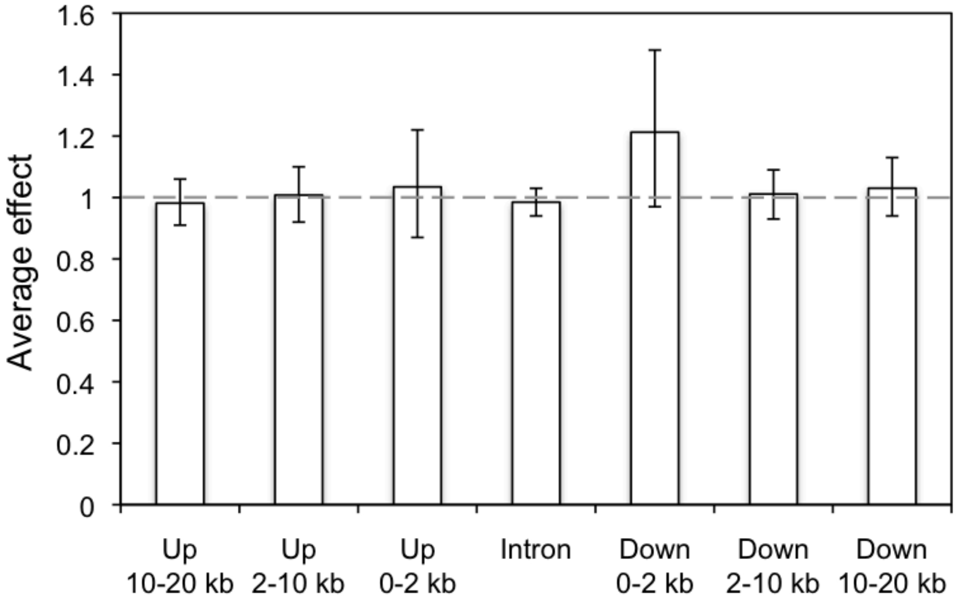


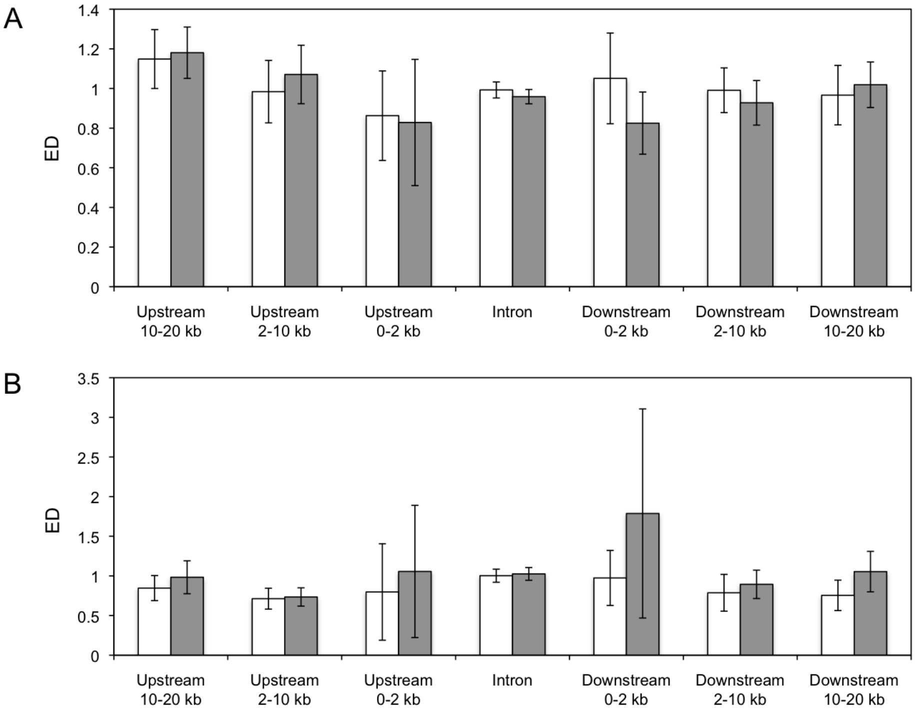
Figure. 2.4. The relative effect of TE insertion on ED. The relative effect was calculated as the ratio between the mean ED values for genes with and without species-specific TEs, divided by the ratio of the mean ED values for genes with or without shared TEs. The bars represent 95% confidence intervals obtained by bootstrapping.

clude that TE activity has not contributed to the genome-wide evolution of gene expression levels in humans and chimpanzees.

Although we find no evidence that new TE insertions increase ED in the analysis above, it is possible that this is due to a lack of power. We therefore sought to test whether TEs affect ED using a complementary approach. For genes with a new TE insertion in humans we compared the ED between human and macaque to the ED between chimpanzee and macaque. We also performed the corresponding analysis for genes with a TE in chimpanzees. If TEs affect ED, we predict that genes with a human-specific TE insertion will show higher ED between human and macaque than between chimpanzee and macaque, with the converse being the case for genes with a chimpanzee-specific insertion. To perform the analysis, we only considered genes that had one or more insertions in one species, but none in the other. We analysed microarray data for 3747 genes in the prefrontal cortex of 39 humans, 14 chimpanzees and 9 rhesus macaques (Somel et al. 2009). The presence of an out-group in this dataset allowed us to assess changes in ED on a lineage-specific basis. To do so, we calculated human-macaque and chimpanzee-macaque ED as the Euclidean distance between the means of the log-transformed expression values for each species. Because human and chimpanzee share a common history, these ED values represent the sum of a species-specific component as well as a shared component that accounts for all ED between rhesus macaque and the human-chimpanzee ancestor. Any difference between the human-macaque and chimpanzee-macaque ED values can therefore be directly attributed to human-specific or chimpanzee-specific events.

On average, chimpanzee-macaque ED is higher than human-macaque ED in this dataset. Consequently, if we test for an increase in ED for the lineage with TE insertions, the test would be too conservative for human-specific TEs and too liberal for chimpanzee-specific TEs. To allow for an unbiased test, we normalised all ED values by dividing the ED for each gene by the mean ED for that species pair. Note, the fact that ED between chimpanzee and macaque is higher than that between human and macaque does not necessarily imply accelerated evolution along the chimpanzee lineage. Rather, it might be best explained by the higher variance among chimpanzee individuals in this dataset, especially considering previous work indicating that ED in the brain is higher along the human lineage (Khaitovich et al. 2005).

Consistent with our previous analysis, we find no evidence, in any of the regions examined, that a lineage-specific TE tends to increase ED in that species relative to ED in the other species (Figure 2.5). This is true even if we combine probabilities across introns and flanking regions ($p = 0.32$ for human-specific TEs and $p = 0.13$ for chimpanzee-specific TEs; Mann-Whitney U test and Z transformation). Because the samples used to generate the expression data were taken from individuals of varying ages (Somel et al. 2009), we repeated the analysis separately for samples from pre-pubertal and post-pubertal individuals, in order to reduce age-related variation. Again, the results were not



Genes with TEs	
U 10-20	102
U 2-10	71
U 0-2	18
Intron	843
D 0-2	19
D 2-10	62
D 10-20	91

Genes with TEs	
U 10-20	30
U 2-10	26
U 0-2	3
Intron	180
D 0-2	6
D 2-10	18
D 10-20	16

Figure 2.5. The effect of TE insertions on lineage-specific ED. A. Mean human-macaque ED (white) and chimpanzee-macaque ED (gray) for genes with human-specific TE insertions. B. The same for genes with chimpanzee-specific TE insertions. The number of genes with human-specific or chimpanzee-specific TE insertions is listed. Standard errors are indicated as bars.

significant (combined probabilities, pre-pubertal individuals: p = 0.42 for human TEs, p = 0.17 for chimpanzee TEs; post-pubertal individuals: p = 0.84 for human TEs, p = 0.13 for chimpanzee TEs), providing further support for the hypothesis that recent TE insertions have not acted to increase ED between humans and chimpanzees.

During our analysis of species-specific TEs, we observed that upstream insertions were more frequent than downstream insertions. In total, we identified 561 genes with at least one new TE within 20 kb upstream of the transcription start site in either human or chimpanzee and 496 genes with at least one new TE downstream of the transcribed region. The difference is just significant (p = 0.049, two-tailed binomial test) and upstream insertions are also more common if we only consider TEs within 10 kb or 2 kb upstream or downstream of genes, although the overrepresentation is not significant (p = 0.075 and p = 0.047, respectively). This enrichment of upstream insertions is surprising, since we might expect that TEs inserted upstream would be more likely to disrupt transcriptional regulatory elements and therefore tend to be selected against, although it has previously been noted that TE insertions in the 3′ flanking region of rodent genes tend to show bigger effects on ED than those in the 5′ region (Pereira, Enard, and Eyre-Walker 2009). Another explanation is that TEs are preferentially inserted upstream of genes, as is the case for P elements in *Drosophila melanogaster* (Spradling et al. 1995), where it is presumed to be linked to the altered chromatin structure around the transcription start site of active genes (Kelley et al. 1987; Voelker et al. 1990). If the same is true for hominid

TEs, then we would expect an enrichment of upstream TE insertions for genes that are expressed in the germ line, but not for other genes. Based on expression data downloaded from Ensembl (see Section 2.2), we categorised all genes as active or inactive in the germ line and compared the number of upstream and downstream insertions for active and inactive genes. When we considered all recent TE insertions together, we found that the inactive genes had approximately the same number of upstream and downstream insertions, whereas active genes had significantly more upstream insertions ($p = 0.003$, χ^2 test). The pattern is contributed mainly by Alu and, to some extent, SVA elements, whereas L1 elements appear unaffected (Table 2.1).

Although species-specific TEs have not affected ED between human and chimpanzee, they may still have had an influence on other aspects of gene expression evolution, such as transcript diversity. As described above, we established that genes with recent TE insertions have a significantly higher number of annotated transcripts than genes without such insertions. Since both Alu and L1 elements can be involved in processes such as alternative promoter usage (Nigumann et al. 2002; Faulkner et al. 2009) and alternative splicing (Makalowski, Mitchell, and Labuda 1994; Sorek, Ast, and Graur 2002; Belancio, Hedges, and Deininger 2006; Lev-Maor et al. 2008), which act to increase transcript diversity, we speculated that the TE insertions themselves might in part explain why the affected genes tended to produce more transcripts. The differences in annotation quality between the human and chimpanzee transcriptomes makes a direct comparison of transcript numbers difficult. Instead,

Table 2.1. Number of recent upstream and downstream TE insertions in genes that are active or inactive in the germline.

	Active Upstream	Active Downstream	Inactive Upstream	Inactive Downstream	p value
Total	255	181	306	315	p = 0.003
Alu	169	129	206	215	p = 0.04
L1	19	19	38	32	Not significant
SVA	51	33	53	55	Not significant

we reasoned that if TEs increase transcript diversity, then human genes should have more transcripts on average if they contained a human-specific TE, than if their chimpanzee orthologue contained a chimpanzee-specific TE. Conversely, we would expect chimpanzee genes with chimpanzee-specific TEs to produce more transcripts than chimpanzee genes where the human equivalent had undergone TE insertion.

We calculated the number of transcripts in the release 54 of the Ensembl database (Flicek et al. 2010) for human and chimpanzee genes that contained recently inserted Alu, L1 or SVA insertions (Figure 2.5). Before correction for multiple tests, there was only one significant result; human genes with a new SVA insertion have significantly more transcripts than human genes with a new SVA insertion in chimpanzees. However, this result is not significant after correction for multiple tests and we do not see a similar pattern for chimpanzee genes. Of course it should be noted that the lack of observed effect of TEs on transcript diversity could be due to insufficient annotation of alternative isoforms.

2.4. Discussion

TEs have previously been proposed as important contributors to the evolution of gene regulation (Britten and Davidson 1971; Feschotte 2008). In contrast to

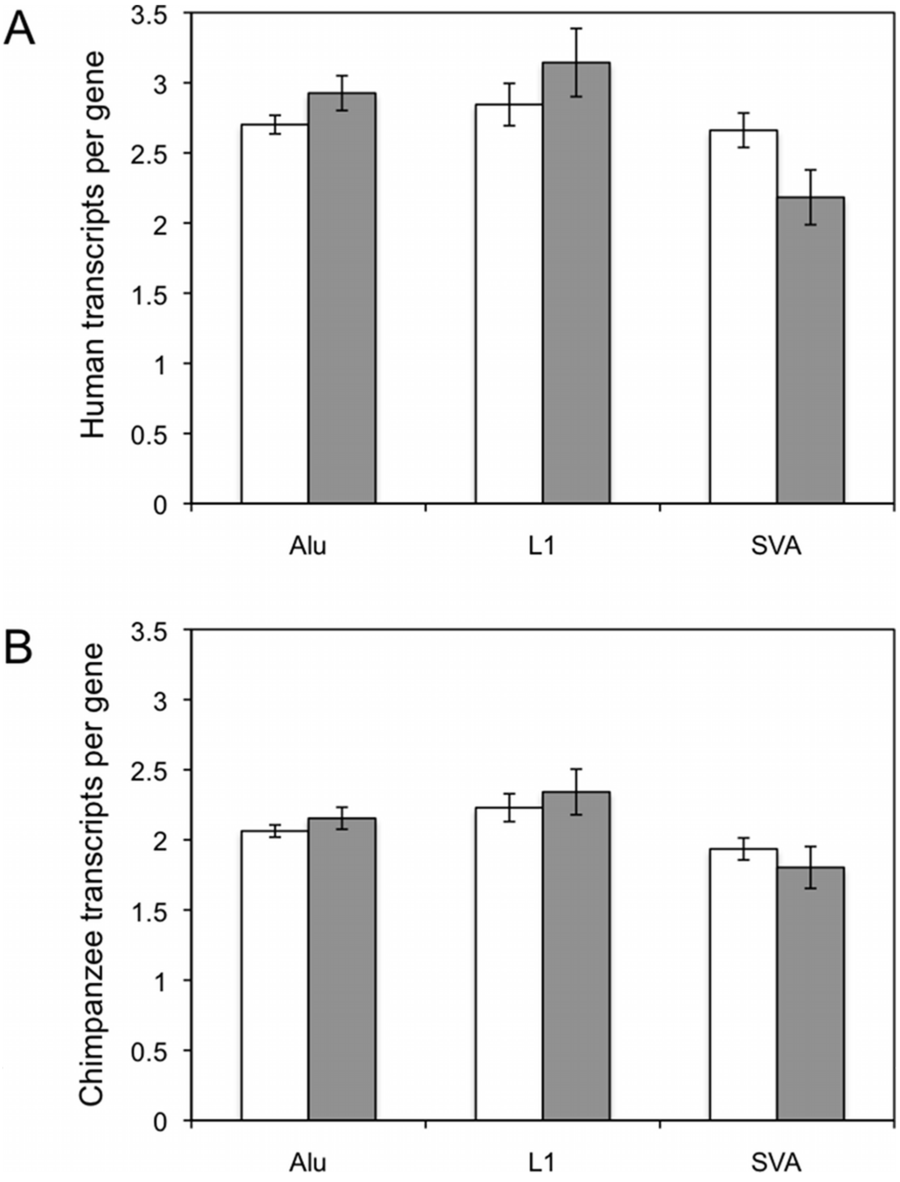


Figure 2.6. The association between transcript diversity and lineage-specific TEs. Figure gives the average number of annotated alternative transcripts in humans (A) and chimpanzees (B) which have a lineage-specific insertion in either humans (white) or chimpanzees (grey).

this, our results show that recent TE activity has not had a detectable effect on ED between human and chimpanzee, suggesting that while TEs may contribute occasionally to gene expression divergence in hominids, they are not a major source of regulatory change.

Our results are consistent with those of Urrutia, Ocana and Hurst (2008), but are surprising considering previous results in mouse and rat, in which it was estimated that 20% of all ED was due to the insertion of new SINE and LTR elements (Pereira, Enard, and Eyre-Walker 2009). The discrepancy between hominids and rodents might be due to qualitative differences in TE activity in the two groups. In rodents the TEs with strongest apparent influence on ED were LTRs and SINEs, however new LTR insertions are rare in the human and chimpanzee genomes, and SINEs, although common, are represented mainly by the primate-specific Alu element (Mills et al. 2006). Pereira, Enard and Eyre-Walker (2009) also attempted to establish causality between ED and new TE insertions by considering the correlation between ED and TE insertions shared by mouse and rat, but these shared TEs were potentially much older than those we have used here and may therefore have been an imperfect control if the pattern of TE insertion had changed over time.

The results presented here are not consistent with a model where TEs affect gene expression by disrupting existing sequences or providing "ready-to-use" regulatory elements. In particular, we find no indications that intronic L1 insertions affect ED, as might have been expected considering that *in vitro* assays have shown that such insertions can attenuate reporter gene expression by reducing elongation efficiency (Han, Szak, and Boeke 2004). On the other hand, although a few candidate cases exist (Schwahn et al. 1998; Yajima et al. 1999), it has yet to be shown that this form of regulation is used *in vivo* (Han and Boeke 2005).

It has been argued that TEs initially may only have a weak impact on gene expression and that this regulatory function is subsequently refined by selection (Faulkner and Carninci 2009). Possibly, the short time scale of this study might therefore not allow us to gauge the full impact of TEs on gene expression, however the findings of two previous studies argue against this: Firstly, at least for Alu elements, recent insertions do not appear to be under selection (Cordaux et al. 2006) and secondly, there is no conclusive evidence that Alu elements have contributed to gene expression evolution along the primate lineage (Urrutia, Ocana, and Hurst 2008). This is not to say that decaying TEs may not provide sequence material in which functional elements can later evolve. There are several examples of human enhancers that have arisen in this way (Britten 1994; Ackerman et al. 2002; Medstrand et al. 2005). Nevertheless, the presence of TE-derived regulatory sequences might best be explained by the abundance of TEs in the genome. Considering that 45% of the human genome has been contributed by TEs (Lander et al. 2001), it stands to reason that these sequences would harbour a fair share of regulatory modules.

It should also be appreciated that while we find no evidence for TEs contributing to differences in gene expression between hominid species, it is still possible that they contribute to variation within a single species. For example, it may be that TEs in general cause mutations of large effect, which rarely are beneficial or neutral and therefore never become fixed between species. Such large effect mutations, providing that they are not lethal, can contribute substantially to variation in fitness and phenotypes, even if they are very deleterious (Eyre-Walker 2010). Thus, while the contribution of TEs to gene expression evolution might be negligible, their impact on human gene regulation could still be of great interest from a medical perspective (Belancio, Hedges, and Deininger 2008).

In a recent study, it was shown that human genes are more likely to be expressed at high levels and in broad patterns if their promoters are rich in TEs, which might indicate that TEs are used to modify chromatin structure upstream of the transcription start site (Huda et al. 2009). Our results, showing that TEs preferentially insert upstream of genes that are transcribed in the germ line, suggest insertion bias as a possible alternative explanation of these results. The same process might also have contributed to the overall enrichment of SINEs in upstream sequences previously observed by Medstrand et al. (2005). Interestingly, it seems that it is primarily Alu elements and, to some extent, SVA elements that experience insertion bias, whereas L1 elements appear to be unaffected. This is surprising, considering that Alus and SVAs are non-autonomous elements that do not encode proteins necessary for transposition, but instead parasitise the L1 machinery (Dewannieux, Esnault, and Heidmann 2003; Ostertag et al. 2003). Although there are some mechanistic differences between Alu and L1 insertions (Kroutter et al. 2009), it is unclear how this might contribute to the observed bias.

The distribution of TEs in the human genome is non-random and correlates with various aspects of gene expression, such as expression levels, transcript diversity and activity in the germ line. Importantly, as illustrated in this study, a correlation does not necessarily imply causality. When studying the contributions of TEs to gene expression evolution it is therefore crucial to apply proper controls in order to disentangle any real effects from the background.

3.

A McDonald-Kreitman-type test for positive selection on gene expression

3.1. Introduction

It has long been suggested that differences between species are often due to alterations in gene expression (Britten and Davidson 1969; King and Wilson 1975; Wray 2007). It would therefore be of great interest to be able to estimate the proportion of expression divergence that is due to positive selection.

If the regulatory regions are already known, a number of sequence analysis tools can be used to test for positive selection acting on the relevant sequences (Jenkins, Ortori, and Brookfield 1995; Kohn, Fang, and Wu 2004; Rockman et al. 2005; Gaffney, Blekhman, and Majewski 2008), however this is a rare situation. While expression quantitative loci (eQTLs) may be used to detect very recent cases of positive selection (Kudaravalli et al. 2009), the use of sequence analysis methods on a larger scale generally relies on assumptions regarding which sequences are involved in regulation (Andolfatto 2005; Haygood et al. 2007; Holloway et al. 2007; Torgerson et al. 2009; Babbitt et al. 2010) and will therefore exclude currently unidentified regulators, such as many distant-acting elements, in spite of their potentially substantial contribution to gene regulation (Visel, Rubin, and Pennacchio 2009). Furthermore, the positively selected changes that are identified using these methods do not necessarily have an effect on gene regulation. A more desirable solution would therefore be to infer adaptive evolution directly from gene expression data, without requiring knowledge of regulatory sequences.

Many methods have been proposed to this end (Fay and Wittkopp 2008), although none has been generally adopted. Firstly, it has been suggested that different theoretical models of gene expression could be further developed to serve as null hypotheses in tests for positive selection (Khaitovich, Paabo, and Weiss 2005; Bedford and Hartl 2009), but the necessary framework is currently lacking. Secondly, in the absence of such quantitative models, Fraser, Moses and Schadt (2010) argued that positively selected eQTLs that affect the same gene should tend to change expression in the same direction and used this qualitative information to estimate the proportion of adaptive expression evolution in yeast. However, in its present form, this approach relies on genetic crosses between strains, making it unsuitable for the study of human evolution.

A third strategy for the detection of positive selection on gene expression has been to list genes that have evolved in a pattern consistent with adaptive evolution, by identifying genes that either have changed their expression in one lineage, while remaining stable in others (Gilad et al. 2006; Blekhman et al. 2008; Blekhman et al. 2010), or that have an unusually high ratio between their between-species and within-species expression variance (Nuzhdin et al. 2004). The underlying assumption is that, although no formal tests are performed, these lists will nonetheless be enriched for positively selected genes, although this will naturally depend on the amount of adaptive evolution that has occurred in the species of interest, as genes under relaxed negative selection might exhibit similar patterns. To reduce the number of false positives, Whitehead and Crawford (2006) performed a similar analysis to that of Nuzhdin et al. (2004) to identify positively selected genes in fish populations, but additionally required that the part of the among-population variation, which could not be explained by genetic distance, should regress with an ecological variable. However, while this approach may be useful under some circumstances, it is restricted to cases where a single environmental parameter is predicted to have a large biological influence.

A more general test for positive selection would however be possible if the ratio of between-species to within-species expression variance could be estimated for neutrally evolving genes. Rifkin, Kim and White (2003) attempted to provide such a cut-off point based on an estimate of the mutational variance in gene expression, i.e., the increase in variance per generation that is caused by new mutations. For model organisms with short generation times, mutation accumulation lines may be used to experimentally estimate this quantity (Denver et al. 2005; Rifkin et al. 2005), but for most species such estimates would be based on speculation. Expressed pseudogenes have been proposed as an alternative neutral standard (Khaitovich et al. 2004b), but it is questionable whether they fulfil the requirement of being non-functional (Svensson, Arvestad, and Lagergren 2006) and they are not common. Instead of estimating the mutational variance, Lemos et al. (2005) therefore based their cut-off point on the mutational heritability, which had previously been determined for various characters (Lynch 1988). However, both methods to obtain a threshold value for positive selection rely on assumptions about population size and other factors, which might explain their different results: while Rifkin, Kim and White (2003) concluded that 25% of the investigated genes in a comparison of *Drosophila melanogaster* and *D. simulans* had undergone positive selection, Lemos et al. (2005) compared the same species without identifying a single positively selected gene.

Thus, although a number of methods have been devised to investigate the contribution of positive selection to gene expression evolution, there is no straightforward procedure for estimating the proportion of adaptive evolution directly from human data. Here, we will outline how the McDonald-Kreitman test, which is frequently used to estimate levels of positive selection in sequence data (McDonald and Kreitman 1991; Fay, Wyckoff, and Wu 2001; Eyre-Walker et al. 2002) can be extended to gene expression data. The resulting test is easy to perform and takes the evolutionary history of each gene into account. We hope that it will serve as a standard tool to make studies of positive selection on gene expression levels comparable across species and experiments.

3.2. Materials and methods

We describe a new test for positive selection on gene expression levels, based on the McDonald-Kreitman (MK) test of positive selection in DNA sequence data. In the MK test the number of synonymous (P_s) and non-synonymous (P_n)

polymorphisms are compared to the numbers of synonymous (D_s) and non-synonymous (D_n) substitutions. Under a neutral model in which mutations at synonymous sites are neutral and mutations at non-synonymous sites are neutral or strongly deleterious, then $D_n/D_s = P_n/P_s$. In contrast, if some non-synonymous mutations are advantageous $D_n/D_s > P_n/P_s$, and if some are slightly deleterious $D_n/D_s < P_n/P_s$ (McDonald and Kreitman 1991).

We can formulate an MK test for gene expression divergence as follows: Let us assume that mutations that affect gene expression are either neutral or strongly deleterious, and that a proportion, *f*, of mutations are neutral. Let us also assume that the evolution of gene expression over a short time follows that of a random walk. If *X(t)* is the expression level at time *t*, then

$$(X(t) - X(0))^2 = \mu ft\sigma^2$$

where μ is the mutation rate and σ^2 is the increase of gene expression per neutral mutation (Khaitovich, Paabo, and Weiss 2005). Hence the squared difference in expression between two individuals, be they of the same or different species, is

$$E(t) = (X_1(t) - X_2(t))^2 = 2\mu ft\sigma^2.$$

The squared difference is expected to increase linearly with time, i.e., the variance in gene expression between individuals is expected to increase linearly with time (Khaitovich, Paabo, and Weiss 2005; Pereira, Waxman, and Eyre-Walker 2009). This is expected to be true over the shorter time scale, but there will eventually be limits as to how high or low expression can evolve (Bedford and Hartl 2009).

Let us split the divergence between the two individuals into three time periods: t_b, the time between the most recent common ancestors in each species for the locus in question; t_{wi}, the expected time to coalescence for two randomly chosen lineages in species *i*; and t_{ci}, the difference between t_{wi} and the time at which all lineages coalesce (Figure 3.1). For a recombining sequence each of these times will be the average across sites within the locus in question. The expected expression divergence between species, E_b, is therefore expected to be equal to $E(t_b)$ and the average expression divergence between pairs of individuals within a species, E_w, is expected to be $E(t_w)$. Let us also define $E_c = E(t_c)$.

We can make a similar argument for sequence divergence: If mutations are strongly deleterious or neutral, then the divergence between individuals are linearly related to the time that separates them:

$$S(t) = 2\mu t$$

so the ratio of the divergence between species, S_b, is expected to equal $S(t_b)$ and the divergence between individuals of the same species, S_w, is expected to be $S(t_w)$. Hence we expect under strict neutrality to have $E_b/E_w = S_b/S_w$. This may be rearranged analogously to the MK test above: $E_b/S_b = E_w/S_w$.

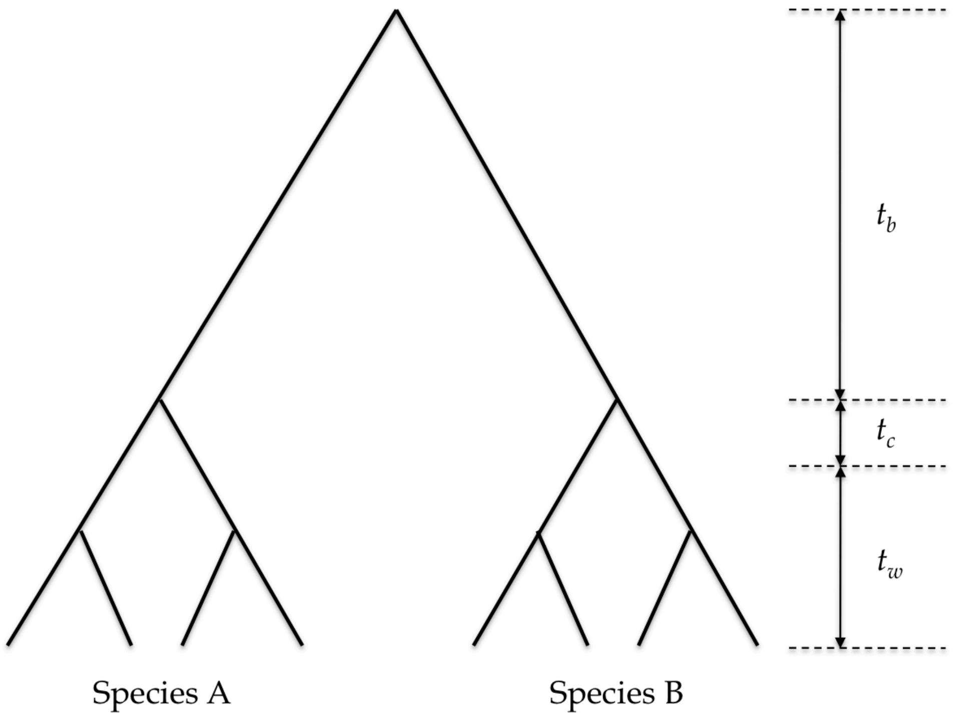


Figure 3.1. Tree illustrating the time between the most recent common ancestors of each species (t_b), the time to expected time to coalescence for two randomly chosen lineages within a given species (t_w) and the difference between t_w and the time at which all lineages coalesce (t_c).

If we assume that some expression mutations are advantageous, then we expect E_b/S_b > E_w/S_w because advantageous mutations contribute more to divergence than they do to polymorphism. If we assume that the advantageous mutations are rare, but strongly selected, then we can ignore their contribution to polymorphism, as an advantageous mutation contributes at most twice the nucleotide diversity of a neutral mutation (Kimura 1969). We then have

$$E_w = 2\mu f t_w \sigma^2$$

and

$$E_b = (2\mu f t_b \sigma^2)/(1 - \alpha_e)$$

where α_e is the proportion of the expression divergence that is driven by positive selection. Hence

$$\alpha_e = 1 - E_w S_b /(E_b S_w).$$

This is analogous to the method for estimating the proportion of substitutions driven by positive selection (Fay, Wyckoff, and Wu 2001; Smith and Eyre-Walker 2002).

We need to estimate the variance in expression (E_b) and between individuals within a species (E_w). This can be accomplished by using a nested analysis of variance (ANOVA), in which the variance between individuals can be divided into error variance, the variance between individuals and the variance between species (Nuzhdin et al. 2004). The variance within individuals, V_w, is an estimate of E_w, and the variance between species, V_b, is an estimate of E_b + E_c. Similarly, we can consider the average divergence between individuals within a species, the nucleotide diversity, π, to be an estimate of S_w, and the average divergence between individuals of different species, dS, to be an estimate of S_b + S_c + S_w. If we assume that t_c is small relative to t_b, we can ignore E_c and S_c and estimate α_e as

$$\alpha_e = 1 - \overline{V}_w (dS - \overline{\pi}) / (V_b \overline{\pi})$$

where the averages are across species. If expression or sequence data is not available for both species then we suggest that we assume that the within-species expression variance and nucleotide diversity in the species with missing data is the same as in the species for which we have data.

3.2.1. Simulations

To evaluate the performance of our method, we simulated expression data according to the model

$$y_{ijk} = \mu_i + I_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the $\log_2$ expression value for species i, individual j and replicate k, I_{ij} represents the variation between individuals and is drawn from a normal distribution with a variance of 10^n, where n is drawn from a uniform distribution between -4 and 0, and ε_{ijk} is the measurement error, drawn from a normal distribution with a variance of 10^{-2}. Thus, our simulations incorporate a relatively high error variance, which can be up to 100 times larger than the within-species variance. The species mean μ_i was drawn from a normal distribution with a variance equal to the within-species variance multiplied by the ratio between the average sequence divergence and heterozygosity, in the case of neutral genes, or the neutral variance multiplied by some factor in the case of positively selected genes. The absolute variance values are not important in this and the following simulations, as it is only the ratio between V_b and V_w, which is evaluated. Sample sizes and other parameters are provided in the description of each simulation below.

The neutral sequences in our simulations were based on human and chimpanzee data (see further details in Section 3.2.2). We generated the sequence divergence and nucleotide diversity for each of these sequences by sampling from a binomial distribution such that the expected value would equal the average number of substitutions between humans and chimpanzees, and the average number of polymorphisms per site in humans. For the polymorphism data we then sampled from the site frequency spectrum obtained from 1000 Genomes data on the human CEU population (The 1000 Genomes Project Consortium et al. 2010) that we used in our analysis (see below). For example, the number of polymorphisms per site in the 1000 Genomes data is 0.0037, so after binomial sampling we might have 37 polymorphisms in a sequence of 10000 bp. We then sampled allele frequencies from the corrected site frequency spectrum (see below) of the 1000 Genomes dataset and used these to calculate the average nucleotide diversity for our simulated locus. In our simulation we are effectively assuming that all loci are similar and that there is free recombination.

In our first simulation, we created datasets of 100, 1000 and 10000 genes, where each dataset contained expression levels from five individuals from each species, with two replicates per individual and where each gene was associated with 10000 bp of neutral sequences. For every set of parameters, we calculated α_e for 1000 datasets. We generated our data under five evolutionary scenarios: no adaptive evolution of gene expression, 10% of genes experiencing adaptive evolution (with an overall α_e of 10% or 50%), or all genes experiencing adaptive evolution (again with α_e set to 10% or 50%). It may appear counter-intuitive to have an α_e of 50%, when only 10% are under positive selection. However, as α_e estimates the proportion of the between-species expression variance that is due to adaptive evolution, it may exceed the proportion of positively selected genes, if these genes have changed their expression to a considerable degree. In similar fashion, the MK test as applied to sequence data estimates the proportion of sites that have been positively selected, but these sites could be evenly distributed among genes or concentrated to only a handful.

Secondly, we generated datasets of 100 genes where more data were available for each gene. Compared to the simulation described above, we either increased the associated neutral sequences to 100000 bp, extended the expression dataset to 100 individuals times 5 replicates per species, or both. Again, we analysed

1000 experiments for each setting and used this to calculate the mean estimated α_e together with a 95% confidence interval.

To test whether our method could identify individual genes under positive selection, we generated two datasets, each with 9000 neutral and 1000 positively selected genes and with an overall α_e of 10% and 50%, respectively. Each gene was associated with 100000 bp of neutrally evolving sequences. We created 95% confidence intervals for our estimates of the overall α_e by bootstrapping per gene (1000 repeats). For each gene we calculated a one-sided 95% confidence interval for the ratio $V_b\overline{\pi}/(\overline{V}_w(dS-\overline{\pi}))$ by bootstrapping the expression data per individual and the sequence data per site.

3.2.2. Data analysis

To estimate V_w and V_b from experimental data, we used a previously published expression dataset from human and chimpanzee lymphoblastoid cell lines, measured on the human-specific Affymetrix U133A microarray (Choy et al. 2008). We masked the data by removing all probes that did not have a single perfect match in the chimpanzee genome. Probe sets with less than four remaining probes were discarded, as smaller probe sets tend to give unreliable results (Lu et al. 2007). Expression values were calculated with the robust microchip average (RMA) method as implemented in Bioconductor (Irizarry et al. 2003a; Irizarry et al. 2003b; Gentleman et al. 2004). For genes with multiple probe sets on the array, we chose a single probe set at random to represent that gene.

The dataset from Choy et al. (2008) included cell lines derived from 5 chimpanzees and 46 humans, of which 13 were of European descent (CEU), 19 of Han Chinese or Japanese descent (CHB/JPT) and 14 of Yoruba descent (YRI). For each human sample, two replicates were available, whereas three or four replicates were available for the chimpanzee samples. To achieve a balanced experimental design, five individuals were randomly chosen from each of the human populations, and two replicates were randomly chosen for each chimpanzee individual. The between-species, within-species and error variance components were then estimated by nested ANOVA of the log-transformed expression values, with the modification that we calculated separate estimates for the human and chimpanzee within-species and error variances. To verify that our variance estimates were unbiased even in cases with unequal variances, we used the same method to analyse simulated expression datasets with known variances. These datasets were based on the same model as described above, but with set variances from Table 3.2.

Estimates of π and dS for each gene were obtained as follows. We extracted the intron coordinates of all human autosomal protein-coding genes in Ensembl release 56 (Flicek et al. 2010), as mammalian introns are essentially neutral (Gaffney and Keightley 2006). To further ensure that we were working with purely neutral sequences, we removed any sequences that were within 50 bp of a splice junction or that overlapped with exons from other genes. We also removed conserved elements identified by the phastCons program (Siepel et al. 2005) by excluding all sequences that featured in the "Primate El" table of the Conservation track for the human genome release hg18 in the UCSC Genome Browser (Rhead et al. 2010). The SNP frequency spectra for these neutral sequences in the CEU, CHB/JPT and YRI populations were taken from low coverage pilot data from the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2010). To correct for the limited power to detect very rare variants, we divided the number of observed SNPs at different frequencies by the power to detect SNPS at that frequency (estimates of detection power were kindly provided by Adam Auton). To estimate the degree of sequence divergence, we downloaded blastz alignments (Schwartz et al. 2003b) of the human and chimpanzee genomes (released hg18 and panTro2, respectively) from the UCSC Genome Browser (International Human Genome Sequencing Consortium et al. 2001; Chimpanzee Sequencing and Analysis Consortium 2005; Rhead et al. 2010). We excluded sites where the human sequence was unknown ("N") or where the chimpanzee sequence had a quality score of 40 or below, as judged from the Quality Scores track in the UCSC Genome Browser.

In our correction of *dS*, we approximated the chimpanzee average heterozygosity by its human counterpart. The true chimpanzee value is likely to be larger, which means that our estimate of *dS* is slightly inflated and will cause our test to be somewhat conservative. To test whether this had a major influence on our results, we repeated the analysis, assuming that the chimpanzee average heterozygosity was 10-fold larger than the one found in humans.

3.3. Results

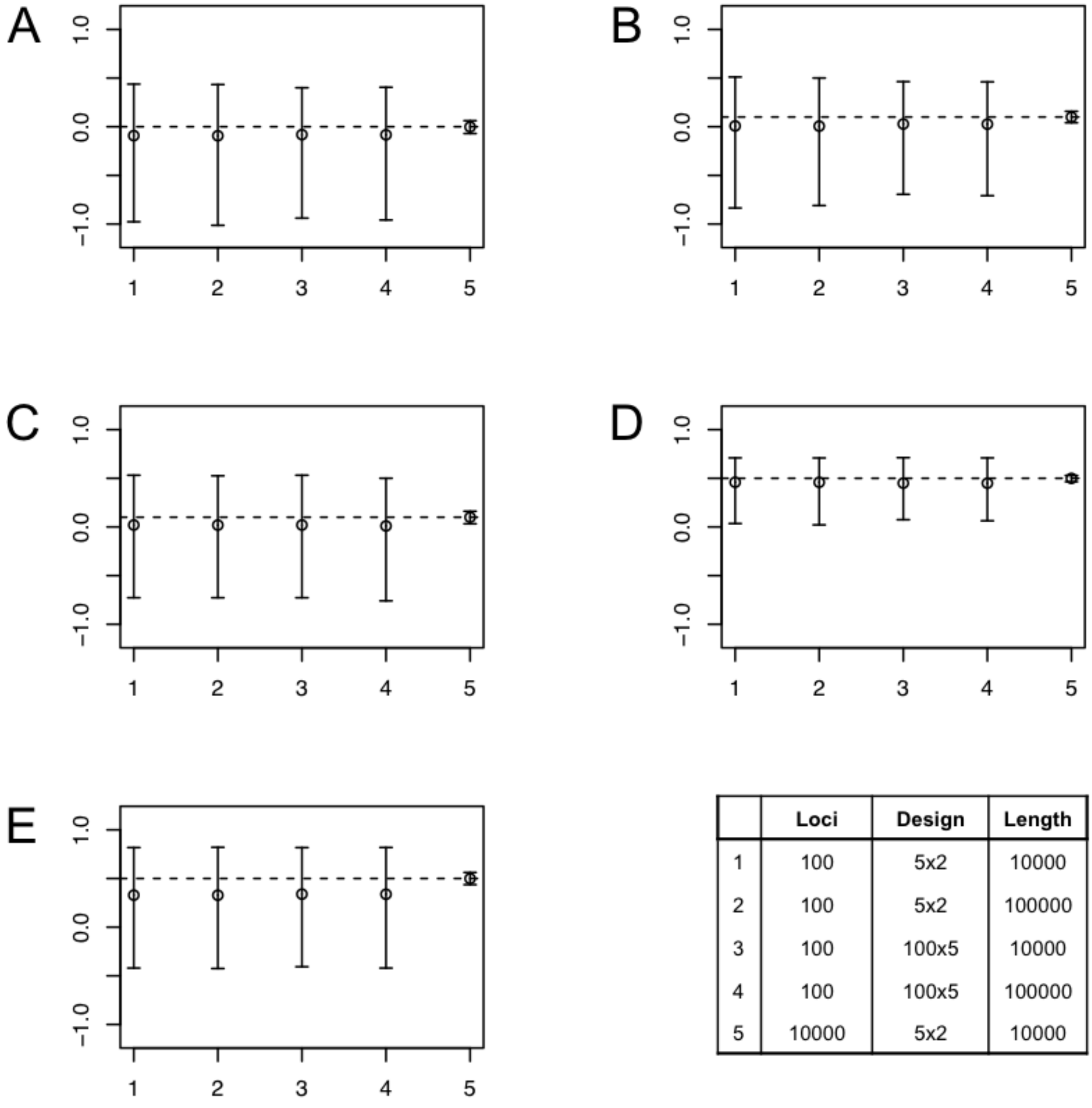
Here we propose a test, analogous to the well-established McDonald-Kreitman test for sequence data (McDonald and Kreitman 1991), of whether expression divergence has been subject to positive selection, and if so, to estimate the proportion of expression divergence that can be attributed to adaptive evolution. The method contrasts the expression divergence between and within species to the level of neutral sequence divergence between and within species. Suitably measured expression divergence is expected to increase linearly with time, just as we expect for neutral sequence evolution.

To investigate the performance of our method we performed a series of simulations. First, we generated expression datasets of 100, 1000 or 10000 genes that had experienced different levels of adaptive evolution (see Section 3.2.1). The datasets were of moderate size, with five individuals per species and two replicates per individuals, and they were relatively noisy, with an error variance that could be up to 100-fold larger than the within-species variance. We further let each gene be associated with 10000 bp of neutrally evolving sequences with the same expected divergence and heterozygosity as in human and chimpanzee intronic sequences. For each set of parameters, we simulated 1000 experiments and calculated the mean estimated α_e, together with the standard error and standard deviation. The simulations confirmed that our method gives an essentially unbiased of α_e when the sample size is above 1000 genes (Table 3.1). For smaller datasets, and especially when the true α_e is small, the estimates are biased downward. This should not be a general problem as datasets of 10000 genes or more are easily obtained using microarrays or RNA sequencing. However, it may make it more difficult to determine α_e for subsets of genes that are of special interest, unless they have been heavily targeted by positive selection. We therefore wanted to see whether we could compensate for reduced sample size by adding more data for the genes in question. However, adding more neutral sequences and/or expression data for more individuals only had a negligible effect on the confidence intervals associated with our estimates (Figure 3.2). This points to that the main obstacle to estimate α_e for small groups of genes is the inherent difficulty of estimating the between-species variance based on only two species.

The extended MK test can also be used to search for individual genes that have been positively selected. We therefore generated two datasets, in the same manner as above, where 10% of the genes had experienced adaptive evolution and where the true α_e was either 10% or 50%. To maximise our power to detect positive selection, we generated expression data for 100 individuals per species, with 5 replicates per individual. For the first dataset, we estimated α_e to be 0.11, with a confidence interval of (0.05, 0.17), which we obtained by bootstrapping

Table 3.1. Mean, standard error of the mean and standard deviation of the estimated α_e, when the true α_e is either 0, 0.1, or 0.5 and when either 100% or 10% of all genes in the sample have experienced adaptive evolution of gene expression.

		α_e = 0	**α_e = 0.1 (100% pos.)**	**α_e = 0.1 (10% pos.)**	**α_e = 0.5 (100% pos.)**	**α_e = 0.5 (10% pos.)**
100 genes	*Mean*	-0.092	0.008	0.020	0.459	0.329
	S.e.	0.012	0.011	0.010	0.005	0.106
	S.d.	0.369	0.341	0.331	0.173	0.336
1000 genes	*Mean*	-0.014	0.089	0.085	0.494	0.477
	S.e.	0.003	0.003	0.003	0.002	0.003
	S.d.	0.101	0.095	0.099	0.053	0.100
10000 genes	*Mean*	-0.002	0.099	0.099	0.499	0.499
	S.e.	0.001	0.001	0.001	0.001	0.001
	S.d.	0.034	0.030	0.032	0.016	0.032



	Loci	Design	Length
1	100	5x2	10000
2	100	5x2	100000
3	100	100x5	10000
4	100	100x5	100000
5	10000	5x2	10000

Figure 3.2. Confidence intervals for estimates of α_e based on datasets with different number of loci, different experimental designs for the expression dataset (table gives number of individuals, followed by number of replicates) and different lengths of the associated neutral sequences. A. All genes evolve neutrally, $\alpha_e = 0$. B. All genes under positive selection, $\alpha_e = 10\%$. C. 10% of genes under positive selection, $\alpha_e = 10\%$. D. All genes under positive selection, $\alpha_e = 50\%$. E. 10% of genes under positive selection, $\alpha_e = 50\%$.

the data per gene. For the second dataset, the estimate of α_e was 0.47, with a confidence interval of (0.41, 0.53). Thus, the test gave accurate estimates of α_e in both cases. For each gene, we calculated the ratio $V_b\bar{\pi}/(\overline{V}_w(dS - \bar{\pi}))$, which is an estimate of E_bS_w/E_wS_b. If this ratio is significantly above one, it suggests that expression of the gene has been positively selected. To test for significance, we created confidence intervals for each value using non-parametric bootstrapping (see Section 3.2.1). For smaller expression datasets, parametric bootstrapping could be considered.

We found that although there was an enrichment of positively selected genes among the genes that were called as significant, 87% of the significant genes were false positives when the true α_e was 10% and 82% when α_e was 50% (Figure 3.3). There was also a high rate of false negatives; the proportion of true positively selected genes that showed up as insignificant was 51% and 35%, respectively. The reason for this becomes clear if we consider the distribution of values, which we would get from an ideal experiment in which all measurements were free of error (in other words, where gene expression could be measured for an infinite number of individuals and where each gene was associated with neutral sequences of infinite length). Figure 3.4 shows that while the distribution is shifted to the right for positively selected genes, there is considerable overlap with neutrally evolving genes, both when α_e = 10% and when α_e = 50%. Thus, positively selected genes will frequently be indistinguishable from neutral genes.

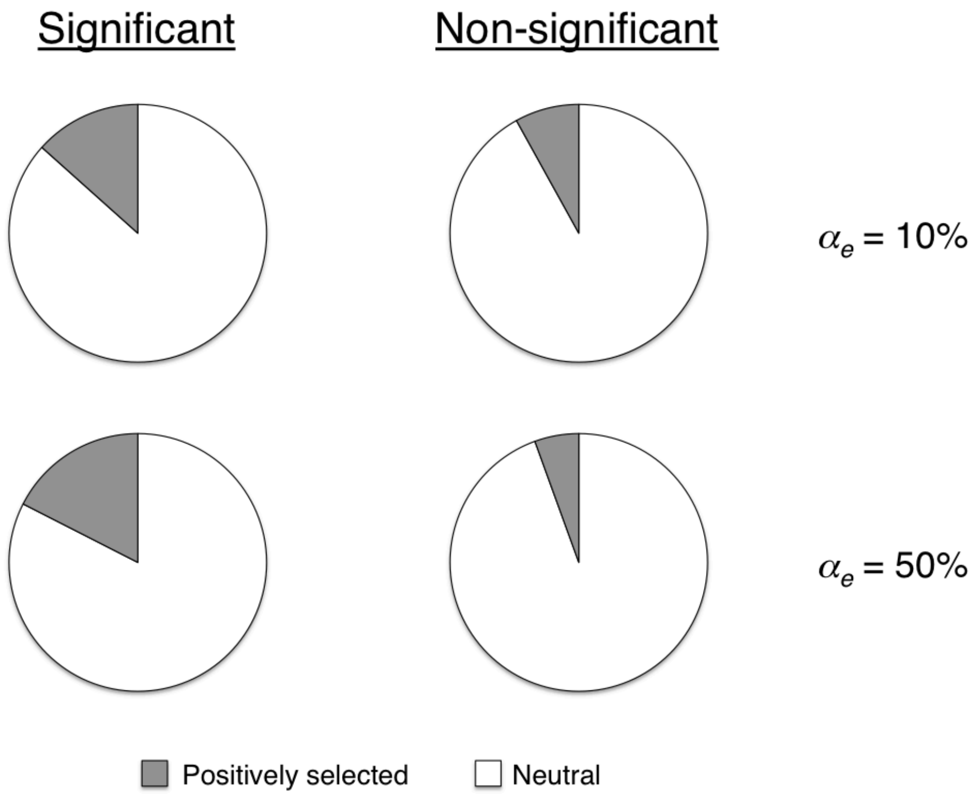


Figure 3.3. Proportion of positively selected and neutral genes where the ratio $V_b\overline{\pi}/(\overline{V}_w(dS-\overline{\pi}))$ was significantly above one, at different values of α_e.

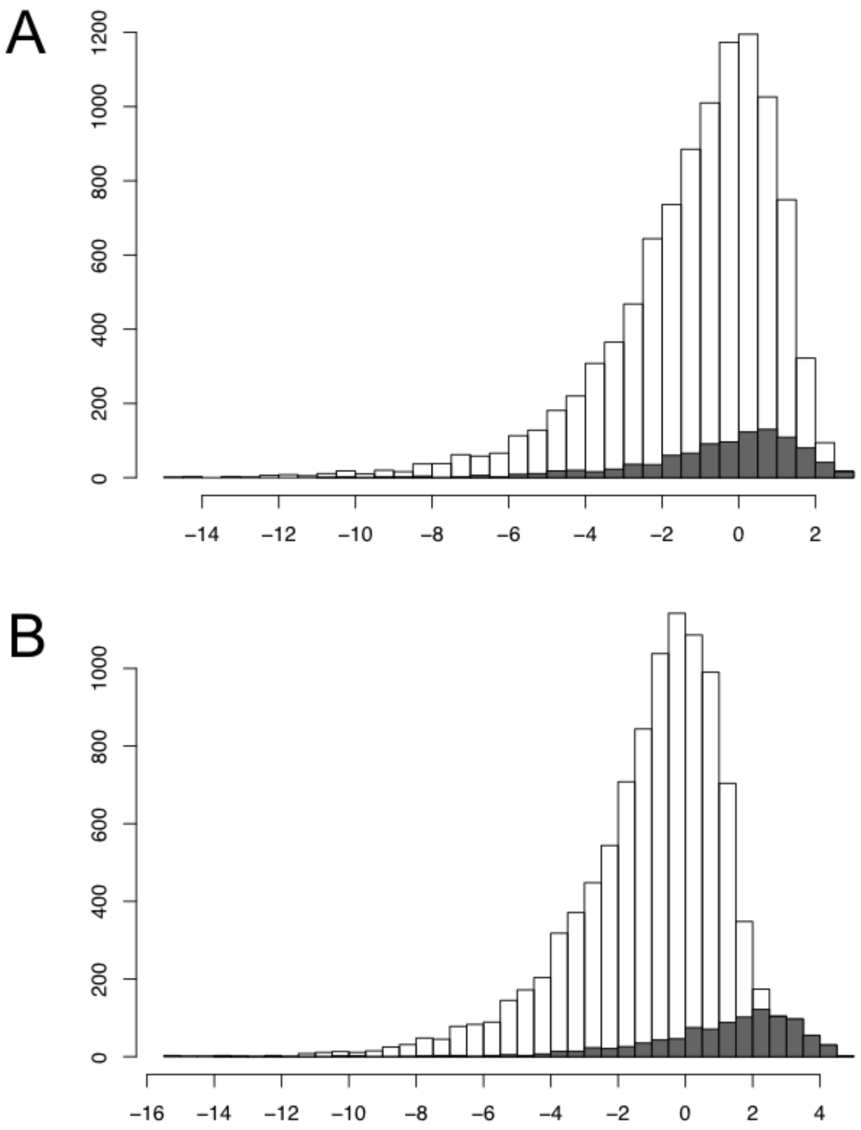


Figure 3.4. A. Distribution of the ratio $V_b\bar{\pi}/\bar{V}_w(dS - \bar{\pi}))$ for neutral genes (white portion of bars) and positively selected genes (grey portion of bars) when α_e = 10%. B. Same for α_e = 50%.

Overall, our simulations indicate that as long as the number of investigated genes in large, a moderately sized expression dataset can be sufficient to estimate the overall proportion of expression divergence that is due to adaptive evolution. However, lists of likely targets of positive selection should be interpreted with caution, even when α_e is significantly above zero. Notably, this applies not only to the method presented here, but to any analysis that estimates the between-species variance from pair-wise comparisons.

Next, we applied our test to experimental data. As the method relies on estimates of the between-species and within-species expression variance, it is necessary to have replicate measurements of each individual, so that the error variance can be removed by nested ANOVA. We therefore chose to apply our method to the dataset of Choy et al. (2008) who measured gene expression in human and chimpanzee lymphoblastoid cell lines. This was also a suitable dataset for two additional reasons: Firstly, humans and chimpanzees are closely related species, where the between-species variation in gene expression has not reached saturation (Khaitovich et al. 2004b). Secondly, cell lines can be grown under more standardised conditions, which may remove much of the environmental variation that could otherwise obscure the results (Somel et al. 2008; Hodgins-Davis and Townsend 2009). It should however be noted that the transformation into cell lines alters the expression of many genes, although most of these changes are minor (Caliskan et al. 2011).

In total, we had expression and sequence data for 7302 genes in chimpanzees and the three human populations CEU, CHB/JPT and YRI (see Section 3.2.2). Nested ANOVA assumes that the experimental design is balanced, that the data is normally distributed and that variances do not differ between groups (Sokal and Rohlf 1995). Before proceeding, we therefore investigated whether the expression data fulfilled these requirements. The design of the original dataset was not balanced, as it contained different numbers of individuals and replicates for the two species. Although methods exist to estimate variance components based on unbalanced designs, they tend to be either cumbersome or give biased results (Sahai and Ojeda 2003). We therefore chose to balance the design by excluding some of the raw data. We therefore randomly selected five individuals and two replicates from chimpanzees and from each of the three human populations represented in the original dataset. After processing the resulting dataset (see Section 3.2.2) we examined the distributions of the standardised log-transformed expression values, which in all cases proved to be approximately normal. However, using single-classification ANOVA to estimate the within-species and error variance for each gene, we found that the variances were not equal: the average human within-species variance was 0.02, while the average chimpanzee within-species variance was 0.05. The difference could be due to the fact that chimpanzees have a higher effective population size than humans do (Eyre-Walker et al. 2002; Hey 2010), or because the sampled chimpanzees were bred in captivity and may therefore belong to different subspecies (Becquet et al. 2007). The mean error variance also differed between humans and chimpanzees, which might reflect variation in the establishment and maintenance of the cell lines. However, unequal variances are only problematic if they introduce bias into the nested ANOVA procedure. To test if this was the case, we simulated datasets of 10000 replicates with differing within-species and error variances, calculated the variance components using nested ANOVA and compared the estimated between-species variance to the set value (Table 3.2). We found that a 10-fold increase in chimpanzee within-species and error variances only had a marginal effect on the between-species variance estimate, which was overestimated by around 3%. In cases with unequal variances our test may therefore give a biased estimate of α_e, but the overall effect is negligible.

We used our method to estimate α_e for the divergence between human and chimpanzee using the polymorphism data from each of the three human populations. In each case, the estimate was negative (Table 3.3). These results relied on the assumption that we could correct *dS* by assuming that the chimpanzee average heterozygosity was equal to the human average heterozygosity. If the true chimpanzee average heterozygosity were larger, this would cause us to underestimate α_e. However, the estimates of α_e remained significantly negative, even when we repeated the analysis assuming a 10-fold higher average heterozygosity in chimpanzees (data not shown).

In principle, α_e should not be able to take on negative values, but if slightly deleterious mutations are segregating in the population, these will cause an increase in the within-species expression variance that is not matched by a

Table 3.2. Nested ANOVA estimates of variance components based on datasets with unequal variances. The variance estimates were averaged across 10000 simulations. The true variances used to generate the data are given in brackets. The first set of simulations were based on the average observed variances in humans and chimpanzee, and the chimpanzee error variance and within-species variances were then increased by a factor of 10.

	Between	Within (human)	Error (human)	Within (chimpanzee)	Error (chimpanzee)
Average	0.061 (0.06)	0.020 (0.02)	0.063 (0.06)	0.051 (0.05)	0.096 (0.10)
Higher V_e	0.061 (0.06)	0.020 (0.02)	0.060 (0.06)	0.046 (0.05)	1.002 (1.00)
Higher V_w	0.062 (0.06)	0.020 (0.02)	0.600 (0.06)	0.492 (0.50)	0.101 (0.10)
Higher V_e and V_w	0.062 (0.06)	0.020 (0.02)	0.060 (0.06)	0.512 (0.50)	0.995 (1.00)

Table 3.3. Estimates of the proportion of between-species expression variance in lymphoblastoid cell lines, which is due to positive selection.

Population	α	95% CI lower limit	95% CI upper limit
CEU	-9.75	-11.39	-8.36
CHB/JPT	-1.07	-1.66	-0.54
YRI	-7.14	-8.34	-6.13

similar increase in the between-species variance. This leads to an underestimation of α_e and in cases with no or very little positive selection, the estimate can be negative (Fay, Wyckoff, and Wu 2001; Charlesworth and Eyre-Walker 2008). From Table 3.3, it seems to be the case that α_e estimate is higher for the CHB/JPT population, which could be taken to indicate that this population either carries fewer slightly deleterious mutations or that it has experienced more positive selection. However, an examination of the average expression variances for the three populations (Table 3.4), suggests that the deviation is instead due to the fact that a markedly higher proportion of the total variance was attributed to error variance. We therefore conclude that although we cannot rule out the possibility of a limited amount of adaptive evolution, there is currently no evidence for adaptive evolution of human gene expression.

3.4. Discussion

We propose an extended McDonald-Kreitman test as a useful tool to evaluate the contribution of positive selection to gene expression evolution in any closely related species pair. As we compare expression data and sequence data from the same genes, we reduce the problem of sampling the neutral standard from a different genomic region to that in which in the regulatory changes are

Table 3.4. Average between-species, within-species and error variances for the three human populations.

	Between species	**Within species**	**Error**
CEU	0.061	0.038	0.044
CHB/JPT	0.058	0.008	0.076
YRI	0.055	0.035	0.045

occurring. We have successfully used our test on simulated data to estimate the proportion of between-species variance that is due to adaptive evolution.

We have also illustrated how individual genes may be tested for signs of positive selection. However, our simulations highlight the inherent difficulty of accurately estimating the between-species variance for single genes and so lists of top-candidate genes should be treated with caution, especially if α_e is low. This applies not only to the test presented here, but to any method where estimates of the between-species variance are based on a single species pair. More precise rankings might be possible if several species or tissues were taken into account, although this introduces the additional problem of non-independence between measurements.

Our analysis of human and chimpanzee lymphoblastoid cell lines gave highly negative estimates of α_e. This is consistent with the segregation of slightly deleterious mutations, which affect expression in humans. These mutations inflate the within-species relative to the between-species expression variance and cause α_e to be underestimated. The same issue is known to affect the original McDonald-Kreitman test and some strategies to correct for this have been developed (Fay, Wyckoff, and Wu 2001; Eyre-Walker and Keightley 2009). However, the effect of slightly deleterious mutations has, to our knowledge, never been incorporated into models of gene expression evolution. Following the method of Eyre-Walker and Keightley (2009), it might be possible to determine the distribution of fitness effects for mutations that affect gene expression and use this to control for the effects of slightly deleterious mutations. It might also be that the negative values of α_e reflect limitations on the evolution of gene expression; expression divergence will not increase forever in a linear fashion because there must be limits to how highly or lowly a gene can be expressed. This seems an unlikely explanation in the current dataset because expression divergence appears to increase linearly across primates (Khaitovich, Paabo, and Weiss 2005).

While it is possible that a modest amount of adaptive evolution is masked by slightly deleterious mutations or limits on how far gene expression can evolve, our results argue against pervasive positive selection along the human lineage since the split from chimpanzees. This is consistent with the results of Kudaravalli et al. (2009), who estimated that 0.1% of human genes have undergone very recent positive selection, as judged from gene expression in lymphoblastoid cell lines from the YRI population. Similarly, Lemos et al. (2005) compared human and chimpanzee expression data from liver and kidney samples, without identifying any likely targets of positive selection. Sequence analyses of potential regulatory regions have given slightly higher estimates: Haygood et al. (2007) found that 4% of human genes had experienced positive selection within 5 kb upstream of the transcript, whereas Torgerson et al. (2009) estimated that 5% of all fixed differences between human and chimpanzee at conserved non-coding sites were adaptive. The extent to which these predicted regulatory changes translate to real differences in gene expression nevertheless remains unclear. In addition to the low levels of positive selection on human protein-coding sequences (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009), we can therefore conclude that there is little evidence for adaptive evolution of gene expression levels in humans.

Does this mean that human evolution has not depended on adaptive changes of gene regulation? It is difficult to answer this question, because a complete understanding of gene expression requires sampling of every cell type at every stage of an organism's lifetime. It could therefore be that we are not seeing signs of positive selection, simply because we are not studying the tissues or developmental time points where adaptations are most likely to occur. For example, perhaps the most famous example of positively selected change in human gene expression, that of lactase persistence (Tishkoff et al. 2007), would not be detected in our analysis, as we are not analysing intestinal samples. On the other hand, lymphoblastoid cell lines are derived from blood cells involved in the body's immune response. Genes with functions in immunity show signs of positive selection on both protein-coding and non-coding sequences (Haygood et al. 2010), so we might expect these cells to be a good starting point in the search for positive selection on gene expression. The lack of signal may therefore be seen as surprising.

It is clear that more extensive expression datasets are needed to settle the question of adaptive regulatory evolution in humans. It will also be of great interest to investigate the role of positive selection in shaping the transcriptomes of many other species. We believe that the framework presented here will aid these investigations by allowing straightforward analysis of gene expression evolution.

4.

The accumulation of gene regulation through time

4.1. Introduction

The upper limit for regulatory complexity in the genome is not known, yet such a limit must exist. Taking alternative splicing as an example, while one might easily imagine a gene that produces 20 splicing isoforms, a gene with 200 000 isoforms appears highly unrealistic, due to the overwhelming amount of regulatory sequences that would be required to avoid aberrant splice variants, which may cause disease (Tazi, Bakkour, and Stamm 2009), and the severe constraints that this would impose on the coding sequence (Parmley et al. 2007). It follows that genes have a maximum capacity for new isoforms and that once this maximum has been reached, the organisational difficulties of adding additional isoforms will completely outweigh the beneficial effects that these isoforms may provide.

The same logic can be extended to the many other mechanisms that control gene expression, such that a single gene can only support a limited level of regulation by transcription factors (TFs), microRNAs (miRNAs) and other processes. While these types of regulation rarely involve coding sequences, they will still be limited by a finite supply of sequences that can house regulatory elements, as well as interference between new and old elements. At saturation, new features can therefore only become fixed if they replace pre-existing ones, or following a gene duplication event.

To what extent have human genes reached their maximum regulatory capacity? This question can be addressed by analysing the level of regulation associated with genes that arose at different evolutionary times. Four potential scenarios are illustrated in Figure 4.1. In the first (Figure 4.1.A), genes are continuously acquiring regulatory features and have not yet reached their maximum capacity. In the second scenario (Figure 4.1.B), older genes are saturated in terms of gene regulation and do not show a further increase in complexity. These two scenarios assume that gene regulatory features accumulated over time. It might however be that different forms of regulation dominate in genes of different age categories (Figure 4.1.C) or that regulation and age are uncorrelated factors (Figure 4.1.D). This last scenario does however appear unlikely, as evolutionary age is known to correlate with aspects of gene architecture, including gene length and intron density (Wolf et al. 2009), as well as with gene expression, such that older genes tend to be expressed in more tissues (Milinkovitch, Helaers, and Tzika 2010) and at higher levels (Wolf et al. 2009) than younger genes.

To distinguish between these scenarios, we have collected information on a variety of regulatory mechanisms operating in the human genome and related this to the evolutionary age of the affected genes. We found that older genes tend to be bound by more TFs, have more conserved upstream sequences, use more alternative transcription start sites (TSSs), produce more alternative splicing isoforms and use more alternative polyadenylation sites. Furthermore, older genes are more likely to be affected by miRNAs, nonsense-mediated

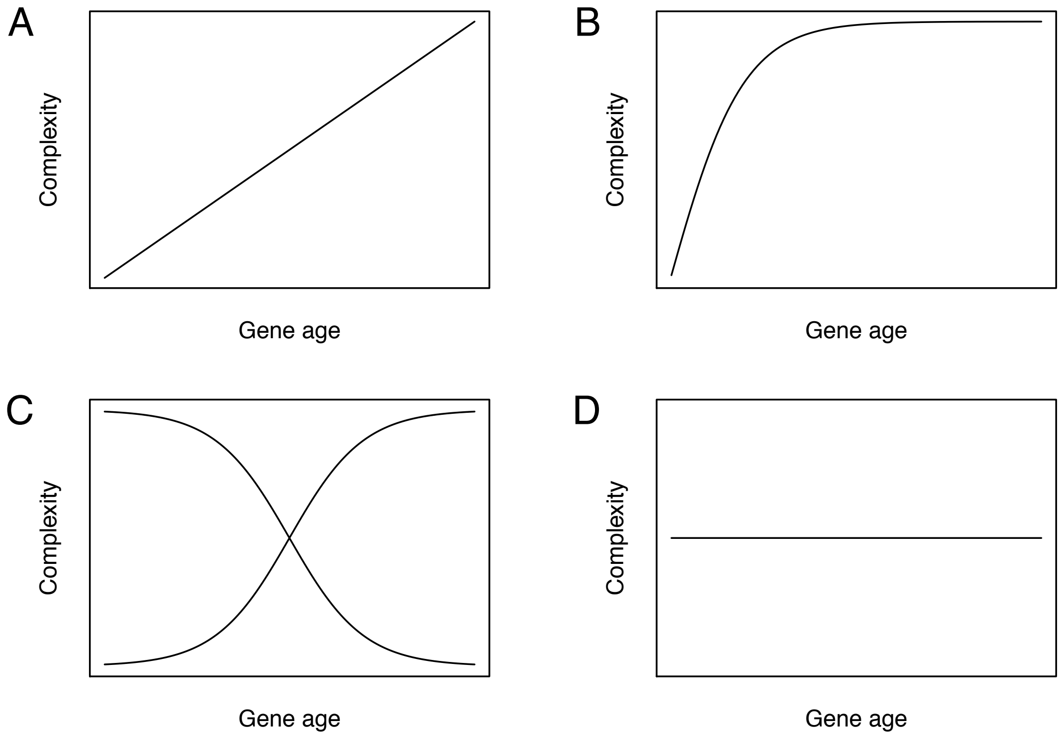


Figure 4.1. Potential relationships between regulatory complexity and gene age. A. Genes continuously increase their regulatory complexity throughout their lifetime. B. Regulatory complexity increases over a time until the maximum capacity is reached. C. Old and young genes tend to be regulated by different regulatory mechanisms. D. Regulatory complexity is independent of gene age.

decay (NMD) and RNA editing. Based on this and the lack of apparent saturation, we draw the conclusion that the majority of human genes could support higher levels of regulation than what we currently observe.

4.2. Materials and Methods

To group human genes according to time of origin, we used the phylostratigraphic classifications established by Domazet-Lošo and Tautz (2010), with the additional requirement that the genes should be represented in release 59 of the Ensembl database (Flicek et al. 2010). We excluded human genes shared by archaea and bacteria from our analysis, as many of the regulatory mechanisms that we consider are specific to eukaryotes. The number of genes for each of the 18 age categories is shown in Table 4.1.

Next, we calculated eight measures of the regulatory complexity of human genes. Firstly, we estimated the complexity of transcriptional regulation for each gene, by counting the number of TFs that bound within 10 kb upstream of the TSS in the human cell line GM12878. This dataset came from ENCODE ChIP-seq experiments performed at the HudsonAlpha Institute (Birney et al. 2007) and was available through the HAIB TFBS track for the human genome (release hg18) in the UCSC Genome Browser (Rhead et al. 2010). The following 20 TFs were analysed: BATF, BCL3, BCL11, EBF, Egr-1, GABP, IRF4, NRSF,

Table 4.1. Human genes classified according to time of origin. Age classifications were taken from Domazet-Lošo and Tautz (2010) and time estimates from Hedges, Dudley and Kumar (2006). In cases where the time estimates did not match the phylogeny (marked with an asterisk), the divergence time was interpolated from those of the surrounding taxa.

Category	*Time of origin (mya)*	*Taxon*	*Number of genes*
1	77.5	Primates	163
2	91	Euarchontoglires	24
3	97.4	Boreoeutheria	84
4	104.7	Eutheria	294
5	176.1	Mammalia	213
6	324.5	Amniota	121
7	361.2	Tetrapoda	73
8	454.6	Euteleostomi	455
9	568.8 *	Craniata	394
10	682.9 *	Olfactores	33
11	797	Chordata	168
12	842	Deuterostomia	52
13	910	Bilateria	728
14	1036	Eumetazoa	1770
15	1237	Metazoa	341
16	1302.5 *	Holozoa	281
17	1368	Opisthokonta	449
18	1628	Eukaryota	4906

p300, PAX5c, PAX5n, Pbx3, POU2F, Sin3A, SP1, SRF, TAF1, TCF12, USF-1 and ZBT33. As a second measure of transcriptional regulation, we calculated the degree of conservation of sequences within 10 kb upstream of the TSS, as the proportion of bases that were identified as conserved within primates by the phastCons program (Siepel et al. 2005). This information was taken from the Conservation track in the UCSC Genome Browser.

Our next three complexity measures were based on the number of transcripts that are generated due to alternative use of TSSs, alternative splicing and alternative polyadenylation, respectively. To distinguish between these mechanisms we evaluated the exon coordinates, downloaded from Ensembl release 59 (Flicek et al. 2010), for all transcripts produced by genes for which we had age information. From the same database, we also downloaded a list of transcripts that were predicted to undergo NMD. Finally, we considered the degree of miRNA regulation based on the experimentally verified miRNA targets in TarBase v5.0.1 (Papadopoulos et al. 2009), as well as the number of sites that undergo RNA editing, taken from the DARNED database (Kiran and Baranov 2010).

We investigated the relationship between gene age and regulatory complexity for each of our eight measures by calculation the Pearson correlation. This analysis was based on the complexity values of each gene, not the averaged values, which are provided for overview in Figure 4.2.

To examine whether the observed correlations persisted even when we corrected for gene function, we first grouped genes into functional categories based on Gene Ontology terms (Ashburner et al. 2000). To this end, we downloaded GOslim terms for "molecular function" and "biological process" from Ensembl release 59 (Flicek et al. 2010). We then repeated the analysis described above for each functional category, while correcting for multiple tests using the Bonferroni method.

4.3. Results and Discussion

In order to assess whether there is a limit to regulatory complexity, we have examined the accumulation of regulatory complexity in human genes by analysing several aspects of gene expression in genes of different evolutionary ages. To group genes according to time of origin, we used the classifications given by Domazet-Lošo and Tautz (2010). These age estimates rely on orthologue identification by BLAST (Altschul et al. 1997), which could mean that some faster-evolving genes escape detection. However, simulations indicate that overall this strategy is reliable (Albà and Castresana 2007). In total, human genes were divided into 18 age categories, with the oldest category including human genes that were present in the eukaryote ancestor and the youngest category consisting of primate-specific genes (Table 4.1). Divergence times for the different categories were taken from the TimeTree database

(Hedges, Dudley, and Kumar 2006), except in cases of contradictory estimates, where instead we interpolated the divergence time from the surrounding categories by taking the average time (Table 4.1). Qualitatively similar results were obtained when we excluded these categories, as well as when we performed the analysis using the category numbers rather than the time estimates. Our conclusions are therefore robust to errors in the estimated divergence times.

We calculated eight measures of regulatory complexity, based on publicly available data (see Section 4.2). To estimate the level of transcriptional regulation, we analysed sequences within 10 kb upstream of the TSS. Firstly, we counted the number of TFs that bind to this region in the human lymphoblastoid cell line GM12878. To exclude non-expressed genes, only genes that were bound by at least one TF were included in the analysis. Figure 4.2.A shows the average number of TFs that bind to genes of different ages, with a clear increase in TF binding for old relative to young genes. As the data is rather noisy and some of the age categories contain relatively few genes (Table 4.1), differences between individual age categories should be interpreted with caution in this and the following graphs. A list of means and standard errors for all investigated regulatory mechanisms is provided in Table 4.2. Analysis confirmed that evolutionary age is significantly correlated with TF binding diversity, such that older genes are typically associated with more types of TFs ($p = 2 \times 10^{-16}$, $r = 0.12$, Pearson correlation, note all correlations are performed on the raw data, not the means shown in the figures). To estimate the magnitude of the increase in diversity, we fitted a linear model to the data,

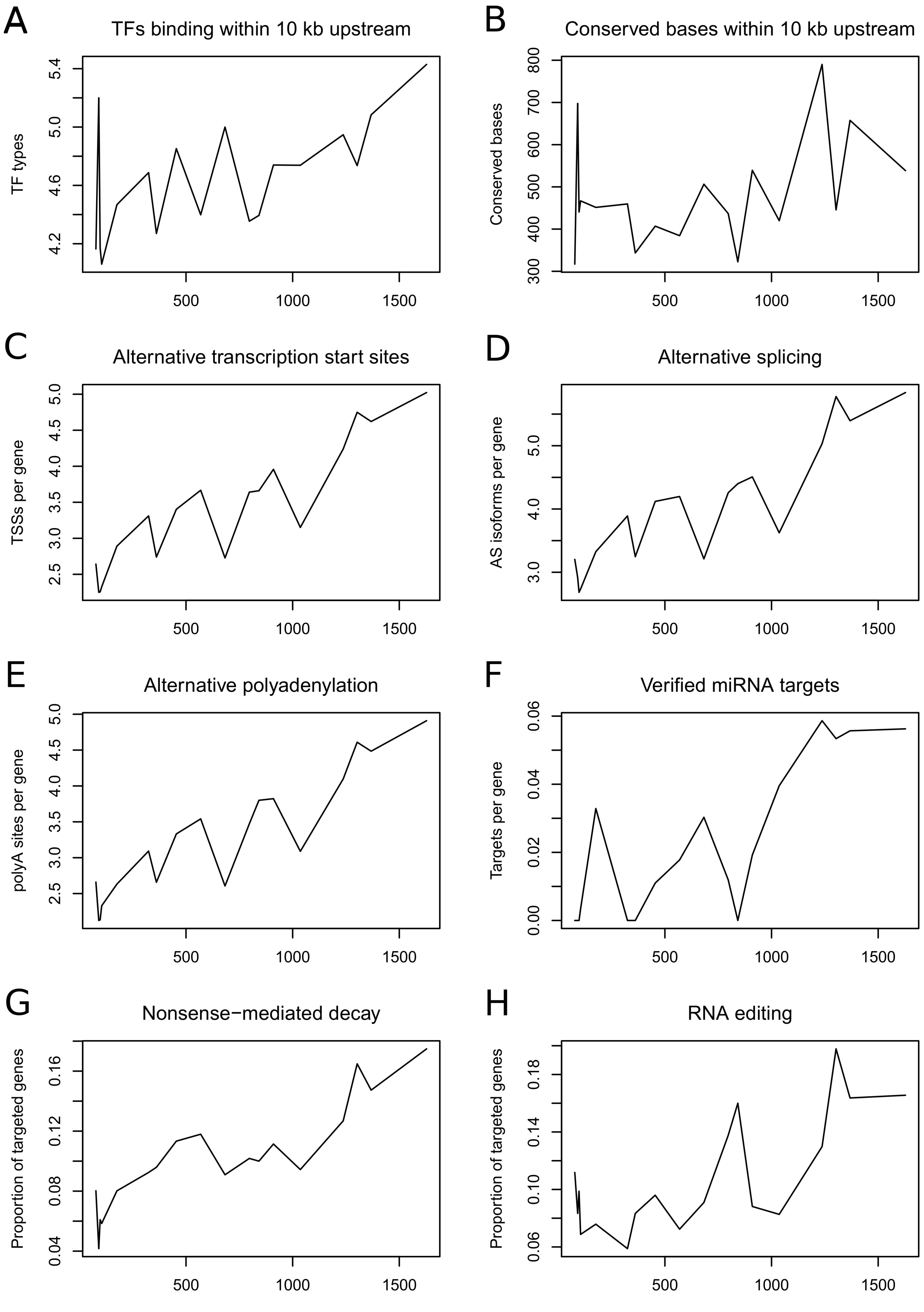

A
TFs binding within 10 kb upstream
TF types
4.2
4.6
5.0
5.4
500
1000
1500
B
Conserved bases within 10 kb upstream
Conserved bases
300
400
500
600
700
800
500
1000
1500
C
Alternative transcription start sites
TSSs per gene
2.5
3.0
3.5
4.0
4.5
5.0
500
1000
1500
D
Alternative splicing
AS isoforms per gene
3.0
4.0
5.0
500
1000
1500
E
Alternative polyadenylation
polyA sites per gene
2.5
3.0
3.5
4.0
4.5
5.0
500
1000
1500
F
Verified miRNA targets
Targets per gene
0.00
0.02
0.04
0.06
500
1000
1500
G
Nonsense−mediated decay
Proportion of targeted genes
0.04
0.08
0.12
0.16
500
1000
1500
H
RNA editing
Proportion of targeted genes
0.06
0.10
0.14
0.18
500
1000
1500

Figure 4.2. Evolution of regulatory complexity. A. Average number of TFs binding within 10 kb upstream of genes. B. Average number of conserved bases within 10 kb of the TSS. C. Average number of TSSs per gene. D. Average number of splicing isoforms per gene. E. Average number of polyadenylation sites per gene. F. Average number of verified miRNA targets per genes. G. Proportion of genes that are targeted by NMD. H. Proportion of genes that are RNA edited. The age of the gene categories in million years is on the x axis. Note that these are averages per age categories, whereas the statistical analysis described in the text was performed on raw data.

Table 4.2. Average values per age category and regulatory mechanism. Standard errors are given in brackets. The category numbers refer to Table 4.1.

Category	*TF binding*	*Conservation*	*Alt.TSSs*	*Alt. splicing*
1	4.16 (0.41)	317 (47)	2.64 (0.23)	3.20 (0.36)
2	5.20 (0.96)	698 (198)	2.25 (0.38)	2.92 (0.54)
3	4.17 (0.46)	441 (69)	2.26 (0.26)	2.68 (0.34)
4	4.06 (0.29)	467 (53)	2.33 (0.19)	2.74 (0.24)
5	4.47 (0.32)	451 (55)	2.89 (0.25)	3.33 (0.36)
6	4.69 (0.46)	460 (59)	3.31 (0.35)	3.89 (0.42)
7	4.27 (0.53)	343 (70)	2.74 (0.33)	3.25 (0.38)
8	4.85 (0.21)	407 (34)	3.40 (0.15)	4.12 (0.19)
9	4.40 (0.22)	384 (29)	3.67 (0.17)	4.20 (0.20)
10	5.00 (0.82)	506 (105)	2.73 (0.31)	3.21 (0.63)
11	4.35 (0.32)	437 (52)	3.64 (0.24)	4.26 (0.30)
12	4.39 (0.57)	323 (68)	3.66 (0.54)	4.40 (0.65)
13	4.74 (0.15)	539 (29)	3.96 (0.16)	4.51 (0.19)
14	4.74 (0.11)	420 (15)	3.15 (0.08)	3.62 (0.10)
15	4.95 (0.21)	790 (51)	4.24 (0.23)	5.03 (0.28)
16	4.74 (0.24)	446 (35)	4.75 (0.28)	5.78 (0.33)
17	5.08 (0.18)	658 (42)	4.62 (0.23)	5.40 (0.26)
18	5.43 (0.05)	538 (11)	5.02 (0.07)	5.84 (0.08)

Category	*Alt. polyA*	*miRNAs*	*NMD*	*RNA editing*
1	2.66 (0.24)	0 (0)	0.080 (0.021)	0.112 (0.025)
2	2.13 (0.37)	0 (0)	0.042 (0.042)	0.083 (0.058)
3	2.13 (0.27)	0 (0)	0.061 (0.027)	0.099 (0.033)
4	2.33 (0.17)	0.003 (0.003)	0.058 (0.014)	0.069 (0.015)
5	2.63 (0.24)	0.033 (0.015)	0.080 (0.014)	0.076 (0.018)
6	3.09 (0.32)	0 (0)	0.092 (0.027)	0.059 (0.022)
7	2.66 (0.32)	0 (0)	0.096 (0.035)	0.083 (0.033)
8	3.33 (0.15)	0.011 (0.007)	0.113 (0.015)	0.096 (0.014)
9	3.54 (0.16)	0.018 (0.007)	0.118 (0.016)	0.072 (0.013)
10	2.61 (0.46)	0.030 (0.030)	0.091 (0.051)	0.091 (0.051)
11	3.47 (0.24)	0.012 (0.008)	0.102 (0.023)	0.138 (0.027)
12	3.80 (0.57)	0 (0)	0.100 (0.043)	0.160 (0.052)
13	3.82 (0.16)	0.020 (0.007)	0.111 (0.012)	0.088 (0.011)
14	3.09 (0.08)	0.040 (0.008)	0.094 (0.007)	0.083 (0.007)
15	4.10 (0.23)	0.059 (0.016)	0.127 (0.018)	0.130 (0.018)
16	4.61 (0.27)	0.053 (0.016)	0.165 (0.022)	0.198 (0.024)
17	4.48 (0.22)	0.056 (0.012)	0.147 (0.017)	0.164 (0.018)
18	4.91 (0.07)	0.056 (0.004)	0.175 (0.005)	0.166 (0.005)

which showed that genes in the youngest category are typically bound by 4.1 TFs, whereas the oldest genes are bound by 5.4 TFs (Table 4.3).

Secondly, we assessed the level of conservation of upstream sequences, by counting the number of bases within 10 kb of the TSS that were identified as conserved among primates by the phastCons program (Siepel et al. 2005). Again, we found a significant correlation with age, where older genes tend to have more conserved upstream sequences than younger genes ($p = 1 \times 10^{-10}$, $r = 0.06$, such that the upstream regions of the oldest genes contain almost 40% more conserved bases, compared to younger genes (Table 4.3). Thus, both TF binding and upstream conservation show a highly significant correlation with evolutionary age.

We then considered complexity in terms of alternative isoforms generated by differential use of TSSs (Figure 4.2.C), splice sites (Figure 4.2.D) and polyadenylation sites (Figure 4.2.E). For each of these mechanisms we found significant, positive correlations with gene age (alternative TSSs: $p < 2 \times 10^{-16}$, $r = 0.18$; alternative splicing: $p < 2 \times 10^{-16}$, $r = 0.18$; alternative polyadenylation: $p < 2 \times 10^{-16}$, $r = 0.18$). Compared to the youngest genes in our dataset, the oldest genes have gained 2.57 alternative start sites, 2.96 alternative splicing isoforms and 2.54 alternative polyadenylation sites (Table 4.3). This is consistent with the recent results of Roux and Robinson-Rechavi (2011), who also showed an accumulation in alternative splicing isoforms over time.

Table 4.3. Differences in complexity between the youngest and oldest age categories. The estimates were obtained by fitting a linear model to the data.

Category	*Youngest genes (primates)*	*Oldest genes (eukaryotes)*	*Ratio*
TF binding sites	4.12	5.38	1.31
Conserved bases upstream	396	547	1.38
Transcription start sites	2.35	4.92	2.09
Splicing isoforms	2.76	5.72	2.07
Polyadenylation sites	2.26	4.80	2.12
miRNA sites	0.0017	0.0573	33.7
NMD proportion	0.058	0.168	2.90
RNA editing proportion	0.052	0.161	3.10

Notably, the patterns for these last three mechanisms are highly similar. This is to be expected, since they are frequently coupled (for example, a gene with two potential last exons will need to accommodate at least two polyadenylation sites and produce at least two alternative splicing isoforms). However, the similarity could also be a sign of ascertainment bias: if some genes have been more intensely studied, we might expect more alternative isoforms, of all three types, to have been identified in these genes. To exclude biased identification as an explanation, we analysed cases where one of the three mechanisms acts independently of the others. Thus, we identified alternative TSSs and poly-adenylation sites that occur within a single exon and therefore cannot be directly associated with an increase in splicing. We also counted the number of alternative coding sequences generated from each gene, as this is not coupled directly to changes in UTR structure. As seen in Figure 4.3, the three resulting distributions of alternative events are distinct from each other, as we would expect for unbiased data. Remarkably, the correlations between complexity and age remained positive and significant (alternative TSSs: $p = 1 \times 10^{-5}$, $r = 0.04$; alternative splicing: $p < 2 \times 10^{-16}$, $r = 0.16$; alternative polyadenylation: $p = 3 \times 10^{-5}$, $r = 0.05$), even though this analysis was performed on limited datasets (number of genes with multiple isoforms of a given type was 2655, 6547 and 3028, respectively).

Next, we investigated the distribution of verified miRNA binding sites across the 18 categories (Figure 4.2.F) and found that older genes are enriched in this type of regulation ($p < 5 \times 10^{-11}$, $r = 0.06$), with the average number of miRNA

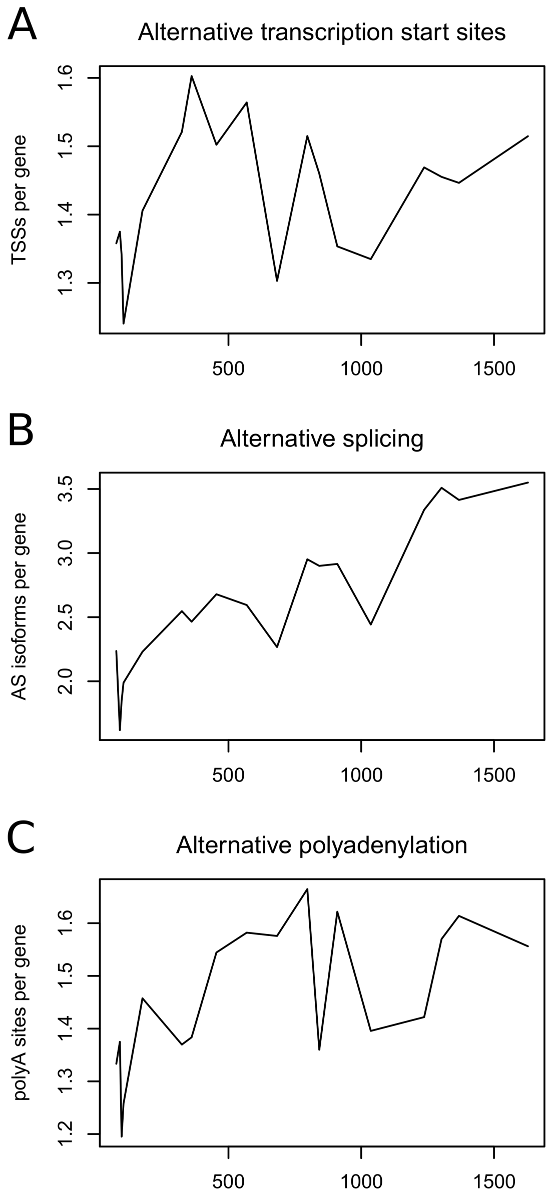


Figure 4.3. Alternative isoforms arising from independent mechanisms. The average number of isoforms that are due to TSSs within a single exon (A), splicing of coding sequences (B) and polyadenylation sites (C) within a single exon for genes of different ages. The x axis shows gene age in million years.

targets per gene increasing more than 30-fold from 0.0017 to 0.0573. We also found significant, positive correlations between gene age and the likelihood for genes to be targeted by the less common regulatory mechanisms NMD ($p < 2 \times 10^{-16}$, r = 0.10) and RNA editing ($p < 2 \times 10^{-16}$, r = 0.10). For both of these mechanisms, around 5% of the youngest genes are affected, whereas the proportion among the oldest genes is three times larger.

In theory, the results described above could be influenced by an uneven distribution of gene functions among the age categories. If "early" genes predominantly are of a functional type that requires a certain level or mode of regulation, whereas "late" genes have other functions and therefore different regulatory needs, then we might see a superficial correlation between age and regulatory complexity. To test this possibility, we further divided our dataset according to Gene Ontology terms (Ashburner et al. 2000) and repeated the analysis for a number of functional categories (see Section 4.2). In the vast majority of cases, the correlations between complexity and gene age remained positive even for functional subsets of genes (Appendix A), showing that the positive correlations that we obtained for the full dataset are not due to functional bias.

Based on these results, we can exclude the two last possibilities shown in Figure 4.1, (no increase in complexity with time and certain types of complexity being associated with particular time periods) as all forms of regulatory complexity investigated here show a significant increase over time. We are therefore left to determine whether the oldest human genes have reached regulatory saturation, i.e., whether the pace at which genes accumulate new features has slowed down for older genes. To do this, we performed a regression analysis involving a quadratic term. However, in all eight cases, this term was either not significant or it indicated that the pace is higher for older genes. Thus, we have not found any evidence to suggest that human genes have reached saturation or that the rate with which they increase in regulatory complexity slows down over time. This partially contradicts the results of Roux and Robinson-Rechavi (2011), who showed that for non-duplicated genes, the rate of splicing isoform acquisition decreases as genes grow older. For duplicated genes, they found a linear relationship, consistent with our results, but argued that the linearity may be due to biased duplication.

Wolf et al. (2009) recently showed that the ratio of the rate of non-synonymous to synonymous substitution (*dN/dS*) decreases with gene age, indicating that older genes are under stronger constraint. Rather than being the cause of the observed correlations, the decrease in *dN/dS* might be a consequence of the increase in the complexity of gene regulation, as regulatory elements within protein-coding sequences would be expected to constrain both non-synonymous and synonymous sites, but might affect non-synonymous sites more, as they also need to encode the protein sequence. However, even if due to some currently unknown mechanism the increase in constraint with evolutionary age was the cause of the increase in complexity, this does not alter the fact that regulatory complexity accumulates through time.

To summarise, we have demonstrated that older genes tend to be bound by more TFs, have more conserved upstream sequences, use more alternative TSSs, produce more alternative splicing isoforms, use more alternative poly-adenylation sites, contain more miRNA binding sites and that they are also more likely targets of NMD and RNA editing. The differences between young and old genes are of such a magnitude that they could have a substantial impact on gene function. Furthermore, we have shown that the accumulation of new regulatory features has been an ongoing process over the past 1.5 billion years of eukaryote evolution. Therefore, although human gene regulation is a highly elaborate process, it has not reached its peak and human genes would thus be able to become even more complex in the future.

5.

Ultraconserved elements in the *Drosophila Hox* gene *Ultrabithorax*

5.1. Introduction

Sequence conservation is frequently used to predict functional genomic elements (Kellis et al. 2003; Stark et al. 2007). At the extreme end of the conservation spectrum lie the so-called ultraconserved elements (UCEs), originally defined as orthologous regions of at least 200 bp that are identical in the human, mouse and rat genomes (Bejerano et al. 2004). Around 500 UCEs are shared between these species (Bejerano et al. 2004), whose last common ancestor lived 91 million years ago (Hedges, Dudley, and Kumar 2006). Some of these elements function as developmental enhancers (Pennacchio et al. 2006), whereas others constitute non-coding RNAs (Calin et al. 2007) or are associated with alternative splicing (Bejerano et al. 2004; Lareau et al. 2007; Ni et al. 2007). However, although functions have been identified for many UCEs, no currently known molecular mechanism can fully explain the preservation of these sequences over long evolutionary distances, as most functional genomic elements, including enhancers, are typically not conserved to this degree (Visel et al. 2008).

In theory, apparent ultraconservation can be due to regionally lowered mutation rates rather than intense selection, but the distribution of single nucleotide polymorphisms (SNPs) within UCEs speaks against this, as the allele frequency spectrum is shifted towards rarer alleles in UCEs, indicating that most derived alleles are removed by selection before reaching higher frequencies (Drake et al. 2006; Katzman et al. 2007). Negative selection acting on UCEs has been estimated to be three times stronger than that acting on nonsynonymous sites (Katzman et al. 2007) and elements of at least 100 bp that are perfectly conserved between primates and dog are more than 300 times less likely to have been lost in rodents, compared to neutral DNA (McLean and Bejerano 2008).

Considering that UCEs are under powerful negative selection, it might be expected that mutations within these elements would dramatically compromise genome function. Paradoxically, simultaneous deletion of four UCEs from the mouse genome did not produce any major phenotypic changes, even though the deleted elements were verified enhancers located close to genes for which inactivation or expression changes had previously been shown to lead to distinct phenotypes (Ahituv et al. 2007). While these results might be explained by insufficient testing for phenotypes or redundancy of regulatory elements in the genome (in which case, however, the need for ultraconservation is not obvious), they suggest that UCEs play a subtler role than was initially thought. It has also been observed that human individuals can be homozygous for derived SNP alleles within UCEs and still be phenotypically normal (Drake et al. 2006; Chen, Wang, and Cohen 2007). In light of this conflicting evidence, further study of the processes underlying ultraconservation is clearly warranted.

Although most analyses of ultraconservation have focussed on the genomes of mammals and other vertebrates, the phenomenon is not limited to this taxon. It has been known for some time that the genome of the fruit fly *Drosophila melanogaster* harbours many constrained non-coding sequences (Bergman and Kreitman 2001) and a comparison between *D. melanogaster* and *D. pseudoobscura* established the presence of over 23000 ultraconserved sequences of at least 50 bp, some of which are conserved even in the more distantly related mosquito *Anopheles gambiae* (Glazov et al. 2005). The lineages leading to *D. melanogaster* and *D. pseudoobscura* split 54 million years ago, while the split between *Drosophila* and *Anopheles* is estimated to have occurred around 470 million years ago, which corresponds to the time of the split between terrestrial vertebrates and bony fishes (Hedges, Dudley, and Kumar 2006). Other studies have identified UCEs shared by *D. melanogaster* and *D. virilis* (Papatsenko et al. 2006), as well as highly conserved (but not ultraconserved) elements in the genomes of *D. melanogaster*, *D. yakuba*, *D. pseudoobscura* and *A. gambiae* (Siepel et al. 2005). A common theme emerging from these studies is that non-coding UCEs are frequently associated with developmental genes, which is consistent with the distribution of UCEs in vertebrates (Bejerano et al. 2004).

When comparative methods are used to predict functional genomic elements, the choice of study organisms influences what type of elements can be detected (Boffelli, Nobrega, and Rubin 2004). Comparisons between distantly related organisms allow the identification of highly constrained sequences, but exclude any elements that originated after the chosen organisms diverged from each other. Conversely, the use of closely related species allows identification of more recent elements, but is complicated by the higher overall sequence similarity. In their comparison of the *D. melanogaster* and *D. pseudoobscura* genomes, Glazov et al. (2005) narrowed down their list of 23000 UCEs by focussing on longer elements and those UCEs that are shared with *A. gambiae*.

Here, we use a complementary approach to that of Glazov et al. (2005) and identify UCEs that are shared by twelve sequenced *Drosophila* genomes. The most distantly related species in our dataset diverged 62 million years ago (Hedges, Dudley, and Kumar 2006), meaning that we are able to detect both ancient and more recently derived UCEs. Furthermore, the use of multiple species improves resolution and we can therefore also identify shorter elements. Following our genome-wide analysis, we perform an in-depth survey of elements located within the *Ultrabithorax* (*Ubx*) gene, which is enriched in UCEs. *Ubx* belongs to the *Hox* genes, a family of key developmental regulators that are present throughout the animal kingdom and that are also associated with UCEs in vertebrates (Sabarinadh et al. 2004; Sandelin et al. 2004; Lampe et al. 2008; Lin, Ma, and Nei 2008). In *Drosophila*, alternative splicing of the *Ubx* gene yields functionally distinct isoforms by differential inclusion of the “mI” and “mII” exons (Mann and Hogness 1990; Subramaniam, Bomze, and Lopez 1994; Gebelein et al. 2002; Reed et al. 2010). It has previously been shown that both exons are highly conserved at the nucleotide level (Bomze and Lopez 1994). Here, we show that the mI exon is embedded within a UCE and go on to test how synonymous mutations within this short exon affects the *Ubx* alternative splicing pattern, thereby providing a link between the general phenomenon of ultraconservation and the biologically relevant regulation of a gene with a well-established role in development.

5.2. Materials and methods

The genome assemblies (as of June 2008) for the following species were downloaded from the UCSC Genome Browser (Rhead et al. 2010): *D. ananassae, D. erecta, D. grimshawi, D. melanogaster, D. mojavensis, D. persimilis, D. pseudoobscura, D. sechellia, D. simulans, D. virilis, D. willistoni* and *D. yakuba*. The number of UCEs shared by these genomes had previously been estimated using an alignment-free method (Warnefors 2007): A sliding window of 50 bp, moved 1 bp at a time, was applied along each genome to divide it into overlapping fragments and identical fragments shared by all twelve genomes were identified as putatively ultraconserved. Position coordinates were then obtained by a BLAT search against the *D. melanogaster* genome (Kent 2002) and any fragments with overlapping coordinates were reassembled into longer sequences. Here, this initial analysis was further refined by validating each potential UCE by visual inspection of the corresponding multiple alignment in the UCSC Genome Browser. Notably, the algorithm used to generate this alignment is not set to maximise UCE size and some elements may therefore appear shorter than their actual length.

Genome coordinates for the Glazov dataset (Glazov et al. 2005) were updated from the dm1 to the dm3 *D. melanogaster* genome assembly using the liftOver tool from the UCSC Genome Browser (Rhead et al. 2010). Overlap between UCEs and protein-coding and non-coding genes downloaded from Ensembl release 59 (Flicek et al. 2010) was determined using intersectBed (Quinlan and Hall 2010). Enrichment of Gene Ontology terms (Ashburner et al. 2000) was calculated using the goseq R package (Young et al. 2010) while correcting for the lengths of the exonic, intronic and intergenic regions belonging to each gene. We used a false discovery rate (FDR) of 5% as our cut-off to consider terms significant.

A multiple alignment of the genomic *Ubx* sequences (not including the UTRs) from the twelve species was prepared with the MultiPipMaker tool (Schwartz et al. 2003a). All nucleotide stretches with complete identity across the twelve species were counted and classified according to size. As this approach is alignment-based, there are some discrepancies in UCE length between these results and those from the genome-wide analysis. The decay pattern of identical blocks under neutral evolution was simulated by randomly shuffling the positions of the *Ubx* alignment, while keeping the number of gaps intact. This was repeated ten times and each of the resulting alignments was analysed in terms of completely conserved sequence stretches.

The Ubx.4 plasmid was a gift from Manuel de la Mata. Mutations were introduced into the Ubx.4 plasmid by splicing PCR-driven overlap extension

(Heckman and Pease 2007). The PCR fragment was cloned into the pGEM-T Easy vector (Promega) and the plasmid was sequentially digested with AflII and PmlI to release a 255 nt fragment, which was cloned into Ubx.4 to create the derivative construct Ubx.4_mutA.

5.3. Results

Here, we have identified UCEs as sequences of at least 50 bp that are perfectly conserved in twelve *Drosophila* species, whose last common ancestor lived approximately 62 million years ago (Hedges, Dudley, and Kumar 2006). Putative UCEs were identified by extracting sequence fragments of exactly 50 nt that were present in all twelve *Drosophila* genomes, followed by assembly of the fragments into longer sequences and visual inspection to confirm their status as UCEs (see Methods). Using this approach, we identified 1557 *Drosophila* UCEs.

The majority (59%) of the UCEs identified here are located in intergenic sequences (Figure 5.1), but many are located within the introns (25%) and exons (12%). Only a small proportion (4%) overlaps known non-coding RNAs, which indicates that the identification of UCEs shared between multiple species can lead to the identification of novel functional elements. In sharp contrast to this, 74% of previously identified UCEs shared between *D. melanogaster* and its distant relative *A. gambiae* (Glazov et al. 2005) overlap with known RNAs,

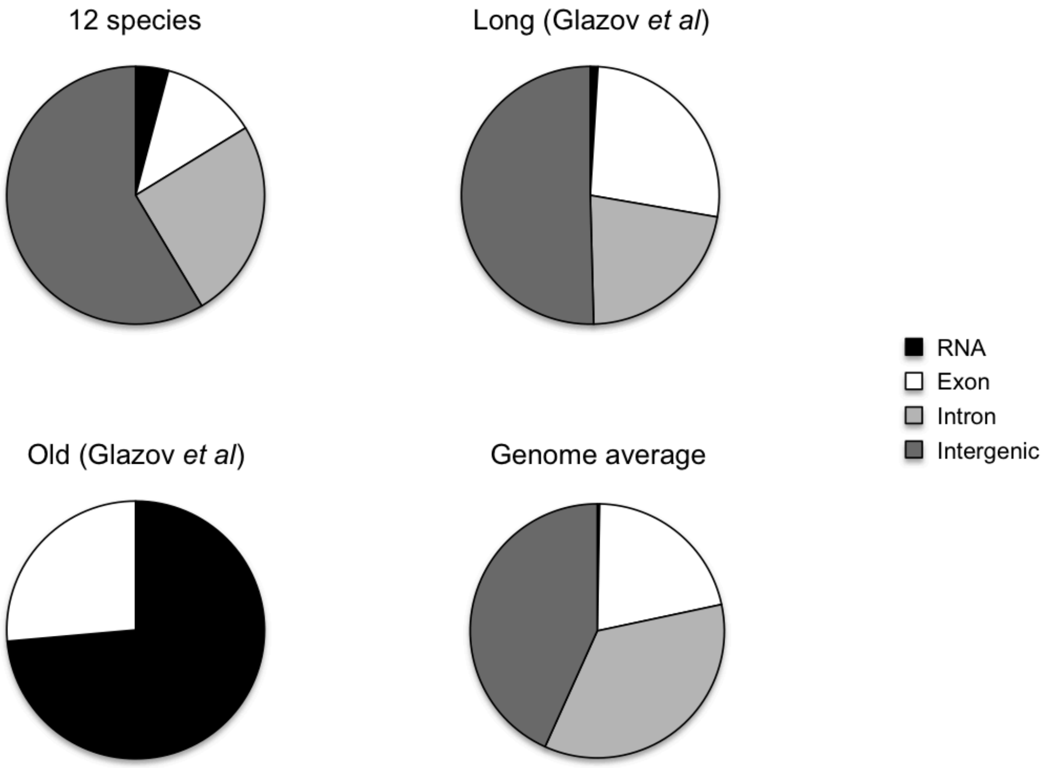


Figure 5.1. Genomic distribution of the UCEs identified here and those identified in Glazov et al. (2005). UCEs were categorised according to whether they overlapped with non-coding RNAs (black), exons in protein-coding genes (white), introns in protein-coding genes (light grey) or if they did not overlap with any known transcripts (dark grey). The charts represent the complete set of UCEs identified here ("12 species"), UCEs of at least 80 bp from the Glazov dataset ("Long"), UCEs from the Glazov dataset that were shared between *D. melanogaster* and the distantly related *A. gambiae* ("Old") and the proportion of bases belonging to the different categories in the dm3 genome assembly, excluding sequences on "chromosome Uextra".

suggesting that the inclusion of a distant species is of limited value in terms of detecting regions of unexpected conservation that would be candidates for further functional analysis. The distribution of the UCEs identified here is relatively similar to that of the longest UCEs (at least 100 bp) shared by *D. melanogaster* and *D. pseudoobscura* (Glazov et al. 2005), with the main difference being that the Glazov UCEs are more likely to overlap exons. This is not surprising as these UCEs were identified in a two-species comparison and therefore are more likely to occur by chance in protein-coding regions, even if there is no strong selection on synonymous sites. While the UCE dataset presented here includes a smaller proportion of exonic UCEs, these elements are more likely to represent true cases of extreme conservation.

To test whether ultraconservation is more common for certain types of genes, we calculated the enrichment of Gene Ontology terms describing molecular function (Ashburner et al. 2000) for genes with UCEs. Intergenic UCEs were assigned to the nearest neighbouring gene, irrespective of distance. Because longer genes and genes without close neighbours would be more likely to be assigned UCEs, we performed the enrichment analysis using the goseq package (Young et al. 2010), which takes length bias into account. At a false discovery rate (FDR) of 5%, no terms were significantly enriched for genes with exonic UCEs, possibly due to the small sample size. Among the highest-ranking terms were functions related to ion transport, which is consistent with the results obtained for the Glazov dataset. Genes with intronic and intergenic UCEs were significantly enriched for terms related to transcription factor activity, which is consistent with previous results in both insects and vertebrates (Bejerano et al. 2004; Glazov et al. 2005). However, although these results suggest that many UCEs might be important for transcriptional regulation, it is in general not possible to pinpoint a hypothetical function, which could then be tested experimentally. To allow a more detailed functional analysis, we therefore turned to UCEs located within the *Hox* gene *Ubx*. The first *Drosophila* mutant related to this gene was identified in 1915 and continued study has resulted in a rich collection of information on *Ubx* regulation (Maeda and Karch 2009). Furthermore, *Ubx* is significantly enriched for UCEs, as it contains 13 UCEs, whereas the expected number for a sequence of similar length (~75 kb) would be 0.89 ($p < 10^{-10}$, Poisson).

The *Ubx*-UCEs are unique sequences that cannot be found elsewhere in the genome, showing that they are not made up of repetitive sequences or transposable elements. We also tried to trace the origin of the elements by examining the multi-species Multiz alignment provided through the UCSC Genome Browser (Blanchette et al. 2004). Apart from the twelve *Drosophila* species, this alignment also includes *Anopheles gambiae*, the flour beetle *Tribolium castaneum* and the honeybee *Apis mellifera*. However, no sequences from these species matched the *Ubx*-UCEs, suggesting that the *Ubx*-UCEs originated after the initial dipteran radiation. As more insect genomes become available, it would be of interest to search for *Ubx*-UCEs in intermediate dipterans.

Although the extreme conservation of the *Ubx*-UCEs is unlikely to be compatible with neutral evolution, the elements themselves do not necessarily constitute discrete functional units. It might be the case that the entire *Ubx* locus is conserved to a high degree and that the *Ubx*-UCEs merely represent random aggregations of perfectly conserved positions. Indeed, when we aligned the genomic *Ubx* sequences from the the twelve *Drosophila* species using MultiPipMaker (Schwartz et al. 2003a), we found a large number of shorter regions of perfect conservation. Distinguishing between these two possibilities is crucial if we want to use ultraconservation as a means to understand genome function, as the mechanisms causing high levels of conservation over large genomic distances would presumably be different from those causing ultraconservation within a discrete region. To test whether the *Ubx*-UCEs are conserved units in their own right or whether they are a product of high local conservation levels, we simulated ten sequences of the same length as *Ubx* with the same number of indels and overall conservation level (Figure 5.2A). In no case did we see conserved stretches of 50 bp or more. Thus, we expect the *Ubx*-UCEs to represent distinct functional elements.

The simulation of *Ubx* sequences further indicates that already ultraconserved blocks of 20 bp or more are highly unlikely to occur by chance, yet a large number of such blocks exist within the *Ubx* locus. Like the *Ubx*-UCEs, they are relatively evenly distributed throughout the *Ubx*, without obvious clustering (Figure 5.2B). The *Ubx*-UCEs therefore do not seem to constitute a separate class of elements, but rather represent the extreme end of a continuum. The

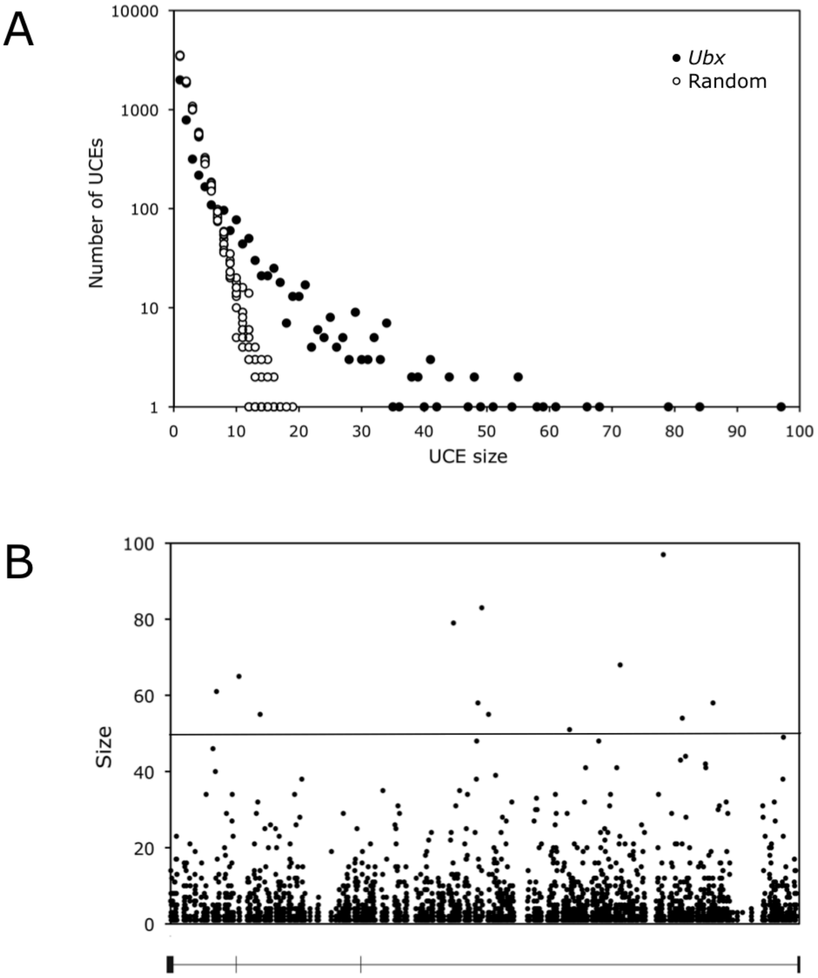


Figure 5.2. A. Distribution of UCEs within *Ubx* compared to ten neutrally evolving sequences of the same length and overall conservation level. B. Position of shorter UCEs within *Ubx*. UCE size is somewhat underestimated in this graph as UCEs were identified based on a multispecies alignment. Positions of the *Ubx* exons are indicated below the graph. The gene measures 75 kb between the start and the stop codons.

conclusions that we draw based on our study of *Ubx*-UCEs should therefore be relevant to understanding these shorter elements.

To link our set of *Ubx*-UCEs to potential functions, we searched for positional overlap between these sequences and previously reported functional elements within the *Ubx* locus (Figure 5.3). Firstly, we looked for overlaps between *Ubx*-UCEs and known protein-coding and non-coding transcripts. One element, *Ubx*-UCE-2, overlaps with the coding *Ubx* exon mI and will be discussed in more detail below. No other *Ubx*-UCEs overlap with *Ubx* coding sequences or with the CG31498 gene, which is located within the *Ubx*locus. The region does not contain any reported non-coding RNAs, but there is some evidence for additional, protein-coding transcripts (Hild et al. 2003) and *Ubx*-UCE-11 overlaps one of these putative genes (BK002585). For *Ubx*-UCE-2 and *Ubx*-UCE-11 selection on amino acid sequences might therefore contribute to the observed level of conservation.

As many vertebrate UCEs are known to drive gene expression, we went on to consider overlap between the *Ubx*-UCEs and transcriptional enhancers that are located within the *Ubx* introns. Two such enhancers are known: *bithorax* (*bx*) and *anterobithorax* (*abx*), both of which regulate aspects of *Ubx* expression (Peifer and Bender 1986; Maeda and Karch 2006). Perhaps surprisingly, the *bx* enhancer does not overlap with any *Ubx*-UCEs. The *abx* region, on the other hand, overlaps with both *Ubx*-UCE-8 and *Ubx*-UCE-9. The *Ubx*-UCE-8 is of

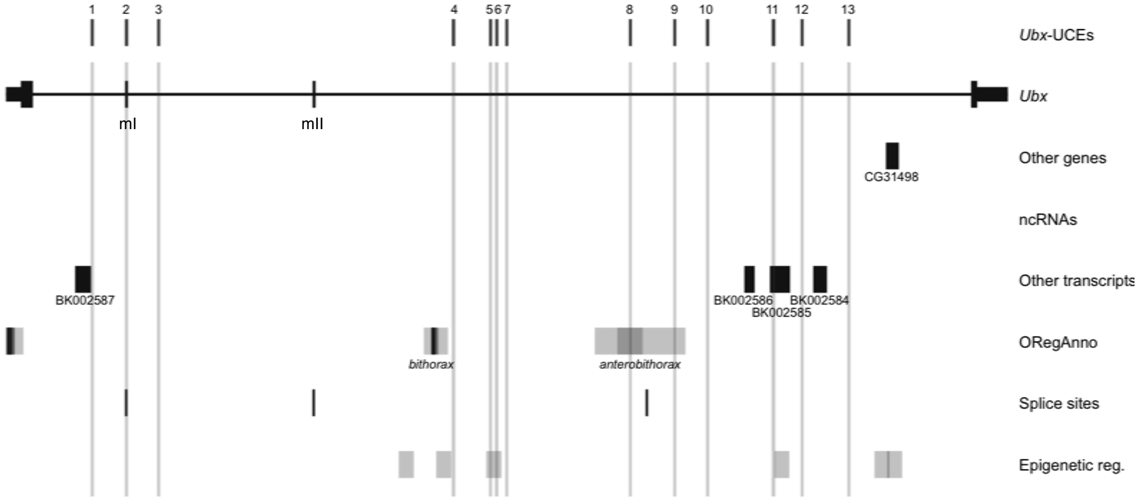


Figure 5.3. Overlap between *Ubx*-UCEs and known functional elements. See the text for further details.

particular interest as it sits within a 1.7 kb fragment that can drive reporter gene expression in an *abx*-like pattern (Simon et al. 1990).

Next, we considered *Ubx*-UCEs located in the vicinity of splice junctions, as many of the longest elements in the Glazov dataset overlapped with splice sites (Glazov et al. 2005). The *Ubx* contains two short, alternatively spliced exons: mI and mII. Unusually, the gene also contains an intronic splice site, which is used to subdivide the largest intron (Burnette et al. 2005). The *Ubx*-UCE-2 extends into the intronic sequence on both sides of the mI exon, meaning that it overlaps with both the upstream and downstream splice site. The mII exon, although well-conserved (Bomze and Lopez 1994), does not overlap with any *Ubx*-UCEs and neither does the intronic splice site.

Following the initial activation of *Drosophila Hox* genes, their expression state (active or inactive) is maintained by proteins of the Polycomb and Trithorax groups (Ringrose and Paro 2004). The binding site for five of these proteins within the *Ubx* locus have been determined by ChIP-chip (Beisel et al. 2007). *Ubx*-UCE-5 and *Ubx*-UCE-6 both overlap with one of the regions that was enriched for these regulators (PCR fragment 20287), however, this region only shows enrichment for one protein and the enrichment is only present in one of the two tested cell types. Thus, there does not appear to be a strong correlation between *Ubx*-UCEs and epigenetic regulation.

To summarise, we have some reason to believe that *Ubx*-UCE-2 may play a role in alternative splicing and that *Ubx*-UCE-8 is part of the *abx* enhancer. Furthermore, it is possible that *Ubx*-UCE-5 and *Ubx*-UCE-6 are involved in epigenetic regulation, that *Ubx*-UCE-9 is an additional component of the *abx* enhancer and that *Ubx*-UCE-11 is transcribed as part of the predicted BK002585 gene. For the remaining seven *Ubx*-UCEs, no potential functions were identified through our literature review. Notably, several functional regions of *Ubx*, such as the mII exon and the *bx* enhancer, were not connected with any *Ubx*-UCEs.

Based on this functional overview, we decided to further investigate the causes of ultraconservation within the mI exon. Our observation that mI resides within a *Ubx*-UCE adds to previous work showing that the nucleotide sequence of this exon is identical in four *Drosophila* species (Bomze and Lopez 1994). Our analysis also shows that *Ubx*-UCE-2 is 71 bp long and extends into the introns on both sides of the mI exon, which is 51 bp long. This already suggests that coding constraints cannot be the only cause of this case of ultraconservation, but to test this formally, we compared the mI exon to the *Ubx* homeodomain, which encodes a DNA-binding protein domain that is identical on the amino acid level in all twelve *Drosophila* species. For both sequences we counted the number of synonymous sites and the number of changes that had occurred at those sites. All 15 sites within mI are identical, whereas 29 out of 57 sites within the homeodomain have changed, showing that the conservation pattern of mI is significantly different from that of the homeodomain ($p = 0.0002$, Fisher's exact test). We only considered the third position of each codon for this analysis, although some synonymous mutations within the homeodomain have occurred at the first position of sixfold degenerate codons. Including these substitutions yields qualitatively similar results.

Splicing of the *Ubx* gene has been successfully studied using the Ubx.4 minigene construct, which reproduces the tissue-specific alternative splicing pattern of *Ubx* (Subramaniam, Bomze, and Lopez 1994). Previous experiments using the Ubx.4 minigene have established that changes in the mI nucleotide sequence can affect exon inclusion levels in *Drosophila* SL2 cell culture (Hatton, Subramaniam, and Lopez 1998). However, since these constructs contained a combination of synonymous and non-synonymous changes within mI, the potential causes of ultraconservation become hard to disentangle: If the observed changes in splicing pattern depend mainly on one or more non-synonymous mutations, the ultraconservation at those positions might be due to selection on the amino acid sequence, selection for correct splicing or both. It might even be the case that the two selection pressures are opposed to each other, for example such that the need to encode a specific amino acid overrides potentially beneficial changes in splicing regulation. Here, we therefore wished to extend these previous results by testing whether purely synonymous mutations have an impact on *Ubx* splicing.

Towards this end, we produced the Ubx.4 derivative construct Ubx.4-mutA (Figure 5.4). Mutations were introduced at all synonymous positions within the mI exon, except within 5 bp of the exon borders, to avoid interference with

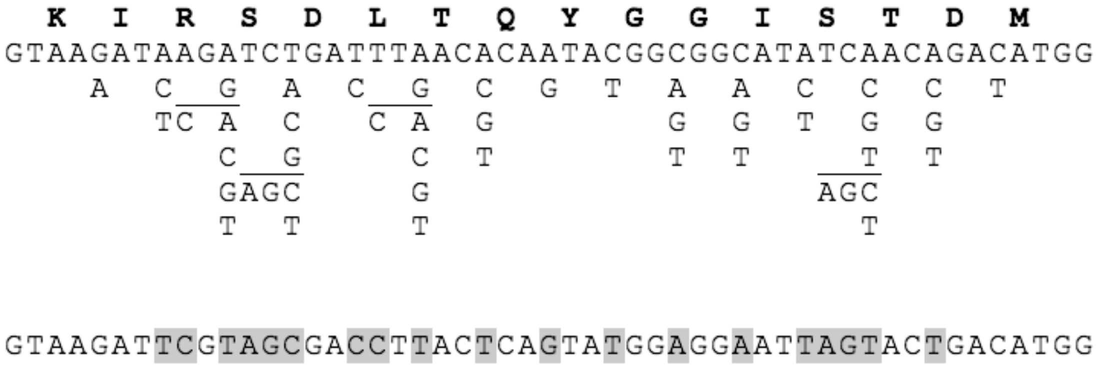

K I R S D L T Q Y G G I S T D M
GTAAGATAAGATCTGATTTAACACAATACGGCGGCATATCAACAGACATGG
A C G A C G C G T A A C C C T
TC A C C A G G G T G G
C G C T T T T T
GAGC G AGC
T T T T
GTAAGATTCGTAGCGACCTTACTCAGTATGGAGGAATTAGTACTGACATGG

Figure 5.4. Comparison of the wildtype mI exon and the mutated exon in the Ubx.4-mutA construct. The amino acid sequence is shown at the top, followed by the wildtype coding sequence, with all possible synonymous mutations indicated below. At the bottom is the Ubx.4-mutA mI sequence with substitutions highlighted in grey.

basic splice site recognition. Thus, although the mutated mI exon encodes the same amino acids, any splicing signals within the exonic RNA sequence are likely to have been altered.

5.4. Discussion

We have identified 1557 sequences of at least 50 bp that are identical in the genomes of twelve *Drosophila* species. To explore potential functions of these elements, we turned to the *Hox* gene *Ubx*, which contains an unexpectedly large number of UCEs. We surveyed the literature on *Ubx* regulation, which has accumulated over several decades, to search for positional overlap between the 13 *Ubx*-UCEs and known functional elements. This analysis suggested plausible roles for two *Ubx*-UCEs in alternative splicing and transcription, respectively, and indicated that some of the other *Ubx*-UCEs might be transcribed or involved in epigenetic regulation. However, in spite of the rich literature on the *Ubx* locus, we were not able to find any links with the pre-existing literature for seven *Ubx*-UCEs. Thus it seems that many features of *Ubx* biology have so far escaped detection and that comparative sequence analysis can serve as a tool to further our understanding of this biologically important locus.

To study the mechanisms that underlie ultraconservation of a single element, we focussed on *Ubx*-UCE-2, which overlaps with the short mI exon, the alternative splicing of which has important consequences in *Drosophila* development. Further to previous experiments showing that nucleotide changes within mI can affect the *Ubx* alternative splicing pattern (Hatton, Subramaniam, and Lopez 1998), experiments by Britta Hartmann (unpublished observations) have established that an effect on splicing is observed for purely synonymous nucleotide substitutions. This suggests that the extreme conservation of mI and the immediately surrounding intronic sequences is likely due to a mixture between selection on the coding sequence and selection on splicing signals. Notably, many human UCEs overlap with alternative exons (Bejerano et al. 2004) indicating that this type of dual selection pressure might be a common phenomenon. Possibly, ultraconservation in general might be often explained by multiple functions being imposed on a single sequence, for example enhancers that also function as silencers (Pennacchio et al. 2006) or coding sequences that contain transcriptional regulators (Lampe et al. 2008). Analysis of mutations within single UCEs provides a means to tease apart these different contributions.

6.

General discussion

In this thesis, I have investigated four aspects of gene expression evolution. In Chapter 2, we saw that TEs have not had a noticeable effect on differences in expression levels and transcript diversity between humans and chimpanzees. In addition, it appears that TEs do not accumulate upstream of genes due to a presumed role in gene regulation, but rather because they are preferentially inserted upstream of genes that are active in the germ line.

In Chapter 3, I proposed a test for positive selection acting on gene expression levels. Contrary to previous methods, this test estimates the rate of neutral evolution directly from experimental data. Applied to expression data from human and chimpanzee lymphoblastoid cell lines, the test indicated that slightly deleterious mutations are segregating in humans and that they overshadow any effects of adaptive evolution. Although exact quantification will have to await accurate modelling of the distribution of fitness effects associated with expression mutations, it seems that humans have undergone little or no adaptive evolution in terms of expression levels.

Chapter 4 dealt with the evolution of regulatory complexity in the human genome. I showed that older genes are more extensively regulated than younger genes and that the rate of increase in complexity does not slow down over time. Therefore, the evolution of gene regulation does not appear to be curbed by difficulties of organising very complex genes.

Finally, in Chapter 5, the phenomenon of ultraconservation, which has been most extensively analysed in vertebrates, was studied in twelve *Drosophila* species. Although functions have been found for some UCEs, it is still not clear why these elements are conserved to such an extreme degree. The use of an easily manipulated model organism, such as *Drosophila melanogaster*, will allow careful dissection of UCEs to reveal their full responsibilites.

How has this thesis contributed to our understanding of gene expression evolution? I will begin, in Section 6.1, by considering the importance of gene regulation in human evolution, in light of the data presented here. Next, in Section 6.2, I will discuss how the results of each chapter may extend to other species an to within-species variation. In Section 6.3, I will then outline some upcoming challenges in the field and will finish, in Section 6.4, with some concluding remarks on the lessons learnt from this thesis.

6.1. The role of gene expression in human evolution

How have changes in gene regulation affected the evolution of our own species? In Section 6.1.1, I will discuss human evolution in the short term, dating back to the split from our closest living relatives, the chimpanzees. In Section 6.1.2, I will consider gene regulation over longer time periods.

6.1.1. Adaptive and non-adaptive evolution of gene regulation

The conclusion of Chapter 3, that there is no evidence for pervasive positive selection on gene expression in humans, is consistent with a previous study by Lemos et al. (2005). In both cases, the analyses were based on comparisons between humans and chimpanzees. In a complementary approach, Kudaravalli et al. (2009) analysed variation within human populations and estimated that 0.1% or less of their identified eQTLs showed signs of positive selection. It should be emphasised that the observed selection signal was not necessarily due to the eQTLs themselves, but could be caused by selection on nearby sequences that do not have an impact on gene expression. So far, there is therefore very little evidence that recent human evolution has been heavily influenced by adaptive changes in gene expression.

This is not to say that there might not be individual cases of important human adaptations that have relied on differences in gene expression. Many critical differences might for example occur in early development and would therefore not have been visible in the above studies, which focussed on adult tissues and cell lines. The role of positive selection on human gene expression is therefore a question that remains to be finally settled. The extended McDonald-Kreitman test, presented in Chapter 3, will hopefully provide a useful framework for this purpose.

The relative contributions of regulatory and protein-coding mutations to phenotypic evolution are a matter of debate (Hoekstra and Coyne 2007). Should the apparent lack of adaptive gene expression evolution in humans be taken as evidence that structural changes have been more influential? Not necessarily, because positive selection acting on protein-coding sequences appears to have been much less efficient in humans, where estimates of the proportion of positively selected sites lie in the range of 0-20% (Boyko et al. 2008; Eyre-Walker and Keightley 2009), compared to other mammals such as mice, where it is above 50% (Halligan et al. 2010). This is at least in part due to variations in effective population size; while the human effective population size is around 10000 (Eyre-Walker et al. 2002), the mouse equivalent is 580000 (Halligan et al. 2010). A smaller effective population size means that it is easier for neutral and nearly neutral mutations to reach fixation. We must therefore take into account that many of the changes that made us human, whether regulatory or structural, might have been fixed due to random processes rather than because of their adaptive value.

In principle, TEs could drive extensive remodelling of a genome in the absence of selection (Feschotte 2008). We might even expect a larger impact of TEs in humans compared to other species, as the reduced effective population size would make it more difficult for negative selection to weed out slightly deleterious insertions (Lynch 2007b). However, as we saw in Chapter 2, TEs have not contributed expression differences between humans and chimpanzees. Future studies will have to determine whether this is because TEs do not significantly influence gene expression or whether their effect is typically too large to be tolerated (see Section 6.2.1).

6.1.2. Turn-over of regulatory elements

The evolution of gene expression does not only depend on forces that influence the creation and fixation of new elements, but also on forces that cause them to be removed. In Chapter 4, I showed that genes tend to accumulate regulatory elements, as they grow older. To what extent is this regulatory complexity beneficial?

The observation that complexity increases over time does not necessarily indicate that it has a purpose (Lynch 2007a). In theory, we might observe the same phenomenon for completely non-functional elements, if for some reason they were easier to add than to remove. It is of course not true that all regulation is unproductive; careful control of gene expression plays a critical role for survival and, as our knowledge of gene regulation in humans and other species grows, our appreciation of this will likely increase. However, it is worth keeping in mind that we are not yet in a position where we can explain what the regulatory information available from genome-wide surveys means in terms of organism function. Some signal will be due to technical and biological noise and some might represent regulatory elements that have become obsolete, but which are difficult to remove without disturbing gene function. However, at least for alternative splicing, the increase in complexity is not explained by a larger proportion of non-functional isoforms, as the trend persists even when only isoforms with a confirmed protein product are included in the analysis (Roux and Robinson-Rechavi 2011).

6.2. Generality of the results of this thesis

To what degree can the observations presented here teach us something about gene expression evolution in general? In this section I will discuss how the results of each chapter may extend to other species and, where applicable, to variation between human individuals.

6.2.1. Transposable elements

We saw in Chapter 2 that TEs have not made a detectable contribution to gene expression in humans and chimpanzees. This is consistent with a previous study where it was shown that human Alu elements are more common around genes that are expressed in many tissues, but that they do not themselves cause genes to increase their expression breadth (Urrutia, Ocana, and Hurst 2008). However, it contrasts with results from *Arabidopsis* (Hollister and Gaut 2009), rice (Naito et al. 2009) and rodents (Pereira, Enard, and Eyre-Walker 2009) in which TEs were shown to contribute to differences in gene expression, although in the last case the contribution was relatively minor. Furthermore, comparisons of embryonic stem cells from humans and mice have identified changes in transcriptional regulation that are linked to the activity of a rare type of TE known as endogenous retrovirus 1 (ERV1) (Kunarso et al. 2010). Together these observations argue for a lineage-specific effect of TEs, where the types of TEs that are present might be more important than the overall TE activity.

As discussed in Chapter 2, very recent TE insertions may cause variation in gene expression between human individuals, but be too deleterious to ever reach fixation. Identifying such insertions might therefore be interesting from a medical perspective. Data has recently become available to test this hypothesis, as a number of studies have identified human polymorphic TEs (Beck et al. 2010; Huang et al. 2010; The 1000 Genomes Project Consortium et al. 2010), which could be contrasted with expression data from the same individuals.

6.2.2. Positive selection on gene expression levels

It is likely that adaptive evolution of gene expression has been more prominent in other species than it has in humans, as this is what has been seen for protein-coding sequences (see Section 6.1.1). As more sequence and expression data become available, this could easily be tested using the extended McDonald-Kreitman test from Chapter 3. It will be particularly interesting to see whether positive selection on expression changes will be tightly linked to positive selection on structural changes or whether under some evolutionary scenarios, new adaptations are more likely to be expression-based and vice versa.

6.2.3. Accumulation of regulatory complexity

How has regulatory complexity evolved in other species? It seems plausible that the increase to complexity over time, shown in Chapter 4, is a general phenomenon that applies to many other lineages. However, it would be interesting to see whether organisms with slimmer genomes, such as *D. melanogaster* or the yeast *Saccharomyces cerevisiae,* are more constrained in terms of gene regulation and might therefore show slower rates of increase for older genes. It should be emphasised that these trends concern the average behaviour and that the variance in complexity can be large between individual genes. For example, the *Drosophila* genome contains the spectacular *Dscam* gene, which contains 95 alternative exons and could theoretically give rise to over 38000 isoforms (Park and Graveley 2007).

6.2.4. Ultraconservation

Ultraconservation is a phenomenon that exists both in *Drosophila* and vertebrates, and most likely also in many other clades for which we currently do not have sufficient genome information. Studying UCE function in *Drosophila* has both advantages and disadvantages if the aim is to understand ultraconservation in the human genome. Vertebrate UCEs tend to be longer and older; in some cases they can be traced back to cartilaginous fishes (Wang et al. 2009). It is therefore unlikely that all lessons learned from *Drosophila* will be directly transferable. On the other hand, analysis of an independent set of UCEs, such as those in *Drosophila,* will give us more power to elucidate the general principles that underlie ultraconservation.

6.3. Future studies

The field of human molecular genetics has undergone a revolution in recent years: The human genome has been sequenced (Lander et al. 2001; Venter et al. 2001), the contributions of regulatory mechanisms such as alternative splicing and miRNAs has been re-evaluated (Pasquinelli et al. 2000; Pan et al. 2008) and vast amounts of gene expression data have been collected (Parkinson et al. 2011). However, many central questions remain largely unanswered: Which sequences are involved in gene regulation? How do these sequences differ between humans and other species? What impact have these differences had on gene expression and, ultimately, fitness? Although it will undoubtedly take time to exhaustively address these topics, new developments will allow us further insights in the near future.

6.3.1. Improved and extended gene expression measurements

The ongoing ENCODE project is an effort to catalogue all functional elements in the human genome and the pilot phase of the project has already provided maps of transcribed sequences, transcription factor binding sites, histone modifications and many other features in 1% of the genome (ENCODE Project Consortium et al. 2007). Having access to this type of information is vital, if we are to fully decipher our genome. For example, it should increase our understanding of the mechanisms underlying ultraconservation, as well as allow informed choices about which sequences to analyse for signs of positive selection that might affect gene expression.

Another area where we might expect significant progress within the next few years is proteomics: For technical reasons, many studies of gene expression have focussed on RNAs rather than proteins, as reliable measurement of protein levels have proven notoriously difficult (Bell et al. 2009), especially for genome-wide assessments. However, in a pioneering study, Schwanhäusser et al. (2011) recently measured both protein and mRNA levels for thousands of genes in mammalian cells and showed that the regulation of translation is the most important factor for controlling protein abundance.

6.3.2. Gene expression in model and non-model organisms

Much effort is also being made to systematically investigate gene regulation in other species, which should give important evolutionary clues. The mouse ENCODE project has been designed to be largely analogous to its' human counterpart, so that the data can be used for comparative analysis, but will also extend the human data by using techniques that cannot be applied to humans (Raney et al. 2011). There are also similar projects underway for the fruit fly *D. melanogaster* (The modENCODE Consortium et al. 2010) and the nematode *C. elegans* (Gerstein et al. 2010).

From a human perspective, it is also of great interest to study gene regulation in primates, as this allows us to investigate changes that may have contributed to human-specific characteristics. While these animals are not as easily manipulated as model organisms such as mouse, the sequencing of several primate genomes, including gorilla, baboon and marmoset (Flicek et al. 2011), along with the previously published chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005), orang-utan (Locke et al. 2011) and rhesus macaque (Gibbs et al. 2007) genomes allow detailed comparative analysis of this clade and also enables the use of RNA sequencing and other techniques that require genome mappings. Other methods may also be available for these species, as, in some cases, procedures established in humans may be directly transferable; for example, Cain et al. (2011) studied histone modifications in chimpanzees and macaques using ChIP-seq with a human antibody.

Access to data from multiple species improves analyses by providing greater power to detect lineage-specific differences and by putting findings into context. For example, it has been reported that certain genes related to metabolism are upregulated in human relative to chimpanzee brain tissue, which could indicate that increased energy supply was a crucial step towards the enhanced cognitive ability seen in humans (Khaitovich et al. 2008). However, as lack information on expression levels of metabolic genes in other primate species, it is not yet possible to say whether this is truly a human-specific pattern. Subsequent analysis may therefore require us to revise this and other hypothesis about human evolution.

6.3.3. Gene expression in context

Gene expression is not an end goal in itself, but a means to build up a whole organism, whose fitness will determine its' survival. Therefore, we should aim to integrate the study of gene expression into a larger biological framework. In the first instance, this could mean relating gene expression data to biochemical processes in the cell, for example by correlating expression levels to the concentrations of different metabolites (Fu et al. 2011). Following on from this, we will also want to know how gene expression influences how cells specialise and interact with each other, for example by studying gene expression from a developmental perspective (Venkataraman et al. 2008). Finally, because dynamic gene expression influences how animals interact with the outside world (Warren et al. 2010), molecular genetics also has links to psychology, ecology and other biological sciences. Thus, many future discoveries on the role of gene expression in human evolution are likely to come from interdisciplinary studies (Varki, Geschwind, and Eichler 2008).

6.4. General conclusions

Understanding gene expression evolution is not only a matter of collecting data. As shown in this thesis, naïve interpretation of genomic information can easily lead to erroneous conclusions regarding the selective regimes operating on a given sequence. For example, non-random distributions of genomic elements can sometimes arise through neutral processes. This was illustrated in Chapter 2, where the enrichment of TEs upstream of protein-coding genes could be explained by biased insertion, without the need to invoke selection. Further, the degree of sequence similarity cannot be taken as a direct indication of functional importance, as shown by the study of *Drosophila* UCEs in Chapter 5; while future studies might provide functional explanations for all of these elements, it is nonetheless striking that they only rarely overlap with known regulatory sequences. On the other side of the spectrum, as discussed in Chapter 3, large sequence divergence may or may not be indicative of emerging new functions, as it may arise from either positive selection or reduced constraint. In this context, one can speculate that many new features, such as TE insertions (Chapter 2) or new regulatory elements (Chapter 4) might be passed on to further generations, not because of their functional significance, but because they are selective neutral. Based on these considerations, the overarching message of this thesis is therefore the absolute need to complement genome-wide maps of gene regulation with thorough evolutionary analysis and not to assume that observed patterns are the result of natural selection until all neutral alternatives have been exhausted.

References

Ackerman, H., I. Udalova, J. Hull, and D. Kwiatkowski. 2002. Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. Mol Biol Evol **19**:884-890.

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal, L. A. Pennacchio, and E. M. Rubin. 2007. Deletion of ultraconserved elements yields viable mice. PLoS Biol **5**:e234.

Albà, M. M., and J. Castresana. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol **7**:53.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**:3389-3402.

An, J. J., K. Gharami, G. Y. Liao, et al. (11 co-authors). 2008. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. Cell **134**:175-187.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature **437**:1149-1152.

Ashburner, M., C. A. Ball, J. A. Blake, et al (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**:25-29.

Babbitt, C. C., O. Fedrigo, A. D. Pfefferle, A. P. Boyle, J. E. Horvath, T. S. Furey, and G. A. Wray. 2010. Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. Genome Biol Evol **2**:67-79.

Bai, L., and A. V. Morozov. 2010. Gene regulation by nucleosome positioning. Trends Genet **26**:476-483.

Baumann, M., J. Pontiller, and W. Ernst. 2010. Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. Mol Biotechnol **45**:241-247.

Beck, C. R., P. Collier, C. Macfarlane, M. Malig, J. M. Kidd, E. E. Eichler, R. M. Badge, and J. V. Moran. 2010. LINE-1 retrotransposition activity in human genomes. Cell **141**:1159-1170.

Becquet, C., N. Patterson, A. C. Stone, M. Przeworski, and D. Reich. 2007. Genetic structure of chimpanzee populations. PLoS Genet **3**:e66.

Bedford, T., and D. L. Hartl. 2009. Optimization of gene expression by natural selection. Proc Natl Acad Sci U S A **106**:1133-1138.

Beisel, C., A. Buness, I. M. Roustan-Espinosa, B. Koch, S. Schmitt, S. A. Haas, M. Hild, T. Katsuyama, and R. Paro. 2007. Comparing active and repressed expression states of genes controlled by the Polycomb/Trithorax group proteins. Proc Natl Acad Sci U S A **104**:16615-16620.

Bejerano, G., C. B. Lowe, N. Ahituv, B. King, A. Siepel, S. R. Salama, E. M. Rubin, W. J. Kent, and D. Haussler. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature **441**:87-90.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. Science **304**:1321-1325.

Belancio, V. P., D. J. Hedges, and P. Deininger. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res **18**:343-358.

Belancio, V. P., D. J. Hedges, and P. Deininger. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res **34**:1512-1521.

Bell, A. W., E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis, S. Sechi, T. Nilsson, and J. J. Bergeron. 2009. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods **6**:423-430.

Bergman, C. M., and M. Kreitman. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res **11**:1335-1345.

Birney, E.J. A. Stamatoyannopoulos, A. Dutta, et al. (310 co-authors). 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**:799-816.

Blanchette, M., W. J. Kent, C. Riemer, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res **14**:708-715.

Blekhman, R., J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad. 2010. Sex-specific and lineage-specific alternative splicing in primates. Genome Res **20**:180-189.

Blekhman, R., A. Oshlack, A. E. Chabot, G. K. Smyth, and Y. Gilad. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet **4**:e1000271.

Blow, M. J., D. J. McCulley, Z. Li, et al. (17 co-authors). 2010. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet **42**:806-810.

Boffelli, D., M. A. Nobrega, and E. M. Rubin. 2004. Comparative genomics at the vertebrate extremes. Nat Rev Genet **5**:456-465.

Bomze, H. M., and A. J. Lopez. 1994. Evolutionary conservation of the structure and expression of alternatively spliced Ultrabithorax isoforms from Drosophila. Genetics **136**:965-977.

Bourque, G., B. Leong, V. B. Vega, et al. (11 co-authors). 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res **18**:1752-1762.

Boyko, A. R., S. H. Williamson, A. R. Indap, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet **4**:e1000083.

Britten, R. J. 2010. Transposable element insertions have strongly affected human evolution. Proc Natl Acad Sci U S A **107**:19945-19948.

Britten, R. J. 1994. Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proc Natl Acad Sci U S A **91**:5992-5996.

Britten, R. J., and E. H. Davidson. 1969. Gene regulation for higher cells: a theory. Science **165**:349-357.

Britten, R. J., and E. H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol **46**:111-138.

Burnette, J. M., E. Miyamoto-Sato, M. A. Schaub, J. Conklin, and A. J. Lopez. 2005. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics **170**:661-674.

Cain, C. E., R. Blekhman, J. C. Marioni, and Y. Gilad. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics **187**:1225-1234.

Calin, G. A., C. G. Liu, M. Ferracin, et al. (27 co-authors). 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell **12**:215-229.

Caliskan, M., D. A. Cusanovich, C. Ober, and Y. Gilad. 2011. The effects of EBV transformation on gene expression levels and methylation profiles. Hum Mol Genet.

Carninci, P., T. Kasukawa, S. Katayama, et al. (194 co-authors). 2005. The transcriptional landscape of the mammalian genome. Science **309**:1559-1563.

Carninci, P., A. Sandelin, B. Lenhard, et al. (41 co-authors). 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet **38**:626-635.

Carroll, S. B., J. Grenier, and S. Weatherbee. 2004. From DNA to Diversity. Blackwell Publishing, Malden.

Charlesworth, J., and A. Eyre-Walker. 2008. The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol **25**:1007-1015.

Chen, C. T., J. C. Wang, and B. A. Cohen. 2007. The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet **80**:692-704.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Choy, E., R. Yelensky, S. Bonakdar, et al. (17 co-authors). 2008. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet **4**:e1000287.

Collas, P. 2010. The Current State of Chromatin Immunoprecipitation. Molecular Biotechnology **45**:87-100.

Cordaux, R., J. Lee, L. Dinoso, and M. A. Batzer. 2006. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene **373**:138-144.

Costa, V., C. Angelini, I. De Feis, and A. Ciccodicola. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol **2010**:853916.

de Magalhaes, J. P., and J. Costa. 2009. A database of vertebrate longevity records and their relation to other life-history traits. J Evol Biol **22**:1770-1774.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet **37**:544-548.

Dewannieux, M., C. Esnault, and T. Heidmann. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet **35**:41-48.

Domazet-Lošo, T., and D. Tautz. 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol **8**:66.

Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601-603.

Drake, J. A., C. Bird, J. Nemesh, et al. (11 co-authors). 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet **38**:223-227.

ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, et al. (311 co-authors). 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**:799-816.

Eyre-Walker, A. 2010. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A **107 Suppl 1**:1752-1756.

Eyre-Walker, A., and P. D. Keightley. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol **26**:2097-2108.

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.

Farajollahi, S., and S. Maas. 2010. Molecular diversity through RNA editing: a balancing act. Trends Genet **26**:221-230.

Farre, D., and M. M. Alba. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. Mol Biol Evol **27**:325-335.

Faulkner, G. J., and P. Carninci. 2009. Altruistic functions for selfish DNA. Cell Cycle **8**:2895-2900.

Faulkner, G. J., Y. Kimura, C. O. Daub, et al. (22 co-authors) 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet **41**:563-571.

Fay, J. C., and P. J. Wittkopp. 2008. Evaluating the role of natural selection in the evolution of gene regulation. Heredity **100**:191-199.

Fay, J. C., G. J. Wyckoff, and C. I. Wu. 2001. Positive and negative selection on the human genome. Genetics **158**:1227-1234.

Feschotte, C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet **9**:397-405.

Flicek, P., B. L. Aken, B. Ballester, et al. (57 co-authors). 2010. Ensembl's 10th year. Nucleic Acids Res **38**:D557-562.

Flicek, P., M. R. Amode, D. Barrell, et al. (52 co-authors). 2011. Ensembl 2011. Nucleic Acids Res **39**:D800-806.

Fraser, H. B., A. M. Moses, and E. E. Schadt. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc Natl Acad Sci U S A **107**:2977-2982.

Friedman, R. C., K. K. Farh, C. B. Burge, and D. P. Bartel. 2009. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res **19**:92-105.

Fu, N., I. Drinnenberg, J. Kelso, J. R. Wu, S. Paabo, R. Zeng, and P. Khaitovich. 2007. Comparison of protein and mRNA expression evolution in humans and chimpanzees. PLoS One **2**:e216.

Fu, X., P. Giavalisco, X. Liu, et al. (13 co-authors). 2011. Rapid metabolic evolution in human prefrontal cortex. Proc Natl Acad Sci U S A **108**:6181-6186.

Gaffney, D. J., R. Blekhman, and J. Majewski. 2008. Selective constraints in experimentally defined primate regulatory regions. PLoS Genet **4**:e1000157.

Gaffney, D. J., and P. D. Keightley. 2006. Genomic selective constraints in murid noncoding DNA. PLoS Genet **2**:e204.

Gebelein, B., J. Culi, H. D. Ryoo, W. Zhang, and R. S. Mann. 2002. Specificity of Distalless repression and limb primordia development by abdominal Hox proteins. Dev Cell **3**:487-498.

Gentleman, R. C., V. J. Carey, D. M. Bates, et al. (25 co-authors). 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5**:R80.

Gerstein, M., B.Z. J. Lu, E. L. Van Nostrand, et al. (131 co-authors). 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science **330**:1775-1787.

Gibbs, R., A.J. Rogers, M. G. Katze, et al. (136 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**:222-234.

Gilad, Y., A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440**:242-245.

Gilad, Y., S. A. Rifkin, P. Bertone, M. Gerstein, and K. P. White. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res **15**:674-680.

Glazov, E. A., M. Pheasant, E. A. McGraw, G. Bejerano, and J. S. Mattick. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res **15**:800-808.

Goodrich, J. A., and R. Tjian. 2010. Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. Nat Rev Genet **11**:549-558.

Griffin, N. M., J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol, and J. E. Schnitzer. 2010. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol **28**:83-89.

Gry, M., R. Rimini, S. Stromberg, A. Asplund, F. Ponten, M. Uhlen, and P. Nilsson. 2009. Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genomics **10**:365.

Gu, X. 2004. Statistical framework for phylogenomic analysis of gene family expression profiles. Genetics **167**:531-542.

Gu, X., Z. Zhang, and W. Huang. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci U S A **102**:707-712.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell **130**:77-88.

Halligan, D. L., F. Oliver, A. Eyre-Walker, B. Harr, and P. D. Keightley. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet **6**:e1000825.

Han, J. S., and J. D. Boeke. 2005. LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? Bioessays **27**:775-784.

Han, J. S., S. T. Szak, and J. D. Boeke. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature **429**:268-274.

Hatton, A. R., V. Subramaniam, and A. J. Lopez. 1998. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell **2**:787-796.

Haygood, R., C. C. Babbitt, O. Fedrigo, and G. A. Wray. 2010. Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci U S A **107**:7853-7857.

Haygood, R., O. Fedrigo, B. Hanson, K. D. Yokoyama, and G. A. Wray. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet **39**:1140-1144.

Heckman, K. L., and L. R. Pease. 2007. Gene splicing and mutagenesis by PCR-driven overlap extension. Nat Protoc **2**:924-932.

Hedges, D. J., and P. L. Deininger. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res **616**:46-59.

Hedges, S. B., J. Dudley, and S. Kumar. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics **22**:2971-2972.

Hey, J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. Mol Biol Evol **27**:921-933.

Hild, M., B. Beckmann, S. A. Haas, et al. (12 co-authors). 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. Genome Biol **5**:R3.

Hodgins-Davis, A., and J. P. Townsend. 2009. Evolving gene expression: from G to E to GxE. Trends Ecol Evol **24**:649-658.

Hoekstra, H. E., and J. A. Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution **61**:995-1016.

Hollister, J. D., and B. S. Gaut. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res **19**:1419-1428.

Holloway, A. K., M. K. Lawniczak, J. G. Mezey, D. J. Begun, and C. D. Jones. 2007. Adaptive gene expression divergence inferred from population genomics. PLoS Genet **3**:2007-2013.

Huang, C. R., A. M. Schneider, Y. Lu, et al. (14 co-authors). 2010. Mobile interspersed repeats are major structural variants in the human genome. Cell **141**:1171-1182.

Huda, A., L. Marino-Ramirez, D. Landsman, and I. K. Jordan. 2009. Repetitive DNA elements, nucleosome binding and human gene expression. Gene **436**:12-22.

Huminiecki, L., and K. H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res **14**:1870-1879.

Huntzinger, E., and E. Izaurralde. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet **12**:99-110.

Innan, H., and F. Kondrashov. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet **11**:97-108.

International Human Genome Sequencing Consortium, E. S. Lander, L. M. Linton, et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. 2003a. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res **31**:e15.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**:249-264.

Irizarry, R. A., D. Warren, F. Spencer, et al. (25 co-authors). 2005. Multiple-laboratory comparison of microarray platforms. Nat Methods **2**:345-350.

Jenkins, D. L., C. A. Ortori, and J. F. Brookfield. 1995. A test for adaptive change in DNA sequences controlling transcription. Proc Biol Sci **261**:203-207.

Jordan, I. K., I. B. Rogozin, G. V. Glazko, and E. V. Koonin. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet **19**:68-72.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet **16**:418-420.

Kamal, M., X. Xie, and E. S. Lander. 2006. A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci U S A **103**:2740-2745.

Katzman, S., A. D. Kern, G. Bejerano, G. Fewell, L. Fulton, R. K. Wilson, S. R. Salama, and D. Haussler. 2007. Human genome ultraconserved elements are ultraselected. Science **317**:915.

Kazazian, H. H., Jr., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature **332**:164-166.

Kelley, M. R., S. Kidd, R. L. Berg, and M. W. Young. 1987. Restriction of P-element insertions at the Notch locus of Drosophila melanogaster. Mol Cell Biol **7**:1545-1548.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423**:241-254.

Kelso, J., J. Visagie, G. Theiler, et al. (12 co-authors). 2003. eVOC: a controlled vocabulary for unifying gene expression data. Genome Res **13**:1222-1230.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome Res **12**:656-664.

Khaitovich, P., W. Enard, M. Lachmann, and S. Paabo. 2006. Evolution of primate gene expression. Nat Rev Genet **7**:693-702.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850-1854.

Khaitovich, P., H. E. Lockstone, M. T. Wayland, et al. (12 co-authors). 2008. Metabolic changes in schizophrenia and human brain evolution. Genome Biol **9**:R124.

Khaitovich, P., B. Muetzel, X. She, et al. (15 co-authors). 2004a. Regional patterns of gene expression in human and chimpanzee brains. Genome Res **14**:1462-1473.

Khaitovich, P., S. Paabo, and G. Weiss. 2005. Toward a neutral evolutionary model of gene expression. Genetics **170**:929-939.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. 2004b. A neutral model of transcriptome evolution. PLoS Biol **2**:E132.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**:893-903.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107-116.

Kiran, A., and P. V. Baranov. 2010. DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics **26**:1772-1776.

Kohn, M. H., S. Fang, and C. I. Wu. 2004. Inference of positive and negative selection on the 5' regulatory regions of Drosophila genes. Mol Biol Evol **21**:374-383.

Kroutter, E. N., V. P. Belancio, B. J. Wagstaff, and A. M. Roy-Engel. 2009. The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. PLoS Genet **5**:e1000458.

Kudaravalli, S., J. B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, and J. K. Pritchard. 2009. Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol **26**:649-658.

Kunarso, G., N. Y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y. S. Chan, H. H. Ng, and G. Bourque. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet **42**:631-634.

Lampe, X., O. A. Samad, A. Guiguen, C. Matis, S. Remacle, J. J. Picard, F. M. Rijli, and R. Rezsohazy. 2008. An ultraconserved Hox-Pbx responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. Nucleic Acids Res **36**:3214-3225.

Lander, E. S., L. M. Linton, B. Birren, et al. (254 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Lareau, L. F., M. Inada, R. E. Green, J. C. Wengrod, and S. E. Brenner. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature **446**:926-929.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics **11**:733-739.

Lemos, B., C. D. Meiklejohn, M. Caceres, and D. L. Hartl. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution **59**:126-137.

Lettice, L. A., T. Horikoshi, S. J. Heaney, et al. (21 co-authors). 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A **99**:7548-7553.

Lev-Maor, G., O. Ram, E. Kim, N. Sela, A. Goren, E. Y. Levanon, and G. Ast. 2008. Intronic Alus influence alternative splicing. PLoS Genet **4**:e1000204.

Li, B., M. Carey, and J. L. Workman. 2007. The role of chromatin during transcription. Cell **128**:707-719.

Liao, B. Y., and J. Zhang. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol **23**:530-540.

Lin, Z., H. Ma, and M. Nei. 2008. Ultraconserved coding regions outside the homeobox of mammalian Hox genes. BMC Evol Biol **8**:260.

Locke, D. P., L. W. Hillier, W. C. Warren, et al. (101 co-authors). 2011. Comparative and demographic analysis of orang-utan genomes. Nature **469**:529-533.

Lowe, C. B., G. Bejerano, and D. Haussler. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A **104**:8005-8010.

Lu, J., J. C. Lee, M. L. Salit, and M. C. Cam. 2007. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. BMC Bioinformatics **8**:108.

Lu, T., Y. Pan, S. Y. Kao, C. Li, I. Kohane, J. Chan, and B. A. Yankner. 2004. Gene regulation and DNA damage in the ageing human brain. Nature **429**:883-891.

Lynch, M. 2007a. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A **104 Suppl 1**:8597-8604.

Lynch, M. 2007b. The origins of genome architecture. Sinauer, Sunderland, MA.

Lynch, M. 1988. The rate of polygenic mutation. Genet Res **51**:137-148.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151-1155.

Maeda, R. K., and F. Karch. 2006. The ABC of the BX-C: the bithorax complex explained. Development **133**:1413-1422.

Maeda, R. K., and F. Karch. 2009. The bithorax complex of Drosophila: an exceptional Hox cluster. Pp. 1-33. Curr Top Dev Biol.

Makalowski, W., G. A. Mitchell, and D. Labuda. 1994. Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet **10**:188-193.

Mann, R. S., and D. S. Hogness. 1990. Functional dissection of Ultrabithorax proteins in D. melanogaster. Cell **60**:597-610.

Margaritis, T., and F. C. Holstege. 2008. Poised RNA polymerase II gives pause for thought. Cell **133**:581-584.

Martens, J. H., R. J. O'Sullivan, U. Braunschweig, S. Opravil, M. Radolf, P. Steinlein, and T. Jenuwein. 2005. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J **24**:800-812.

Martin, K. C., and A. Ephrussi. 2009. mRNA localization: gene expression in the spatial dimension. Cell **136**:719-730.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature **351**:652-654.

McLean, C., and G. Bejerano. 2008. Dispensability of mammalian DNA. Genome Res **18**:1743-1751.

Medstrand, P., L. N. van de Lagemaat, C. A. Dunn, J. R. Landry, D. Svenback, and D. L. Mager. 2005. Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res **110**:342-352.

Milinkovitch, M. C., R. Helaers, and A. C. Tzika. 2010. Historical constraints on vertebrate genome evolution. Genome Biol Evol **2**:13-18.

Millevoi, S., and S. Vagner. 2010. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic Acids Res **38**:2757-2774.

Mills, R. E., E. A. Bennett, R. C. Iskow, C. T. Luttig, C. Tsui, W. S. Pittard, and S. E. Devine. 2006. Recently mobilized transposons in the human and chimpanzee genomes. Am J Hum Genet **78**:671-679.

Naito, K., F. Zhang, T. Tsukiyama, H. Saito, C. N. Hancock, A. O. Richardson, Y. Okumoto, T. Tanisaka, and S. R. Wessler. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature **461**:1130-1134.

Ni, J. Z., L. Grate, J. P. Donohue, C. Preston, N. Nobida, G. O'Brien, L. Shiue, T. A. Clark, J. E. Blume, and M. Ares, Jr. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev **21**:708-718.

Nicholson, P., H. Yepiskoposyan, S. Metze, R. Zamudio Orozco, N. Kleinschmidt, and O. Muhlemann. 2010. Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. Cell Mol Life Sci **67**:677-700.

Nigumann, P., K. Redik, K. Matlik, and M. Speek. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics **79**:628-634.

Noonan, J. P., and A. S. McCallion. 2010. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet **11**:1-23.

Nuzhdin, S. V., M. L. Wayne, K. L. Harmon, and L. M. McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol Biol Evol **21**:1308-1317.

Oakley, T. H., Z. Gu, E. Abouheif, N. H. Patel, and W. H. Li. 2005. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. Mol Biol Evol **22**:40-50.

Okamura, K., and E. C. Lai. 2008. Endogenous small interfering RNAs in animals. Nat Rev Mol Cell Biol **9**:673-678.

Orgel, L. E., and F. H. Crick. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604-607.

Ostertag, E. M., J. L. Goodier, Y. Zhang, and H. H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet **73**:1444-1451.

Ozsolak, F., P. Kapranov, S. Foissac, S. W. Kim, E. Fishilevich, A. P. Monaghan, B. John, and P. M. Milos. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell **143**:1018-1029.

Ozsolak, F., J. S. Song, X. S. Liu, and D. E. Fisher. 2007. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol **25**:244-248.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40**:1413-1415.

Papadopoulos, G. L., M. Reczko, V. A. Simossis, P. Sethupathy, and A. G. Hatzigeorgiou. 2009. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res **37**:D155-158.

Papatsenko, D., A. Kislyuk, M. Levine, and I. Dubchak. 2006. Conservation patterns in different functional sequence categories of divergent Drosophila species. Genomics **88**:431-442.

Park, J. W., and B. R. Graveley. 2007. Complex alternative splicing. Adv Exp Med Biol **623**:50-63.

Parkinson, H., U. Sarkans, N. Kolesnikov, et al. (22 co-authors). 2011. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res **39**:D1002-1004.

Parmley, J. L., A. O. Urrutia, L. Potrzebowski, H. Kaessmann, and L. D. Hurst. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol **5**:e14.

Pasquinelli, A. E., B. J. Reinhart, F. Slack, et al. (19 co-authors). 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature **408**:86-89.

Peifer, M., and W. Bender. 1986. The anterobithorax and bithorax mutations of the bithorax complex. EMBO J **5**:2293-2303.

Pennacchio, L. A., N. Ahituv, A. M. Moses, et al. (19 co-authors). 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature **444**:499-502.

Pereira, V., D. Enard, and A. Eyre-Walker. 2009. The effect of transposable element insertions on gene expression evolution in rodents. PLoS One **4**:e4321.

Pereira, V., D. Waxman, and A. Eyre-Walker. 2009. A problem with the correlation coefficient as a measure of gene expression divergence. Genetics **183**:1597-1600.

Ponting, C. P., P. L. Oliver, and W. Reik. 2009. Evolution and functions of long noncoding RNAs. Cell **136**:629-641.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**:841-842.

Raney, B. J., M. S. Cline, K. R. Rosenbloom, et al. (23 co-authors). 2011. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res **39**:D871-875.

Reed, H. C., T. Hoare, S. Thomsen, T. A. Weaver, R. A. White, M. Akam, and C. R. Alonso. 2010. Alternative splicing modulates Ubx protein function in Drosophila melanogaster. Genetics **184**:745-758.

Reimers, M. 2010. Making informed choices about microarray data analysis. PLoS Comput Biol **6**:e1000786.

Rhead, B., D. Karolchik, R. M. Kuhn, et al. (23 co-authors). 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res **38**:D613-619.

Rifkin, S. A., D. Houle, J. Kim, and K. P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438**:220-223.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet **33**:138-144.

Ringrose, L., and R. Paro. 2004. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu Rev Genet **38**:413-443.

Robinson, M. D., and T. P. Speed. 2007. A comparison of Affymetrix gene expression arrays. BMC Bioinformatics **8**:449.

Rockman, M. V., M. W. Hahn, N. Soranzo, F. Zimprich, D. B. Goldstein, and G. A. Wray. 2005. Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol **3**:e387.

Roux, J., and M. Robinson-Rechavi. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res.

Sabarinadh, C., S. Subramanian, A. Tripathi, and R. K. Mishra. 2004. Extreme conservation of noncoding DNA near HoxD complex of vertebrates. BMC Genomics **5**:75.

Sahai, H., and M. M. Ojeda. 2003. Analysis of variance for random models, Vol. II: Unbalanced data. Pp. 496 p. Birkhäuser, Boston, Mass.

Sandelin, A., P. Bailey, S. Bruce, P. G. Engstrom, J. M. Klos, W. W. Wasserman, J. Ericson, and B. Lenhard. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics **5**:99.

Saunders, A., L. J. Core, and J. T. Lis. 2006. Breaking barriers to transcription elongation. Nat Rev Mol Cell Biol **7**:557-567.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270**:467-470.

Schmidt, D., M. D. Wilson, B. Ballester, et al. (13 co-authors). 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science **328**:1036-1040.

Schwahn, U., S. Lenzner, J. Dong, et al. (16 co-authors). 1998. Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nat Genet **19**:327-332.

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. 2011. Global quantification of mammalian gene expression control. Nature **473**:337-342.

Schwartz, S., L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, E. D. Green, R. C. Hardison, and W. Miller. 2003a. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res **31**:3518-3524.

Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. 2003b. Human-mouse alignments with BLASTZ. Genome Res **13**:103-107.

Sharp, P. A. 1987. Splicing of messenger RNA precursors. Science **235**:766-771.

Shatkin, A. J. 1976. Capping of eucaryotic mRNAs. Cell **9**:645-653.

Siepel, A., G. Bejerano, J. S. Pedersen, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15**:1034-1050.

Simon, J., M. Peifer, W. Bender, and M. O'Connor. 1990. Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. EMBO J **9**:3945-3956.

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

Sokal, R. R., and F. J. Rohlf. 1995. Biometry : the principles and practice of statistics in biological research. Freeman, New York.

Somel, M., H. Creely, H. Franz, U. Mueller, M. Lachmann, P. Khaitovich, and S. Paabo. 2008. Human and chimpanzee gene expression differences replicated in mice fed different diets. PLoS One **3**:e1504.

Somel, M., H. Franz, Z. Yan, et al. (15 co-authors). 2009. Transcriptional neoteny in the human brain. Proc Natl Acad Sci U S A **106**:5743-5748.

Sonenberg, N., and A. G. Hinnebusch. 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell **136**:731-745.

Sorek, R., G. Ast, and D. Graur. 2002. Alu-containing exons are alternatively spliced. Genome Res **12**:1060-1067.

Spradling, A. C., D. M. Stern, I. Kiss, J. Roote, T. Laverty, and G. M. Rubin. 1995. Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. Proc Natl Acad Sci U S A **92**:10824-10830.

Stark, A., M. F. Lin, P. Kheradpour, et al. (44 co-authors). 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature **450**:219-232.

Stewart, M. 2010. Nuclear export of mRNA. Trends Biochem Sci **35**:609-617.

Stranger, B. E., M. S. Forrest, M. Dunning, et al. (17 co-authors). 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science **315**:848-853.

Subramaniam, V., H. M. Bomze, and A. J. Lopez. 1994. Functional differences between Ultrabithorax protein isoforms in Drosophila melanogaster: evidence from elimination, substitution and ectopic expression of specific isoforms. Genetics **136**:979-991.

Svensson, O., L. Arvestad, and J. Lagergren. 2006. Genome-wide survey for biologically functional pseudogenes. PLoS Comput Biol **2**:e46.

Tazi, J., N. Bakkour, and S. Stamm. 2009. Alternative splicing and disease. Biochim Biophys Acta **1792**:14-26.

The 1000 Genomes Project Consortium, R. M. Durbin, G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. A map of human genome variation from population-scale sequencing. Nature **467**:1061-1073.

The modENCODE Consortium, S. Roy, J. Ernst, et al. (97 co-authors). 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science **330**:1787-1797.

Thomas, M. C., and C. M. Chiang. 2006. The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol **41**:105-178.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, et al. (19 co-authors). 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet **39**:31-40.

Toleno, D. M., G. Renaud, T. G. Wolfsberg, M. Islam, D. E. Wildman, K. D. Siegmund, and J. G. Hacia. 2009. Development and evaluation of new mask protocols for gene expression profiling in humans and chimpanzees. BMC Bioinformatics **10**:77.

Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell **10**:1453-1465.

Torgerson, D. G., A. R. Boyko, R. D. Hernandez, et al. (11 co-authors). 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet **5**:e1000592.

Urrutia, A. O., L. B. Ocana, and L. D. Hurst. 2008. Do Alu repeats drive the evolution of the primate transcriptome? Genome Biol **9**:R25.

van de Lagemaat, L. N., J. R. Landry, D. L. Mager, and P. Medstrand. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet **19**:530-536.

Varki, A., D. H. Geschwind, and E. E. Eichler. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. Nat Rev Genet **9**:749-763.

Venkataraman, S., P. Stevenson, Y. Yang, L. Richardson, N. Burton, T. P. Perry, P. Smith, R. A. Baldock, D. R. Davidson, and J. H. Christiansen. 2008. EMAGE--Edinburgh Mouse Atlas of Gene Expression: 2008 update. Nucleic Acids Res **36**:D860-865.

Venter, J. C., M. D. Adams, E. W. Myers, et al. (273 co-authors). 2001. The sequence of the human genome. Science **291**:1304-1351.

Verdugo, R. A., C. F. Deschepper, G. Munoz, D. Pomp, and G. A. Churchill. 2009. Importance of randomization in microarray experimental designs with Illumina platforms. Nucleic Acids Res **37**:5610-5618.

Visel, A., J. A. Akiyama, M. Shoukry, V. Afzal, E. M. Rubin, and L. A. Pennacchio. 2009. Functional autonomy of distant-acting human enhancers. Genomics **93**:509-513.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin, and L. A. Pennacchio. 2008.

Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet **40**:158-160.

Visel, A., E. M. Rubin, and L. A. Pennacchio. 2009. Genomic views of distant-acting enhancers. Nature **461**:199-205.

Voelker, R. A., J. Graves, W. Gibson, and M. Eisenberg. 1990. Mobile element insertions causing mutations in the Drosophila suppressor of sable locus occur in DNase I hypersensitive subregions of 5'-transcribed nontranslated sequences. Genetics **126**:1071-1082.

Walsh, C. P., J. R. Chaillet, and T. H. Bestor. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet **20**:116-117.

Walsh, C. T., S. Garneau-Tsodikova, and G. J. Gatto, Jr. 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl **44**:7342-7372.

Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature **456**:470-476.

Wang, J., A. P. Lee, R. Kodzius, S. Brenner, and B. Venkatesh. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. Mol Biol Evol **26**:487-490.

Warnefors, M. 2007. MSc thesis: Evolution of alternative splicing in the *Hox* gene *Ultrabithorax*. Lund University.

Warren, W. C., D. F. Clayton, H. Ellegren, et al. (81 co-authors). 2010. The genome of a songbird. Nature **464**:757-762.

Whitehead, A., and D. L. Crawford. 2006. Neutral and adaptive variation in gene expression. Proc Natl Acad Sci U S A **103**:5425-5430.

Whitlock, M. C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol **18**:1368-1373.

Wolf, Y. I., P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman. 2009. Inaugural Article: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A **106**:7273-7280.

Wray, G. A. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet **8**:206-216.

Yajima, I., S. Sato, T. Kimura, K. Yasumoto, S. Shibahara, C. R. Goding, and H. Yamamoto. 1999. An L1 element intronic insertion in the black-eyed white (Mitf[mi-bw]) gene: the loss of a single Mitf isoform responsible for the pigmentary defect and inner ear deafness. Hum Mol Genet **8**:1431-1441.

Yanai, I., D. Graur, and R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8**:15-24.

Yang, N., and H. H. Kazazian, Jr. 2006. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol **13**:763-771.

Young, M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol **11**:R14.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol **18**:292-298.

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

Zhang, P., Z. Gu, and W. H. Li. 2003. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol **4**:R56.

GOslim	Mechanism	Genes	r	p value	
Binding (MF)	*TFBS*	4492	0.11	$1.9 * 10^{-10}$	***
	Cons.	6308	0.06	$2.3 * 10^{-3}$	**
	Alt. prom.	6263	0.13	$7.9 * 10^{-14}$	***
	Alt. splic.	6263	0.12	$7.9 * 10^{-14}$	***
	Alt. polyA	6263	0.13	$7.9 * 10^{-14}$	***
	miRNA	6308	0.05	$5.6 * 10^{-2}$	
	NMD	6263	0.07	$3.6 * 10^{/-5}$	***
	RNA ed.	6248	0.09	$2.2 * 10^{9}$	***
Cellular process (BP)	*TFBS*	3222	0.12	$3.5 * 10^{-9}$	***
	Cons.	4438	0.04	1	
	Alt. prom.	4412	0.13	$7.9 * 10^{-14}$	***
	Alt. splic.	4412	0.12	$1.6 * 10^{-13}$	***
	Alt. polyA	4412	0.13	$7.9 * 10^{-14}$	***
	miRNA	4438	0.04	1	
	NMD	4412	0.09	$5.4 * 10^{-6}$	***
	RNA ed.	4398	0.08	$1.5 * 10^{-4}$	***
Protein binding (MF)	*TFBS*	3256	0.13	$1.4 * 10^{-10}$	***
	Cons.	4382	0.05	$2.2 * 10^{-1}$	
	Alt. prom.	4362	0.15	$7.9 * 10^{-14}$	***
	Alt. splic.	4362	0.14	$7.9 * 10^{-14}$	***
	Alt. polyA	4362	0.15	$7.9 * 10^{-14}$	***
	miRNA	4382	0.05	$1.4 * 10^{-1}$	
	NMD	4362	0.09	$6.9 * 10^{-7}$	***
	RNA ed.	4348	0.11	$3.3 * 10^{-10}$	***
Regulation of biological process (BP)	*TFBS*	2804	0.10	$5.2 * 10^{-5}$	***
	Cons.	4380	0.11	$2.6 * 10^{-10}$	***
	Alt. prom.	4354	0.18	$7.9 * 10^{-14}$	***
	Alt. splic.	4354	0.17	$7.9 * 10^{-14}$	***
	Alt. polyA	4354	0.17	$7.9 * 10^{-14}$	***
	miRNA	4380	0.04	1	
	NMD	4354	0.10	$3.6 * 10^{-9}$	***
	RNA ed.	4345	0.11	$1.3 * 10^{-11}$	***
Multicellular organismal process (BP)	*TFBS*	1324	0.13	1	
	Cons.	2512	0.14	$1.4 * 10^{-1}$	
	Alt. prom.	2500	0.20	$3.0 * 10^{-6}$	***
	Alt. splic.	2500	0.20	$8.8 * 10^{-7}$	***
	Alt. polyA	2500	0.19	$1.5 * 10^{-5}$	***
	miRNA	2512	0.05	1	
	NMD	2500	0.08	1	
	RNA ed.	2495	0.11	0.66	
Metabolic process (BP)	*TFBS*	1695	0.11	$1.6 * 10^{-3}$	**
	Cons.	2226	0.00	1	
	Alt. prom.	2214	0.11	$2.3 * 10^{-4}$	***
	Alt. splic.	2214	0.10	$7.9 * 10^{-3}$	**
	Alt. polyA	2214	0.11	$1.6 * 10^{-4}$	***
	miRNA	2226	0.05	1	
	NMD	2214	0.07	0.21	
	RNA ed.	2210	0.02	1	
Nucleic acid binding (MF)	*TFBS*	1523	0.07	1	
	Cons.	2052	0.02	1	
	Alt. prom.	2037	0.03	1	
	Alt. splic.	2037	0.02	1	
	Alt. polyA	2037	0.02	1	
	miRNA	2052	0.02	1	
	NMD	2037	-0.01	1	
	RNA ed.	2033	-0.00	1	
Response to stimulus (BP)	*TFBS*	1099	0.12	$3.3 * 10^{-2}$	*
	Cons.	1978	0.15	$6.0 * 10^{-9}$	***
	Alt. prom.	1966	0.23	$7.9 * 10^{-14}$	***
	Alt. splic.	1966	0.22	$7.9 * 10^{-14}$	***
	Alt. polyA	1966	0.23	$7.9 * 10^{-14}$	***
	miRNA	1978	0.09	$5.4 * 10^{-2}$	
	NMD	1966	0.14	$3.8 * 10^{-8}$	***
	RNA ed.	1959	0.14	$4.6 * 10^{-8}$	***
Macromolecule metabolic process (BP)	*TFBS*	1310	0.09	$5.0 * 10^{-1}$	
	Cons.	1657	-0.01	1	
	Alt. prom.	1653	0.10	$2.2 * 10^{-2}$	*
	Alt. splic.	1653	0.09	$1.3 * 10^{-1}$	
	Alt. polyA	1653	0.09	$4.1 * 10^{-2}$	*
	miRNA	1657	0.04	1	
	NMD	1653	0.08	$6.6 * 10^{-1}$	
	RNA ed.	1650	0.01	1	

GOslim	Mechanism	Genes	r	p value	
Catalytic activity (MF)	*TFBS*	1069	0.05	1	
	Cons.	1449	0.08	$4.7 * 10^{-1}$	
	Alt. prom.	1432	0.09	$2.4 * 10^{-1}$	
	Alt. splic.	1432	0.07	1	
	Alt. polyA	1432	0.08	1	
	miRNA	1449	0.06	1	
	NMD	1432	0.04	1	
	RNA ed.	1428	0.00	1	
Signal transducer activity (MF)	*TFBS*	641	0.02	1	
	Cons.	1419	0.17	$6.8 * 10^{-8}$	***
	Alt. prom.	1404	0.15	$8.0 * 10^{-6}$	***
	Alt. splic.	1404	0.14	$8.3 * 10^{-5}$	***
	Alt. polyA	1404	0.14	$1.2 * 10^{-4}$	***
	miRNA	1419	0.06	1	
	NMD	1404	0.03	1	
	RNA ed.	1402	0.07	1	
Transport (BP)	*TFBS*	954	0.13	$3.8 * 10^{-2}$	*
	Cons.	1307	-0.00	1	
	Alt. prom.	1303	0.13	$1.2 * 10^{-3}$	**
	Alt. splic.	1303	0.12	$6.4 * 10^{-3}$	**
	Alt. polyA	1303	0.14	$9.1 * 10^{-5}$	
	miRNA	1307	0.07	1	
	NMD	1303	0.07	1	
	RNA ed.	1301	0.10	$1.3 * 10^{-1}$	
Transcription regulator activity (MF)	*TFBS*	829	0.04	1	
	Cons.	1173	0.03	1	
	Alt. prom.	1166	0.02	1	
	Alt. splic.	1166	0.01	1	
	Alt. polyA	1166	0.02	1	
	miRNA	1173	-0.00	1	
	NMD	1166	-0.01	1	
	RNA ed.	1164	0.01	1	
Multicellular organismal development (BP)	*TFBS*	1006	0.09	1	
	Cons.	1673	0.09	$1.4 * 10^{-1}$	
	Alt. prom.	1666	0.14	$3.0 * 10^{-6}$	***
	Alt. splic.	1666	0.13	$8.8 * 10^{-5}$	***
	Alt. polyA	1666	0.13	$1.5 * 10^{-5}$	***
	miRNA	1673	0.05	1	
	NMD	1666	0.04	1	
	RNA ed.	1662	0.08	$6.6 * 10^{-1}$	
Receptor activity (MF)	*TFBS*	467	-0.03	1	
	Cons.	1163	0.15	$4.1 * 10^{-5}$	***
	Alt. prom.	1148	0.05	1	
	Alt. splic.	1148	0.04	1	
	Alt. polyA	1148	0.03	1	
	miRNA	1163	0.09	$4.4 * 10^{-1}$	
	NMD	1148	0.00	1	
	RNA ed.	1146	0.04	1	
Cell differentiation (BP)	*TFBS*	600	0.09	1	
	Cons.	965	0.10	$6.2 * 10^{-1}$	
	Alt. prom.	962	0.10	$7.3 * 10^{-1}$	
	Alt. splic.	962	0.08	1	
	Alt. polyA	962	0.10	1	
	miRNA	965	0.05	1	
	NMD	960	0.04	1	
	RNA ed.	962	0.07	1	
Nucleob, nucleos., nucleot. and nucleic acid metab.proc. (BP)	*TFBS*	680	0.06	1	
	Cons.	817	-0.03	1	
	Alt. prom.	814	0.06	1	
	Alt. splic.	814	0.05	1	
	Alt. polyA	814	0.06	1	
	miRNA	817	0.01	1	
	NMD	814	0.06	1	
	RNA ed.	811	-0.00	1	
Biosynthetic process (BP)	*TFBS*	560	0.06	1	
	Cons.	702	-0.04	1	
	Alt. prom.	697	0.06	1	
	Alt. splic.	697	0.04	1	
	Alt. polyA	697	0.07	1	
	miRNA	702	0.01	1	
	NMD	697	0.05	1	
	RNA ed.	696	-0.00	1	

GOslim	Mechanism	Genes	r	p value	
Hydrolase activity (MF)	*TFBS*	482	0.02	1	
	Cons.	682	0.13	$1.8 * 10^{-1}$	
	Alt. prom.	671	0.11	1	
	Alt. splic.	671	0.11	1	
	Alt. polyA	671	0.06	$6.3 * 10^{-1}$	
	miRNA	682	0.12	1	
	NMD	671	0.07	1	
	RNA ed.	668	0.05	1	
Enzyme regulator activity (MF)	*TFBS*	350	0.14	1	
	Cons.	524	0.05	1	
	Alt. prom.	522	0.23	$4.5 * 10^{-5}$	***
	Alt. splic.	522	0.22	$9.9 * 10^{-5}$	***
	Alt. polyA	522	0.23	$4.4 * 10^{-5}$	***
	miRNA	524	-0.03	1	
	NMD	522	0.19	$4.3 * 10^{-3}$	**
	RNA ed.	519	0.19	$5.7 * 10^{-3}$	**
Transporter activity (MF)	*TFBS*	335	0.12	1	
	Cons.	490	0.05	1	
	Alt. prom.	489	0.10	1	
	Alt. splic.	489	0.06	1	
	Alt. polyA	489	0.12	1	
	miRNA	490	0.07	1	
	NMD	489	0.02	1	
	RNA ed.	489	0.07	1	
Cell communication (BP)	*TFBS*	291	0.11	1	
	Cons.	491	0.17	$4.3 * 10^{-2}$	*
	Alt. prom.	489	0.18	$2.0 * 10^{-2}$	*
	Alt. splic.	489	0.19	$1.0 * 10^{-2}$	**
	Alt. polyA	489	0.17	$6.9 * 10^{-2}$	*
	miRNA	491	-0.00	1	
	NMD	489	0.10	1	
	RNA ed.	488	0.11	1	
Multi-organism process (BP)	*TFBS*	280	0.23	$3.1 * 10^{-2}$	*
	Cons.	432	0.14	1	
	Alt. prom.	431	0.25	$7.6 * 10^{-5}$	***
	Alt. splic.	431	0.23	$3.1 * 10^{-5}$	
	Alt. polyA	431	0.24	$9.4 * 10^{-5}$	
	miRNA	432	0.06	1	
	NMD	431	0.12	1	
	RNA ed.	429	0.13	1	
Transferase activity (MF)	*TFBS*	313	0.09	1	
	Cons.	425	-0.02	1	
	Alt. prom.	420	-0.05	1	
	Alt. splic.	420	-0.10	1	
	Alt. polyA	420	-0.09	1	
	miRNA	425	0.02	1	
	NMD	420	-0.02	1	
	RNA ed.	420	-0.09	1	
Cell death (BP)	*TFBS*	300	0.02	1	
	Cons.	398	0.02	1	
	Alt. prom.	395	0.18	$8.1 * 10^{-2}$	*
	Alt. splic.	395	0.20	$5.6 * 10^{-2}$	*
	Alt. polyA	395	0.16	$3.9 * 10^{-1}$	
	miRNA	398	0.14	1	
	NMD	395	0.06	1	
	RNA ed.	393	0.22	$3.4 * 10^{-3}$	**
Catabolic process (BP)	*TFBS*	294	0.12	1	
	Cons.	369	-0.01	1	
	Alt. prom.	369	0.14	1	
	Alt. splic.	369	0.14	1	
	Alt. polyA	369	0.16	1	
	miRNA	369	0.07	1	
	NMD	369	0.14	1	
	RNA ed.	369	0.10	1	
Cellular component movement (BP)	*TFBS*	176	0.18	1	
	Cons.	281	0.19	$4.5 * 10^{-1}$	
	Alt. prom.	280	0.19	$4.0 * 10^{-1}$	
	Alt. splic.	280	0.18	1	
	Alt. polyA	280	0.18	$9.4 * 10^{-1}$	
	miRNA	281	0.08	1	
	NMD	280	0.08	1	
	RNA ed.	279	0.11	1	

GOslim	Mechanism	Genes	r	p value	
Behavior (BP)	*TFBS*	164	0.13	1	
	Cons.	280	0.28	$9.2 * 10^{-4}$	***
	Alt. prom.	280	0.30	$8.0 * 10^{-5}$	***
	Alt. splic.	280	0.29	$3.8 * 10^{-4}$	***
	Alt. polyA	280	0.30	$1.7 * 10^{-4}$	***
	miRNA	280	0.05	1	
	NMD	280	0.17	1	
	RNA ed.	280	0.14	1	
Ion transmembrane transporter activity (MF)	*TFBS*	166	0.09	1	
	Cons.	236	0.05	1	
	Alt. prom.	236	0.07	1	
	Alt. splic.	236	0.02	1	
	Alt. polyA	236	0.10	1	
	miRNA	236	0.02	1	
	NMD	236	0.06	1	
	RNA ed.	236	0.04	1	
Ligase activity (MF)	*TFBS*	173	-0.06	1	
	Cons.	208	0.09	1	
	Alt. prom.	208	0.03	1	
	Alt. splic.	208	0.01	1	
	Alt. polyA	208	0.03	1	
	miRNA	208	0.05	1	
	NMD	208	0.09	1	
	RNA ed.	206	-0.05	1	
Structural molecule activity (MF)	*TFBS*	132	0.24	1	
	Cons.	208	0.06	1	
	Alt. prom.	207	0.12	1	
	Alt. splic.	207	0.12	1	
	Alt. polyA	207	0.11	1	
	miRNA	208	0.01	1	
	NMD	207	0.06	1	
	RNA ed.	206	0.08	1	
Secretion (BP)	*TFBS*	111	0.10	1	
	Cons.	175	-0.01	1	
	Alt. prom.	174	0.23	$9.5 * 10^{-1}$	
	Alt. splic.	174	0.23	$7.2 * 10^{-1}$	
	Alt. polyA	174	0.25	$3.5 * 10^{-2}$	*
	miRNA	175	0.11	1	
	NMD	174	0.17	1	
	RNA ed.	174	0.23	$8.2 * 10^{-1}$	
Channel activity (MF)	*TFBS*	106	0.08	1	
	Cons.	167	0.03	1	
	Alt. prom.	167	0.05	1	
	Alt. splic.	167	-0.01	1	
	Alt. polyA	167	0.06	1	
	miRNA	167	-0.07	1	
	NMD	167	0.03	1	
	RNA ed.	167	0.04	1	
Oxidoreductase activity (MF)	*TFBS*	107	0.03	1	
	Cons.	141	0.16	1	
	Alt. prom.	138	0.19	1	
	Alt. splic.	138	0.19	1	
	Alt. polyA	138	0.21	1	
	miRNA	141	0.05	1	
	NMD	138	-0.01	1	
	RNA ed.	138	-0.17	1	
Kinase activity (MF)	*TFBS*	89	0.26	1	
	Cons.	126	0.01	1	
	Alt. prom.	123	-0.02	1	
	Alt. splic.	123	-0.06	1	
	Alt. polyA	123	-0.06	1	
	miRNA	126	0.03	1	
	NMD	123	-0.00	1	
	RNA ed.	123	-0.03	1	
Motor activity (MF)	*TFBS*	48	0.09	1	
	Cons.	69	0.00	1	
	Alt. prom.	63	-0.03	1	
	Alt. splic.	63	-0.11	1	
	Alt. polyA	63	-0.08	1	
	miRNA	69	0.03	1	
	NMD	63	0.07	1	
	RNA ed.	62	-0.24	1	

GOslim	Mechanism	Genes	r	p value	
Protein transporter activity (MF)	*TFBS*	56	0.13	1	
	Cons.	62	0.17	1	
	Alt. prom.	62	-0.03	1	
	Alt. splic.	62	0.03	1	
	Alt. polyA	62	0.02	1	
	miRNA	62	0.15	1	
	NMD	62	-0.11	1	
	RNA ed.	62	0.05	1	
Cellular amino acid and derivative metabolic process (BP)	*TFBS*	43	-0.03	1	
	Cons.	60	-0.14	1	
	Alt. prom.	59	0.11	1	
	Alt. splic.	59	0.06	1	
	Alt. polyA	59	0.14	1	
	miRNA	59	n/a	n/a	
	NMD	59	0.07	1	
	RNA ed.	59	-0.04	1	
Extracellular structure organization (BP)	*TFBS*	23	-0.33	1	
	Cons.	50	-0.13	1	
	Alt. prom.	50	0.22	$1.2 * 10^{-2}$	*
	Alt. splic.	50	0.29	$5.9 * 10^{-1}$	
	Alt. polyA	50	0.26	$5.7 * 10^{-2}$	
	miRNA	50	0.10	1	
	NMD	50	0.09	1	
	RNA ed.	50	0.02	1	
Electron carrier activity (MF)	*TFBS*	27	0.23	1	
	Cons.	34	0.31	1	
	Alt. prom.	34	0.03	1	
	Alt. splic.	34	0.07	1	
	Alt. polyA	34	0.07	1	
	miRNA	34	0.16	1	
	NMD	34	0.07	1	
	RNA ed.	33	0.03	1	
Cellular membrane fusion (BP)	*TFBS*	24	0.09	1	
	Cons.	31	0.23	1	
	Alt. prom.	31	0.02	1	
	Alt. splic.	31	-0.00	1	
	Alt. polyA	31	0.03	1	
	miRNA	31	0.11	1	
	NMD	31	-0.15	1	
	RNA ed.	31	-0.34	1	
Lyase activity (MF)	*TFBS*	10	0.03	1	
	Cons.	11	-0.47	1	
	Alt. prom.	11	0.48	1	
	Alt. splic.	11	0.48	1	
	Alt. polyA	11	0.42	1	
	miRNA	11	n/a	n/a	
	NMD	11	-0.07	1	
	RNA ed.	11	0.27	1	
Isomerase activity (MF)	*TFBS*	9	0.38	1	
	Cons.	11	0.21	1	
	Alt. prom.	11	-0.09	1	
	Alt. splic.	11	-0.12	1	
	Alt. polyA	11	-0.22	1	
	miRNA	11	n/a	n/a	
	NMD	11	-0.67	1	
	RNA ed.	11	0.15	1	
Translation regulator activity (MF)	*TFBS*	9	0.39	1	
	Cons.	11	0.36	1	
	Alt. prom.	11	0.11	1	
	Alt. splic.	11	0.20	1	
	Alt. polyA	11	0.12	1	
	miRNA	11	0.23	1	
	NMD	11	0.17	1	
	RNA ed.	11	0.23	1	
Antioxidant activity (MF)	*TFBS*	4	-0.56	1	
	Cons.	7	0.11	1	
	Alt. prom.	7	-0.06	1	
	Alt. splic.	7	-0.05	1	
	Alt. polyA	7	-0.14	1	
	miRNA	7	n/a	n/a	
	NMD	7	-0.41	1	
	RNA ed.	7	n/a	n/a	

GOslim	Mechanism	Genes	r	p value	
Pathogenesis (BP)	*TFBS*	0	n/a	n/a	
	Cons.	0	n/a	n/a	
	Alt. prom.	0	n/a	n/a	
	Alt. splic.	0	n/a	n/a	
	Alt. polyA	0	n/a	n/a	
	miRNA	0	n/a	n/a	
	NMD	0	n/a	n/a	
	RNA ed.	0	n/a	n/a	
Helicase activity (MF)	*TFBS*	0	n/a	n/a	
	Cons.	0	n/a	n/a	
	Alt. prom.	0	n/a	n/a	
	Alt. splic.	0	n/a	n/a	
	Alt. polyA	0	n/a	n/a	
	miRNA	0	n/a	n/a	
	NMD	0	n/a	n/a	
	RNA ed.	0	n/a	n/a	