Effect of gene structure changes on the rate of protein sequence evolution

by

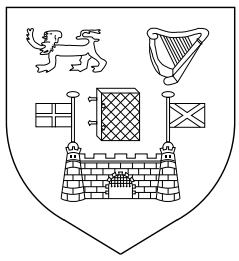
Brian Cusack

B.Sc. M.Res.

A Thesis submitted to
The University of Dublin
for the degree of

Doctor of Philosophy

Department of Genetics
Trinity College
University of Dublin



October, 2006

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Brian Cusack.

Signature of Author .....

Brian Cusack

31st October, 2006

Acknowledgements

Ken - thanks for your patient supervision and encouragement through the well-judged application of both stick and carrot.

Thanks to all current and past members of the Wolfe Lab for providing a great working environment. Thanks to Devin, Gavin, Jeff, Jonathan, Kevin, Marie, Matt, Meg, Nadia and Nora for their good humour and willingness to help. Thanks to Gavin for help with his like-tri-test software, to Marie for help with LaTeX and to Meg for knocking my grammar into shape. Many thanks to Andrew for inspiring the work in Chapter 2.

I am particularly grateful to my mother and brother for their constant support.

To my parents.

Contents

List of Figures

List of Tables

Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	base pairs
CAI	Codon Adaptation Index
cDNA	complementary DNA
DPE	Downstream Promoter Element
E-value	Expectation value
ESE	Exon Splicing Enhancer
ESS	Exon Splicing Silencer
EST	Expressed Sequence Tag
kb	kilobase
Mb	megabase
Mya	Million years ago
Myr	Million years
NMD	Nonsense-mediated decay
ORF	Open Reading Frame
PTC	Premature Termination Codon
rRNA	ribosomal RNA
UTR	Untranslated region

"The race is not always to the swift, nor the battle to the strong...
but time and chance happen to them all."

Ecclesiastes 9:11

Summary

The elaborate architecture of the genes of multicellular eukaryotes is likely to underpin the unique complexity of eukaryotic gene functions. The structure of eukaryotic genes differs from that of prokaryotes and represents an assemblage of coding exons, introns that are spliced out of precursor mRNAs, extended UTRs and complex regulatory regions. It is likely that these features provided a platform for the evolution of the complex traits that typify metazoans including alternative splicing and complex gene regulation.

Here I performed genome-wide studies of the association between the rate of protein sequence evolution and the modification of gene structures that can result from the processes of gene duplication and alternative splicing. By considering recent gene duplicates in rodents I investigated genomic relocation following duplication and gene structure alteration by retrotransposition as possible determinants of evolutionary rate differences between duplicates. I found evidence that retrotransposition frequently results in asymmetric evolution of gene duplicates and that functional retrogenes consistently accelerate relative to their paralogs. Although the act of relocating a gene duplicate by transposition explains part of this effect my results show that the mechanism of retrotransposition makes an independent contribution to this acceleration. This is likely to reflect the fact that duplicates created by retrotransposition violate the assumption common to most theoretical models that gene duplicates are born equal. My results further suggest that the rate acceleration of functional retrogenes is likely to be mediated by changes in their expression.

Alternative splicing is a parallel route to the generation of functional diversity that is also associated with changes in the exon-intron structure of genes. The effect of changes in alternative splicing on evolutionary rate can be assessed by comparing evolutionary pattterns in genes where alternative splicing is species-specific to genes where it is conserved. I show that the existence of species-specific alternative exons in human and mouse orthologs is a result of recent gain of these exons. The gene structure alterations associated with

these gains have resulted in an acceleration in the rate of sequence evolution of constant regions of the encoded protein. Moreover, this effect is shown to strongly correlate with the frequency of incorporation of these new exons. I argue that this correlation reflects a causative relationship between these variables and demonstrates the impact on constitutive parts of proteins of the acquisition of functional alternative splice forms.

Finally I present evidence from a single gene study supporting the intuition that alternative splicing and gene duplication can be parallel and complementary routes to the generation of functional diversity. I describe a gene fusion event that created a bifunctional gene coding for two proteins by alternative splicing. This chimeric gene persists in the mangrove genome but has duplicated in poplar and undergone subfunctionalisation to re-form its constituent genes through the complementary degeneration of its exons. This example is a clear illustration of the partitioning of alternative splice forms by subfunctionalisation at the level of gene structure. I also discuss evidence that accelerated protein sequence evolution occurred simultaneously with the gene structure changes corresponding to the initial gene fusion and the subsequent gene fission following duplication.

These results support the assertion that modifications of eukaryotic gene structure are frequently accompanied by an increase in the rate of protein sequence evolution.

Chapter 1

Introduction

1.1 Preface

In the first part of this introduction I describe the state of the field in the study of protein sequence evolution and the ongoing quest for the determinants of the evolutionary rate of proteins. In the second part I address the impact on the rate of protein evolution of the processes of gene duplication and alternative splicing. This section also outlines the research chapters that investigate the impact on evolutionary rate of the changes in gene structure that are frequently associated with both of these phenomena.

1.2 Causes of variation in the rate of protein sequence evolution

Since the foundations for the theory of molecular evolution were laid over thirty years ago the concept that different proteins, and the genes that encode them, evolve at characteristic rates has become concrete. In a landmark study Zuckerkandl and Pauling demonstrated evidence not only for a molecular clock of protein evolution but also showed that this linear rate of accumulation of amino-acid changes differs among proteins (Zuckerkandl and Pauling, 1965). This study established that the slow evolution of cytochrome *c* was "spectacularly at variance" with the relatively fast evolving haemoglobin. Current perspectives on the variability of protein rates place histones and actins among the slowest evolving protein sequences while relaxins and the fibrinopeptides reside at the opposite pole in the spectrum of rates (Li, 1997).

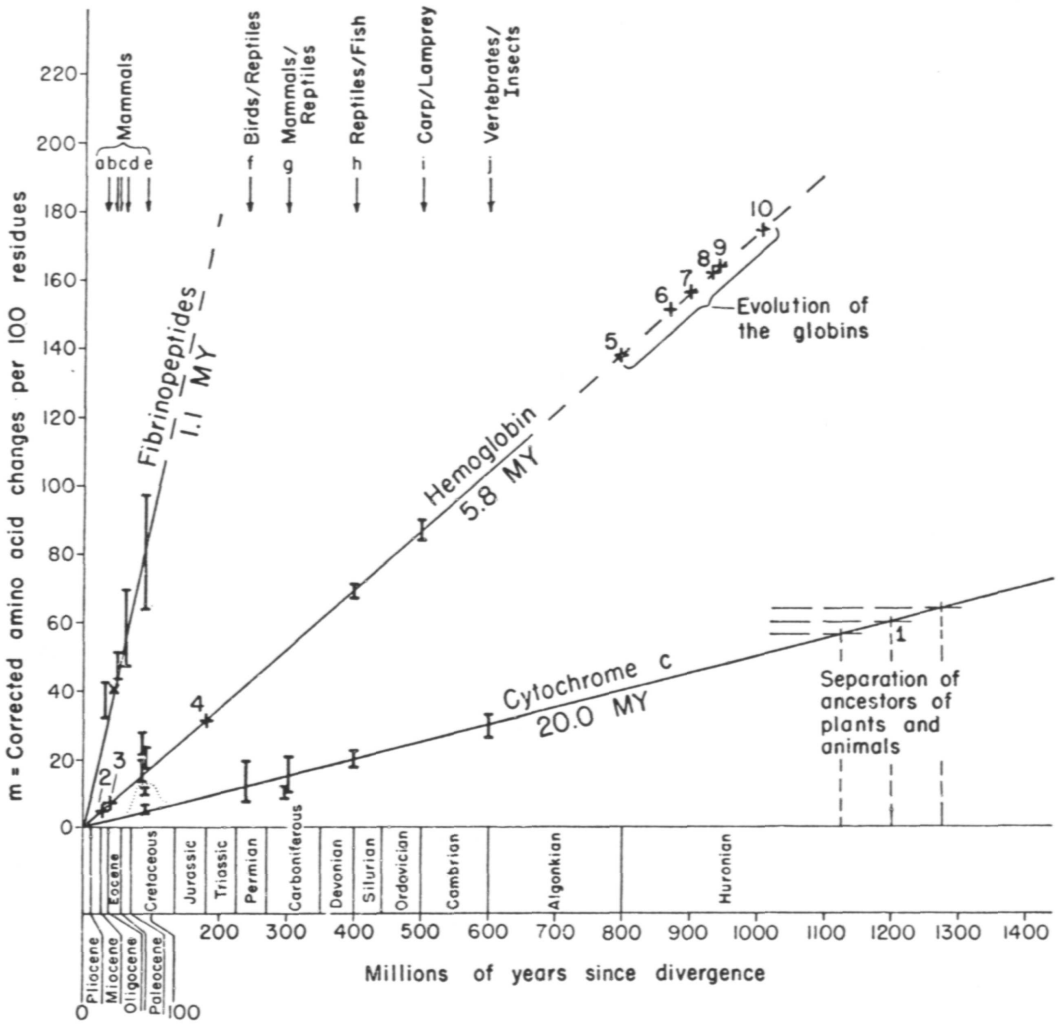


Figure 1-1: Rates of amino acid substitution in fibrinopeptides, haemoglobin, and cytochrome c. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point 1 represents a date of 1,200 ± 75 Myr for the separation of plants and animals, based on a linear extrapolation of the cytochrome c curve. Points 2-10 refer to events in the evolution of the globin family. The δ/β separation is at point 3, γ/β is at 4, and α/β is at 500 Myr (carp/lamprey). Reproduced from Dickerson (1971).

With the advent of the neutral theory of molecular evolution, Kimura proposed that advantageous mutations occur so infrequently as to make no significant contribution to rates of molecular evolution. This provided a rationale for the molecular clock by relegating the importance of advantageous mutations in favour of neutral changes that accumulate at a speed determined by the constant mutation rate. This result also provided a framework for understanding the characteristic rates of different proteins (Figure 1-1): the rate of evolution of a protein should simply reflect the relative proportions of constrained amino-acid sites

(where mutations are deleterious) and sites at which changes are neutral (with no effect on fitness) (Dickerson, 1971). Under the neutral theory (Kimura, 1983) the substitution rate per site (k) simply equals the neutral mutation rate per site (v_0). Furthermore, if a certain fraction (f_0) of mutations are neutral or nearly neutral and the rest are deleterious, then

$$k = v_0 = v_T f_0 \tag{1.1}$$

where v_T is the total rate of mutation. Under this model f_0 is a measurement of selective constraint on a sequence. Greater values of f_0 indicate that mutations at most sites are not selected against and are fixed at a faster rate. This predicts that less important proteins should evolve at faster rates (have greater values of k) because f_0 should be greater for less important proteins. This model explains the observation that pseudogenes, which are assumed to have no function, show the highest rates of nucleotide substitution because they are free of selective constraint ($f_0 = 1$) (Graur and Li, 2000).

Therefore, proteins that are functionally less important are assumed to evolve at faster rates reflecting the low level of selective constraint operating on them. It would appear reasonable to turn this statement around and use observed rates of sequence substitution to infer the intensity of selective constraint operating on a gene and therefore infer its functional importance. Despite the circularity of this logic (Graur and Li, 2000) the application of this principle has become common practice in molecular biology where sequence conservation is routinely used as a measure of functional importance. It has been suggested, for example, that the fast evolution of proteins such as fibrinopeptides may be due to the 'acceptability' of virtually any amino acid change in the protein sequence (Kimura and Ohta, 1974).

1.2.1 Early approaches to explaining protein rate variation

Is there a single factor that can explain the approximately 1,000-fold variation (Graur and Li, 2000) in the rate of evolution of different proteins? Many current hypotheses for the determinants of the protein rate are implicitly grounded in Zuckerkandl's concept of "functional density". This term considers the number (n_s) of amino acid sites in a protein that are involved in specific functions and cannot easily be substituted. Therefore the functional density of a protein (F) can be expressed simply as

$$F = n_s/N \tag{1.2}$$

where N is the total number of sites in the protein.

Intuitively this quantity should reflect the ratio of constrained to neutral amino acids for a given protein which should be directly proportional to its rate of sequence evolution. More recent work has led to an extension of this concept and the proposal of the term "fitness density" (see section 1.2.10, page 34).

In a pioneering study Dickerson (1971) suggested that the surface residues of a protein should be constrained by the protein's interactions with its partners. There are potentially many surface residues that could engage in such interactions relative to the handful of sites concerned with an enzyme's catalytic activity. Therefore these "contact functions" were proposed to make a relatively large contribution to the functional density of a protein. This assumption finds a contemporary echo in the proposal that proteins with high connectivity in protein-protein interaction (PPI) networks (i.e. high densities of contact functions) should evolve slowly (see section 1.2.6, page 29).

Tests of the impact of functional density on protein evolution are hindered by the absence of direct measurements of F (such as those provided by saturation mutagenesis). For those proteins for which functional density has been experimentally determined there is a rough negative correlation between F and the rate of protein evolution, k (Graur and Li, 2000). However, most work has attempted to explain variation in evolutionary rate using variables that are assumed to be adequate surrogates of functional density, such as expression level, pleiotropy, gene essentiality and gene dispensability.

One of the implications of Kimura's neutral theory of evolution is the prediction that important genes (those making the largest contributions to organismal fitness) should be subject to the strongest purifying selection. Wilson et al. (1977) therefore proposed that in addition to "functional density" the other major determinant of protein evolution is "dispensability" as formulated in the expression

$$k = PQ \tag{1.3}$$

where P is the probability that a substitution is compatible with the function of the protein and Q is the probability that the organism can survive and reproduce without the

protein, reflecting protein dispensability. In other words, P is a measure of the change in function of the mutant protein relative to the wild-type and Q scales this functional impact by the overall importance of the protein (i.e., its dispensability).

Therefore, predicting the effect of selection on the protein as a whole requires knowledge not only of the fraction of sites engaged in protein function but also of the impact of deleterious mutations of those sites on organismal survival. In modern biology (at least for unicellular organisms) a gene's dispensability is quantified using the reduction of growth rate relative to the wild-type to approximate the fitness effect associated with deletion of the gene. An alternative discrete classification distinguishes between essential and non-essential genes depending on whether deletion of the gene is lethal or not.

1.2.2 Codon-based models of protein evolution

Genome projects have allowed the evolution of proteins to be studied from the perspective of the nucleotide sequences that encode them. Codon-based analyses of protein-coding sequences treat the codon as the unit of evolution and distinguish between synonymous and nonsynonymous rates of evolution. Synonymous mutations yield a different codon without changing the encoded amino-acid and therefore do not affect the protein sequence. Nonsynonymous mutations, on the other hand, result in replacement of one amino-acid with another. This distinction enables the calculation of two substitution rates: d_S, the number of synonymous substitutions per synonymous site and d_N, the number of nonsynonymous substitutions per nonsynonymous site (Goldman and Yang, 1994; Muse and Gaut, 1994).

By distinguishing between synonymous (d_S) and nonsynonymous substitution rates (d_N) it is possible to draw inferences regarding the nature of the selection operating on the protein-coding sequence. In particular, the ratio of these rates (d_N/d_S) is commonly used to estimate ω (the amino acid selection pressure) corrected for π (the background nucleotide mutation rate). This follows from the fact that, because synonymous changes are silent at the protein level, synonymous sites are typically regarded as neutrally evolving (ignoring selection on codon usage). Therefore, the synonymous rate is dependent on the nucleotide mutation rate, π and not on amino acid selection pressure, ω. Nonsynonymous sites, on the other hand, evolve at a rate determined by both these processes.

In a neutrally evolving protein-coding sequence nonsynonymous mutations are as likely to be fixed as synonymous mutations (i.e., d_N/d_S is expected to equal one). This fact

has led to the common use of the ratio d_N/d_S to estimate the nature and magnitude of different types of amino acid selection pressure. Values of $d_N/d_S < 1$ indicate the operation of purifying selection in causing a reduction in the fixation rate of amino acid changes that are deleterious relative to the silent synonymous rate. Positive selection for beneficial amino acid changes is frequently inferred when $d_N/d_S > 1$.

Estimates of these rates are commonly derived in a maximum likelihood framework that starts with an explicit model of codon substitution and searches for the combination of parameter values that best describes the observed data. This approach accounts for unequal substitution rates for nucleotide transitions compared to transversions (the transition/transversion rate ratio, κ) as well as differences in codon frequencies. The model parameters estimated from the data include κ, the time t and the d_N/d_S ratio ω. This allows subsequent derivation of the rates d_N and d_S. The procedure simultaneously corrects for the occurrence of multiple substitutions at the same site and performs a realistic weighting of alternative pathways of change between codons (Yang and Bielawski, 2000).

The advantages of a codon-based perspective on the rate of protein evolution are threefold. First, when comparing evolutionary rates between different genes, the ratio d_N/d_S has the property of controlling for regional variation in mutation rates. Second, by considering the structure of the genetic code, the codon-based nonsynonymous divergence measure (d_N) accounts for the fact that some amino acid replacements can only be achieved with multiple nonsynonymous substitutions. This is not the case for measurements of protein sequence divergence which, as a result, are prone to underestimating divergence. Third, because amino acids with similar chemical properties are encoded by similar codons (Woese, 1965), codon-based methods will at least partially account for the differential probabilities of conservative and radical amino acid changes. Unweighted measurements of protein sequence divergence, on the other hand, regard all amino acid changes equally. These last two properties imply that the measure of non-synonymous divergence (d_N) for codon sequences should be superior to unweighted estimates of protein sequence divergence (k). However, weighted estimates of protein sequence divergence based on empirical amino acid substitution models do attempt to model differences in amino acid exchangeability (Dayhoff et al., 1972; Henikoff and Henikoff, 1992; Muller et al., 2002).

1.2.3 The impact of functional and comparative genomics

The development of high-throughput functional genomics methods in the recent past has enabled the re-appraisal of some early predictions in molecular evolution that were formulated largely from anecdotal examples. This has had a particularly significant impact on studies of the determinants of protein evolution, expanding on the early work of Zuckerkandl, Dickerson and Wilson. The benefits of this wealth of genomic data are however partly offset by the hidden cost of experimental noise. For example, measurements of gene expression are particularly noisy reflecting the combined effects of measurement inaccuracy and biological variability across growth conditions and strains (Coghlan and Wolfe, 2000; Drummond et al., 2006).

Furthermore, the new wealth of genomics data is not taxonomically well spread. Even among model organisms the unicellular budding yeast *Saccharomyces cerevisiae* has amassed the greatest variety and quantity of data. Accordingly, before attempting to explain the heterogeneity of protein rates in higher eukaryotes it is instructive to consider the extent to which protein rate variation can be explained using genomic approaches in yeast.

1.2.4 Pitfalls in interpreting genomic correlations

Notwithstanding the axiom that correlation does not equal causation, the obstacles to interpretation of genome-wide trends are both biological and statistical. First, measurements of the pair of variables under consideration are vulnerable to experimental limitations. Thus, a gene that is essential *in vivo* may be classified as 'non-essential' *in vitro* as a consequence of the limitations of laboratory growth conditions in mimicking conditions in the wild. In addition, pairwise comparisons of genomic variables may be less powerful when these variables are measured on incomparable phylogenetic timescales. This problem of phylogenetic scale is highlighted by the imperfect correlation between the short-term effect of deletion of a gene and its propensity for loss during evolution (a measure of the gene's long term dispensability) (Krylov et al., 2003; Wolf, 2006) and might reflect the evolutionary variability of a protein's importance (Zhang and He, 2005). Similar considerations may apply to quantifying gene expression levels. Expression data from exponentially growing yeast cells may have limited validity to growth under ecological conditions. Furthermore, if gene expression is itself an evolving trait then assays of current gene expression in yeast may be inadequate compared to measures such as codon adaptation index (which reflects long-term

expression-mediated selection on nucleotide substitutions).

A further major source of error is that an observed strong pairwise correlation may be induced as a trivial consequence of the mutual dependency of each variable on a third, confounding, variable. In this context, the deluge of genomics data has brought with it the paradoxical side-effect that large numbers of data points can suggest highly significant associations between variables that are only weakly correlated. In such a situation the task becomes one of disentangling the primary, evolutionarily relevant associations from secondary, induced, correlations (Koonin and Wolf, 2006).

A recent, far-reaching, suggestion is that approaches that try to remove the confounding effect of expression (e.g., partial correlation analysis) fail to do so when measurements of expression level are noise-prone (Drummond et al., 2006). The authors argued that techniques such as partial correlation analysis and multiple linear regression are inapplicable to situations where the variables under study intercorrelate (are "collinear") and are further undermined by measurement noise. Simulations showed that highly significant but entirely spurious partial correlations can be detected between unrelated variables when analysing noisy data and crucially this might underlie the significant partial correlation between the rate of protein evolution and dispensability (Hirsh and Fraser, 2001; Pal et al., 2003; Wall et al., 2005) that remains after attempting to control for noise-prone measurements of expression level. An alternative approach advocated by Drummond *et al.* is that of principal component regression (PCR) (Drummond et al. (2006); see section 1.2.8, page 32).

1.2.5 The controversy surrounding gene-dispensability

Of all the potential candidates that might determine the rate of protein evolution, essentiality and dispensability would seem to come closest to capturing the essence of a gene's 'importance'. The impact of gene essentiality on protein evolution should therefore be unequivocal: we would expect genes that are essential to organism survival (or fertility) to evolve slowly, reflecting the strong selective constraints on their function. However, the pitfalls described above beset the proposed association between the rate of protein evolution and any candidate explanatory variable. This is clearly illustrated by the controversy that has centred on the value of dispensability in explaining evolutionary rate, with the debate foundering on several sources of error.

The first study to use comparative genomics data to address this question reached the

surprising conclusion that, in mammals, there is no association between the fitness effect of a gene's deletion and its evolutionary rate once positively selected genes were excluded (Hurst and Smith, 1999). Although subsequent studies did claim to establish a connection, the association was found to be surprisingly weak (Hirsh and Fraser, 2001; Jordan et al., 2002). In fact, even this marginal effect was diluted in the light of evidence that expression level is a major predictor of evolutionary rate in yeast (Pal et al., 2001) and following use of partial correlation analysis to remove expression's confounding influence (Pal et al., 2003). More recent studies (Wall et al., 2005; Zhang and He, 2005; Drummond et al., 2006) have attempted not only to account for the confounding effect of expression level but also to address the problem of experimental noise that causes observed measurements to deviate from real values of the underlying biological variables. Two of these studies concluded that gene dispensability, although weak, is a significant and independent correlate of evolutionary rate once expression level is controlled for (Wall et al., 2005; Zhang and He, 2005). Moreover, it was suggested that the true association between dispensability and rate of protein evolution could only be uncovered when measuring sequence divergence on short evolutionary time scales which better approximate the instantaneous rate of protein evolution (thus illustrating the problem of phylogenetic scale (Herbeck and Wall, 2005)). However, the issue remains unresolved since by modelling the impact of noise on expression data one of these studies concluded that the apparent correlation between gene dispensability and evolutionary rate is spurious and results purely from noise in the measurement of expression level (Drummond et al., 2006).

1.2.6 Quantifying pleiotropy in yeast: protein interaction data

Pleiotropic mutations are those having multiple phenotypic effects. By extension pleiotropic genes are inferred to be multifunctional since their mutation may affect multiple phenotypic traits.

Pleiotropy should affect the rate of protein evolution in two distinct ways. The pattern of purifying selection on both functions of a bi-functional gene may be such that mutations improving one function might have deleterious consequences for the other function. However, not all pleiotropic scenarios are antagonistic in this way. In some cases the trade-offs between different functions are surprisingly low as has been demonstrated recently in directed evolution studies of multifunctional enzymes (Aharoni et al., 2005).

For multi-functional genes pleiotropic mutations will incurr a fitness cost amplified by the number of affected traits leading to stronger selective constraint on these mutations. Secondly, pleiotropy is thought to impede the process of adaptive evolution by reducing the likelihood that a mutation is advantageous (Fisher, 1930).

An interesting theoretical study implicates pleiotropy as a possible determinant of evolutionary rate. This study suggests that when many characters are affected by a mutation this leads to the predominance of a single optimal gene sequence. This leads to a reduction of within-population variation with a resultant lowering in substitution rate (Waxman and Peck, 1998).

Although pleiotropy is an important biological phenomenon an adequate measurement has proven elusive. There are several variables that might serve as proxies of pleiotropy and for which large-scale genomics data is available in yeast. Among these, the number of interactions in which a protein participates may be particularly informative. Therefore, proteins with many interaction partners ("hubs") might be considered to be multifunctional and are expected to show high levels of pleiotropy. However, the search for an independent correlation between protein evolutionary rate and the number of interaction partners has become mired in technical problems similar to those encountered in studies of the role of protein dispensability (Fraser et al., 2003; Jordan et al., 2003).

Despite these difficulties an appealing distinction has recently been drawn between protein-interaction hubs engaging in multiple, simultaneous interactions (intramodule "party" hubs) and those that interact with different partners at different times (intermodule "date" hubs). It was suggested that date hubs (having low coexpression with their interactors) are more pleiotropic than party hubs (exhibiting high coexpression with their interactors) because of their transient interactions with many, functionally semi-autonomous, modules (Fraser, 2005). However, the observation that party hubs are, in fact, more conserved than date hubs is contrary to expectation given the proposed difference in their pleiotropic level. Moreover, recent work has cast doubt on the meaningfulness of this distinction in hub types (Batada et al., 2006).

At first sight, this result would appear to relegate the importance of pleiotropy in protein evolution. However, the observation may be more readily explicable with reference to Dickerson's suggestion that the surfaces of proteins are highly constrained by the interactions in which they participate. Party hubs should exemplify this notion since by engaging

in multiple simultaneous interactions a large proportion of their surface residues is expected to be involved in interactions (i.e. the density of contact functions is high) with a resultant increase in the strength of purifying selection (Drummond et al., 2006; Rocha, 2006). Date hubs on the other hand may interact with their many partners through repeated interaction at the same site and are therefore likely to be less conserved, by definition.

Alternative approaches to quantifying pleiotropy have used the number of biological processes annotated for a gene to approximate the number of phenotypic traits it affects. However, less than 1% of the variation in selective constraint (measured by d_N/d_S) of yeast genes seems to be explained by this variable (Salathe et al., 2006). A similar result was obtained using the effects on growth of yeast mutants in 21 different conditions to quantify pleiotropy (Salathe et al., 2006). A parallel study found a similarly weak, although significant, association between a protein's evolutionary rate (measured by d_N) and the number of biological processes in which it participates. However, no correlation was found between protein conservation and other potential measurements of pleiotropy (e.g. number of annotated molecular functions and number of protein domains) (He and Zhang, 2006).

It seems, therefore, that even in well-studied model organisms such as yeast, an adequate description of pleiotropy remains tantalisingly out of reach.

1.2.7 Evolutionary rate and protein structure: the "designability" of proteins

According to the conventional view of protein activity the existence of a correctly folded three-dimensional structure is a prerequisite for protein function. However, protein structures differ with respect to their "designability", i.e., the number of possible sequences that can fold into that structure (Li et al., 1996; Koehl and Levitt, 2002). Highly designable structures are determined by a large "neighbourhood" of such sequences and this reflects their robustness to random mutations. It is therefore reasonable to expect that highly designable proteins evolve at faster rates than less designable proteins.

A recent study presented evidence of a positive association between protein designability and the rate of protein evolution (Bloom et al., 2006). This study used contact density (the average number of intramolecular contacts per residue) as a proxy for protein designability. A weak positive correlation between contact density and d_N was interpreted as meaning that roughly 5% of the variation in the rate of protein evolution is explicable by variation

in protein designability. This positive correlation could be considered as being at odds with Zuckerkandl's supposition that the contact density of proteins should correlate negatively with their evolutionary rate. The apparent contradiction may be explicable by the fact that Bloom *et al's* study only considered intramolecular contacts in calculating contact density. Therefore, the possibility of a negative correlation between the density of intermolecular contacts and rate of protein evolution is not rejected by this result.

It is possible that protein structural constraints will better explain variation in evolutionary rates among sites within a given protein, than rate differences between proteins. This is suggested by the fact that non-synonymous rates correlate with the solvent accessibility of residues, and are twice as fast on the surface of globular proteins than in buried regions (Goldman et al., 1998).

1.2.8 Most variation in rate of yeast protein evolution is explained by a single determinant

Expression level is frequently observed to be one of the strongest predictors of protein evolutionary rate. Techniques such as partial correlation analysis or multiple linear regression have been used in an attempt to reveal the primary association between protein rate and the focal variable by subtracting the secondary effect of expression. However, until recently, most studies did not seek to explain what underlies the recurrent association between expression level and the rate of protein evolution.

A recent study came to the striking conclusion that expression-related measures are by far the strongest predictor of evolutionary rate in yeast proteins (Drummond et al., 2006). Drummond *et al.* used Principal Component Regression (PCR) to examine the associations between evolutionary rate and seven variables that have previously been reported to independently predict protein evolutionary rate: gene expression level, protein abundance, codon adaptation index (CAI), dispensability, gene length, number of protein interaction partners and interaction network centrality. Following PCR analysis, the first principal component was found to be a composite variable that consists mostly of equal contributions from three related input variables: gene expression level, protein abundance and CAI. Strikingly, this principal component explains nearly half of the variance in the rate of yeast protein evolution. In contrast, all remaining components were each found to explain less than 1% of the variance in protein rate. Furthermore, this study rejected an independent

role for protein dispensability in protein evolution.

This final result is in striking contrast to a second study that used different methodology to address the same problem of measurement inaccuracy on partial correlation analysis (Wall et al., 2005). Using a structural equation model Wall *et al.* proposed that gene dispensability makes a small but significant contribution to the rate of protein evolution.

The fact that roughly half of the variability in protein rate remains to be explained suggests that other, unconsidered, causative variables may account for a significant degree of protein rate variation. This possibility is largely discounted by Drummond and coworkers on the grounds that the correlations they describe are necessarily underestimates due to the inherent stochasticity of the evolutionary process, attenuation by measurement noise and the possible non-linearity of the relationships between the predictors and evolutionary rate. However, given better surrogates of functional density and dispensability, these variables might be found to account for some fraction of the residual protein rate variation yet to be explained (Rocha, 2006).

1.2.9 Translational Robustness

The existence for each protein structure of a neighbourhood of compatible protein sequences was discussed above in the context of "protein designability" at the genotypic level. Parts of this neighbourhood are also explored at the phenotypic level as a consequence of errors in the translation of the genotype into the phenotype. The ribosome's error rate is estimated to cause the mistranslation of 20% of proteins and in many cases these mistranslated proteins may misfold (Drummond et al., 2005). However, some protein sequences reside in the middle of the "neighbourhood" of sequences that can each correctly determine the protein's native structure. As a result, when these "translationally robust" protein sequences are mistranslated, misfolding is avoided.

The cellular burden imposed by the toxicity and aggregation of misfolded proteins provides selective pressure for translational robustness. In fact the fitness cost of a misfolded protein is predicted to be proportional to its frequency of translation. Therefore the translational robustness hypothesis predicts that highly expressed proteins evolve slowly because they are under intense purifying selection to preserve those relatively rare sequences that are robust to mistranslation (Drummond et al., 2005).

This hypothesis provides a convincing explanation for the observation that expression-

related variables are the most important determinants of evolutionary rate in yeast. The underlying phenomenon captured by these variables is likely to be the frequency of translation of each gene. Therefore the production rate of yeast proteins appears to determine their evolutionary rate.

The paradoxical implication of this hypothesis is that "translationally robust" protein molecules are encoded by "mutationally fragile" genes. Thus while a considerable fraction of highly conserved sites in the primary sequence can be mutated (e.g. in site directed mutagenesis) with no inactivating effect on protein function, these mutations will be selected against to preserve translational robustness. This may explain the observation that genetic studies of the slowly evolving and highly abundant plant enzyme Rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase) have revealed very few inactivating mutations (Drummond et al., 2005). Therefore the sequence conservation of Rubisco to a large extent reflects translational robustness and not functional fragility.

1.2.10 Fitness density versus functional density

A fundamental consequence of the translational robustness hypothesis is that selection not directly related to protein function can also constrain the evolution of protein sequences. Therefore, in addition to the selective constraint operating on specific residues to conserve protein function (contributing to functional density) selection also operates on a sequence-wide background of residues not directly constrained by function to conserve translational robustness. Collectively these sites contribute to the "fitness density" of a protein i.e., the proportion of residues in a protein constrained by natural selection with each site weighted by the fitness effect of mutation (Pal et al., 2006). According to the definition of Pal *et al.*, fitness density is a measure of the change in fitness of the mutant *protein* (relative to the wildtype molecule). To determine the fitness difference of the individual mutant *organism* (relative to wildtype individuals) this measure must be scaled by the overall importance of the protein to the organism (Pal et al., 2006). Accordingly, the most important determinants of protein evolution should be fitness density and dispensability. However, as highlighted earlier the role of dispensability remains the subject of vigorous debate.

The difference between fitness density and functional density is best illustrated with reference to functionally similar pairs of paralogous genes. The yeast duplicates *URA5* and

URA10 (orotate phosphoribosyltransferases 1 and 2) differ more than 60-fold in expression level and six-fold in evolutionary rate with *URA5* being the more highly expressed and slower-evolving. Given the similar functions of these proteins a similar fraction of their residues are expected to be constrained by function (i.e., their functional densities should be equivalent). However, the higher expression level of *URA5* should increase selective constraint on the remaining residues to ensure correct folding in the event of mistranslation. Selection for translational robustness, therefore, increases fitness density of *URA5* compared to *URA10* while their functional densities should remain comparable (Drummond et al., 2005).

1.2.11 Determinants of evolutionary rate of mammalian proteins

Studies of the causes of protein rate variation in yeast may provide only a limited first approximation to explain the variability of rates in multicellular organisms such as mammals. The difficulties inherent in any extrapolation over broad phylogenetic distances are amplified by three sorts of evolutionary transition.

First, at a fundamental level the transition from large effective population sizes in unicellular eukaryotes to smaller effective population sizes in metazoans is expected to influence the efficiency of selection against deleterious mutations. Second, compared to unicellular organisms, the mammalian genome shows considerable heterogeneity with respect to both mutation rate and fixation rate. Third, the emergence of tissue and organ differentiation is likely to be associated with selective constraints unique to metazoans (e.g., mammals) compared to unicellular organisms (e.g., budding yeast). These last two evolutionary transitions may contribute to intra-genomic variability in the rates of mammalian protein evolution and will be considered in turn.

1.2.12 Heterogeneity of the mammalian genome

In mammals there is considerable genomic variability of both of the variables that dictate the neutral rate of protein evolution according to Kimura's formulation (equation 1.1). In other words different parts of the genome can differ both in their rate of mutation (v_T) and in their rate of fixation of mutations (f_0).

Genomic heterogeneity in mammalian mutation rate was originally observed as a variation in the synonymous substitution rates between mammalian genes (Wolfe et al., 1989)

and this has been corroborated using other measures of the rate of neutral substitution (Matassi et al., 1999; Lercher et al., 2001; Smith et al., 2002). What causes this regional variation of mutation rate? One possible explanation lies in the observation that GC content varies considerably across the mammalian genome, contributing to a genomic landscape of long ($>$ 300kb) regions of homogenous GC content ("isochores" (Eyre-Walker and Hurst, 2001)). Moreover, the neutral substitution rate is likely to positively correlate with GC content according to a non-equilibrium isochore model under which the
GC $\rightarrow$ AT rate is higher than the AT $\rightarrow$ GC rate (and both rates are constant across the genome) (Piganeau et al., 2002). Therefore, it has been suggested that the mutation rate of genes located in GC-rich regions should be greater than that of genes in GC-poor regions (Smith et al., 2002). A second possible explanation for intra-genomic variability in mutation rate is provided by variation in recombination rate in the mammalian genome (Kong et al., 2002). Because recombination is mutagenic in mammals (Hellmann et al., 2003; Lercher and Hurst, 2002) genes in highly recombining regions should have higher mutation rates than those residing in regions of low recombination rate.

Genomic heterogeneity is also seen in the fixation rate of mutations. This is a consequence of genome-wide variation in the balance between the efficiency of selection on the one hand and the power of genetic drift on the other. In fact the regional variation in recombination rate described above also plays a role in this type of within-genome variability. The efficiency of selection is greatest in highly recombining regions because the disruption of genetic linkage by recombination allows selection to act on single alleles without interference from alleles at neighbouring loci (i.e., Hill-Robertson effects are reduced). Therefore, purifying selection will be at its most efficient in regions of high recombination. If most mutations are deleterious this should mean that genes in highly-recombining regions should evolve more slowly than those in regions of low recombination.

It is currently unclear to what extent such regional genomic properties can explain the differences observed in the rates of mammalian protein evolution. With regard to variation in mutation rate, there is evidence from studies of neighbouring genes that local similarities in synonymous substitution rate are mirrored in similarities of non-synonymous rate (Williams and Hurst, 2000). On the other hand the fact that nonsynonymous rate variability is roughly four-fold greater than synonymous rate variability (Waterston et al., 2002) suggests that mutation rate differences have only limited potential to explain the

diversity of protein rates (but see Wyckoff et al. (2005)).

1.2.13 The transition to tissue differentiation

A major implication of the emergence of tissue differentiation is that the expression of a mammalian gene must be described not only in terms of its level but also in terms of the "breadth" of its tissue distribution, i.e. the number of tissues in which it is expressed.

The multiplicity of mammalian cell-types underlies an extraordinary diversity of highly differentiated tissues that adds two additional dimensions to the concept of gene pleiotropy. First, the developmental timing of gene expression during tissue differentiation might correlate with pleiotropy. According to the "hourglass model" intermediate developmental stages are highly conserved while earlier and later stages show greater evolutionary plasticity (Raff, 1996). Mutations in proteins expressed at intermediate stages in development are therefore expected to have greater pleiotropic effects and these proteins should evolve more slowly as a consequence. There is some support for this prediction in the case of mouse development (Castillo-Davis et al., 2004). The second potential correlate of pleiotropy in mammals is the tissue breadth of a gene's expression. Specifically, a situation of antagonistic pleiotropy might result if a new allelic variant that benefits a gene's function in one tissue is deleterious to its function in a different tissue. These mutations are expected to be eliminated efficiently by purifying selection leading to slower protein evolution.

1.2.14 Impact of breadth of expression on protein evolution in mammals

Early studies of the impact of gene expression pattern on the rate of amino acid substitution concluded, firstly, that the rate of a protein's evolution depended on which tissue it is expressed in (Kuma et al., 1995; Hughes, 1997) and, secondly, that tissue-specific proteins are more rapidly evolving than those expressed in a broad range of tissues (Hastings, 1996). There is therefore an apparent effect on the rate of protein evolution of "tissue identity" (the particular tissue in which a gene is expressed) in addition to "tissue breadth" (the count of tissues in which a gene is expressed). The latter association was explained in terms of an increase in functional constraint on proteins expressed in a range of diverse cellular environments. Therefore, the slow evolution of a ubiquitously expressed protein could be due to an increase in the functional density of a sequence resulting from the dual requirement that the protein should function under a wide range of physicochemical

conditions where it encounters a wide range of molecular interaction partners.

These early observations were extended by a genome-wide study of the relationship between the rate of protein evolution of 2400 human-rodent orthologs and their breadth of expression determined using expressed sequence tag (EST) data from 19 tissues (Duret and Mouchiroud, 2000). This study drew two major conclusions regarding protein rate variability. First, with regard to the effect of tissue breadth, it was shown that tissue-specific proteins evolve up to three times faster than ubiquitously expressed proteins. Second, the influence of tissue identity was reflected in the roughly 2.5 fold variation in the rate of protein evolution among genes having similar breadths but different tissue-specificities.

Duret and Mouchiroud (2000) proposed that the first of these differences is of larger magnitude than could be explained by Hasting's suggestion of increased functional constraint on broadly expressed genes due to inter-tissue variation in cellular environment. This led to the alternative explanation that the fitness effect of a mutation that is slightly deleterious to a gene's function is multiplied by the number of tissues in which the gene is expressed. Thus, Duret *et al.* attributed the slower evolution of ubiquitously expressed genes to an increase in the strength of selection proportional to the number of tissues in which a gene is expressed. This echoes the intuition that, in multicellular organisms, the breadth of expression of a gene should correlate with the gene's pleiotropic level. Therefore, the slower evolution of ubiquitously expressed compared to tissue-specific genes in mammals (Duret and Mouchiroud, 2000; Zhang and Li, 2004), should reflect an increase in constraint associated with increased pleiotropy.

Strikingly, Duret and Mouchiroud's second observation demonstrating the influence of tissue-identity implies that, for tissue-specific genes, inter-tissue differences account for much variation in the rate of protein evolution. However, they suggested that the slower evolution of brain-specific compared to liver-specific genes reflects the relatively peripheral role of the liver compared to the brain rather than reflecting inter-tissue variability in cellular environment. Thus, the more central role of the brain is expected to manifest itself in greater fitness effects of sequence changes among brain-specific proteins.

At first sight, it could be argued that the impact of tissue identity on rate provides an alternative explanation for Duret and Mouchiroud's primary observation that broadly expressed genes are more slowly evolving than tissue-specific genes. In this context, the slower rate of ubiquitous genes might not be determined by their broad expression *per se*

but could be a simple consequence of the gene's expression in a single rate-determining tissue (e.g. brain). Duret and Mouchiroud's two major results were borne out by a more recent study performed by Zhang and Li (2004). This study found a nearly two fold increase in the rate of non-synonymous divergence of tissue-specific genes compared to ubiquitously expressed "housekeeping" genes defined on the basis of microarray data. The large effect of tissue identity was confirmed by the finding that lung-specific proteins evolve on average nearly three times faster than muscle-specific genes. However, Zhang and Li also demonstrated that tissue-specific genes in the slowest evolving categories (i.e. brain and muscle) were significantly faster evolving than broadly expressed genes thus negating the possibility of a "rate-determining tissue". This result therefore supports the concept of an additive pleiotropic effect of expression breadth on the evolutionary rate of mammalian proteins.

1.2.15 Expression breadth versus tissue-specificity

Previous studies have claimed that the level of a gene's expression is highly correlated with its expression breadth (Lercher et al., 2002; Subramanian and Kumar, 2004). This is believed to reflect the assumption that housekeeping genes tend to be highly expressed (Vinogradov, 2004). However, the term "housekeeping gene" has occasionally been applied loosely (Lercher et al., 2002) and in a manner that has not always accorded with the strict definition of housekeeping genes as those genes that are always expressed in every tissue to maintain cellular functions (Watson et al., 1965). A more recent working definition has defined housekeeping genes as "those genes critical to the activities that must be carried out for successful completion of the cell cycle" (Warrington et al., 2000). Interestingly, this definition also encapsulates the concept of gene essentiality, highlighting the interrelatedness of ubiquitous expression and essentiality. Recent refinements of the conventional housekeeping gene concept have followed from two whole genome expression studies that have demonstrated that (i) housekeeping genes are not necessarily the most highly expressed genes in all tissues and (ii) the expression of housekeeping genes can be variable across tissues (Warrington et al., 2000; Hsiao et al., 2001).

It should be noted that part of the correlation between the level and breadth of a gene's expression is artefactual and stems from the use of an arbitrary cutoff to derive a measure of expression breadth by assigning a gene as either 'on' or 'off' in a given tissue. Cutoffs

have been applied to measure breadth of expression in the context of both microarray (Zhang and Li, 2004) and EST-based studies (Duret and Mouchiroud, 2000) leading to an intrinsic dependence of measured expression breadth on expression level (Liao and Zhang, 2006). For microarray data this dependency results from the use of signal intensity cutoffs whereas for expressed sequence-based measures it is a function of the sampling depth of EST libraries. This raises the possibility that previous observations of an association between the evolutionary rate of a protein and its tissue-specificity may have arisen due to the confounding influence of expression level (Duret and Mouchiroud, 2000). This may be particularly pertinent given the fact that, in yeast, expression level is the strongest predictor of the rate of protein evolution (Drummond et al., 2006).

This problem can be addressed using a recently proposed alternative measure of the tissue-distribution of a gene's expression. This "tissue-specificity index" (Yanai et al., 2005) does not rely on the use of expression cut-offs to distinguish between presence or absence of expression. Interestingly, this measure of tissue-specificity is found not to correlate with gene expression level, thus apparently overturning the long-standing assumption that housekeeping genes are expressed at high levels and in agreement with more recent results (Warrington et al., 2000; Hsiao et al., 2001). The lack of dependence of this measure on gene expression level allows the effect of tissue specificity on protein evolution to be assessed independently of the confounding influence of expression level. In fact, a statistically significant association was found between tissue-specificity index and both the rate of protein evolution (measured by d_N) and the strength of selective constraint (measured by d_N/d_S) (Liao et al., 2006). Therefore, previous claims that the evolutionary rate of a protein is correlated with its tissue-specificity remain robust. This has been separately confirmed using a partial correlation analysis approach: expression breadth and rate of mammalian protein evolution remain significantly correlated once expression level is controlled for (Martin Lercher, personal communication). However, the magnitude of this association appears to be small. At most 3% of the ordinal variation in protein rate in explained by ordinal variation in tissue-specificity (Liao et al., 2006).

1.2.16 Tissue-specificity and protein secretion

As a group, tissue-specific genes evolve at a faster rate than genes that are expressed ubiquitously. However, this poses the question of whether tissue-specificity alone is the

primary determinant of this effect or whether other possible properties distinguishing these groups of genes could account for the difference in evolutionary rates. In other words, does a classification of genes with respect to tissue specificity introduce a hidden bias with respect to some other potential determinant of the rate of protein evolution? For example, tissue-specific genes are likely to function more frequently in cell-cell communication and signal transduction roles compared to the more common metabolic activity of housekeeping genes.

Therefore, the unequivocal demonstration that tissue-specificity alone is responsible for accelerating the rate of mammalian protein evolution (e.g., through a reduction in pleiotropy) would require the comparison of proteins that differ only with respect to their breadth of expression but share all other relevant properties (e.g., have a common biochemical function).

One approach to disentangle the effects on protein evolution of tissue-specificity and functional differences is to consider evolutionary rates within gene families. According to this approach, if two paralogous genes that differ in their rate of evolution also differ in their expression breadth then the rate difference can be solely attributed to the difference in their breadth of expression. The common-ancestry (and presumed common biochemical function) of members of a gene family provides a control for the impact of functional differences on rate. An early study of this nature found that among 15 studied gene families, 14 showed a pattern of evolutionary rate consistent with the effect of expression breadth (Hastings, 1996). In these 14 families the slowest evolving member was found to be expressed in the broadest range of tissues.

More recent work has exposed one potential correlate of tissue-specificity that may explain some of the observed association between the rate of a protein's evolution and its expression profile. Two genome-wide studies have examined the effect of protein subcellular localisation on mammalian protein evolution. Part of the analysis of the completed mouse genome included an examination of the effect of subcellular location on protein evolution. Over 500 domain families were classified as secreted, cytoplasmic or nuclear and the strength of selective constraint operating on their sequences was estimated using d_N/d_S. Higher values of d_N/d_S among secreted domains revealed that these are subject to either weaker purifying selection or increased positive selection or both. Moreover, Winter et al. (2004) established a relationship between tissue-specificity and protein secretion by

showing that the most tissue-specific genes in human and mouse are more than three times more likely to be secreted than the most broadly expressed genes based on signal peptide predictions.

The study of Winter et al. (2004) enabled a reassessment of the effects on protein evolution of both tissue breadth and tissue identity in the light of the frequent secretion of tissue-specific proteins. When secreted and non-secreted proteins were considered separately two different trends in tissue breadth emerged. First, for secreted proteins a positive correlation between tissue-specificity and d_N/d_S (measuring selective constraint) confirmed that expression breadth does influence protein evolution independently of the effects of secretion (but see Julenius and Pedersen (2006)). In fact the converse was also true: secretion and d_N/d_S are correlated irrespective of tissue-specificity. However, the second observation that for non-secreted proteins there is no correlation between tissue-specificity and d_N/d_S suggests that whatever the impact of expression breadth on protein sequence conservation it is restricted to influencing the evolution of secreted proteins only.

In addition, the effect of tissue identity was also found to be independent of protein-secretion effects. As a case in point, among secreted proteins, those specific to brain were found to evolve more slowly than proteins specific to other tissues.

1.2.17 Is the translational robustness hypothesis phylogenetically robust?

Provisionally, the translational robustness hypothesis (section 1.2.9, page 33) represents the most compelling explanation for protein rate variation in yeast. However, it is an open question whether mammalian protein evolution is also governed by this phenomenon.

As the above discussion of studies of mammalian protein evolution indicates, the breadth of a gene's expression is a recurrent correlate of evolutionary rate that is independent of expression level. However, it is uncertain whether the reverse is true: partial correlation analysis suggests that expression level does not correlate with protein rate when expression breadth is controlled for (Martin Lercher, personal communication) although one study suggests otherwise (Subramanian and Kumar, 2004). In this context it is likely that the gene expression level of a unicellular organism and the average (or peak) gene expression level across tissues in a multicellular organism might not be strictly comparable measurements.

If the level of mammalian gene expression does not independently correlate with the

rate of non-synonymous evolution this implies that, in mammals, there is weaker selection against protein aggregation due to misfolding. However, protein aggregation is known to be associated with significant pathology in humans as evidenced by Alzheimer's disease and Huntington's disease. Although these examples are late-onset diseases with little effect on organismal fitness, they point to the potential for protein misfolding to impose significant selective costs in mammals. However, even if there is effective selective pressure for translational robustness on a given yeast protein it is possible that its mammalian ortholog may not be evolving under such selection. This may be true even if the selective coefficients for alleles with reduced robustness are of the same magnitude in yeast and mammals because the reduced effective population size of mammals relative to yeast will result in less efficient selection against slightly deleterious alleles. Therefore, an estimation of the magnitude of selective coefficients (or strength of selection, s) against mutations deleterious to translational robustness will enable an assessment of the likely importance of this hypothesis to mammalian protein evolution.

However, there is at least one possible obstacle to the assumption that the strength of selection associated with translational robustness is equal in orthologous proteins in yeast and mammals. One source of complication lies in the potential differences in chaperone function between these taxa. If mammalian chaperones are more efficient in rescuing misfolded proteins this might alleviate the selective pressure in mammals for translationally robust proteins.

Furthermore, any test of the translational robustness hypothesis in mammals is complicated by the plethora of mammalian cell types. As there is extensive variation in cell volumes between mammalian tissues, the quantity of interest is likely to be a protein's expression level in the tissue where it attains its highest cellular concentration. This quantity will determine the propensity for a protein to impose a cellular burden through misfolding.

In summary, despite the seductive simplicity of the translational robustness hypothesis as an explanation for protein rate variation in yeast it remains to be seen whether this model will be upheld in mammals.

1.3 Impact of gene duplication on rates of molecular evolution

Mutation is the sole source of the genetic variation on which natural selection operates. The constant supply of point mutations of single base pairs, insertions, deletions and duplications of DNA sequence and recombination between sequences is grist to the mill of natural selection. However, of all the raw materials that mutation provides for selection, the duplication of entire genes constitutes a highly valuable source of pre-tested molecular prototypes. Although the duplication of a gene is often characterised as opening an avenue for the emergence of functional novelty it may be more precise to consider it as a chance for molecular derivations on a pre-existing theme. Herein lies the major benefit of gene duplication to the organism: as long as the ancestral gene function is guaranteed by one copy, experimentation with its duplicate can proceed unchecked by selection. Therefore, gene duplication allows natural selection to engage in radical experimentation without forfeiting its innate conservatism.

The role of gene duplication in the origin of new gene functions implicates it as a potentially significant determinant of the rate of protein evolution.

1.3.1 The broad spectrum of gene duplications

Duplication of genetic material can occur on multiple physical scales extending from the largest to the smallest. At the most global level, duplication of entire genomes (polyploidisation) has the potential to maintain relative gene dosage and may be a route to the copying of entire pathways. In a landmark work Susumo Ohno proposed that polyploidisation therefore provides a unique opportunity to be seized by evolution (Ohno, 1970). At an intermediate scale, segmental chromosomal duplications can span multiple neighbouring genes and their regulatory sequences (Bailey et al., 2002). The duplication of single genes in their entirety can occur by DNA-based tandem duplication and duplicative transposition. However, some duplication events do not generate perfect facsimiles of their progenitor. Gene duplication by RNA-based retrotransposition faithfully copies exonic sequences but alters gene structure by creating duplicates lacking introns and most of their ancestral promoter sequences (Soares et al., 1986; Boer et al., 1987). However, the retrotransposition of genes whose promoters contain downstream promoter elements (DPEs) located 3' of the

transcription start site is likely to create a retrocopy with a promoter capable of initiating basal transcription (Arkhipova, 1995). Moreover, when the unit of duplication (the "duplication span") does not encompass an entire gene a partial gene duplicate is generated (Katju and Lynch, 2003). Finally, at the intra-genic level the tandem duplication of exons has been described and is frequently associated with alternative splicing (Letunic et al., 2002; Kondrashov et al., 2002).

The heterogeneity of the mammalian genome (section 1.2.12, page 35) and the fact that the duplication of single genes is frequently associated with the transposition of one of the daughter copies suggests that the differentiation of gene duplicates may be related to differences in their mutational or selective contexts. This question has been addressed from the perspective of both recombination rate differences (Zhang and Kishino, 2004) and epigenetic differences between duplicates (Rodin and Riggs, 2003). Moreover, it has been suggested that the X-chromosome plays a disproportionate role in both the generation and recruitment of retrogenes (Emerson et al., 2004). Given the differing selective regimes of the X-chromosome and the autosomes, differences in the selective context of retroduplicates and their source genes might be especially great.

In Chapter 2 of this thesis I explore the inequality in evolutionary rate between duplicate gene copies formed by retrotransposition and contrast this with the rate inequality between duplicates formed by DNA-based gene duplication. I show that retrogenes have a consistently faster rate of evolution than their source copies, but that the act of gene relocation associated with any mechanism of duplicative transposition also contributes significantly to the increased rate.

1.3.2 Birth and death of duplicate genes

The significance of gene duplication to genome evolution is illustrated by the estimated high birth rate for gene duplicates (Lynch and Conery, 2000). Lynch and Conery used a demographic approach based on the age distributions of extant gene duplicates in a range of eukaryote genomes to produce a conservative estimate of the birth rate of gene duplicates that averages 0.01 per gene per million years. Notably, this rate is comparable to the rate of point mutation per nucleotide site. This implies that changes in gene content are at least as important as changes in gene sequence to phenotypic evolution.

Therefore, the duplication of single genes appears to frequently open a window of oppor-

tunity for evolutionary innovation. However, this birth rate is set against a relatively high mortality rate with the implication that most gene duplicates are destined to be lost from the genome. This is true even of genes duplicated by polyploidisation where the likelihood of duplicate gene retention is expected to be greatest because relative gene dosage is maintained (Lynch and Conery, 2000). The modern yeast genome, for example, has retained only 11% of all the gene duplicates created following a whole genome duplication event about 100 Mya (Byrne and Wolfe, 2005). However, even this apparently low survival rate exceeds that expected from the rate of gene loss estimated by Lynch and Conery. They estimated that duplicate gene loss proceeds as an exponential decay function. Therefore, most gene duplicates are expected to have a short evolutionary life-span with an average half life of 5 Myr.

1.3.3 Mechanisms for duplicate gene preservation

The life of most gene duplicates is cut short by their silencing and ultimate loss from the genome. However, there are several mechanisms that increase the survival chances of a newly formed gene duplicate. These mechanisms make differing predictions about whether or not functional divergence is a characteristic of duplicate gene preservation and about the nature of that functional divergence when it occurs.

The frequent loss of gene duplicates is consistent with Ohno's classical model under which gene duplication creates two paralogous genes that are functionally redundant (Ohno, 1970). This redundancy implies that one of the paralogs can become invisible to natural selection and evolve free from selective constraints. Under Ohno's model the ultimate fate of this unconstrained paralog is determined by the neutral accumulation of mutations that were previously forbidden by selection. Given the abundance of degenerative mutations, the most likely outcome of this neutral phase of evolution is the fixation of a null allele that results in the nonfunctionalisation and ultimate loss of the gene duplicate. The functional redundancy of the duplicate means that selection is indifferent to its loss.

According to the classical model the only escape route from gene silencing for a duplicate gene is the creation of a new function by the rare fixation of beneficial mutations. This process is referred to as neofunctionalisation and can occur by the fixation of mutations in the duplicate gene's protein-coding or regulatory sequences. Under this model the very property that makes possible the evolution of a new function in a gene duplicate (i.e.,

redundancy) is an Achilles heel that generally leads to its loss by nonfunctionalistion. A characteristic of the neofunctionalisation mechanism is that the survival of a newly formed gene duplicate is guaranteed by its gain of a novel function that differentiates it both from its sister duplicate and from the ancestral single copy gene. Thus, retention of the second gene copy and its change of function are achieved by the same mutation.

An alternative, more recently described, model proposes that gene duplicates can be retained without functional innovation and adaptation. The subfunctionalisation model (Force et al., 1999) provides a neutral explanation for the retention of duplicate copies of a multifunctional gene by degenerative mutations that lead to the loss of different subfunctions in each duplicate. A crucial requirement is the independent mutability of the subfunctions concerned. The complementary pattern of subfunction loss ensures that both duplicates are required to perform the ancestral set of subfunctions and therefore both must be retained in the genome. The outcome of this process is the functional divergence of duplicates both from the ancestral gene and from each other, with each retaining a subset of the functions of the ancestral gene.

There is an alternative, adaptive, route for the preservation of duplicate copies of a multifunctional gene. The starting point of the "adaptive conflict" model is a bifunctional gene in which conflicting selective pressures exist between the gene's different roles (Piatigorsky and Wistow, 1991; Hughes, 1994). Duplicating such a gene provides a chance to eliminate these negative pleiotropic constraints, and sets each duplicate on a path enabling the refinement of each subfunction by positive selection.

Finally, there are two alternative ways in which gene duplicates can be maintained without divergence of function. Natural selection may sometimes view the duplication of a gene as providing a beneficial increase in dosage of a gene product. For example, this mechanism explains the retention of multiple duplicate copies of the rRNA genes in almost all organismns to enable increased ribosome synthesis (Seoighe and Wolfe, 1999). The second potential mechanism for the preservation of duplicate genes without functional divergence is based on the premise that the genetic redundancy resulting from possession of a paralog allows the buffering of null mutations that impair gene function. However, population genetic considerations suggest that the maintenance of genetic redundancy for the purpose of mutational robustness is feasible only in organisms with very large population sizes (Wagner, 2000b).

Therefore, of the various potential preservational events promoting retention of gene duplicates only neo-functionalisation involves true functional innovation. However, this fact does not diminish the importance of gene duplication as a source of new functions. It should be noted that gene duplicates retained in the genome provide considerable molecular substrate for the later development of evolutionary novelty. Therefore the preservation of a pair of gene duplicates either through increased protein dosage or as a result of subfunctionalisation is compatible with later acquisition of novel functions (Force et al., 1999; Kondrashov et al., 2002; He and Zhang, 2005).

The relative importance of these mechanisms in the retention of duplicate genes is currently unresolved. On the one hand, subfunctionalisation might be thought to occurr relatively rarely since the probability of subfunctionalisation is strongly dependent on the number of independently mutable ancestral subfunctions capable of being divided between duplicates (Force et al., 1999). Moreover, population genetic considerations show that this process is only likely to occur in organisms with relatively small effective population sizes. On the other hand, subfunctionalisation might be expected to be more prevalent than neofunctionalisation because it occurs through the neutral accumulation of frequently occurring degenerative mutations. Lastly, there is mounting evidence that gene duplications are often imperfect, and lead to gene structure differences between "unequal" gene duplicates created for example by retrotransposition or partial gene duplication. If an incomplete duplicate is missing one subfunction at birth then the duplicate pair is likely to be propelled towards subfunctionalisation from the outset (Averof and Ferrier, 1996; Lynch, 2004).

1.3.4 Gene duplicate preservation and its impact on evolutionary rate

The mutations that ensure the long-term preservation of gene duplications in the genome can occur in either protein-coding regions or in regulatory sequences. The resultant divergence in protein function or in expression profile is likely to underlie the acceleration in evolutionary rate that frequently follows gene duplication (see below).

1.3.4.1 Duplicate preservation by divergence in protein function

In the early phases of gene duplication a period of neutral evolution accompanies the lifting of selective constraint on the coding sequence of one or both duplicates. This is likely to lead to a short-term acceleration in the rate of coding sequence evolution of the duplicates.

However, duplication also has implications for the longer term evolutionary rate of duplicates that become established in the genome. This follows from the fact that surviving gene duplicates are likely be functionally differentiated either in the short term (as a result of the early events that preserve duplicates) or in the long term (due to functional innovations of established duplicates). As outlined earlier, protein function and the rate of molecular evolution are believed to be closely coupled. It is therefore to be expected that alterations in evolutionary rate frequently accompany the differentiation in protein function that follows gene duplication events.

Gene duplication can result in the divergence in protein function, firstly, of one duplicate from the other and secondly, of both duplicates from their progenitor gene. The potential rate acceleration associated with gene duplication can be considered from both these perspectives.

First, following gene duplication the function of one member of a duplicate pair may diverge relative to that of its sister duplicate. Therefore, when one member of a pair undergoes neofunctionalisation, positive selection for this new function will result in a rate acceleration of the protein relative to its paralog. An implicit assumption of the neofunctionalisation model is that the second duplicate performs the ancestral gene's function and continues to evolve at the same rate as its parent.

Rate differences between duplicates can be detected when an outgroup sequence is used. This outgroup is typically a single copy ortholog whose evolutionary rate approximates that of the ancestral progenitor of the ingroup sequences. Small scale studies using this approach produced very different estimates of the frequency of rate inequality between gene duplicates. A study of 17 duplicate pairs in *Xenopus laevis* showed that both copies are under strong purifying selection (Hughes and Hughes, 1993). In contrast, of 26 zebrafish duplicate pairs roughly one half were found to evolve asymmetrically (Van de Peer et al., 2001). The differing conclusions of these studies might be attributable to biases in these small datasets. Genome-wide analyses using different datasets and methods have tended to report higher incidences of rate inequality between the amino-acid sequences of gene duplicates (20-60% of cases asymmetric, (Conant and Wagner, 2003; Zhang and Li, 2004; Kellis et al., 2004)) with some exceptions (<5% of pairs asymmetric, (Kondrashov et al., 2002)). A limitation common to all approaches is that only the cumulative pattern of substitutions is observable and thus early phases of rate acceleration associated with neofunctionalisation are likely to

be obscured by subsequent purifying selection once the gene has taken on its new function (Van de Peer et al., 2001).

Second, when functional divergence of both duplicates occurs with respect to the ancestral gene this should result in the acceleration of both gene duplicates. This situation might arise following the partial loss of ancestral subfunctions of duplicate genes preserved by subfunctionalisation: the reduction of pleiotropic constraints as the duplicates partition ancestral gene functions between them can be expected to cause acceleration of both copies. A similar argument applies to the reduction in negative pleiotropy that follows the resolution of adaptive conflict. In this case, however, the expected rate acceleration is a consequence of positive selection to optimise previously constrained subfunctions.

So far there is little direct evidence for an increase in the evolutionary rate of both duplicates following gene duplication. Among the most suggestive results in this context are measurements which show an increase in the average evolutionary rate of both duplicates relative to the ancestral rate of their single copy parent. Typically paralogs are seen to accumulate more amino acid changes than orthologs of comparable divergence time. However, estimates of the magnitude of this effect vary from, at the lower end, an increase of one-third to, at the upper end, a fourfold increase (Nembaware et al., 2002; Kondrashov et al., 2002; Huminiecki and Wolfe, 2004). Needless to say, measurements of the average acceleration of a duplicate pair will to some extent reflect the acceleration in rate of only one member of the pair as described above. It is an open question, therefore, how often the more slowly evolving paralog also experiences an increase in rate.

1.3.4.2 Duplicate preservation by divergence in gene expression

The preceding examples show how the preservation of gene duplicates by divergence in protein function can directly impact on rates of protein evolution. However, the maintenance of duplicate genes by neofunctionalisation or subfunctionalisation can also proceed by divergence in gene expression profile. In this case, the immediate target of these preservational processes is the non-coding sequence responsible for gene expression regulation rather than the protein-coding sequence. The modularity of non-coding regulatory sequences makes them independently mutable and especially prone to the complementary degenerative mutations characteristic of subfunctionalisation.

It is likely that the preservation of a duplicate pair by expression profile divergence will

also influence each duplicate's rate of protein sequence evolution, albeit indirectly. This effect can be mediated by the various influences of gene expression on coding sequence evolution outlined in the previous section. In an analogous way to divergence in protein function, gene expression changes following duplication can result in the divergence (and acceleration) of one duplicate with respect to the other and of both duplicates from their progenitor gene.

For example, it has been suggested that the large rate differences still apparent between some yeast whole-genome duplicates could be a simple side-effect of their divergence in expression (Drummond et al., 2005). The long-term persistence of rate differences between duplicates of this age is unlikely to be due to the short-term effects of neo-functionalisation. It was suggested that a more likely explanation is that this effect is caused by the differing levels of selection for translational robustness on these duplicates. This would explain the striking observation that in gene duplicate pairs where gene expression level differs more than two-fold the more highly expressed paralog evolves more slowly in more than 90% of cases (Drummond et al., 2006). This suggests that the expression-mediated impact of gene duplication on the rate of molecular evolution should be generally detectable in the coupling of coding sequence evolution with expression divergence. However, this relationship is not very apparent in studies that relate pairwise measures of divergence in sequence and expression (Wagner, 2000a; Li et al., 2005). A more meaningful comparison in this context is to relate the *asymmetry* of sequence evolution to *asymmetry* in functional divergence, instead of using pairwise divergence measures. However, this approach can be limited by the availability of a preduplication outgroup sequence. Moreover, there is a current lack of functional data for outgroup species. Nevertheless, a recent study in yeast has provided compelling evidence that the rate differences between yeast proteins can be partly explained by asymmetric expression divergence between them (Kim and Yi, 2006). The variables associated with coding sequence divergence of duplicate genes to a large extent parallel those variables discussed earlier as potential determinants of the evolutionary rate of the sequences of single genes. Furthermore, it was frequently observed that within a duplicate pair the slower evolving member is more abundant in yeast cells, is less dispensable and engages in more interactions with other proteins (Kim and Yi, 2006). This in good agreement with expectation given the functional parameters thought to determine the rate of evolution of single proteins.

Expression divergence may also govern the sequence divergence of gene duplicates in multicellular eukaryotes. In mammals, population genetic considerations suggest that subfunctionalisation is more likely to occur because of their small population sizes. As outlined above, the vulnerability of *cis*-regulatory sequences to degenerative mutations means that expression patterns are particularly susceptible to subfunctionalisation. In fact, the partitioning of ancestral expression patterns can proceed both quantitatively (by a division of the ancestral expression level between duplicates so that their summed expression is required to fulfill ancestral function (Ferris and Whitt, 1979)), spatially (by division of the constituent tissues of the ancestral expression domain among the duplicates (McClintock et al., 2002)) or temporally (by division of expression at different developmental stages among the duplicates (Yan et al., 2005)). In each case the divergence in expression is expected to result in an increase in evolutionary rate of the duplicates relative to the ancestral gene. For example, the association between expression breadth and evolutionary rate described in the previous section is likely to mediate an increase in the rate of protein divergence following the narrowing of gene expression of duplicates retained by spatial subfunctionalisation. Moreover, if the allocation of tissue subfunctions between duplicates is unequal then this relationship predicts an increase in evolutionary rate of the more tissue-specific duplicate relative to its sister. Although the generality of this prediction has not yet been tested, anecdotal examples support it. For example, of the duplicates of the triose phosphate isomerase gene in zebrafish the faster copy has inherited fewer ancestral subfunctions (Merritt and Quattro, 2001). In summary, it seems likely that asymmetry in expression divergence between mammalian gene duplicates has the potential to explain some of their asymmetry in sequence divergence.

1.4 Impact of alternative splicing on rates of molecular evolution

The discovery of alternative splicing has overthrown the traditional "one gene-one protein" interpretation of the central dogma. By establishing another mechanism for evolution to "seek new solutions without destroying the old" (Gilbert, 1978) it challenges the convention that gene duplication holds a monopoly on the emergence of new functions. Early estimates suggested that this strategy was available to only 5% of all genes, but there is mounting

evidence that up to 80% of all human genes are alternatively spliced (Kampa et al., 2004). The new wealth of genomics data has led to the recognition that alternative splicing rivals gene duplication as a well-trodden evolutionary path to increasing the ensemble of protein functions. This potential is illustrated by the extraordinary diversity generated by the alternative splicing of the *Drosophila DSCAM* gene whose protein products function in axon guidance. This gene consists of four arrays of alternative exons and the combinatorial potential of this arrangement in theory allows one gene to encode 38,016 different proteins (Black, 2000).

Further evidence for alternative splicing's potential in expanding the proteome's functional repertoire comes from the observation that alternative exons frequently coincide with complete protein domains (Kriventseva et al., 2003). Furthermore, the fact that the encoded domains are enriched for protein-interaction functions led to the suggestion that alternative splicing may allow the modification of linkages in protein interaction networks (Resch et al., 2004b).

1.4.1 Alternative splicing is associated with gene structure changes

It is increasingly apparent that alternative splicing is strongly associated with the creation of new exons by processes that include intragenic exon duplication, exaptation of transposable elements and exonization of unique intronic material.

Given the conceptual parallels existing between alternative splicing and the duplication of genetic material it is not surprising to find that these are not unrelated processes at the level of gene structure change. The intersection of these evolutionary strategies is illustrated by the association between the intragenic duplication of exons and the alternative splicing of these tandem exons by exon-skipping (Letunic et al., 2002). A frequent consequence of the creation of new exons in this manner is that the duplicate exons are spliced into mature transcripts in a mutually exclusive manner to avoid frameshifts (Kondrashov et al., 2002). This configuration has the potential to set these alternative exons onto independent evolutionary paths and allows the functional divergence of the encoded isoforms in an manner analogous to the duplication of entire genes.

The association between alternative splicing and the creation of new exons is further illustrated by the discovery that transposable elements, such as *Alu* elements, can contribute sequence to human proteins (Makalowski et al., 1994). It is now apparent that

the infiltration of genes by *Alu* elements can only be tolerated if the *Alu*-containing exons are alternatively spliced (Sorek et al., 2002). This implicates alternative splicing in the relaxation of negative (purifying) selection against changes in gene structure.

Alternative splicing may, therefore, be a mechanism that enables evolution to experiment with newly created exons. The obvious correspondence with gene duplication was made explicit by a study documenting differences in gene structure between human and mouse orthologs (Modrek and Lee, 2003). This study coined the term "internal paralog" to describe an alternative transcript spliced at low frequencies ("minor-form transcript") and whose alternative exons are generally not conserved between human and mouse. The predominant isoform of a gene ("major-form transcript") consists of exons that are conserved between human and mouse. By fulfilling the gene's ancestral function, the major form transcript should enable the relaxation of selective constraint on the minor-form transcript thus allowing it to evolve free of selective constraint. Modrek and Lee demonstrated that alternative splicing is likely to have facilitated many of the changes that have occurred in the structure of genes since the human/mouse divergence. This was illustrated by the finding that species-specific exons are 10 times more likely to be alternatively spliced than conserved exons. Notably, this facilitation is dependent on the low frequency incorporation of these species-specific exons into transcripts. Alternatively spliced exons specific to human or mouse are nearly eight times more likely to be spliced at low frequencies (i.e., as the minor form) than alternative exons conserved between these mammals (Modrek and Lee, 2003). It is this low-frequency expression, by alternative splicing, of species-specific "internal paralogs" that is likely to shield newly formed exons from selection while ensuring that the gene's ancestral function is not compromised.

Modrek and Lee's study did not determine whether the species-specific exons they described represented recent exon losses or gains. However, subsequent analysis suggests that a large proportion of these exons have been recently created (Cusack and Wolfe, 2005; Wang et al., 2005) (see Chapter 3). Moreover, since most of these gained exons are unique genomic sequences they are likely to have emerged by exonization of intronic sequences rather than by exon duplication or transposable element exaptation (Wang et al., 2005).

1.4.2 Differing selective pressures associated with alternative splicing

Evidence is emerging for the existence of two contrasting selective pressures operating on alternatively spliced exons.

On the one hand, alternative splicing is associated with an apparent relaxation of negative selection. This is not solely restricted to increased plasticity of gene structure described in the previous section. There is also evidence that purifying selection on amino acid changes (as measured by d_N/d_S) is up to seven-fold weaker in alternatively spliced exons (Xing and Lee, 2005). Furthermore, purifying selection against the inclusion of premature termination codons (PTCs) is significantly reduced in minor compared to major-form exons. The incorporation of PTCs into mRNAs results in their degradation by nonsense mediated decay (NMD) (Maquat, 2004) implying that PTC-containing alternative transcripts do not directly contribute to protein function. However these transcripts may not represent an evolutionary "dead-end" since they might function in post-transcriptional gene regulation (Lewis et al., 2003) (but see Pan et al. (2006)) or acquire new functions through subsequent mutations (Lynch and Kewalramani, 2003).

On the other hand, alternative splicing subjects other genic properties to increased selective constraint. Alternatively spliced exons are observed to be under stronger selection to preserve reading frame (Resch et al., 2004a) and to have fewer single nucleotide polymorphisms (Yeo 2005). Strikingly, a more than six-fold reduction in synonymous site divergence is seen among minor-form exons compared to constitutive exons (Xing and Lee, 2005). Ordinarily, such a pattern would be indicative of biased codon usage for higher translational efficiency (Akashi and Eyre-Walker, 1998; Akashi, 2001, 2003; Duret, 2002). However, this explanation is unlikely in this case for two reasons. First, the fact that, by definition, alternative exons are translated less frequently than constitutively spliced exons means that they are expected to have weaker codon usage biases (Iida and Akashi, 2000). Second, codon usage bias can not account for the observation that in the intronic sequence immediately flanking minor-form exons nucleotide substitutions are two-fold less frequent than in the flanks of constitutive exons (Xing and Lee, 2005). The most likely explanation for the strong selective pressure on synonymous sites in alternative exons relates to the preservation of splicing regulatory motifs (Pagani and Baralle, 2004). These motifs include exon splicing enhancers (ESEs) and exon splicing silencers (ESSs) that, respectively, promote and inhibit exon recognition. Polymorphism and divergence data supply evidence of

purifying selection on ESEs (Fairbrother et al., 2002; Carlini and Genut, 2006; Parmley et al., 2006) However, this may be only part of the picture since suppression of synonymous site changes extends outside of predicted ESE motifs in alternative exons and these exons show no increase in predicted ESE density compared to constitutive exons (Xing and Lee, 2006a).

In summary, compared to constitutive exons, the relaxed selective regime on amino acid changes in alternatively spliced exons contrasts with a regime of strong purifying selection at the RNA level. In Chapter 3 of this thesis I explore the effect of alternative splicing on evolutionary rates of genes and show that even the constitutive regions of alternatively spliced genes show increased rates of protein sequence evolution following the acquisition of alternatively spliced exons.

1.4.3 Heterogeneity in intragenic sequence evolution due to alternative splicing

The relaxation of selective constraint on amino acid changes associated with alternative splicing is reminiscent of the early evolution of duplicated genes. In the case of gene duplicates this effect contributes to differences in evolutionary rate between paralogs. However, the rate differences resulting from alternative splicing are observable on the intragenic scale. Therefore, the selective differences between alternative and constitutive exons of alternatively spliced genes are a source of significant within-gene variability in synonymous and nonsynonymous rates. Strikingly, rates of nonsynonymous evolution are found to be up to twice as fast in alternative exons compared to constitutive exons (Xing and Lee, 2005; Wang et al., 2005; Chen et al., 2006). It is interesting to interpret these "hot-spots" of increased rates of protein evolution *within* genes in the light of the suspected determinants of protein rate differences *between* genes discussed earlier. Thus the lower expression level, greater tissue-specificity and phylogenetic youth of alternative exons are consistent with their increased nonsynonymous rate.

Equally striking patterns of within-gene rate variation are exhibited by synonymous site evolution in alternatively spliced genes. The classic illustration of this is the occurrence of a "cold spot" of synonymous divergence in the *BRCA1* gene (Hurst and Pal, 2001; Orban and Olah, 2001). Neither the presence of a CpG island nor codon usage bias could explain the dramatic reduction in d_S in this region which is more likely to reflect the fact

that this part of the gene is alternatively spliced and coincides with two putative ESEs. Moreover, the alternative region of the *BRCA1* sequence also displays a dramatic increase in the value of d_N/d_S above the neutral expectation of one. Intuitively, the elevation of d_N/d_S might be thought to be a consequence of the local suppression of d_S (Chamary et al., 2006). This would imply that the d_N/d_S metric of selective constraint on amino acid changes could be invalidated when synonymous site evolution deviates from neutrality due to RNA-level selection on splicing regulatory motifs. However, it has been suggested that measurements of d_N/d_S will only be distorted in this way if RNA-level selection specifically targets synonymous sites (e.g., due to the "synonymous phasing" of splicing motifs (Xing and Lee, 2006b)).

1.4.4 Complementarity of alternative splicing and gene duplication

The potential complementarity of alternative splicing and gene duplication in generating proteomic diversity is hinted at by two recent studies describing a genome-wide association between these phenomena (Kopelman et al., 2005; Su et al., 2006). Specifically, these studies described a weak negative correlation between gene family size and the number of alternative splice forms. Therefore, singleton genes (without paralogs) tend to encode more alternative splice forms than the members of gene families. Although weak, this correlation is consistent with the occurrence of subfunctionalisation of alternative splicing leading to the loss of alternative splice variants in parallel with increases in gene family size. These observations are in agreement with both theoretical expectation and anecdotal evidence that ancestral alternative splice forms can be partitioned by subfunctionalisation following gene duplication.

The first theoretical requirement for subfunctionalisation is that the ancestral gene giving rise to duplicates should fulfil multiple discrete roles that are independently mutable and therefore can be subdivided among the duplicates. In the case of alternative splicing by exon-skipping the alternative transcripts can be regarded as distinct subfunctions that can be resolved among duplicates by the simple mechanism of differential exon loss by degenerative mutations. Moreover, in mammals the high frequency of gene duplication by retrotransposition (Vinckenbosch et al., 2006) of spliced mRNA provides an alternative way to partition isoforms among gene duplicates.

Theoretical considerations suggest that the probability of subfunctionalization depends

on the amount of 'mutational target' that is vulnerable to null mutations (Force et al., 1999). The sequence of the alternative exon itself provides a significant mutational target that may be exposed to degenerative mutations following gene duplication. Additional mutational target is provided by the extensive range of *cis*-acting elements in both intronic and exonic sequence that mediate the complex regulation of alternative splicing.

Anecdotal evidence for the subfunctionalisation of alternative splice variants is provided by two examples in fish. The microphthalmia-associated transcription factor (*MITF*) gene is alternatively spliced and single copy in mammals and has undergone gene duplication followed by degeneration of alternative exons in teleost fish (Altschmied et al., 2002). The single-copy synapsin gene in human encodes two isoforms that have apparently been subdivided among its Fugu co-orthologs *SYN2a* and *SYN2b* by complementary degenerative mutations (Yu et al., 2003).

In Chapter 4 of this thesis I report a further example of the fixation of two alternatively spliced transcripts following gene duplication, and show that the gene pair has been preserved by subfunctionalisation.

Chapter 2

Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates

The research described in this chapter has been published in *Molecular Biology and Evolution* (Cusack and Wolfe, 2007). Soon after submission of this work an independent study was published describing the evolution of ten recently duplicated retrogenes in mice and confirming many of the results in this work (Gayral et al., 2007).

2.1 Abstract

Duplicated genes frequently evolve at different rates. This asymmetry is evidence of natural selection's ability to discriminate between the two copies, subjecting them to different levels of purifying selection, or even permitting adaptive evolution of one or both copies. However, if gene duplication creates pairs of protein coding sequences that are initially identical, this raises the question of how selection tells the two copies apart. Here we investigated asymmetric sequence divergence of recently duplicated genes in rodents and related this to two possible sources of such asymmetry: gene relocation as a consequence of duplication, and retrotransposition as a mechanism of gene duplication. We found that most young rodent duplicates that have been relocated were created by retrotransposition. The degree of rate

asymmetry in gene pairs where one copy has been relocated (either by retrotransposition or DNA-based duplication) is greater than in pairs formed by local DNA-based duplication events. Furthermore, by considering the direction of transposition for distant duplicates, we found a consistent tendency for retrogenes to undergo accelerated protein evolution relative to their static paralogs, whereas DNA-based transpositions showed no such tendency. Finally, we demonstrate that the faster sequence evolution of retrogenes correlates with the profound alteration of their expression pattern that is precipitated by retrotransposition.

2.2 Introduction

Several genome-wide studies of gene duplication have sought to investigate the processes by which the fates of the paralogs generated by a duplication event become uncoupled (Kondrashov et al., 2002; Zhang et al., 2003; Kellis et al., 2004). This question is relevant to determining the relative importance of different processes governing the fixation of duplicates (Lynch, 2004). The period immediately following gene duplication is commonly described as one of genetic redundancy between functionally equivalent copies. This situation was originally thought to be resolved by one of two alternative mechanisms: non-functionalisation resulting from loss of a superfluous copy, or neo-functionalisation in which a gain of function in one copy leads to the retention of both copies (Ohno, 1970). More recently, the subfunctionalisation model for preservation of duplicate genes has received much attention. This model proposes that pairs are retained if they partition the subfunctions of the ancestral gene between them (Force et al., 1999).

The non-functionalisation and neo-functionalisation processes predict accelerated evolution of the deconstrained copy compared to the paralog that remains constrained by purifying selection on the ancestral function. Unequal rates of duplicate gene divergence are therefore expected under both scenarios. In contrast, if duplicates are retained by equally strong purifying selection on both copies through a gene dosage effect, duplicate divergence should be roughly symmetrical (Lynch, 2004), with any observed rate asymmetry resulting purely from stochastic effects. Finally, the subfunctionalisation model does not make any prediction about the asymmetry (or otherwise) of sequence evolution because the ancestral subfunctions could be divided equally or unequally between duplicates. In this context, recent work has begun to shed light on the extent to which asymmetric sequence divergence is explained by asymmetric functional divergence (Kim and Yi, 2006).

Although previous studies have reached conflicting conclusions as to whether asymmetry in sequence divergence following gene duplication is common (Zhang et al., 2003; Conant and Wagner, 2003) or relatively rare (Kondrashov et al., 2002), rate differences between duplicates are often interpreted as evidence that natural selection can somehow differentiate between gene copies that were initially identical and thus functionally interchangeable. However, these studies have largely ignored the mechanism by which genes are duplicated, an issue that is relevant to the assumption that all duplicates are equal at birth (Lynch, 2004; Katju and Lynch, 2006).

Single gene duplications in metazoan genomes can be formed by two mechanisms – DNA-based and RNA-based – that probably differ in their likelihood of generating identical paralogous copies of an ancestral gene. DNA-based duplication (copying of segments of chromosome) is expected to create two copies that are indistinguishable, with conservation of both the exon-intron structure and the regulatory sequences (provided that the entire gene and its promoter are contained within the duplication span (Katju and Lynch, 2003)). Furthermore, if a DNA-based duplication occurs by the tandem duplication of a single gene, or by segmental duplication of a group of linked genes, the duplication will not cause any extensive disruption of synteny. In contrast, gene duplication by retrotransposition (Soares et al., 1985; Boer et al., 1987) creates a new duplicate that differs from its parent in a number of respects. The retrocopy is created by reverse transcription of a spliced messenger RNA, typically creating a single-exon copy of a multi-exon parental gene. In addition, since only the transcribed sequence is duplicated the retrocopy becomes detached from the ancestral promoter that controlled expression of the parental gene. Only if a new promoter is acquired by the retrocopy is it likely to survive as a functional retrogene. Furthermore, new genes formed by retrotransposition are usually not physically linked to their parents so synteny is disrupted. The newly created retrogene is deposited in a novel chromosomal environment with a different set of gene neighbours.

In this study we investigated the impact of two potential causes of rate asymmetry in duplicated mammalian genes: the genomic relocation that may occur as a consequence of gene duplication, and the mechanism of duplication (via DNA or RNA). We hypothesised that if syntenic context is an important aspect of gene function (*e.g.,* due to the chromosomal clustering of co-regulated genes (Lercher et al., 2002; Singer et al., 2005)) then gene duplications that result in gene relocation may create duplicates that are not functionally

equivalent. This might be expected to increase rate asymmetry among relocated duplicates. Similarly, duplicates created by retrotransposition might be expected to show asymmetrical rates of evolution due to the almost inevitable regulatory changes associated with retrotransposition, even if the protein sequence is unaltered by the duplication event itself. In order to focus on recently duplicated genes, we consider genes that have become duplicated since the divergence of mouse and rat. Although it is widely assumed that retrogenes will show fast rates of evolution compared to their progenitors, to our knowledge this has only been demonstrated in one, very recent, study (Gayral et al., 2007).

2.3 Methods

2.3.1 Recent rodent duplicates

We retrieved gene duplicates from the Homolens (version 1) database of automatically inferred phylogenies constructed using Ensembl gene predictions (S. Penel and L. Duret, personal communication; http://pbil.univ-lyon1.fr/databases/homolens.html) and queried using FamFetch (Dufayard et al., 2005). We searched for cases of recent lineage-specific gene duplication in rodents, where a single gene in rat or mouse has exactly two co-orthologs in the second species. Homolens internal identifiers were mapped to Ensembl identifiers which were then used to retrieve map locations. Because Ensembl contains some annotated "introns" that are frameshift corrections, for analyses of intron content we only considered annotated introns that are $\geq$ 50nt and flanked by coding exons.

2.3.2 Gene duplication categories

We categorised recent rodent duplicates on the basis of two criteria: relative location in the genome and mechanism of duplication. We designated all physically linked duplicate pairs with < 5 intervening genes as 'local' duplications. All other duplicates either on the same chromosome or on different chromosomes were classified as 'distant'.

We classified duplicated genes on the basis of duplication mechanism by distinguishing between RNA-mediated retrotranspositions (which typically create a single-exon retrogene from a multi-exon paralog) and DNA-based transpositions (which typically conserve exon-intron structure), using a rigorous set of criteria based on counts of coding exons. For pairs consisting of a single-exon gene with a multi-exon paralog we counted the introns of the

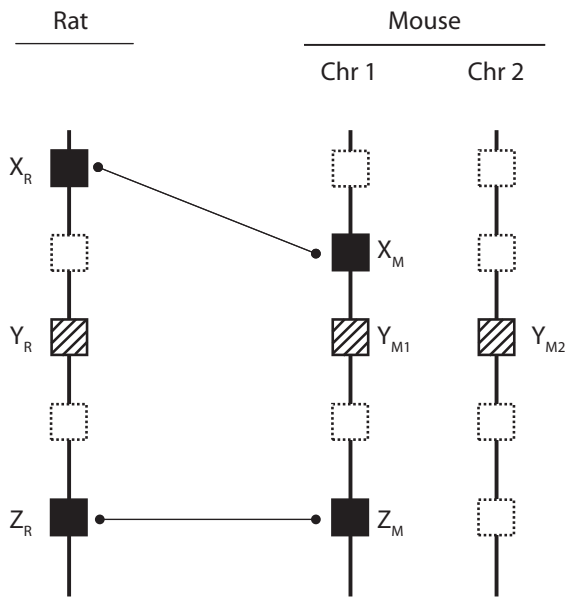
latter gene that lie within the protein alignment of the two duplicates. If ≥ 2 introns were present, we inferred that duplication had occurred by retrotransposition resulting in the loss of these introns in the retrocopy. Since all detected retrogenes have a single coding exon this set excludes retrogenes that have been incorporated into chimeric coding regions following gene-fusion events but potentially includes cases that have acquired non-coding exons *de novo* following retrotransposition (Vinckenbosch et al., 2006). If both members of a duplicate pair contained ≥ 2 exons, we counted introns within the alignment and where there was evidence of no more than one intron loss we inferred that duplication had occurred by DNA-based transposition. Although these strict criteria allow confident inference of the mechanism of duplication for many pairs, they leave some pairs unclassified. For example, when both members of a duplicate pair are single-exon genes we were not able to infer the mechanism. Such unclassified pairs were not used for the analysis of the impact of duplication mechanism on rate asymmetry but were included in the analysis of the effect of duplicate relocation.

2.3.3 Direction of (retro)transposition of distant duplicates

For distant duplicates we established the direction of (retro)transposition, to discriminate between the relocated paralog and the static paralog that remains at the ancestral locus. This was done using a framework of positional anchors consisting of unduplicated single-copy genes for which there is a 1:1:1 orthologous relationship between human, mouse and rat. These singletons were retrieved from Homolens using FamFetch with the query topology ((mouse, rat), human) constrained so that no gene duplication has occurred since the primate-rodent split.

To establish the direction of (retro)transposition of distantly separated mouse duplicates that are co-orthologs of a single rat gene, for example, we located the closest singleton anchors that bracket the rat gene (Figure 2-1A). We then determined the locations of the single mouse orthologs of the rat bracketing genes. When the mouse orthologs of a pair of rat bracketing genes are linked in mouse, and themselves bracket one of the two mouse duplicates, we designated the bracketed mouse duplicate as the static copy and the other mouse duplicate as the relocated duplicate. Assignment of the direction of (retro)transposition by this method was possible for 118 of the 147 distant duplicates in mouse, and for 106 of the 137 distant duplicates in rat.

A



B

	Local	Distant
DNA transposition	44	18
Retro-transposition	0	36

Figure 2-1: (A) Determining the direction of transposition for distantly separated duplicates: Duplication of gene Y in the mouse lineage following mouse-rat speciation created two mouse duplicates (inparalogs Y_{M1} and Y_{M2}) which are co-orthologs of a single rat gene (Y_R). To polarise the direction of transposition in mouse and thus discriminate between the static and transposed duplicates we considered genes X_R and Z_R that flank gene Y_R in rat and have single orthologs in mouse (X_M and Z_M). Since both X_M and Z_M are found to flank Y_{M1} this duplicate can be designated the static copy implying that its paralog (Y_{M2}) has been relocated by retrotransposition. For simplicity exon-intron structures are not depicted. (B) Classification by duplication mechanism of duplicate pairs having branch-specific $d_S > 0.001$ and $d_N > 0.001$. A resampling strategy was applied to these 98 duplicates to separately determine the effect of duplicate relocation and retrotransposition on rate asymmetry measured by RN.

2.3.4 Measures of sequence evolution

For each sequence triplet consisting of a single-copy gene in one rodent species and its two coorthologs in the second rodent species, we aligned the Ensembl protein sequences using CLUSTALW (Thompson et al., 1994) and back-translated it to create a codon-based alignment. These alignments were used as input to the program **like-tri-test** (Conant and Wagner, 2003) to estimate branch-specific rates of synonymous (dS) and nonsynonymous divergence (dN) as well as branch-specific estimates of dN/dS.

For the branches leading to each duplicate, we quantified the magnitude of asymmetry for estimates of dS, dN and ω ($= dN/dS$). In cases where the branch-specific estimate of dS is very small for either duplicate an artefactually large asymmetry in dN/dS may result, so in this analysis we used only those cases for which branch-specific estimates of both dS and dN for both duplicates were > 0.001. For the branches leading from the internal node

to each duplicate we used the absolute (unsigned) normalized difference in divergence as a measure of asymmetric evolution. For example, using the notation of (Kim and Yi, 2006), we quantified the asymmetry in synonymous evolution between duplicates as

$$RS = \left| \frac{dS_1 - dS_2}{dS_1 + dS_2} \right|,$$

where dS_1 and dS_2 are the synonymous divergences estimated for the branches leading from the internal node to duplicates 1 and 2 respectively. Thus RS values of 0.33 and 0.60 correspond to rate differences of 2-fold and 4-fold, respectively. Absolute normalized differences in nonsynonymous divergence (RN) and strength of selective constraint ($R\omega$) were calculated similarly.

For distant duplicates for which we could discern the direction of transposition, we derived a measure of signed (directional) asymmetry in nonsynonymous divergence,

$$SRN = \left(\frac{dN_r - dN_s}{dN_r + dN_s} \right),$$

(Kim and Yi, 2006), where dN_s and dN_r refer to nonsynonymous substitutions on the terminal branches leading to the static and relocated duplicates respectively. Thus, when $SRN > 0$ the relocated duplicate has accelerated at the amino acid level compared to its paralog at the ancestral locus.

2.3.5 Prevalence of significantly asymmetric sequence divergence

We examined the prevalence of significantly asymmetric sequence divergence using a likelihood-based approach. For each pair of duplicates we tested whether a model of unconstrained evolution on the branches leading to each duplicate gave a significantly better fit to the data than a null model in which the duplicates were constrained to evolve symmetrically. We used **like-tri-test** (Conant and Wagner, 2003) to test three null models representing symmetry between duplicates with respect to synonymous divergence ($dS_1 = dS_2$), nonsynonymous divergence ($dN_1 = dN_2$), and strength of selective constraint ($\omega_1 = \omega_2$). For each of these tests we compared the likelihoods of the alternative models of constrained and unconstrained evolution. When twice the difference in log likelihoods exceeded 3.84 (χ^2 test $P \leq 0.05$) the null model of symmetric divergence was rejected and duplicate gene divergence was classed as asymmetric. Otherwise, the divergence of the duplicates was

designated symmetric for that measure. The purpose of this analysis was to calculate the relative prevalence of asymmetry between different types of duplicate, and not to determine whether sequence divergence was significantly asymmetric for an individual pair of duplicates. Therefore, we did not perform a multiple testing correction. Because this approach relies on the likelihood ratio test to assign duplicates as either symmetric or asymmetric but does not quantify the magnitude of sequence asymmetry we did not impose the filter used in the previous section that excludes duplicates with branch-specific values of dS and $dN < 0.001$.

2.3.6 Gene expression information

For each recent duplicate in mouse we looked for evidence of expression by using the predicted transcript as a Megablast (Zhang et al., 2000) query to mouse ESTs and cDNAs. We did not study gene expression in rat duplicates because this species has much lower overall EST coverage than mouse. Starting with all hits with $E <$ 1e-20, any ESTs with >75% of their sequence aligned with >97% nucleotide identity to only one of the two duplicates were assigned to that duplicate. Any ESTs matching both duplicates by these criteria were aligned to them using CLUSTALW. We then considered diagnostic sites at which the EST sequence shares an identical base with only one of the two duplicates, and where all three sequences are well aligned (*i.e.,* no gap occurs within 2 nt). Only if all diagnostic sites group the EST with the same duplicate did we assign the EST to that gene.

We assigned ESTs to tissues using the TissueInfo database (Skrabanek and Campagne, 2001), discarding ESTs from cancerous sources and keeping only those from normal unpooled tissues. For each tissue we quantified a gene's expression frequency using the count of its ESTs from that tissue expressed as a fraction of all ESTs derived from that tissue. We then used the highest expression frequency for a gene among all tissues to represent its peak expression (P). We quantified the asymmetry in peak expression between a given pair of duplicates using the absolute (unsigned) normalized difference in peak expression, $RP = |\frac{P_1-P_2}{P_1+P_2}|$, where P_1 and P_2 are the peak expression levels of each duplicate. For unlinked duplicates for which the direction of (retro)transposition could be determined we also quantified the direction of change in expression using the signed normalized difference in expression peak, $SRP = (\frac{P_r-P_s}{P_r+P_s})$, where P_s and P_r are the peak expression levels of the static and relocated duplicate, respectively. Similarly, we defined expression breadth (B) as

the number of distinct tissues represented among the ESTs assigned to the gene. Because retrogenes are sparsely sampled with ESTs (see Results) we could not reliably quantify the expression breadth of individual retrogenes. Thus if a retrogene is expressed ubiquitously but at a very low level its expression may appear tissue-specific purely as a consequence of low EST coverage. Although this prevented us from estimating asymmetry in expression breadth for individual pairs of retroduplicates (analogous to the measures of asymmetry in expression peak, RP and SRP) we were able to test whether the expression breadth of retrogenes as a group is significantly different to that of their static progenitor paralogs (see Results).

2.4 Results

We measured the magnitude of asymmetric sequence divergence among a set of 147 pairs of recent rodent duplicates (post-dating the rat/mouse divergence) that show at least minimal sequence divergence (branch-specific $dS > 0.001$ and $dN > 0.001$). We classified these pairs as either local (with < 5 intervening genes, $n = 62$ pairs) or distant duplicates ($n = 85$ pairs). Where possible, we categorised the mechanism of gene duplication as either DNA-based duplication (62 pairs) or retrotransposition (36 pairs). For many ($n = 54$) of the distant pairs, we were able to identify which gene copy was at the ancestral location and which was at a novel (transposed) location by comparison to the other rodent species (Figure 2-1A). In addition, we used a likelihood approach to investigate the prevalence of significant sequence asymmetry in a larger set of 81 local and 200 distant duplicate pairs (without the requirement $dS > 0.001$ and $dN > 0.001$, see Methods). This set included 91 DNA-based duplications and 110 retroduplications.

2.4.1 Asymmetry in d_N is greater among relocated duplicates and duplicates created by retrotransposition.

We first examined whether gene relocation and duplication mechanism are each individually related to asymmetric sequence divergence. Both variables are significantly associated with asymmetry in nonsynonymous evolution (RN). Distant duplicates show a more than two-fold increase in RN compared to local duplicates (Table 2.1). Duplication by retrotransposition is associated with a similarly large increase in RN relative to duplication by DNA-based transposition. Thus, on average, duplication by retrotransposition precip-

Table 2.1: Magnitude of relative asymmetry in d_S (RS), d_N (RN) and ω (Rω) between diverged ($d_S > 0.001$, $d_N > 0.001$) rodent duplicates categorised by location and mechanism of duplication. [a]

	N	Pairwise d_S [b]	RS [c]	RN [d]	$R\omega$ [e]
Duplicate location					
Local	62	0.090	0.302	0.284	0.259
Distant	85	0.070^{*}	0.290^{ns}	0.609^{***}	0.550^{***}
Distant (age-matched)	65	0.090^{ns}	0.274^{ns}	0.636^{***}	0.588^{***}
Duplication mechanism					
DNA-based transposition	62	0.075	0.300	0.316	0.303
Retro-transposition	36	0.059^{*}	0.272^{ns}	0.619^{***}	0.550^{ns}

[a] The Wilcoxon rank-sum test was used to determine the significance of differences between local and distant duplicates and between transposed (DNA-based) and retrotransposed (RNA-based) duplication categories. (ns: $P > 0.05$, $* : P < 0.05$, $** : P < 0.01$, $*** : P < 0.001$).
[b] Median pairwise d_S between duplicates.
[c] Median normalized difference in branch-specific d_S between duplicates (see Methods).
[d] Median normalized difference in branch-specific d_N between duplicates (see Methods).
[e] Median normalized difference in branch-specific d_N/d_S between duplicates (see Methods).

itates a more than 4-fold difference in rate between duplicates (median $RN = 0.619$). In contrast, for DNA-based duplicates there is a less than 2-fold difference in rates (median $RN = 0.316$). To determine whether the asymmetry of dN reflects imbalanced selective constraint between duplicates, we considered the relative asymmetry in ω. This measure ($R\omega$) is similarly increased among distant duplicates and among duplicates created by retrotransposition (Table 2.1).

Notably, we found no similar association between asymmetry in synonymous divergence (RS) and either duplicate relocation or duplication mechanism (Table 2.1). Therefore, the increase in RN associated with relocation and retrotransposition cannot be explained as resulting from mutational differences between duplicates. Moreover, this result suggests that if some of the gene duplicates have been created by transposition between different isochores, equilibration of silent sites in the transposed duplicate to local GC content has not led to a general increase in synonymous asymmetry.

In mammals, pairwise dS provides an approximate measure of divergence time. We

noticed that distant duplicates tend to be younger than tandem pairs created by local duplication (Table 2.1). This age difference alone might underlie the result above if a high degree of nonsynonymous asymmetry is a characteristic of the initial stages of duplicate gene differentiation. In order to exclude this possibility we used a subset of the oldest relocated duplicates whose median age matched that of the set of local duplicates (median pairwise dS= 0.09). Comparing these age-matched categories confirms our initial observation of increased asymmetry among distant compared to local duplicates (Table 2.1).

The impact of relocation on significantly asymmetric divergence as determined by the likelihood ratio test (Table 2.2) confirms our observations based on normalised difference measures of the magnitude of sequence asymmetry. Relocated duplicates more frequently show significant asymmetry in nonsynonymous divergence and selective constraint than local duplicates, but relocation is not associated with more frequent significant synonymous asymmetry.

Similarly, relating the mechanism of duplication to the occurrence of statistically significant asymmetry broadly supports the results from normalised difference measures. Retrotransposition leads more frequently to significant asymmetry in nonsynonymous divergence and selective constraint than does DNA-based duplication (although only for selective constraint is the difference significant; Table 2.2). Interestingly, we found a weak tendency for significant asymmetry in synonymous divergence to occur more often following DNA-based duplication than following retrotransposition ($p > 0.05$).

2.4.2 Separating relocation from retrotransposition.

We were able to confidently discern the mechanism of gene duplication as either DNA-mediated or RNA-mediated for 98 (67%; Table 2.1) of the 147 duplicate pairs with branch-specific $dS > 0.001$ and $dN > 0.001$ (for which we could quantify RN). This classification revealed a tight association between relocation and retrotransposition: two-thirds of the distant duplicates were formed by retrotransposition whereas none of the local ones were (Figure 2-1B). This raises the question of whether gene duplication mechanism and genomic relocation exert independent effects on rate asymmetry between duplicates.

To test this we partitioned the dataset in Figure 2-1B by distinguishing distant from local duplicates. This partition revealed a 138% increase in RN among distant duplicates ($n = 54$) compared to local duplicates ($n = 44$), similar to the results in the larger dataset

in the upper part of Table 2.1. We then tested whether an additional partition of the data based on the mechanism of duplication could explain any further variation in rate asymmetry. We introduced this second partition by comparing local duplicates created by DNA-based transposition ($n = 44$) and distant duplicates created by retrotransposition ($n = 36$). This revealed a 145% increase in RN in the latter category. We tested whether the P-value associated with this comparison (reflecting the significance of the impact of relocation on asymmetry in combination with the effect of retrotransposition) was more significant than a random partition of 36 genes derived from the set of distant duplicates (reflecting the significance of the impact of relocation alone on asymmetry). The observed P-value (for the comparison 'local, DNA-transpositions' versus 'distant, retrotranspositions') was lower than the equivalent P-values in 92.5% of comparable random partitions. Thus the increase in the magnitude of rate asymmetry between duplicates owing to retrotransposition is marginally significant ($p = 0.075$) once relocation has been accounted for. Using the same approach, we found a significant effect of genomic relocation on rate asymmetry independent of the mechanism of gene duplication ($p = 0.026$).

2.4.3 Directional sequence asymmetry: retrogenes accelerate relative to their paralogs.

The above results show that the magnitude of rate asymmetry between duplicated genes is strongly affected by the distance of duplication (distant versus local) and only marginally affected by the mechanism (RNA-based versus DNA-based duplication). We hypothesised that the duplication mechanism should however affect the direction of the asymmetry: for retrotransposed genes we would expect that the decoupling of the gene from its original promoter would cause altered (probably lower and narrower) expression and make the relocated copy (the retrocopy) more likely to accelerate than its static paralog. In contrast, for DNA-based duplication we might not expect sequence acceleration to be consistently associated with either the static or the relocated paralog. To investigate this hypothesis we introduced signed measures of sequence asymmetry that take account of the direction of transposition in distant duplicates.

For DNA-mediated duplicates we found no consistent tendency for either copy to accelerate; the median value of $SRN = -0.002$ for these duplicates was not significantly different from zero (Wilcoxon signed rank test: $p = 0.96$, $n = 12$). In contrast, following

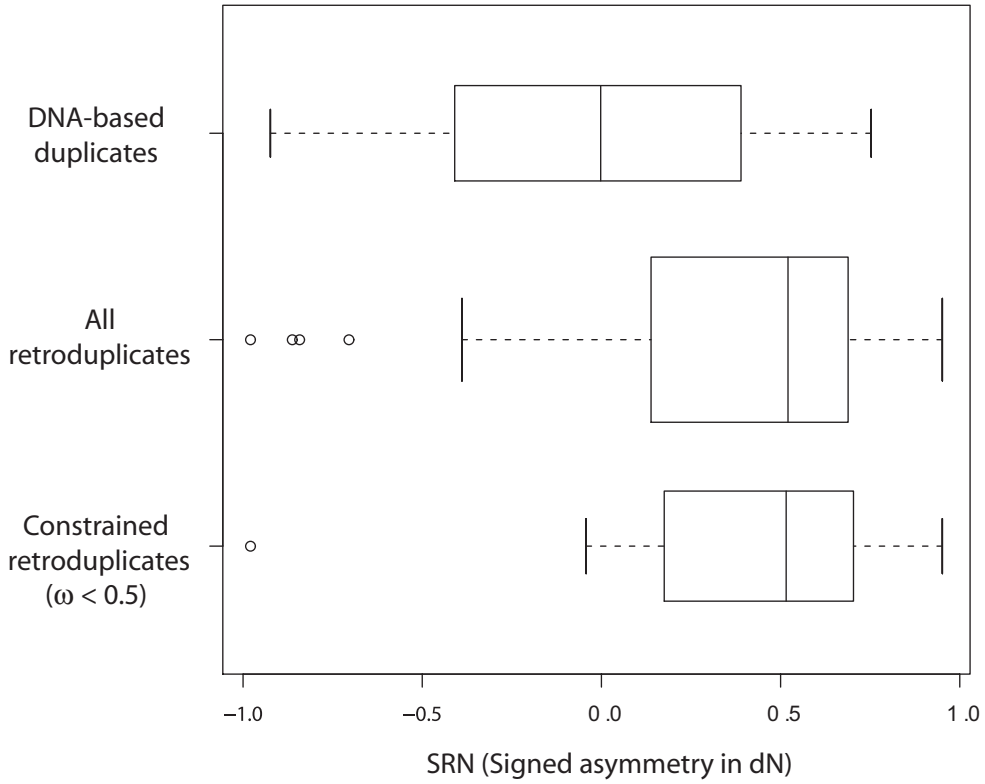


Figure 2-2: Signed nonsynonymous sequence asymmetry (SRN) among distant duplicates for which the direction of transposition is known. Duplicates were categorised as created by DNA-based transposition ($n = 12$) or RNA-based retrotransposition (retroduplicates, $n = 36$). A subset of retroduplicates under selective constraint ('constrained retroduplicates' with pairwise $\omega < 0.5$, $n = 16$) is enriched for putatively functional retrogenes and is likely to be depleted of retropseudogenes. The left and right edges of each box depict the first and third quartiles, respectively, whilst the vertical line within each box corresponds to the median. The left and right whiskers extend to the most extreme data point within 1.5 times the interquartile range of the first and third quartiles, respectively. The height of each box is proportional to the square root of each sample size.

duplication by retrotransposition there is a highly significant tendency for the relocated retrogene to accelerate relative to its static paralog (median $SRN = 0.52$, Wilcoxon signed rank test: $p = 0.001$, $n = 36$) (Figure 2-2). Relocated retrogenes also exhibited relaxed selective constraint relative to their static paralogs (median $SR\omega = 0.43$, Wilcoxon signed rank test: $p = 0.004$, $n = 36$) whereas transposed DNA-based duplicates showed no such tendency (median $SR\omega = 0.19$, Wilcoxon signed rank test: $p = 0.57$, $n = 12$).

One possible artefact that might affect the preceding result is the accidental inclusion of retropseudogenes, which would show strong and directional acceleration of sequence divergence due to their loss of selective constraint. Although all of the retrogenes in our dataset have intact ORFs, this is not in itself unequivocal evidence that they are functional (Marques et al., 2005). Therefore, we attempted to enrich our dataset for duplicate pairs

subject to relatively strong purifying selection, reasoning that these are likely to be functional. We made use of the fact that in a pairwise comparison of putatively protein-coding sequences, $\omega < 0.5$ indicates that selective constraint is operating on both sequences (Emerson et al., 2004). After application of this conservative filter to the set of 36 retrotranspositions with directional information, a total of 16 pairs of duplicates remained. Accelerated evolution of retrogenes was still seen in this subset of the data, as revealed by values of median SRN (0.51, Wilcoxon signed rank test: $p = 0.009$, $n = 16$) and median $SR\omega$ (0.50, Wilcoxon signed rank test: $p = 0.03$, $n = 16$) that are similar to those in the unfiltered dataset (Figure 2-2).

2.4.3.1 Greater expression asymmetry among distant duplicates due to lowering and narrowing of retrogene expression.

Both breadth of tissue distribution and peak gene expression level are known predictors of a gene's evolutionary rate (Duret and Mouchiroud, 2000; Pal et al., 2001; Zhang and Li, 2004; Subramanian and Kumar, 2004; Drummond et al., 2006). We therefore tested whether the increased rate asymmetry of distant gene duplicates reflects changes in expression associated with genomic relocation. For each pair of gene duplicates we used a stringent approach to assign a given EST or cDNA uniquely to a single paralog in each pair (see Methods). The low level of sequence divergence between duplicates meant that only a minority of duplicate pairs had independent expression evidence for both members of the pair.

Compared to local duplicates ($N = 25$), distant duplicates ($N = 29$) showed a marginally significant increase in asymmetry in peak expression (median RP: local duplicates = 0.55, distant duplicates = 0.73, Wilcoxon rank-sum test $p = 0.066$). Thus, distant gene pairs are more asymmetric in both sequence divergence and peak expression than local gene pairs. A large part of this disparity in expression asymmetry may be a consequence of the over-representation of retrogenes among distant duplicates. The hypothesis outlined earlier predicts that retrogenes should show lower and narrower expression relative to their static progenitor paralogs. We studied this using a signed measure of peak expression asymmetry (SRP, see Methods). Among 18 retrotranspositions SRP is significantly less than zero (median SRP = -0.69 , Wilcoxon signed rank test: $p = 0.037$). Because of the sparse EST sampling of retrogenes we were unable to derive a measure of asymmetry in expression breadth analogous to SRP for individual duplicate pairs created by retro-

Table 2.2: Likelihood ratio test: Prevalence of statistically significant asymmetry in d_S, d_N and ω between all rodent duplicates (without the requirement $d_S > 0.001$, $d_N > 0.001$) categorised by location and duplication mechanism. [a]

	N	Frequency of significant asymmetry		
		d_S [b]	d_N [c]	ω [d]
Duplicate location				
Local	81	20%	30%	12%
Distant	200	$16\%^{ns}$	$51\%^{**}$	$29\%^{**}$
Duplication mechanism				
DNA-based transposition	91	23%	36%	15%
Retro-transposition	110	$14\%^{ns}$	$47\%^{ns}$	$28\%^{*}$

[a] Differences in the proportion of significantly asymmetric duplicate pairs between local and distant duplicates and between transposed (DNA-based) and retrotransposed (RNA-based) duplication categories were tested using chi-square tests. (ns: $P > 0.05$; $* : P < 0.05$; $** : P < 0.01$).
[b] Frequency of gene pairs showing significant asymmetry in d_S.
[c] Frequency of gene pairs showing significant asymmetry in d_N.
[d] Frequency of gene pairs showing significant asymmetry in ω.

transposition. However, we found that retrogenes, collectively, show significantly narrower expression breadth compared to their static paralogs (data not shown).

We found a pronounced negative correlation between the signed measure of asymmetry in nonsynonymous rate (SRN) and in peak expression (SRP) ($r^2 = 0.29$, $p = 0.02$, $n = 18$). Thus nearly 30% of the rate acceleration of retrotransposed duplicates is explained by the decrease in their peak expression. Because expression peak and breadth are strongly correlated (Subramanian and Kumar, 2004) we expect part of this association to be mediated by narrowing of their expression breadth. However, for the reason outlined above, we could not estimate the magnitude of this effect and therefore could not exclude the possibility that asymmetry in expression breadth (rather than asymmetry in peak) is the more important determinant of rate asymmetry.

2.5 Discussion

In this study we tested the validity of the conventional view that gene duplication gives rise to redundant and functionally interchangeable paralogs (Ohno, 1970). Our results demon-

strate that the uncoupling of the fates of duplicated gene pairs is facilitated by both their physical relocation and the alteration of gene structure and regulation that follows retrotransposition. Moreover, because more than 60% of rodent duplicates with significantly asymmetric dN are generated by retrotransposition (Table 2.2) this implies that previous reports of frequent sequence asymmetry might be inflated as a result of simple violation of the dogma of 'equality at birth' of duplicated genes (Conant and Wagner, 2003) rather than representing the functional divergence of sister duplicates that are identical twins. Furthermore, the fact that roughly one-third of recent rodent duplicates are retrotranspositions suggests that the assumption of equality is frequently violated.

While it is difficult to disentangle the effects of retrotransposition from those of relocation, our results indicate that gene relocation (by any mechanism) has a strong impact on the asymmetry of protein evolutionary rates. This is consistent with the observation that the nonsynonymous evolution of linked genes occurs at similar rates Williams and Hurst 2000. Genes relocated by retrotransposition show only marginally more rate asymmetry than those relocated by DNA-mediated duplication, but the retrogenes show changes in the expected direction (resulting in narrower expression profiles and accelerated sequence evolution) whereas DNA-mediated distant duplications do not show any consistent direction of rate asymmetry (Figure 2-2).

The effect of duplicate relocation can be appreciated by considering the likely effect of shared genomic context among tandemly duplicated genes. If the span of duplicated DNA includes entire promoters then local tandem duplicates will initially share all their cis-regulatory elements (Katju and Lynch, 2003). Moreover, local duplicates should share the same distal regulatory elements (*e.g.*, locus control regions) in addition to residing in the same chromatin domain and gene neighbourhood. We therefore expect this shared genomic environment to result in co-regulation of local duplicates either as a consequence of selection for co-expression of functionally related genes (Lercher et al., 2002; Singer et al., 2005) or as a neutral consequence of proximity to the same regulatory elements (Spellman and Rubin, 2002; Sémon and Duret, 2006). Conversely, because relocated DNA-based gene duplicates differ in their chromosomal environments they are expected to show a limited degree of co-regulation mediated only by their shared core promoters.

Our results can be interpreted as illustrating the impact of the "scope" of gene duplication (i.e., the degree to which duplication conserves the structure of exons, introns

and promoter regions) on the magnitude of sequence asymmetry between duplicates. This echoes the observation that, at a broader scale, asymmetry in expression is greater among small-scale compared to large-scale (whole genome) duplicates (Casneuf et al., 2006). This is likely to reflect the fact that whole genome duplicates exemplify the concept of equality at birth by not only preserving the structure of gene duplicates but also maintaining their neighbourhood of flanking genes and regulatory sequences. However, it is worth noting that asymmetric divergence in sequence and expression is not exclusive to imperfect small-scale gene duplications but has also been observed among gene duplicates created by polyploidisation (Adams et al., 2003).

The effect of duplication mechanism on sequence asymmetry appears to be mediated by the regulatory changes associated with the process of retrotransposition. Strikingly, we found that for retrogenes nearly 30% of the variation in rate acceleration can be explained by lowering of expression peak. It is likely that the relationship between asymmetry in rate and in peak expression is partly due to narrowing of retrogene expression, although we could not quantify the contribution of expression breadth asymmetry directly. These results are consistent with accumulating evidence that the level and breadth of gene expression are among the most important determinants of the rate of protein evolution (Duret and Mouchiroud, 2000; Drummond et al., 2006) and echo a similar correlation between sequence asymmetry and expression divergence observed in a genome-wide analysis of yeast duplicates (Kim and Yi, 2006). The altered expression of a retrogene compared to its paralog may make it a more permissive target for the fixation of mutations since its lower (and narrower) expression renders some of these mutations less deleterious.

The difficulty in deriving unique expression evidence for each duplicate in combination with the variability in EST coverage between different tissues precludes a detailed *in silico* investigation of duplicate expression profiles. However, we found a general trend for retrogenes collectively to be expressed in a more limited range of tissues than their progenitor paralogs, in broad agreement with previous observations (Marques et al., 2005). Thus, relocated retrogenes show a reduction in both expression level and breadth compared to their static paralogs. These results are consistent with a loss of complex gene regulation accompanying retrotransposition. The survival of a functional retroduplicate is likely to be contingent on its expression. This can happen by acquiring a promoter *de novo*, co-opting a neighbouring gene's promoter, or through fortuitous integration into a transcribed region.

Moreover, because genes giving rise to retrocopies may tend to have exceptionally broad expression (Goncalves et al., 2000) this will magnify the disparity in breadth. It has been suggested that the expression pattern of retrogenes broadens as they evolve more complex regulatory sequences (Vinckenbosch et al., 2006). If this is correct we would expect a gradual equalisation of nonsynonymous rates between duplicates as the retrogene's expression profile broadens.

It has recently been suggested that following duplicative transposition in bacteria roughly one-third of cases show 'inconsistent' acceleration of the static paralog relative to the transposed copy (Notebaart et al., 2005). We note that Notebaart et al's observation is not based on an assessment of statistically significant asymmetry. However, our results based on the likelihood ratio test for statistically significant asymmetry in rodent duplicates show some support for the proposal that rate acceleration is not always consistent with relocation by transposition. Among 11 rodent DNA-based duplicate pairs with significant rate asymmetry (and with an assigned direction of transposition) 3 pairs (27%) showed 'inconsistent' acceleration of the static paralog. A weaker trend is seen following retrotransposition: among 51 retroduplicate pairs with significant rate asymmetry 7 pairs (14%) showed 'inconsistent' acceleration of the static paralog. Cases of significant acceleration of the static paralog following retrotransposition may reflect rare cases of functional displacement by a retrogene of its static paralog (Marques et al., 2005; Krasnov et al., 2005).

Natural selection can seize the opportunity for evolutionary exploration afforded by gene duplication only if the business of maintaining the ancestral gene function is assumed by one of the duplicates. Our results suggest that this division of labour is conservative: the daughter that inherits most of the ancestral gene features (exon/intron structure, regulatory elements and chromosomal neighbourhood) is likely to take on the parental role by default, while the positional and structural modification of its prodigal twin (in particular by retrotransposition) qualify it to take on the mantle of evolutionary entrepreneur.

2.6 Acknowledgements

We thank Marie Semon for assistance with the Homolens dataset and Meg Woolfit for critical reading of the manuscript. This study was supported by Science Foundation Ireland.

Chapter 3

Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons

The research described in this chapter has been published in
Molecular Biology and Evolution (Cusack and Wolfe, 2005).

3.1 Abstract

Alternative splicing is known to be an important source of protein sequence variation, but its evolutionary impact has not been explored in detail. Studying alternative splicing requires extensive sampling of the transcriptome, but new datasets based on ESTs aligned to chromosomes make it possible to study alternative splicing on a genome-wide scale. Although genes showing alternative splicing by exon skipping are conserved as compared to the genome as a whole, we find that genes where structural differences between human and mouse result in genome-specific alternatively spliced exons in one species show almost 60% greater non-synonymous divergence in constitutive exons than genes where exon skipping is conserved. This effect is also seen for genes showing species-specific patterns of alternative splicing where gene structure is conserved. Our observations are not attributable to an inherent difference in rate of evolution between these two sets of proteins, or to differences

with respect to predictors of evolutionary rate such as expression level, tissue specificity or genetic redundancy. Where genome-specific alternatively spliced exons are seen in mammals, the vast majority of skipped exons appear to be recent additions to gene structures. Furthermore, among genes with genome-specific alternatively spliced exons, the degree of non-synonymous divergence in constitutive sequence is a function of the frequency of incorporation of these alternative exons into transcripts. These results suggest that alterations in alternative splicing pattern can have knock-on effects in terms of accelerated sequence evolution in constant regions of the protein.

3.2 Introduction

Proteome diversity is expanded by the twin evolutionary engines of gene duplication and alternative splicing. Completion of whole genome sequences for a range of eukaryotes has revealed the pervasiveness of gene duplication in evolution. However, an appreciation of the prevalence of alternative splicing has had to await deeper sampling of the transcriptome. Alternative pre-mRNA splicing enables a single gene to encode many different mature mRNA transcripts and potentially several different protein products. Estimates of the fraction of alternative spliced human genes have increased as expressed sequence (EST and cDNA) databases have grown (Kan et al., 2002; Boue et al., 2003) and with the development of new technologies such as exon junction microarrays (Johnson et al., 2003). Current estimates are that at least 70% of human multi-exon genes are alternatively spliced (Johnson et al., 2003). At the same time both the fraction of genes that are alternatively spliced, and the number of isoforms generated per gene, appear to be roughly constant over a broad phylogenetic range of metazoa (Brett et al., 2002; Harrington et al., 2004).

Alternative splicing may result in exon truncation or extension, intron retention, or the inclusion/exclusion of entire exons by exon skipping. Different protein isoforms encoded by a single gene are likely to be variants of a constant protein backbone with the addition or deletion of entire alternative domains (Kriventseva et al., 2003). This enables some alternative isoforms to encode distinct functions, as has been demonstrated for transmembrane domains and protein-protein interactions (Xing et al., 2003; Resch et al., 2004b).

A definitive catalog of the types of alternative splicing occurring in a given organism would require both extensive transcriptome sampling and a finished genome sequence (Modrek and Lee, 2002). These requirements are closest to being fulfilled in human and in mouse.

Mapping ESTs onto genomic sequence (Modrek and Lee, 2002) reduces the contaminating effect of mixing paralogous sequences and other EST artifacts and allows alternatively spliced variants to be assigned to specific gene structures. This genomic-confirmation approach was used to create the ASAP database which provides a high quality platform for the annotation of alternative splicing in human and mouse (Lee et al., 2003).

Despite our growing appreciation of the incidence of alternative splicing in the generation of protein diversity, little is known about its evolutionary impact (Kopelman et al., 2005). This contrasts with the depth of research into the evolutionary impact of the other major mechanism of proteome expansion, gene duplication (Lynch and Conery, 2000; Kondrashov et al., 2002; Nembaware et al., 2002). In a key study Modrek and Lee described an association between alternative splicing and changes in the exon-intron structure of orthologous mouse and human genes resulting from lineage-specific gain or loss of exons (Modrek and Lee, 2003). This work suggested that alternative splicing may be used as a mechanism for evolution to try incorporating novel exons into a minority of a gene's transcripts (so-called "minor form" transcripts). Since the gene's ancestral function is maintained by the "major form" transcripts, this may free the minor form transcript from functional constraint, thus reducing purifying selection. This situation can be likened to the relaxation of constraints on recent gene duplicates, and for this reason minor transcripts generated by alternative splicing have been termed "internal paralogs" (Modrek and Lee, 2003). Evidence for relaxed selection on alternatively spliced sequence regions includes the observations that Alu-containing exons are always alternatively spliced (Sorek et al., 2002; Xing and Lee, 2004), and that a larger proportion of minor-form transcripts contain premature termination codons (PTCs) (Xing and Lee, 2004). Furthermore, it has recently been shown that alternatively spliced exons themselves show relaxation on sequence constraint with respect to amino acid substitutions (Xing and Lee, 2005).

Modrek and Lee's model (2003) predicts that relaxation of selective constraint on the minor transcript isoform will result in faster evolution of the alternatively spliced exon alone, but it makes no predictions about constraints on constitutively translated parts of the gene. Here we investigated whether the generation of an internal paralog through alternative splicing has an impact on selection operating on the entire gene. We considered only alternative splice events in human and mouse that result from exon skipping, and distinguished between conserved alternative splicing and alternative splicing that is specific to

either human or mouse. We show that these "genome-specific" alternatively spliced exons appear to be the result of exon gains following the human-mouse split. We find that although genes showing alternative splicing by exon skipping tend to be slowly evolving, the immediate impact of change in alternative splice pattern is acceleration of sequence evolution in the entire gene. Notably this acceleration is detected in constitutive exon sequence and may be a consequence of amino acid substitutions correlated with the accommodation of an alternatively spliced exon.

3.3 Methods

3.3.1 Human-mouse exon-skip conservation

We downloaded the ASAP dataset (Lee et al., 2003) from `http://www.bioinformatics.ucla.edu/ASAP/` in December 2003. This dataset includes conservation information for alternatively spliced exons (i.e., exon skip events) in human and mouse genes assigned as orthologous using Homologene data (Wheeler et al., 2004). Conservation of an exon skip event is recorded first with respect to sequence conservation of the alternatively spliced exon in the genomic DNA of the ortholog, and second by determining whether expressed sequence information supports both the inclusion and exclusion of the homologous exon from transcripts in the second species (transcriptomic evidence of alternative splicing of the exon). We defined conserved alternatively spliced exons as those having transcriptomic evidence of alternative splicing in both species. We defined an alternatively spliced exon as "genome-specific" when there is transcriptomic evidence for its alternative splicing in one species but no genomic evidence for its presence in the second species (see Figure 3-1).

3.3.2 Orthology mapping

ASAP lists the UniGene identifiers of human and mouse genes. We extracted the HUGO gene name for each human UniGene id in ASAP and mapped this to unique human and mouse LocusLink ids using Homologene (2 versions: Dec 2003, Jan 2004). LocusLink ids were then used as queries for the Ensmart tool (`http://www.ensembl.org/Multi/martview`) to obtain the associated human (NCBI build 34) and mouse (NCBIM build 32) Ensembl gene names and predicted protein and transcript

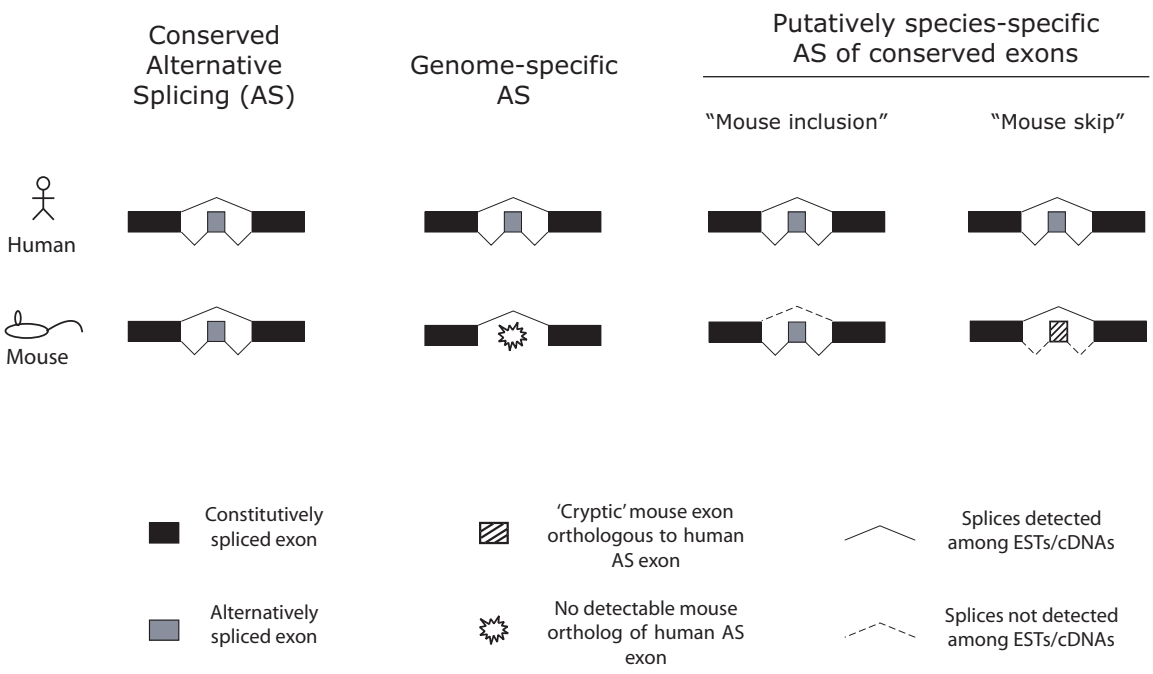


Figure 3-1: Categories of alternative splicing (AS) conservation retrieved from the ASAP database (Lee et al., 2003). For species-specific AS only human-specific events are depicted. See Methods for details.

sequences. Linking to Ensembl using direct Homologene information in this way yielded the sequences of 224 pairs of orthologs. For UniGene ids that we could not map to recent versions of Homologene, we used Ensmart to map the human gene name to a human Ensembl gene identifier and used reciprocal best BLASTP (Altschul et al., 1997) to assign a mouse ortholog. Sequences for a further 56 pairs of orthologs were derived in this step. For those genes that could not be linked to Ensembl via either Homologene or human gene name we used high stringency BLASTP to map a translation product inferred by ASP (Xing et al., 2004) for each gene to a human Ensembl predicted protein, followed by a reciprocal best BLASTP to assign a mouse Ensembl ortholog. This step found an additional 93 pairs of orthologs. Finally, we used BLASTN to verify all assignments of genes to Ensembl ids by ensuring that the Ensembl predicted transcript for a given gene matched its sequence derived from ASAP.

3.3.3 Identification of "representative orthologs" in fish

For each gene showing either conserved alternative splicing or genome-specific alternative splicing we used the human protein as query to detect a reciprocal best hit in fugu and in

zebrafish imposing an e-value < 1e-10 and requiring coverage of at least 50% of the longer sequence. Each pair of fugu and zebrafish orthologs identified is a representative ortholog pair (Davis and Petrov, 2004) belonging to one of two categories, one representing the evolution of genes for which alternative splicing is conserved between human and mouse, and the other representing the evolution of genes with genome-specific alternative splicing where the patterns of alternative splicing differ between human and mouse.

3.3.4 Assessing levels of selective constraint

Human and mouse protein sequences were aligned using CLUSTALW (Thompson et al., 1994) and back-translated to generate a codon-based alignment of transcripts. For each gene we used ASAP annotations to extract the sequences of exons that undergo alternative splicing by exon-skipping. Parts of the transcript alignment corresponding to these exons were masked. We calculated d_N and d_S for the unmasked (constitutive) sequence using the yn00 program in the PAML package (Yang, 1997). For fugu-zebrafish representative orthologs d_N and d_S were calculated based on the entire transcript alignments since no information was available on alternative splicing of exons in these organisms.

3.3.5 Determining alternatively spliced exon presence/absence in the human-mouse ancestor

We used chicken as an outgroup to determine whether a given alternatively spliced exon was present in the human-mouse ancestor. Three strategies were employed to detect homologs of human alternatively spliced exons in either the chicken genome or transcriptome. First, translations of the alternatively spliced exon sequence plus 90 nucleotides from each flanking exon were used as TBLASTN queries against the chicken genome. These were required to hit a stretch of chicken chromosome having an "anchoring" TBLASTN match (E $\leq$ 1e-5) to the Ensembl predicted translation of the human gene. The flanking sequence was used to further "anchor" hits to the chromosome and only hits in which at least two thirds of both flanks were aligned with $\geq$ 50% amino acid similarity were scored. The alternatively spliced exon was scored as "detected' if at least two thirds of its length was aligned with $\geq$ 50% similarity, or as "not detected" otherwise. Second, the alternatively spliced exon sequence alone was used as a BLASTN query against chicken ESTs. Alternatively spliced exons with $\geq$ 80% of their length aligned, $\geq$ 70% identity and E < 0.001 were scored as "detected", or

"not detected" otherwise. The third approach involved a "low stringency" search strategy that did not require the detection of conservation of the alternatively spliced exon sequence itself. The alternatively spliced exon sequence plus 90 nucleotides from each flanking exon was used as a BLASTN query against chicken ESTs. Only hits in which $\geq 50\%$ of both flanks were aligned were scored. The alternatively spliced exon was scored as "detected" if the intervening EST sequence between the aligned flanks was $\geq$10 nt, or "not detected" otherwise. Finally, to identify "high confidence" cases where we expect to see a chicken homolog for an alternatively spliced human exon if it exists we applied a Binomial test as outlined in Kan et al. (2002). We only scored those alternatively spliced exons having sufficiently high splicing frequency in human and for which chicken ortholog EST coverage is deep enough that we expect to detect chicken homologs of these exons.

3.3.6 Influence of frequency of incorporation of alternatively spliced sequence

For each human genome-specific alternatively spliced exon classified as translated and incorporated into productive transcripts according to ASP annotation, we counted the number of ESTs in ASAP supporting each of the two alternative splices: inclusion and exclusion of the exon. We performed the Binomial test employing two threshold cutoffs (we chose $t = 0.03$ and $t = 0.12$ to produce roughly equally sized subdivisions of the data) to categorise each exon skip as belonging to one of three frequency classes as used in Figure 3-3. For example, an alternatively spliced exon whose inclusion frequency satisfied the binomial test at the 95% confidence level (ie. $p < 0.05$) for the lower threshold frequency ($t = 0.03$) but not for the upper threshold frequency ($t = 0.12$) was classified as incorporated at intermediate frequency. We compared each frequency category of human genome-specific alternatively spliced exons with respect to the non-synonymous divergence (d_N) calculated for constitutive sequence in the cognate gene.

3.3.7 Level and breadth of constitutive exon expression

The expression level of the constitutive exons of each gene was approximated using a simple count of all EST/cDNA sequences mapped to that gene by ASAP. Breadth of expression was determined by assigning each EST/cDNA to one of 34 tissue classes using TissueInfo (Skrabanek and Campagne, 2001).

3.3.8 Estimating adequacy of mouse EST sampling in genes with putatively human-specific alternative splicing

Genes with putatively human-specific alternative splicing of conserved exons were identified as those where gene structure is conserved in human and mouse but where alternative splicing is only observed in human. For each gene in this group we determined the number of mouse ESTs as a fraction of the number of human ESTs sampled. This scaling gives a measure of how adequate mouse EST coverage should be in recovering any conserved alternative splicing events under the assumption that alternative splicing occurs with equal frequency in human and mouse. We considered three different measures of depth of EST coverage by counting: (i) all human and mouse ESTs assigned to a gene; (ii) human and mouse ESTs from a set of named tissues only (this excludes ESTs from cancerous sources); (iii) human and mouse ESTs from the tissue(s) in which the putatively human-specific splice event (exon inclusion or skipping) is observed.

3.4 Results

3.4.1 Genes showing exon-skipping are more conserved than the genome average

We downloaded a set of 14,596 human-mouse orthologs with assigned gene names from Ensembl and classified them as either exhibiting exon-skipping (6,580 genes) or as having no evidence of exon-skipping ('control set' of 8,016 genes) based on ASAP annotation (Lee et al., 2003). We compared sequence constraint in the alternatively spliced genes to that of genes in the control set. A commonly used measure of the degree of evolutionary constraint on a sequence is the ratio of non-synonymous substitutions per non-synonymous site (d_N) to synonymous substitutions per synonymous site (d_S). For values of $d_N/d_S \leq 1$, this ratio is generally highest for genes whose sequences are weakly constrained by purifying selection. However, in the case of alternatively spliced sequence the d_N/d_S ratio has to be interpreted with greater caution because the inherent assumption that silent sites in codons are selectively neutral is more likely to be incorrect. The presence of exonic splicing enhancer (ESE) motifs in alternatively spliced exons means that nucleotide changes that disrupt these motifs are likely to be detrimental to function and are therefore subject to purifying selection (Iida and Akashi, 2000; Orban and Olah, 2001). It is not known if the

Table 3.1: Medians and Standard Deviations of d_N/d_S and d_N from human/mouse orthologs showing alternative splicing (AS) by exon-skipping, and from a control set of human/mouse orthologs for which no exon-skipping has been described.

	N	d_N/d_S	d_N
AS genes	6580	0.089 (0.123)	0.053 (0.085)
Control genes	8016	0.117 (0.136)	0.071 (0.096)
		$p < 1e-15$	$p < 1e-15$

NOTE.- Significance was tested using a two tailed Wilcoxon rank sum test.

constraint imposed by ESE motifs is of equal strength at synonymous and non-synonymous sites or whether these motifs have evolved to have minimal impact on the encoded amino acid sequence. For this reason the usefulness of d_N/d_S as a measure of selective constraint on alternatively spliced exons is uncertain (but see Xing and Lee (2005)).

We found that genes undergoing alternative splicing by exon-skipping were more constrained than the control set (Table 3.1). We could also compare d_N for these human-mouse orthologs because they all share a common divergence time. The slower evolution of alternatively spliced genes relative to the genome average is equally striking when we consider d_N alone (Table 3.1). The observed differences are made more conservative by the fact that estimates of both d_N and d_N/d_S for alternatively spliced exons are higher than for constitutive exons (Iida and Akashi, 2000; Xing and Lee, 2005).

3.4.2 Genome-specific alternative splicing is associated with faster protein evolution and weaker selective constraint in constitutive regions

The level of constraint on a protein sequence is likely to differ according to the protein's function. If genes in different functional categories employ alternative splicing to different extents, this could explain why alternatively spliced genes are conserved compared to the genome as a whole. To test the influence of functional bias we focused only on genes that undergo alternative splicing through exon-skipping, and classified them as showing either (i) exon skipping conserved between human and mouse, or (ii) genome-specific exon skipping where the alternatively spliced exon is found in genomic DNA of one species only. The latter group was defined based on failure of a BLASTN search to detect the alternatively spliced exon in genomic DNA of the ortholog. This fact coupled with lack of evidence for

a homologous exon among ESTs in the second species indicates unambiguously that the alternatively spliced exon, and therefore the alternative splicing event, is genome-specific. Throughout this paper, we use the terms "human genome-specific alternative splicing" and "mouse genome-specific alternative splicing" to denote alternative splicing events that are specific to one genome and where the other genome has no ortholog of the alternatively spliced exon (see Figure 3-1). We confirmed using GOstat (Beissbarth et al., 2004) that although genes with exon skips are biased towards certain functional terms, there was no difference in the functions performed by genes with conserved alternatively spliced exons and genes with genome-specific alternatively spliced exons (data not shown). The set of genes with conserved alternatively spliced exons therefore serves as a function-matched control for comparison to genes with genome-specific alternatively spliced exons. Although the classification of genes in these two categories is unambiguous, it should be noted that the groups differ in their degree of gene structure conservation. This potential source of bias was assessed in a later test comparing genes with putatively species-specific alternative splicing patterns but conserved gene structures (see the final section of Results).

Using human/mouse orthologs for the two groups we find that d_N in genes with genome-specific alternatively spliced exons is 33% greater than in genes with conserved alternative splicing (d_N= 0.061 vs. 0.046; Table 3.2). There is also a comparable difference in the d_N/d_S ratio. However, a strict comparison between these groups requires us to account for the possibility of differing selective pressures on alternatively spliced exons compared to constitutive exons (Xing and Lee, 2005). For genes with conserved exon-skipping the conserved alternatively spliced exon is included in the human-mouse alignment and thus contributes to the calculation of d_N and d_S but this is not the case for genome-specific alternatively spliced exons. Omitting the sequence of all alternatively spliced exons from all genes had a negligible effect on the calculated values of d_N, but for genes with conserved alternative splicing it reduced the estimate of d_N/d_S as expected (Table 3.2). Thus when we consider constitutive sequence only we see that d_N/d_S is 48% greater in genes with genome-specific alternative splicing than in genes where alternative splicing is conserved. This result suggests that there is an association between changes in a gene's alternative splicing pattern and an increase in the rate of sequence evolution in the constant part of the protein.

Table 3.2: Medians and Standard Deviations of d_N and d_N/d_S from orthologous comparisons for genes with alternative splicing (AS) conserved between human and mouse compared to genes with genome-specific alternative splicing in human or mouse.

	d_N	d_N/d_S	N
All exons			
Conserved AS	0.046 (0.065)	0.086 (0.099)	68
Genome-specific AS	0.061 (0.089)	0.108 (0.140)	286
	$p < 0.050$	$p < 0.050$	
Constitutive exons			
Conserved AS	0.046 (0.068)	0.075 (0.108)	66
Genome-specific AS	0.060 (0.091)	0.111 (0.148)	285
	$p < 0.050$	$p = 0.055$	
Productive AS, constitutive exons			
Conserved AS	0.049 (0.063)	0.075 (0.091)	51
Genome-specific AS	0.078 (0.102)	0.130 (0.182)	93
	$p < 0.005$	$p < 0.01$	
Fish representative orthologs[a]			
Conserved AS	0.134 (0.108)	0.058 (0.055)	30
Genome-specific AS	0.160 (0.161)	0.061 (0.070)	124
	$p > 0.1$	$p > 0.1$	

NOTE.- Significance was tested using a two tailed Wilcoxon rank sum test.

[a] d_N and d_N/d_S for fugu vs. zebrafish orthologs of mammalian genes that show conserved AS or genome-specific AS in human or mouse

3.4.3 Productive alternative splicing

The inclusion of an alternatively spliced exon may induce a frameshift and introduce a premature termination codon (PTC) into the transcript resulting in transcript degradation by nonsense mediated decay (NMD)(Nagy and Maquat, 1998). Although alternative splicing-coupled NMD can have a regulatory role, these alternative splicing events do not increase the gene's protein-coding potential and we therefore consider them "unproductive". Notably, a greater proportion of non-conserved alternatively spliced exons induce frameshifts than conserved alternatively spliced exons, and many of these are likely to initiate NMD (Sorek et al., 2004).

We used the ASP database (Xing et al., 2004) of predicted alternatively spliced transcript sets inferred from EST and cDNA data for a given human gene and repeated our analysis, excluding all cases generating transcripts with a PTC. We further focused on those

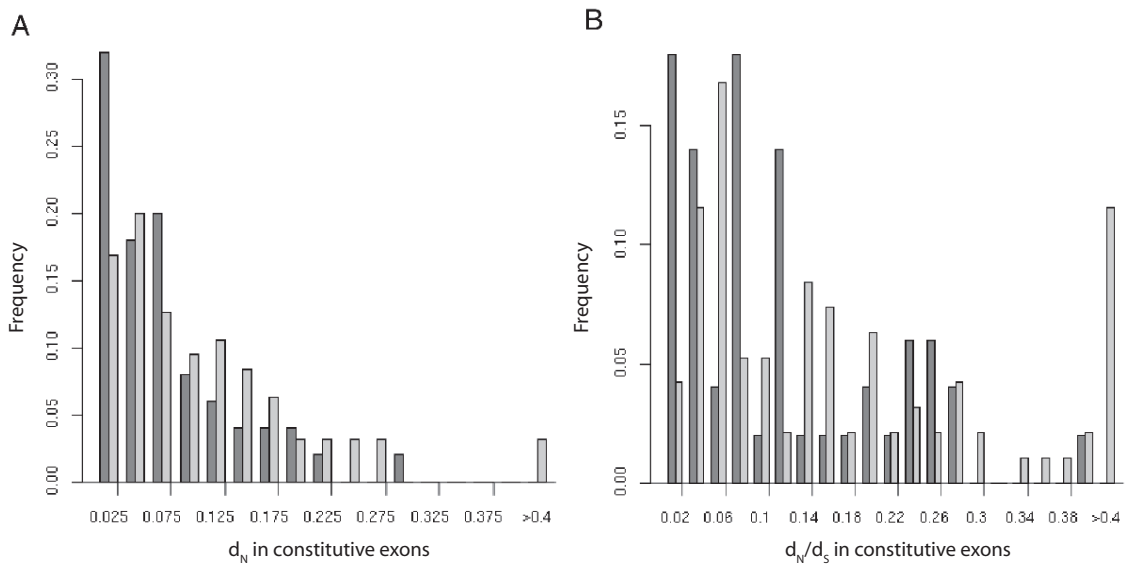


Figure 3-2: Distributions of (A) d_N and (B) d_N/d_S for constitutive exons from human/mouse orthologous comparisons of 50 genes with conserved alternative splicing (dark gray) and 95 genes with human genome-specific alternative splicing (light gray). All alternative splicing events overlap the ORF and generate productive transcripts without PTCs.

cases in which the alternatively spliced exon overlaps the reading frame of the transcript This restricts the analysis to genes undergoing productive alternative splicing that is likely to generate a distinct protein product. Since the ASP database currently contains only inferred transcripts from human we compared genes with conserved alternative splicing in human and mouse (51 genes) to those with human genome-specific alternative splicing (93 genes), considering only productive alternative splicing in both cases. The distributions of d_N and d_N/d_S for constitutive exons from human/mouse orthologous comparisons for both sets of genes are shown in Figure 3-2. Genes with human genome-specific alternative splicing showed a 59% increase in median d_N ($p < 0.005$) and a 73% increase in median d_N/d_S ($p < 0.01$) in their constitutive sequence when compared to genes with conserved alternative splicing (Table 3.2). Therefore the observed increase in evolutionary rate in genes undergoing genome-specific alternative splicing holds for productive alternative splicing events. It is important to note that this increase in d_N was observed in constitutive exons and is distinct from the acceleration reported in alternatively spliced exons (Xing and Lee, 2005).

3.4.4 Differences in strength of selective constraint in mammals are not a reflection of inherent constraint differences

The difference in substitution rates associated with conserved versus genome-specific alternative splicing may lie in an inherent difference between these two classes of genes. Genes under relaxed selective constraint may be more liable both to change their gene structure by gaining or losing an alternatively spliced exon and to have faster rates of sequence evolution.

We addressed this issue by examining the substitution rates in genes independently of the effects of changes in alternative splicing that have emerged during the course of mammalian evolution, by using the "representative orthologs" method of Davis and Petrov (2004). For each pair of human/mouse orthologs we searched for fugu and zebrafish orthologs. We calculated divergence between the two fish species for two groups of genes, according to whether their mammalian orthologs showed conserved alternative splicing or genome-specific alternative splicing. In contrast to the differences seen for the mammalian genes, we found no significant difference in d_N or d_N/d_S between the two groups of fish orthologs (Table 3.2). This is partly a consequence of the smaller size of the representative ortholog sample since we did detect an increase in d_N in fish orthologs for genes that show genome-specific alternative splicing in mammals, but this was less dramatic than the increase seen in the study orthologs (Table 3.2). However, since no difference was seen in d_N/d_S in the fish comparison we conclude that there is no inherent difference in selective constraint between the two classes of alternatively spliced gene. In addition this result suggests that a simple sampling bias does not underlie the difference we observe between these two classes of genes in mammals.

3.4.5 Genes that have changed in alternative splicing pattern have also undergone changes in d_N/d_S ratio

We next looked for indications that changes in alternative splicing pattern have resulted in changes in selective constraint on a gene during the course of mammalian evolution. For a given gene, comparing the d_N/d_S ratio for human/mouse to that for fugu/zebrafish gives an indication of any change in the strength of selective constraint operating on that gene. We considered only those genes for which we were able to calculate d_N/d_S for both the mammal and the fish species pairs. We found that of 124 genes showing genome-specific alternative splicing (i.e., either alternative splicing of an exon in human but not in mouse,

or vice versa), 77 had higher d_N/d_S in mammals than in fish ($p = 0.003$, Binomial test). In contrast, of 29 genes with alternative splicing conserved between human and mouse, only 15 had higher d_N/d_S in mammals than in fish ($p = 0.355$, Binomial test). This test is conservative because it ignores the magnitude of the difference in d_N/d_S. We therefore performed a second comparison considering the distributions of d_N/d_S from human/mouse and fugu/zebrafish orthologs. For the 29 genes in which alternative splicing is conserved between human and mouse, d_N/d_S did not differ significantly in the cross-taxon comparison between mammals (Median $d_N/d_S = 0.066$) and fish (Median $d_N/d_S = 0.055$) (Wilcoxon rank-sum test $p = 0.97$) (Table 3.2). On the other hand, the 124 genes showing alternative splicing in human but not in mouse (or vice versa) are significantly less constrained in mammals (Median $d_N/d_S = 0.086$) than in fish (Median $d_N/d_S = 0.061$) (Wilcoxon rank-sum test $p = 0.003$).

3.4.6 Difference in d_N/d_S ratio is not due to bias with respect to known predictors of evolutionary rate

We looked for alternative explanations of our results using three important predictors of rate of sequence evolution of a gene, namely expression level, breadth of expression, and genetic redundancy. Highly expressed genes are more conserved than genes expressed at low levels (Krylov et al., 2003), and broadly expressed genes are more conserved than genes expressed only in a subset of tissues (Duret and Mouchiroud, 2000; Huminiecki and Wolfe, 2004; Zhang and Li, 2004). It is not known whether there are differences in expression level or breadth between genes with conserved alternative splicing and genes with genome-specific alternative splicing (Resch et al., 2004a) and there is no *a priori* reason to suspect any. However, if these variables cannot be eliminated as possible explanations for our result we do not need to invoke any other, less trivial, explanations. The difference we see in evolutionary rate relates to constitutive exons. So we set out to determine the level and breadth of expression of constitutive exons in each gene by pooling EST information from all its alternative transcript isoforms.

Using the number of assigned ESTs mapped to the genome for each gene as a simple measure of its expression we found no difference in expression levels of genes in the two categories of alternative splicing conservation. In human the median number of ESTs for genes with conserved alternative splicing and genome-specific alternative splicing were 72

and 69, respectively. The corresponding numbers for mouse are 37 and 39. Similarly, there is no difference in breadth of expression for genes in the two alternative splicing conservation categories. The median number of human tissues showing evidence of expression was nine both for genes with conserved alternative splicing and for genes with genome-specific alternative splicing.

Selective constraint can also be affected by presence of a close paralog. Genes that have undergone recent duplication experience relaxation of purifying selection corresponding to a period of functional redundancy (Kondrashov et al., 2002) and this is detected as an increase in d_N/d_S between the paralogs (Lynch and Conery, 2000; Jordan et al., 2004). We tested whether our two categories of genes (conserved alternative splicing and genome-specific alternative splicing) differed with respect to possession of a close paralog. The median value of d_S to the nearest paralog did not differ between categories (data not shown) therefore the difference in orthologous d_N/d_S between categories can not be explained as resulting from different propensities to undergo gene duplication.

3.4.7 Genome-specific alternatively spliced exons are likely to be exon gains

If the association between having a genome-specific alternatively spliced exon and faster protein evolution reflects causation, our observations suggest one of two possibilities. First, genome-specific alternatively spliced exons could be recent gains in one lineage that have had a knock-on effect of speeding up protein sequence evolution. Alternatively, genome-specific alternatively spliced exons could be due to recent exon losses in the sister lineage, which would imply that loss of alternatively spliced sequence accelerates the substitution rate.

We attempted to distinguish between these two possibilities by using chicken (Hillier et al., 2004) as an outgroup species to determine the direction of change. We used human alternatively spliced exons absent from mouse to search both the chicken genome and transcriptome, and compared their detection rate to that of alternatively spliced exons conserved in human and mouse. We did not use mouse-specific alternatively spliced exons for this analysis because the number of cases, and the EST coverage of mouse, is lower.

Direct sequence matches between alternatively spliced exons and chicken chromosomes recovered putative chicken homologs for conserved human-mouse alternatively spliced exons

Table 3.3: Results of searching for homologs of human alternatively spliced (AS) exons in the chicken genome.

	Detected	Not detected	Not scored[a]
TBLASTN Exon + 90bp flanks to chicken genome			
Conserved AS	15 (75%)	5 (25%)	41
Genome-specific AS	1 (1%)	71 (99%)	155
BLASTN Exon to chicken ESTs			
Conserved AS	21 (34%)	40 (66%)	n/a
Genome-specific AS	2 (1%)	225 (99%)	n/a
BLASTN Exon + 90bp flanks to chicken ESTs			
Conserved AS	12 (80%)	3 (20%)	46
Genome-specific AS	1 (4%)	25 (96%)	201
BLASTN Exon + 90bp flanks to genes adequately sampled with chicken ESTs			
Conserved AS	10 (83%)	2 (16%)	49
Genome-specific AS	1 (50%)	1 (50%)	225

NOTE.- 'Conserved' alternatively spliced exons are conserved with respect to human and mouse. 'Genome-specific' alternatively spliced exons are found in a human gene but absent from genomic sequence of the mouse ortholog. Chicken genes with sufficient EST coverage to be confident of recovering homologs for a given human alternatively spliced exon were identified on the basis of a binomial test (Kan et al., 2002). See Methods for details.

[a] Exons without anchoring matches to genomic or EST sequence were not scored. In addition, for the bottom panel exons for which chicken EST coverage was inadequate were not scored.

much more frequently than for alternatively spliced exons that are present in human but not mouse. The detection rate for the latter category was close to zero (Table 3.3). These results point towards exon gain as the source of exons that are alternatively spliced in human but are absent from the mouse genome, lending support to the first possibility above.

The validity of this assertion depends on the assumption that BLAST has the same

power to detect chicken homologs of exons in the two classes (conserved alternative splicing and genome-specific alternative splicing) between human and mouse. This may not be the case if exons in the latter category are faster evolving, in which case failure to detect a BLAST hit in the chicken genome for a given exon cannot be taken as evidence of its absence from chicken. However, we think it is unlikely that a difference in evolutionary rates alone could produce the sort of qualitatively different results for the two classes seen in Table 3.3.

One way to partly account for possible rate differences among exons is to use a low stringency search of the chicken transcriptome for putatively homologous chicken exons without requiring a direct sequence match to the alternatively spliced exon. We did this by searching chicken ESTs with a human query consisting of the alternatively spliced exon plus additional sequence from its flanking exons. Chicken ESTs aligning to the sequence of both flanking exons and which contain a stretch of intervening EST sequence were scored as containing a chicken homolog of the human alternatively spliced exon even in the absence of any detectable sequence similarity to the exon itself. However, this approach is itself based on the assumption that all the alternatively spliced exons in question are spliced at equivalent frequencies, but the two sets of alternatively spliced exons under study here show a significant difference in their frequency of incorporation. Non-conserved alternatively spliced exons are spliced at low frequencies into minor-form transcripts, whereas alternatively spliced exons conserved between human and mouse are generally represented among major-form transcripts (Modrek and Lee, 2003). This means that a given number of chicken ESTs may be sufficient to detect a homolog of a human alternatively spliced exon if it is found in the major-form transcript, but not if it is exclusive to the minor-form.

We attempted to allow for splicing frequency differences by considering only "high confidence cases", i.e., alternatively spliced exons whose splicing frequency in human and EST coverage in chicken is such that we expect to detect homologs in chicken if they do exist (Kan et al., 2002). Since only a small number of such high-confidence cases exist among alternatively spliced exons found in human but not in mouse, we had insufficient evidence from this low stringency strategy to determine the ancestry of many exons. Thus we conclude that it is likely, but not certain, that genome-specific alternative exons are gains.

3.4.8 Influence of frequency of incorporation of alternatively spliced exons

If the gain of an alternatively spliced exon is responsible for increasing the rate of amino acid change in constitutive regions of the gene then we might expect the strength of this effect to be proportional to the frequency at which the alternatively spliced exon is spliced into mRNA. Considering only genome-specific alternatively spliced human exons that are translated and productive we classified each alternatively spliced exon by its frequency of incorporation and binned the alternatively spliced exons into three frequency categories on this basis. A strong correlation was detected between the binned splicing frequency and d_N for constitutive exons (Spearman rank correlation rho = 0.353, $p < 0.001$, $n = 107$). The median values of d_N for genes with genome-specific alternatively spliced exons incorporated at low ($n = 36$), medium ($n = 37$) and high frequency ($n = 35$) were 0.053, 0.086 and 0.122 respectively (Figure 3-3). A different method of classifying exons by splicing frequency is based on the counts of ESTs that either include or exclude the exon and uses inclusion thresholds of 33% and 66% (Resch et al., 2004a) to produce low, medium and high frequency bins. Using this approach gave us a very similar result (not shown). However, the classification of alternative splicing frequency using either approach introduces a bias because low frequency alternative splicing events are more easily detectable in highly expressed genes and gene expression level is a known correlate of evolutionary rate (Krylov et al., 2003).

To establish whether the slower evolution of genes with lower alternative exon inclusion frequency is explained by their higher expression level, we calculated the partial correlation between splicing frequency and d_N controlling for EST coverage (Spearman partial correlation rho = 0.288, $p < 0.01$, $n = 107$). This confirms that there is a positive correlation between frequency of human genome-specific alternative exon inclusion and evolutionary rate in the constitutive parts of the gene, independent of EST coverage level. In contrast no such relationship was found between alternative splicing frequency and d_N among a control set of alternatively spliced exons conserved between human and mouse (not shown).

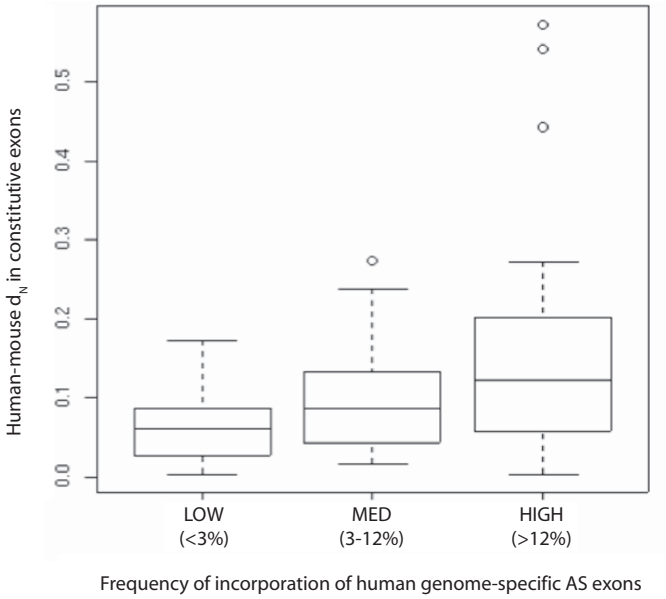


Figure 3-3: Association between frequency of incorporation of human genome-specific alternatively spliced exons that are translated into productive alternatively spliced variants and d_N of constitutive gene sequence between human and mouse. The lower and upper bounds of each box depict the first and third quartiles, respectively, whilst the horizontal line within each box corresponds to the median. The lower and upper whiskers extend to the most extreme data point within 1.5 times the interquartile range of the first and third quartiles, respectively.

3.4.9 Species-specific alternative splicing in genes with conserved exon-intron structure

The results described above indicate a higher rate of protein evolution among genes having genome-specific alternatively spliced exons compared to genes where there is conservation of both the alternatively spliced exon itself and of each alternative splicing event (exon inclusion and exclusion) as detected in human and mouse ESTs. The advantage of contrasting these two groups is that genes in the former group can be unambiguously classified as undergoing species-specific alternative splicing, because there is no genomic evidence of the alternatively spliced exon in the orthologs of these genes. However, these two classes of genes differ not only in the degree of conservation of their alternative splicing patterns but also in the degree of conservation of gene structure. Thus alternative splicing alone may not underlie the described disparity in rates of protein evolution because this may simply reflect faster sequence evolution of genes for which evolution of gene structure is also fast.

In order to account for any effect of changes in gene structure during evolution, we considered genes whose gene structure is conserved between human and mouse but whose

alternative splicing pattern appears to have changed. These genes show evidence of both inclusion and exclusion of an alternatively spliced exon in the first species (e.g., human) but no evidence for alternative splicing of that exon in the second species (e.g., mouse). Classification of these genes is problematic because for some genes conserved alternative splicing may not be detected in one species due to undersampling of ESTs in the second species. Alternatively, some genes in this group may be undergoing truly species-specific alternative splicing. Since this group is likely to be a mixture of both types of gene we consider these genes as having genomically conserved exons whose alternative splicing conservation is "unclassified".

We retrieved two sets of such "unclassified" genes from the ASAP database. Both sets consist of genes showing evidence of alternative splicing in human. In the first set ("mouse skip" set) the mouse ortholog shows no EST evidence of inclusion of the exon, but there is sufficient sequence conservation in the mouse genome to suggest that a cryptic, possibly functional, exon exists. In the second set ("mouse inclusion" set) there is no EST evidence for skipping of the relevant exon in the mouse ortholog.

We consider genes in these two sets to show putatively human-specific alternative splicing. It is important to note that we cannot distinguish truly human-specific alternative splicing in these genes from alternative splicing that is simply more frequent in human than in mouse and has not yet been detected in mouse. In determining the effect on evolutionary rate, these sets of genes are only informative if they can be considered to be enriched for genes having human-specific alternative splicing. We therefore asked whether EST sampling of the mouse genes in these sets is sufficiently deep that we would expect to observe both exon inclusion and exon skipping in the mouse ESTs if alternative splicing were conserved. If a given mouse gene has been adequately sampled with ESTs and we still fail to observe a mouse counterpart for an alternative splicing event seen in human, then we can be more confident that the apparent human-specific alternative splicing in this gene is real. Using a range of approaches we found that mouse EST coverage in the "mouse skip set" and in the "mouse inclusion set" is comparable with or even better than the control set (Table 3.4). It is notable that, even in control genes where mouse EST coverage is adequate and detects conservation of alternative splicing, mouse EST coverage is only half that of human.

On this basis both the mouse skip and mouse inclusion categories can be considered to be enriched for human-specific alternative splicing of conserved exons. This assertion is

Table 3.4: Mouse EST coverage (expressed as the median percentage of human EST coverage for each gene) for genes having alternatively spliced (AS) exons in human but for which the homologous mouse exon is either consistently skipped ("mouse skip") or consistently included ("mouse inclusion").

	Conserved Exon Skip	"Mouse-skips"	"Mouse-inclusion"
All ESTs	53%	47% ($p > 0.1$)	48% ($p > 0.1$)
ESTs from named tissues	41%	58% (NS)	53% (NS)
ESTs from tissues in which human AS exon is:			
Included	29%	36% (NS)	—
Skipped	36%	—	50% (NS)

NOTE.- The median values shown are compared to those of the control group with conserved exon-skipping, where mouse EST coverage is adequate by definition. Coverage in the "mouse-skip" and "mouse-inclusion" groups is not significantly lower than in the control group for any EST set ($P > 0.1$, one-tailed Wilcoxon rank sum test; NS : $P > 0.8$).

supported by a recent study which exploited the fact that most putatively human-specific alternatively spliced exons (corresponding to our mouse inclusion group) have sequence features that can be used to discriminate them from conserved alternatively spliced exons. This led to an estimate that for 89% of such exons alternative splicing is likely to be human-specific (Yeo et al., 2005).

It is therefore meaningful to compare d_N for constitutive sequence between these genes and the control group of genes with conserved alternative splicing. We saw a significant increase in d_N for genes in the mouse-skip category ($n = 163$, median $d_N = 0.068$, $p < 0.05$) and in the mouse-inclusion category ($n = 364$, median $d_N = 0.062$, $p < 0.01$) compared to the control set ($n = 66$, median $d_N = 0.046$): an increase of 48% and 35% respectively. This suggests that genes with species-specific alternative splicing but conserved gene structure also show accelerated protein evolution in constitutive regions.

3.5 Discussion

Our results indicate that gaining an alternatively spliced exon is associated with an increased rate of evolution in the constitutive exons of a gene. We have not attempted to determine the origin of these gained exons. We note that other studies have reported

that tandem exon duplication is one source of alternatively spliced exons (Kondrashov and Koonin, 2001; Letunic et al., 2002), but none of the probable recent exon gains that we identified showed evidence of this. If genome-specific exons are created by tandem duplication then the lack of detectable sequence homology in the orthologous gene must be due to rapid sequence change following duplication. By restricting our comparison to genes undergoing alternative splicing by exon-skipping, and subdividing these into those cases where alternative splicing occurred in the ancestor of human and mouse, and those where alternative splicing emerged in the human or mouse branch only, we have been able to focus on the impact of alternative splicing on recent mammalian sequence evolution. This approach was designed to eliminate the influence of functional differences between genes, unlike the comparison of sequence constraint in alternatively spliced genes to genes in the genome as a whole. Thus, although genes showing alternative splicing by exon-skipping are a slow-evolving subset of the human genome, there is an increased rate of sequence evolution in the immediate aftermath of the appearance of alternative splicing. This result is reminiscent of observations about the evolution of duplicated genes. A number of studies have reported relaxation of sequence constraint in duplicated genes compared to singletons (Lynch and Conery, 2000; Van de Peer et al., 2001; Nembaware et al., 2002; Seoighe et al., 2003), but it has recently been shown that genes that tend to remain duplicated are generally more slowly evolving than genes that are found in single copy (Davis and Petrov, 2004; Jordan et al., 2004). It therefore appears that conserved genes are more likely than faster evolving genes to undergo diversification by either gene duplication or alternative splicing, and that both processes result in an increased rate of sequence change.

Several sources of error are linked to observations of alternative splicing at the genomic level. The primary question is: how reliable is any given observation of alternative splicing? Many EST sequences are derived from cancerous tissue sources and these may exhibit a high rate of aberrant splice events that are not relevant to normal function (Sorek et al., 2004). This is likely to have a disproportionate effect on observations seen in only one species because alternative splicing events conserved across species are more likely to be functional. This may reduce confidence in our observation of a difference in evolutionary rate between the two categories of alternatively spliced genes. However, two sources of evidence reinforce our result. First, if a given alternative splicing event occurs at high frequency we can be more confident that the event is functional (Kan et al., 2002). Our

results show that genes with genome-specific alternative splicing occurring at high frequency (>12%) show the greatest elevation of evolutionary rate. Second, restricting our analysis to include only those alternative splicing events that do not initiate NMD and which encode a distinct translation product shows that the observed rate difference is robust.

The limitations of the analogy between the evolution of gene duplicates and genes undergoing alternative splicing become apparent when we consider that alternatively spliced isoforms are not as free to evolve as paralogs. Nevertheless, we note that a recent study implicitly suggests an evolutionary equivalence between gene duplicates and alternative isoforms (Kopelman et al., 2005). In the case of paralogs the increase in evolutionary rate observed following gene duplication is often explained as resulting from functional redundancy between duplicates because the fates of the two paralogs are uncoupled, thus leading to relaxed selection on one of them (Van de Peer et al., 2001; Nembaware et al., 2002; Seoighe et al., 2003). In contrast, accelerated sequence evolution in the constitutive parts of alternatively spliced genes can not be attributed to simple sequence redundancy. When a gene becomes alternatively spliced, the evolutionary fates of the two transcripts are tightly coupled because some exons remain common to both transcripts. In this case only the alternatively spliced sequence itself would be expected to provide raw material for evolutionary change. This is implied by Modrek and Lee's original model where alternative splicing generates an internal paralog that is shielded from the constraints imposed by purifying selection and has been supported by more recent results (Modrek and Lee, 2003; Xing and Lee, 2005). However, our results show that the constitutive exons shared between transcripts are themselves subject to alteration of sequence constraint following the acquisition of alternative splicing.

The slower evolution we observe in genes undergoing alternative splicing by exon-skipping compared to the average genome-wide rate of evolution is consistent with the classical model of evolutionary constraint accompanying pleiotropy (Fisher, 1930). Thus the fact that an alternatively spliced gene may have multiple roles associated with its multiple isoforms (Xing et al., 2003; Resch et al., 2004b) means that an individual mutation is more likely to be deleterious. On the other hand, we can imagine the constitutive exons in an alternatively spliced gene as being subjected to two distinct selective regimes corresponding to the different functions of its isoforms. This can be likened to a state of adaptive conflict (Piatigorsky and Wistow, 1991) where changes beneficial to one func-

tion may be deleterious to the other. Selective constraint will be imposed by the need to maintain ancestral gene function (encoded by the major-form transcript), which will tend to brake sequence change in the constitutive region of the gene. However, the potential functional innovation associated with an internal paralog (encoded by the minor-form transcript) may demand correlated sequence changes in constitutive regions, thus increasing the rate of sequence evolution in the gene as a whole. These amino acid changes may be fixed if they have an adaptive benefit in the context of the function of the minor isoform while being selectively neutral, or even slightly deleterious, to the function of the major isoform. Piatigorsky and Wistow (1991) proposed that gene duplication can resolve the stalemate between these opposing selective forces. Our results demonstrate that the constitutive exons of alternatively spliced genes possess sufficient plasticity to accommodate the competing functional demands of their isoforms. This is underlined by our observation of a correlation between frequency of alternative exon incorporation and evolutionary rate in constitutive regions. These observations mirror results from a recent directed evolution study which demonstrated that negative trade-offs between different enzyme functions are much weaker than expected (Aharoni et al., 2005).

We should, however, be cautious before interpreting the strong correlation between the apparent gain of genome-specific alternative splicing and the increased rate of protein evolution as reflecting an actual causation. Both variables may be under the influence of some untested variable whereby, following the human-mouse split, a change in selective pressure operating on a gene may manifest itself both as a change in gene structure and in an increased rate of non-synonymous evolution.

3.6 Acknowledgements

This study was supported by Science Foundation Ireland. We thank Meg Woolfit for critical reading of the manuscript.

Chapter 4

When gene marriages don't work out: divorce by subfunctionalisation

This chapter is based on a manuscript recently submitted to *Trends in Genetics* (authors B.P. Cusack and K.H.Wolfe).

4.1 Abstract

We describe how a bifunctional gene, coding for two proteins by alternative splicing, was formed by gene fusion and later broke apart by duplication and complete subfunctionalisation. The bifunctional gene is a chimera that arose when the chloroplast gene *RPL32* integrated into an intron of the nuclear gene *SODcp* in an ancestor of mangrove and poplar trees. Mangrove retains the alternatively spliced chimeric gene, but in poplar it underwent duplication and complementary structural degeneration to re-form separate *RPL32* and *SODcp* genes.

4.2 Introduction

Subfunctionalisation provides an attractive explanation for why so many duplicated genes exist in eukaryotes, without requiring each duplication event to have conferred a selective advantage (Force et al., 1999). For many duplicated genes, however, it has been difficult

to pinpoint different subfunctions of the ancestral gene that were partitioned among the daughter genes. Often, our knowledge of the functions of the ancestral gene is so limited that we might not be able to recognise subfunctionalisation even if it has occurred. Most of the examples of subfunctionalisation reported to date involve changes in gene expression profiles (Force et al., 1999; Lynch, 2004; Cresko et al., 2003; Huminiecki and Wolfe, 2004), and there are only a few reports of duplicate gene pairs that have undergone subfunctionalisation by means of substantial changes in gene structure relative to their common ancestor (Wang et al., 2004; Altschmied et al., 2002; de Souza et al., 2005; Yu et al., 2003). Here we report an example of a structural subfunctionalisation event where the ancestral functions being partitioned among the daughter genes can be readily identified and are clearly distinct.

4.3 Results and Discussion

The gene for chloroplast ribosomal protein L32 (*RPL32*) is located in the chloroplast genome of most flowering plants, but is not present in the chloroplast genomes of two poplar species (*Populus trichocarpa* and *P. alba*; (Steane, 2005; Tuskan et al., 2006), and S. Okumura et al., GenBank accession number AP008956). Loss of *RPL32* from chloroplast DNA occurred after *Populus* (order Malpighiales) diverged from other members of the Eurosid I clade such as cucumber (order Cucurbitales) and legumes (order Fabales). We identified database EST (expressed sequence tag) sequences from a copy of *RPL32* that has become relocated to the nuclear genome in poplar. The *RPL32* coding sequence in this transcript is fused in-frame downstream of a sequence resembling chloroplast Cu/Zn superoxide dismutase (SOD). Further comparisons to ESTs and genomic sequence data from *P. trichocarpa* (Tuskan et al., 2006; Sterck et al., 2005) and *Bruguiera gymnorrhiza* (Miyama et al., 2006) (Burma mangrove, also in the order Malpighiales) enabled us to reconstruct the events that occurred subsequent to the transfer of the gene to the nucleus.

Plants have several isozymes of Cu/Zn SOD, which is an enzyme functioning in redox balance. Some of these isozymes are cytosolic and some are imported into chloroplasts by means of an amino-terminal transit peptide (Schinkel et al., 2001). In the legume *Medicago truncatula* the chloroplast isozyme is encoded by a single nuclear gene (*SODcp*) with eight exons (Figure 4-1). In an ancestor of poplar and mangrove, the *RPL32* sequence from the chloroplast genome was transferred to the nuclear genome where it became inserted into the last intron (intron 7) of *SODcp*. The newly-formed chimeric *SODcp-RPL32 gene* was

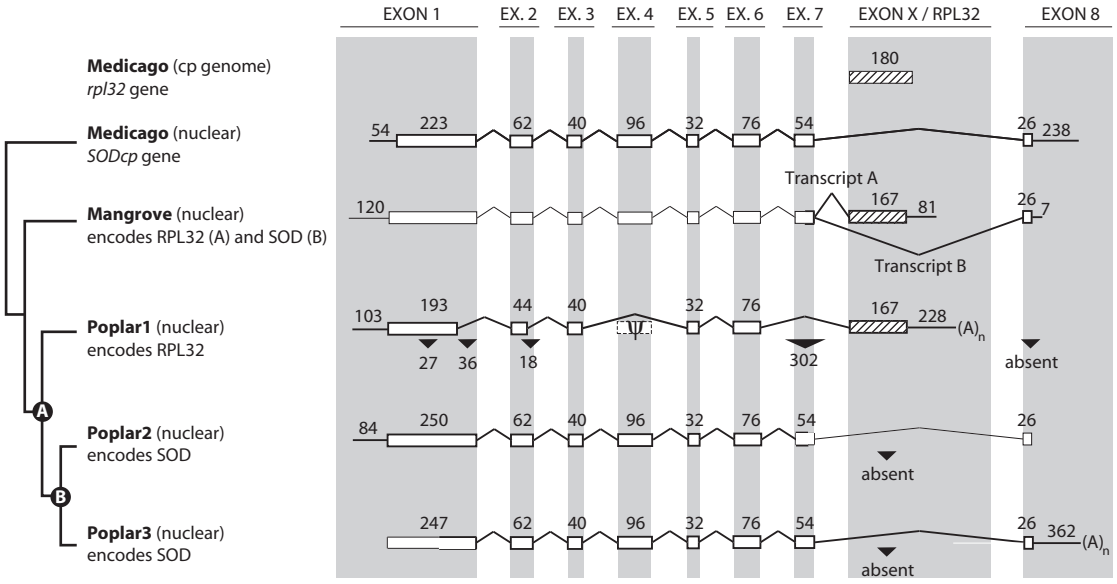


Figure 4-1: Organisation of SODcp, RPL32 and chimeric genes. The tree on the left shows the branching order of the nuclear genes, based on their pairwise d_S values. Nodes A and B represent gene duplications in poplar. Node B corresponds to a large segmental or whole-genome duplication (Sterck et al., 2005; Tuskan et al., 2006) in poplar, because many of the genes neighboring Poplar2 have homologs neighboring Poplar3. In the right panel, boxes represent exons, horizontal lines represent untranslated regions, and the lengths (bp) of some exons are shown. Introns are not drawn to scale. White boxes show SODcp-related exons and hatched boxes show RPL32-related exons. Triangles indicate sequences deleted in the poplar genes (with deletion lengths where known), and ψ indicates the pseudo-exon 4 in Poplar1. Thicker lines show the parts of gene structures that were verified directly by comparing genomic and cDNA or EST sequences from the same species. Thinner lines in poplar show parts of genes for which only genomic sequence is available, and in mangrove show regions where only EST data is available. The intron/exon structure of the 5' part of the mangrove gene is assumed to be the same as in other species. Sources of sequence data are listed in Table 4.1.

alternatively spliced, producing one transcript identical in structure to the original *SODcp* mRNA, and one where exons 1-7 were spliced onto a novel exon (exon X) corresponding almost exactly to the whole *RPL32* coding region, instead of onto the last exon (exon 8) of *SODcp*. This alternatively spliced gene still exists in mangrove, where we identified ESTs corresponding to two types of transcript: one coding for SOD (Transcript B, 219 amino acids), and the other coding for a chimeric protein with residues 1-211 of SOD fused to residues 2-54 of *RPL32* (Transcript A; Figure 4-1 and Figure 4-2). We confirmed that alternative splicing occurs in mangrove by sequencing a genomic PCR product that contains exons 7, X and 8 (Figure 4-1) and perfectly matches the sequences of ESTs of the two types of transcript.

In poplar, after its divergence from mangrove, the chimeric *SODcp-RPL32* gene was

duplicated twice. The first duplication (node A on the phylogenetic tree in Figure 4-1) resulted in subfunctionalisation of the chimeric gene, producing daughter genes that code for either RPL32 (*Poplar1* gene), or SOD, but not both. The SOD-encoding daughter later became duplicated a second time (node B) to produce two genes (*Poplar2* and *Poplar3*) that have virtually identical structures. EST analysis shows that all three poplar genes are transcribed and none of them is alternatively spliced. The *Poplar2* and *Poplar3* genes have lost exon X and code for proteins that can be aligned along their whole length to *Medicago* SOD. Reciprocally, the *RPL32*-encoding copy (*Poplar1*) has retained exon X but has lost exons 4, 7 and 8. Exon 4 of *Poplar1* is a pseudo-exon containing a frameshift mutation and is skipped in all nine database ESTs we identified from the gene. There are also deletions in exons 1 and 2 of *Poplar1* relative to *Poplar2*, *Poplar3* and the *SODcp* genes of other plant species. The *Poplar1* gene still has a continuous open reading frame between the former *SODcp* start codon and the *RPL32* stop codon, and the amino terminus of its protein product is strongly predicted to be a chloroplast transit peptide (Emanuelsson et al., 2000). However, the protein encoded by *Poplar1* cannot be a functional SOD enzyme because it lacks many residues normally conserved in SOD proteins, including all six active site residues (four are deleted and two are substituted; Figure 4-2). In addition to the deletions, the remaining SOD-derived parts of the Poplar1 protein also show deconstrained sequence evolution: in exons 1-6 there is only 60% amino acid sequence identity between Poplar1 and mangrove, lower than for Poplar2 or Poplar3 versus mangrove (both 77% identity). Analysis of nonsynonymous and synonymous nucleotide substitutions shows that the *SODcp*-derived exons of *Poplar1* have been evolving almost free of selective constraint ($d_N/d_S = 0.9$; Figure 4-3). These exons have lost the requirement to specify a functional SOD and instead are constrained only to provide a working transit peptide for the *RPL32* protein.

RPL32's marriage to, and subsequent divorce from, *SODcp* in the poplar lineage provides an unusually graphic example of the partitioning of an ancestral gene's multiple functions among daughter genes formed by duplication. This partitioning process can be categorised as subfunctionalisation because the structural changes in the poplar genes indicate unambiguously that, after the duplication at node A, a complementary loss of subfunctions of the ancestral chimeric gene in its two daughters occurred. The losses of exon X (coding for the *RPL32* subfunction) in the *Poplar2/3* lineage, and of exons 4, 7 and 8

(coding for the SOD subfunction) in *Poplar1*, were caused by degenerative mutations that are likely to have been selectively neutral because in each case the subfunction lost by one gene copy was maintained by the other. As a result, the gene pair was preserved in the genome by subfunctionalisation as envisaged by Lynch and Force (Force et al., 1999; Lynch and Force, 2000).

However, we also find some evidence of adaptive protein sequence changes occurring in the evolutionary history of the *SODcp-RPL32* chimera. There are three branches in Figure 4-3 for which $\omega = \infty$ (two for the *SODcp* part of the gene and one for the *RPL32* part). Although there is insufficient evidence to infer positive selection in each case, it is striking that these branches correspond to the beginning and the end of the *SODcp-RPL32* gene marriage. This may indicate the creation of a state of adaptive conflict following the fusion of these genes and subsequent escape from this conflict after duplication of the chimeric gene. Adaptive conflict is a situation where constitutive parts of a bifunctional gene are placed under conflicting selective pressures by its two subfunctions, resulting in a sequence that is suboptimal for both subfunctions (Piatigorsky and Wistow, 1991). The creation of this 'compromise' sequence following gene fusion is expected to be associated with accelerated sequence evolution. In a short branch corresponding to the gene fusion event (Figure 4-3), we infer that no synonymous substitutions occurred but three and 13 nonsynonymous substitutions occurred in *SODcp* and *rpl32*, respectively. After gene duplication, conflicted regions can show accelerated sequence changes in both daughter copies as they specialise in function. On the branch that re-formed an independent *SODcp* gene in the poplar lineage, we infer that 7 nonsynonymous substitutions, but no synonymous substitutions, occurred between nodes A and B (Figure 4-3). We expect that these changes occurred after the gene pair had been preserved by subfunctionalisation, but we cannot rule out the possibility that the gene duplication was followed immediately by nonsynonymous mutations that relieved a conflict and provided an immediate selective advantage to the duplication, and that the degenerative mutations that resulted in structural subfunctionalisation of the gene occurred later.

Considering that the bifunctional state was imposed on *SODcp* when it was invaded by *RPL32*, and that the two proteins normally have nothing in common except their subcellular targeting to chloroplasts, it seems reasonable to suppose that their short-lived cohabitation in the poplar lineage might have entailed a degree of conflict.

Figure 4-2: Amino acid sequence alignments of SODcp, RPL32 and chimeric genes. Exons in the poplar genes are boxed, residues are coloured using the ClustalW schema (Thompson et al., 1994), and asterisks indicate stop codons. Columns corresponding to the six active site residues in Cu/Zn SOD (five His residues and one Asp) (Banci et al., 2002) are highlighted by gray column shading. The arrow indicates the position of the first amino acid residue in the mature protein after cleavage of the transit peptide, as inferred by comparison to the experimentally determined cleavage sites in Cu/Zn SOD from spinach (Kitagawa et al., 1986) and rice (Komatsu et al., 2004). To maintain continuity of the SOD amino acid sequence, exon 8 is shown upstream of exon X, whereas in the genome it is downstream (Figure 4-1). Sequence alignment was done manually. Sources of sequence data are given in Table 4.1. Poplar3 is the same gene as was cloned as a cDNA by Schinkel et al. (Schinkel et al., 2001) from hybrid aspen (P. tremula x P. tremuloides) and named cp-SOD.

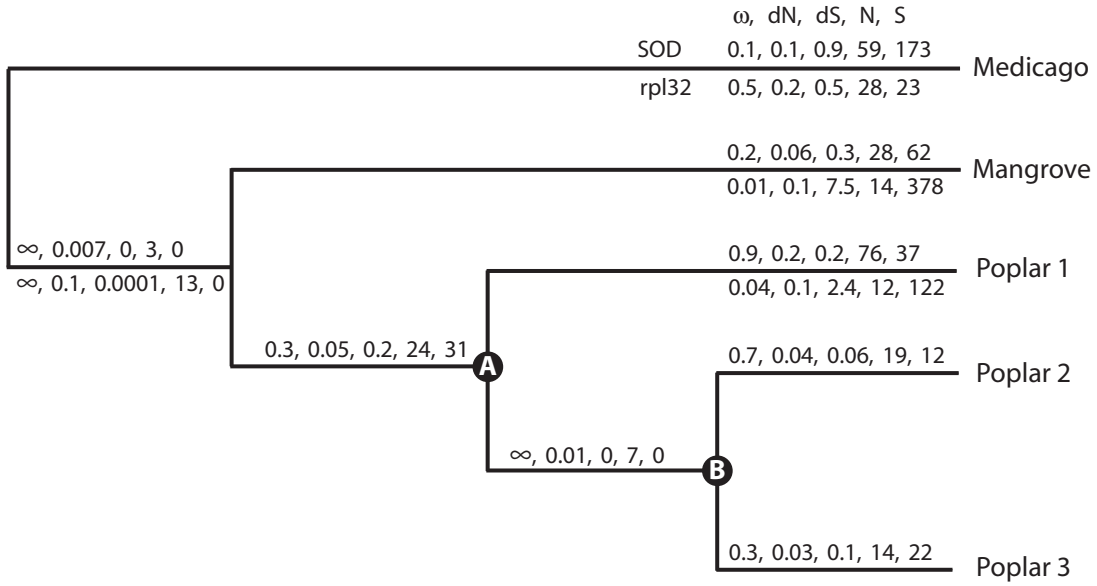


Figure 4-3: Branch specific estimates of levels of nucleotide substitution. Shown are the estimated numbers of nonsynonymous (d_N) and synonymous substitutions (d_S) per site, the d_N/d_S ratio (ω), and the numbers of nonsynonymous (N) and synonymous substitutions (S) for the SOD region and the RPL32 region (above and below each branch, respectively). Nodes A and B represent gene duplications in poplar as described in the legend for Figure 4-1. There are approximately 467 nonsynonymous and 190 synonymous sites in the SOD regions (exons 1-7 and 8) and 132 nonsynonymous and 51 synonymous sites in the RPL32 region (exon X). Substitution estimates were obtained using CODEML (Yang, 1997). The tree was rooted using the sequences of cotton (Gossypium hirsutum) SODcp and RPL32. For each of the three branches for which $\omega = \infty$, we tested for evidence of positive selection. In each case the null hypothesis of neutral evolution could not be rejected according to a likelihood ratio test comparing a free-ratio model to a model with $\omega = 1$ for the tested branch.

4.4 Acknowledgements

We thank M.S. Islam for *B. gymnorrhiza* DNA, G. Butler for PCR amplification, and M. Woolfit and M. Sémon for comments. This study was supported by Science Foundation Ireland.

4.5 Sources of nucleotide sequence data

Poplar (*Populus trichocarpa*)

Genome sequence data (Tuskan et al., 2006) of *Populus trichocarpa* (version 1.0 preliminary draft) was obtained from the DOE Joint Genome Institute website (`http://genome.jgi-psf.org/Poptr1/Poptr1.home.htm`). Table 4.1 lists the genomic coordinates and representative EST accession numbers for the *Poplar1*, *Poplar2* and *Poplar3* genes.

Burma mangrove (*Bruguiera gymnorrhiza*)

GenBank accession numbers for ESTs:

- BP940691 (alternative transcript A)
- BP939979 (alternative transcript B)
- BP938942, BP940900, BP940351, BP943735 (5' region shared by both transcript types)

We amplified and completely sequenced a genomic PCR product of exons 7, X and 8 (GenBank accession number XXXXXX) using *B. gymnorrhiza* genomic DNA generously provided by M. S. Islam (Islam, 2006).

Medicago truncatula

- Genomic BAC clone: AC126007.16 (locus tag: MtrDRAFT_AC126007g11v1; gi GI:92876779).
- Full-length cDNA clone: AF056621.

Gene:	***Poplar1***	
Map location:	LG_IX (441kb)	
EST accessions:	DT473214	
	DT481089	
	DT486400	
	CV282850	
Poplar1 exon coordinates:		
Exon 1	<441208	441503
Exon 2	441918	441961
Exon 3	442064	442103
Exon 5	442404	442435
Exon 6	442586	442661
Exon X	443097	443488
Gene:	***Poplar2***	
Map location:	LG_IX (11Mb)	
EST accessions:	BU879568	
	CK115374	
	CK115705	
Poplar2 exon coordinates:		
Exon 1	<11796579	11796912
Exon 2	11797766	11797827
Exon 3	11797932	11797971
Exon 4	11798105	11798200
Exon 5	11798283	11798314
Exon 6	11798496	11798571
Exon 7	11798913	>11798947
Gene:	***Poplar3***	
Map location:	scaffold_163	
EST accessions:	CV275995	
	DT478877	
Poplar3 exon coordinates:		
Exon 1	<189469	189577
Exon 2	190370	190431
Exon 3	190531	190570
Exon 4	190699	190794
Exon 5	190877	190908
Exon 6	191128	191203
Exon 7	191898	191951
Exon 8	194109	194496

Table 4.1: The genomic coordinates and representative EST accession numbers for the Poplar1, Poplar2 and Poplar3 genes.

Chapter 5

Conclusions

Taken together the three studies above present evidence of an intimate association between alterations in the structure of genes and changes in the evolutionary rate of their encoded proteins. The genome-wide studies in the first two research chapters show that changes in gene structure following both retrotransposition (during which introns and most regulatory sequences are lost, Chapter 2) and the acquisition of alternative splicing (during which new alternative exons are gained, Chapter 3) are associated with a quickening in the pace of protein sequence evolution. The single-gene study in Chapter 4 is primarily a demonstration of the interchangeability of alternative splicing and gene duplication in the evolutionary history of a fused-gene (*SODcp-RPL32*). However, at another level this example serves as an illustration of the effect on evolutionary rate of changes in gene structure that occurred at two stages in the lifetime of this gene. At the first time-point a change in *SODcp's* gene structure occurred equivalent to the gain of an alternatively spliced exon. At the second timepoint, complementary gene structure changes following gene duplication led to the separation of the fused genes by subfunctionalisation. We see some evidence of the speeding-up of protein sequence evolution at both of these time points.

Alterations in gene structure are by no means the only explanation for the described changes in nonsynonymous evolution, however. In Chapters 2 and 3 I also see evidence of acceleration in nonsynonymous rate without gene structure change. The increased rate asymmetry of relocated (as compared to local DNA-based) gene duplications, may illustrate the importance of genomic context to gene function (Chapter 2) and is not a consequence of genomic heterogeneity in mutation rate. Similarly, changes in gene structure do not explain the faster nonsynonymous rate of genes with conserved exon-intron structure but altered

frequencies of alternative splicing (Chapter 3).

The estimates of asymmetric coding sequence divergence following gene duplication presented in Chapter 2 may shed light on the functional divergence that is a prerequisite for the retention of many duplicated genes. However, coding-sequence analysis provides only a partial view of the differentiation of gene duplicates. Although roughly one-half of all duplicates created by retrotransposition (and one-third of DNA-based duplicates) can be said to show asymmetry in their nonsynonymous evolution, this does not imply that the remaining half of retroduplicates (and two-thirds of DNA-based duplicates) are not functionally differentiated. This is because functional divergence is not restricted to changes in the encoded proteins but can also proceed at the level of changes in expression pattern. In this study the attempt to consider divergence in duplicate expression pattern was limited to general observations by the high degree of sequence similarity between duplicates. Nevertheless, as noted in the introduction (section 1.3.4.2, page 50), although measures of nonsynonymous sequence divergence do not assay changes in expression directly they may do so indirectly. From this perspective the acceleration in evolutionary rate shown by retrogenes is consistent with the evidence from other studies that their expression is generally tissue-specific (Marques et al., 2005).

It is interesting to speculate whether the lower frequency of protein sequence asymmetry among DNA-based duplicates (and their possibly lower frequency of divergence in protein function) derives from the fact that a more common mechanism for the preservation of these duplicates is symmetric divergence in their expression pattern. Because DNA-based duplicates should initially share most (if not all) of their regulatory sequences, subfunctionalisation of the ancestral expression pattern has the potential to proceed symmetrically. On the other hand, the probability of symmetric subfunctionalisation of retroduplicates is likely to be negligible. Surviving retroduplicates are more likely to have been preserved by subfunctionalisation than by other mechanisms because the existence of inequalities between gene duplicates at birth is the first step on the path to subfunctionalisation (Averof and Ferrier, 1996). However, when regulatory subfunctionalisation of retroduplicates happens it will probably occur asymmetrically.

It might seem reasonable to expect that the changes in gene structure associated with retrotransposition (loss of introns and regulatory sequences) would condemn retroduplicates to evolutionary oblivion (Brosius, 1991). In fact this assumption has generally permeated

the research literature (Mighell et al., 2000). However, there are at least two mechanisms that can rescue retrocopies from this fate by allowing the emergence of a retroduplicate as a functional retrogene. One possibility relies on the alternative sites of initiation of transcription of the source gene and the partial processing of the resultant mRNA. The frequent usage of alternative first exons allows the incorporation of alternative promoters into first introns (Cooper et al., 2006). A semiprocessed retrogene retaining this first intron is therefore equipped for survival as an expressed retrocopy. One such example is the preproinsulin I retrogene that inherited much of its 5' regulatory sequences from its source gene, preproinsulin 2 (Soares et al., 1985). The second possible survival mechanism is for a retrocopy to fortuitously integrate into a "fertile genomic environment" permitting its expression (Brosius, 1991). In particular, retrotransposition events may frequently lead to the formation of chimeric genes where the intronic sequence of the host gene represents a highly suitable integration site (Long et al., 2003). Such cases constitute a change in exon-intron structure of both the retrogene and the host gene. It is interesting to note that alternative splicing also plays a role in this second mechanism (Vinckenbosch et al., 2006). This point further implicates alternative splicing in facilitating alterations in gene structure and echoes both the observations in Chapter 4 and those of Modrek and Lee (2003).

Notably, because the retrogenes studied in Chapter 2 are annotated single-exon genes, this means they are not likely to have been subject to either of the two preservational mechanisms described in the previous paragraph. Therefore, it may be necessary to reassess the common assumption that the loss of all ancestral regulatory sequences is inevitable following the creation of completely processed retroduplicates. It is possible that reverse-transcribed mRNAs often contain sufficient regulatory sequence in their UTRs to independently promote their transcription. A precedent for this is provided by the example of the promoters of human salivary amylase genes which have recruited regulatory sequence from the UTR of an upstream gamma-actin retropseudogene (Samuelson et al., 1990). More generally we might expect the 5' UTRs of functional retrogenes to contain sequence elements that permit the initiation and regulation of transcription thereby ensuring retrogene survival. Among the candidate regulatory elements for such a role are downstream promoter elements (DPEs) (Arkhipova, 1995). DPEs are distinct 7-nucleotide multicopy core promoter elements that have been shown to direct transcription in TATA-less promoters from *Drosophila* to mammals (Burke and Kadonaga, 1997). In *Drosophila* roughly 40% of all promoters contain

DPEs (Kutach and Kadonaga, 2000). This observation suggests the testable hypothesis that the 5' UTRs of functional retrogenes (and their source genes) should be enriched for DPE elements compared to the genome average.

Within genes the rate of evolution of UTRs and of coding exons has been shown to be correlated (Makalowski and Boguski, 1998). This relates to the fact that both the nonsynonymous rate of a gene and the rate of evolution of its UTR correlate negatively with expression breadth (Duret and Mouchiroud, 2000). Given the tissue-specificity of retrogene expression this observation implies that the UTRs of functional retrogenes should evolve at a fast rate. On the other hand, it might be expected that the UTRs of functional retrogenes should be under strong selective constraint to preserve the regulatory function that is an early determinant of their survival. It is therefore an open question whether the rate of UTR evolution in retrogenes is decoupled from the fast nonsynonymous evolution of their coding sequences.

In chapter 3 I suggested that the relationship between the gain of alternatively spliced exons and the faster rate of constitutive exon evolution is a causative one. Under this scenario newly gained alternative protein regions may require correlated amino-acid substitutions in proportion to their inclusion frequency. Moreover, constitutive exons have sufficient sequence flexibility to accommodate these requirements. An alternative explanation is that no such causation exists but that genes under weak selective constraints are more likely to tolerate changes in gene structure and alternative splicing pattern. However, this hypothesis seems less likely because it requires that fast evolving genes should preferentially accrue alternative exons that are not only spliced at high frequencies but are also productive (i.e., coding for a distinct protein product). The primary findings of this study have been confirmed by subsequent work (Wang et al., 2005; Plass and Eyras, 2006), in particular the conclusion that a large fraction of species-specific exons have been recently created. Although I was able to exclude intra-genic tandem duplication (Kondrashov and Koonin, 2001; Letunic et al., 2002) as the source of these new exons the more recent proposal (Vinckenbosch et al., 2006) that retrotransposition is a source of newly gained alternatively spliced exons remains to be tested.

Chapter 4 presented an example where alternative splicing is a substrate for subfunctionalisation following gene duplication. Therefore the functions of the SODcp and RPL32 proteins appear to be equally well encoded by either a pair of duplicate genes or by alterna-

tive transcripts of the same gene. This example adds the first non-fish example to a short list of anecdotal cases of the subfunctionalisation of alternative splice variants (Altschmied et al., 2002; Yu et al., 2003). However, convincing evidence of the generality of this phenomenon has yet to be demonstrated. It might be naïve to expect that gene duplication and alternative splicing are interchangeable approaches to encoding diversity for all gene-function classes. For example, the generation of protein diversity in the nervous system might be more efficiently achieved through the encoding of multiple alternative transcripts than through the possession of multiple paralogs (Copley, 2004). The question is to what extent the partitioning of splice variants between duplicate genes can advance in an effectively neutral manner. Arguably, the regulation of the relative dosage of two protein isoforms can occur more precisely at the post-transcriptional stage for a single alternatively spliced gene than through the independent regulation of two paralogs. Another relevant consideration is the impact of this type of subfunctionalisation on vulnerability to mutation. On the one hand, the maintenance of two genes to perform two functions that could otherwise have been encoded by a single gene can be considered as a doubling in mutational target. On the other hand, following subfunctionalisation there may be some decrease in mutational vulnerability associated with a reduction in the number of sites involved in splicing regulation.

It is likely that retrotransposition events make a considerable contribution to the preservation of ancestral alternative splice variants by subfunctionalisation. Retrogenes have the potential to "hard-code" one alternative transcript of their parental gene and this may precipitate the complementary loss of this transcript from their multi-exon paralog. Further work is needed to establish the prevalence of this mechanism. However, the potential importance of retrotransposition in this context is hinted at by an early study demonstrating that the expressed retrogene *Zfa* originates from an alternative transcript of its source gene (*Zfx*) (Ashworth et al., 1990) as well as the more recent observation that retropseudogenes provide a "genomic-archive" of alternative splicing activity (Shemesh et al., 2006).

Bibliography

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci U S A **100**:4649–4654.

Aharoni, A., L. Gaidukov, O. Khersonsky, S. McQ Gould, C. Roodveldt, and D. S. Tawfik. 2005. The 'evolvability' of promiscuous protein functions. Nat Genet **37**:73–76.

Akashi, H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev **11**:660–666.

Akashi, H. 2003. Translational selection and yeast proteome evolution. Genetics **164**:1291–1303.

Akashi, H. and A. Eyre-Walker. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev **8**:688–693.

Altschmied, J., J. Delfgaauw, B. Wilde, J. Duschl, L. Bouneau, J.-N. Volff, and M. Schartl. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. Genetics **161**:259–267.

Altschul, S., T. Madden, A. Schaeffer, 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**:3389–402.

Arkhipova, I. R. 1995. Promoter elements in Drosophila melanogaster revealed by sequence analysis. Genetics **139**:1359–1369.

Ashworth, A., B. Skene, S. Swift, and R. Lovell-Badge. 1990. Zfa is an expressed retroposon derived from an alternative transcript of the Zfx gene. EMBO J **9**:1529–1534.

Averof, R. D., M. and D. Ferrier. 1996. Diversification of arthropod Hox genes as a paradigm for the evolution of gene functions. Seminars in Cell and Dev Biol **7**:539–551.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. 2002. Recent segmental duplications in the human genome. Science **297**:1003–1007.

Banci, L., I. Bertini, F. Cramaro, R. Del Conte, and M. S. Viezzoli. 2002. The solution structure of reduced dimeric copper zinc superoxide dismutase. The structural effects of dimerization. Eur J Biochem **269**:1905–1915.

Batada, N. N., L. D. Hurst, and M. Tyers. 2006. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol **2**:e88.

Beissbarth, T., L. Hyde, G. Smyth, C. Job, W. Boon, S. Tan, H. Scott, and T. Speed. 2004. Statistical modeling of sequencing errors in SAGE libraries. Bioinformatics **20 Suppl 1**:I31–I39.

Black, D. L. 2000. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell **103**:367–370.

Bloom, J. D., D. A. Drummond, F. H. Arnold, and C. O. Wilke. 2006. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol **23**:1751–1761.

Boer, P. H., C. N. Adra, Y. F. Lau, and M. W. McBurney. 1987. The testis-specific phosphoglycerate kinase gene pgk-2 is a recruited retroposon. Mol. Cell. Biol. **7**:3107–3112.

Boue, S., I. Letunic, and P. Bork. 2003. Alternative splicing and evolution. Bioessays **25**:1031–4.

Brett, D., H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. 2002. Alternative splicing and genome complexity. Nat Genet **30**:29–30.

Brosius, J. 1991. Retroposons–seeds of evolution. Science **251**:753.

Burke, T. W. and J. T. Kadonaga. 1997. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev **11**:3020–3031.

Byrne, K. P. and K. H. Wolfe. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res **15**:1456–1461.

Carlini, D. B. and J. E. Genut. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol **62**:89–98.

Casneuf, T., S. De Bodt, J. Raes, S. Maere, and Y. Van de Peer. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol **7**:R13.

Castillo-Davis, C. I., F. A. Kondrashov, D. L. Hartl, and R. J. Kulathinal. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res **14**:802–811.

Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7**:98–108.

Chen, F.-C., S.-S. Wang, C.-J. Chen, W.-H. Li, and T.-J. Chuang. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. Mol Biol Evol **23**:675–682.

Coghlan, A. and K. H. Wolfe. 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast **16**:1131–1145.

Conant, G. C. and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. **13**:2052–2058.

Cooper, S. J., N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res **16**:1–10.

Copley, R. 2004. Evolutionary convergence of alternative splicing in ion channels. Trends Genet **20**:171–6.

Cresko, W. A., Y.-L. Yan, D. A. Baltrus, A. Amores, A. Singer, A. Rodriguez-Mari, and J. H. Postlethwait. 2003. Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. Dev Dyn **228**:480–489.

Cusack, B. P. and K. H. Wolfe. 2005. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. Mol Biol Evol **22**:2198–2208.

Cusack, B. P. and K. H. Wolfe. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol **24**:679–686.

Davis, J. and D. Petrov. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol **2**:E55.

Dayhoff, M., R. Schwartz, and B. Orcutt. 1972. A model of evolutionary change in protein. Atlas of Protein Sequences and Structure **5**:345–352.

de Souza, F. S. J., V. F. Bumaschny, M. J. Low, and M. Rubinstein. 2005. Subfunctionalization of expression and peptide domains following the ancient duplication of the proopiomelanocortin gene in teleost fishes. Mol Biol Evol **22**:2417–2427.

Dickerson, R. E. 1971. The structures of cytochrome c and the rates of molecular evolution. J Mol Evol **1**:26–45.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102**:14338–14343.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23**:327–337.

Dufayard, J. F., L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perriere. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics **21**:2596–2603.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev **12**:640–649.

Duret, L. and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol **17**:68–74.

Emanuelsson, O., H. Nielsen, S. Brunak, and G. von Heijne. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300**:1005–1016.

Emerson, J. J., H. Kaessmann, E. Betran, and M. Long. 2004. Extensive gene traffic on the mammalian X chromosome. Science **303**:537–540.

Eyre-Walker, A. and L. D. Hurst. 2001. The evolution of isochores. Nat Rev Genet **2**:549–555.

Fairbrother, W. G., R. F. Yeh, P. A. Sharp, and C. B. Burge. 2002. Predictive identification of exonic splicing enhancers in human genes. Science **297**:1007–1013.

Ferris, S. D. and G. S. Whitt. 1979. Evolution of the differential regulation of duplicate genes after polyploidization. J. Mol. Evol. **12**:267–317.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. Dover, New York.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**:1531–1545.

Fraser, H. B. 2005. Modularity and evolutionary constraint on proteins. Nat Genet **37**:351–352.

Fraser, H. B., D. P. Wall, and A. E. Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol **3**:11.

Gayral, P., P. Caminade, P. Boursot, and N. Galtier. 2007. The evolutionary fate of recently duplicated retrogenes in mice. J Evol Biol **20**:617–626.

Gilbert, W. 1978. Why genes in pieces? Nature **271**:501.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445–458.

Goldman, N. and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**:725–736.

Goncalves, I., L. Duret, and D. Mouchiroud. 2000. Nature and structure of human genes that generate retropseudogenes. Genome Res **10**:672–678.

Graur, D. and W. H. Li. 2000. Fundamentals of Molecular Evolution. Sinauer, Sunderland, MA.

Harrington, E., S. Boue, J. Valcarcel, J. Reich, and P. Bork. 2004. Estimating rates of alternative splicing in mammals and invertebrates. Nat Genet **36**:916–7.

Hastings, K. E. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J Mol Evol **42**:631–640.

He, X. and J. Zhang. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169**:1157–1164.

He, X. and J. Zhang. 2006. Toward a molecular understanding of pleiotropy. Genetics **173**:1885–1891.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet **72**:1527–1535.

Henikoff, S. and J. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. **89**:10915–10919.

Herbeck, J. T. and D. P. Wall. 2005. Converging on a general model of protein evolution. Trends Biotechnol **23**:485–487.

Hillier, L. W., W. Miller, E. Birney, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**:695–716.

Hirsh, A. E. and H. B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **411**:1046–1049.

Hsiao, L. L., F. Dangond, T. Yoshida, et al. 2001. A compendium of gene expression in normal human tissues. Physiol Genomics **7**:97–104.

Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci **256**:119–124.

Hughes, A. L. 1997. Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol Biol Evol **14**:1–5.

Hughes, M. K. and A. L. Hughes. 1993. Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. Mol Biol Evol **10**:1360–1369.

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.

Hurst, L. D. and C. Pal. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet **17**:62–65.

Hurst, L. D. and N. G. Smith. 1999. Do essential genes evolve slowly? Curr Biol **9**:747–750.

Iida, K. and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261**:93–105.

Islam, M. 2006. Development and characterization of ten new microsatellite markers in a mangrove tree species Bruguiera gymnorrhiza (L.). Mol. Ecol. Notes **6**:30–32.

Johnson, J., J. Castle, P. Garrett-Engele, Z. Kan, P. Loerch, C. Armour, R. Santos, E. Schadt, R. Stoughton, and D. Shoemaker. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302**:2141–4.

Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res **12**:962–968.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol **3**:1.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol. Biol. **4**:22.

Julenius, K. and A. G. Pedersen. 2006. Protein evolution is faster outside the cell. Mol Biol Evol **23**:2039–2048.

Kampa, D., J. Cheng, P. Kapranov, et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res **14**:331–342.

Kan, Z., D. States, and W. Gish. 2002. Selecting for functional alternative splices in ESTs. Genome Res **12**:1837–45.

Katju, V. and M. Lynch. 2003. The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. Genetics **165**:1793–1803.

Katju, V. and M. Lynch. 2006. On the Formation of Novel Genes by Duplication in the Caenorhabditis elegans Genome. Mol. Biol. Evol. **23**:1056–1067.

Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428**:617–624.

Kim, S. H. and S. V. Yi. 2006. Correlated asymmetry of sequence and functional divergence between duplicate proteins of Saccharomyces cerevisiae. Mol Biol Evol **23**:1068–1075.

Kimura, M. 1983. The Neutral Theory of Evolution. Cambridge University Press, Cambridge.

Kimura, M. and T. Ohta. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci U S A **71**:2848–2852.

Kitagawa, Y., S. Tsunasawa, N. Tanaka, Y. Katsube, F. Sakiyama, and K. Asada. 1986. Amino acid sequence of copper,zinc-superoxide dismutase from spinach leaves. J Biochem (Tokyo) **99**:1289–1298.

Koehl, P. and M. Levitt. 2002. Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci U S A **99**:1280–1285.

Komatsu, S., K. Kojima, K. Suzuki, K. Ozaki, and K. Higo. 2004. Rice Proteome Database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. Nucleic Acids Res **32**:388–392.

Kondrashov, F. and E. V. Koonin. 2001. Origin of alternative splicing by tandem exon duplication. Hum. Mol. Genet. **10**:2661–2669.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Selection in the evolution of gene duplications. Genome Biol **3**:RESEARCH0008.

Kong, A., D. F. Gudbjartsson, J. Sainz, et al. 2002. A high-resolution recombination map of the human genome. Nat Genet **31**:241–247.

Koonin, E. V. and Y. I. Wolf. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol **17**:481–487.

Kopelman, N. M., D. Lancet, and I. Yanai. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet **37**:588–589.

Krasnov, A. N., M. M. Kurshakova, V. E. Ramensky, P. V. Mardanov, E. N. Nabirochkina, and S. G. Georgieva. 2005. A retrocopy of a gene can functionally displace the source gene in evolution. Nucleic Acids Res. **33**:6654–6661.

Kriventseva, E. V., I. Koch, R. Apweiler, M. Vingron, P. Bork, M. S. Gelfand, and S. Sunyaev. 2003. Increase of functional diversity by alternative splicing. Trends Genet **19**:124–128.

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

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

Kutach, A. K. and J. T. Kadonaga. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol Cell Biol **20**:4754–4764.

Lee, C., L. Atanelov, B. Modrek, and Y. Xing. 2003. ASAP: the Alternative Splicing Annotation Project. Nucleic Acids Res **31**:101–5.

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

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

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

Letunic, I., R. R. Copley, and P. Bork. 2002. Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet **11**:1561–1567.

Lewis, B. P., R. E. Green, and S. E. Brenner. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A **100**:189–192.

Li, H., R. Helling, C. Tang, and N. Wingreen. 1996. Emergence of preferred structures in a simple model of protein folding. Science **273**:666–669.

Li, W. 1997. Molecular Evolution. Sinauer, Sunderland, MA.

Li, W. H., J. Yang, and X. Gu. 2005. Expression divergence between duplicate genes. Trends Genet **21**:602–607.

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

Liao, B. Y. and J. Zhang. 2006. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol **23**:1119–1128.

Long, M., E. Betran, K. Thornton, and W. Wang. 2003. The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. **4**:865–875.

Lynch, M. 2004. Gene duplication and evolution. *In* A. Moya and E. Font, eds., Evolution: From molecules to ecosystems. Oxford University Press, Oxford, pp. 33–47.

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

Lynch, M. and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics **154**:459–473.

Lynch, M. and A. Kewalramani. 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. Mol Biol Evol **20**:563–571.

Makalowski, W. and M. S. Boguski. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A **95**:9407–9412.

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.

Maquat, L. E. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol **5**:89–99.

Marques, A. C., I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. **3**:e357.

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

McClintock, J. M., M. A. Kheirbek, and V. E. Prince. 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. Development **129**:2339–2354.

Merritt, T. J. and J. M. Quattro. 2001. Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. Genetics **159**:689–697.

Mighell, A. J., N. R. Smith, P. A. Robinson, and A. F. Markham. 2000. Vertebrate pseudogenes. FEBS Lett **468**:109–114.

Miyama, M., H. Shimizu, M. Sugiyama, and N. Hanagata. 2006. Sequencing and analysis of 14,842 expressed sequence tags of burma mangrove, Bruguiera gymnorrhiza. Plant Science **171**:234–241.

Modrek, B. and C. Lee. 2002. A genomic view of alternative splicing. Nat Genet **30**:13–19.

Modrek, B. and C. J. Lee. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet **34**:177–180.

Muller, T., R. Spang, and M. Vingron. 2002. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol **19**:8–13.

Muse, S. V. and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11**:715–724.

Nagy, E. and L. Maquat. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci **23**:198–199.

Nembaware, V., K. Crum, J. Kelso, and C. Seoighe. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res **12**:1370–1376.

Notebaart, R. A., M. A. Huynen, B. Teusink, R. J. Siezen, and B. Snel. 2005. Correlation between sequence conservation and the genomic context after gene duplication. Nucleic Acids Res. **33**:6164–6171.

Ohno, S. 1970. Evolution by gene duplication. Springer Verlag, New York.

Orban, T. I. and E. Olah. 2001. Purifying selection on silent sites – a constraint from splicing regulation? Trends Genet **17**:252–253. Letter.

Pagani, F. and F. E. Baralle. 2004. Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet **5**:389–396.

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

Pal, C., B. Papp, and L. D. Hurst. 2003. Genomic function: Rate of evolution and gene dispensability. Nature **421**:496–497.

Pal, C., B. Papp, and M. J. Lercher. 2006. An integrated view of protein evolution. Nat Rev Genet **7**:337–348.

Pan, Q., A. L. Saltzman, Y. K. Kim, C. Misquitta, O. Shai, L. E. Maquat, B. J. Frey, and B. J. Blencowe. 2006. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev **20**:153–158.

Parmley, J. L., J. V. Chamary, and L. D. Hurst. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol **23**:301–309.

Piatigorsky, J. and G. Wistow. 1991. The recruitment of crystallins: new functions precede gene duplication. Science **252**:1078–1079.

Piganeau, G., D. Mouchiroud, L. Duret, and C. Gautier. 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. J Mol Evol **54**:129–133. Letter.

Plass, M. and E. Eyras. 2006. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. BMC Evol Biol **6**:50.

Raff, R. 1996. The Shape of Life. University of Chicago Press, Chicago.

Resch, A., Y. Xing, A. Alekseyenko, B. Modrek, and C. Lee. 2004a. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic Acids Res **32**:1261–1269.

Resch, A., Y. Xing, B. Modrek, M. Gorlick, R. Riley, and C. Lee. 2004b. Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res **3**:76–83.

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

Rodin, S. N. and A. D. Riggs. 2003. Epigenetic silencing may aid evolution by gene duplication. J Mol Evol **56**:718–729.

Salathe, M., M. Ackermann, and S. Bonhoeffer. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol Biol Evol **23**:721–722.

Samuelson, L. C., K. Wiebauer, C. M. Snow, and M. H. Meisler. 1990. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Mol Cell Biol **10**:2513–2520.

Schinkel, H., M. Hertzberg, and G. Wingsle. 2001. A small family of novel CuZn-superoxide dismutases with high isoelectric points in hybrid aspen. Planta **213**:272–279.

Sémon, M. and L. Duret. 2006. Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals. Mol. Biol. Evol. **23**:1715–1723.

Seoighe, C., C. Johnston, and D. Shields. 2003. Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. Mol Biol Evol **20**:484–90.

Seoighe, C. and K. H. Wolfe. 1999. Yeast genome evolution in the post-genome era. Curr Opin Microbiol **2**:548–554.

Shemesh, R., A. Novik, S. Edelheit, and R. Sorek. 2006. Genomic fossils as a snapshot of the human transcriptome. Proc Natl Acad Sci U S A **103**:1364–1369.

Singer, G. A., A. T. Lloyd, L. B. Huminiecki, and K. H. Wolfe. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol. Biol. Evol. **22**:767–775.

Skrabanek, L. and F. Campagne. 2001. TissueInfo: high-throughput identification of tissue expression profiles and specificity. Nucleic Acids Res **29**:E102–2.

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

Soares, M. B., E. Schon, A. Henderson, S. K. Karathanasis, R. Cate, S. Zeitlin, J. Chirgwin, and A. Efstratiadis. 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol **5**:2090–2103.

Soares, M. B., A. Turken, D. Ishii, L. Mills, V. Episkopou, S. Cotter, S. Zeitlin, and A. Efstratiadis. 1986. Rat insulin-like growth factor II gene. A single gene with two promoters expressing a multitranscript family. J. Mol. Biol. **192**:737–752.

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

Sorek, R., R. Shamir, and G. Ast. 2004. How prevalent is functional alternative splicing in the human genome? Trends Genet **20**:68–71.

Spellman, P. T. and G. M. Rubin. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. **1**:5.

Steane, D. A. 2005. Complete Nucleotide Sequence of the Chloroplast Genome from the Tasmanian Blue Gum, Eucalyptus globulus (Myrtaceae). DNA Res **12**:215–220.

Sterck, L., S. Rombauts, S. Jansson, F. Sterky, P. Rouze, and Y. Van de Peer. 2005. EST data suggest that poplar is an ancient polyploid. New Phytol **167**:165–170.

Su, Z., J. Wang, J. Yu, X. Huang, and X. Gu. 2006. Evolution of alternative splicing after gene duplication. Genome Res **16**:182–189.

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

Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**:4673–80.

Tuskan, G. A., S. Difazio, S. Jansson, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313**:1596–1604.

Van de Peer, Y., J. S. Taylor, I. Braasch, and A. Meyer. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J Mol Evol **53**:436–446.

Vinckenbosch, N., I. Dupanloup, and H. Kaessmann. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc. Natl. Acad. Sci. U S A **103**:3220–3225.

Vinogradov, A. E. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet **20**:248–253.

Wagner, A. 2000a. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc Natl Acad Sci U S A **97**:6579–6584.

Wagner, A. 2000b. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. Genetics **154**:1389–1401.

Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A **102**:5483–5488.

Wang, W., H. Yu, and M. Long. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. Nat Genet **36**:523–527.

Wang, W., H. Zheng, S. Yang, et al. 2005. Origin and evolution of new exons in rodents. Genome Res **15**:1258–1264.

Warrington, J. A., A. Nair, M. Mahadevappa, and M. Tsyganskaya. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol Genomics **2**:143–147.

Waterston, R. H., K. Lindblad-Toh, E. Birney, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Watson, J. D., N. H. Hopkins, R. J. W., S. J. A., and W. A. M. 1965. Molecular Biology of the Gene, vol. 1. Benjamin/Cummings, Menlo Park, CA.

Waxman, D. and J. R. Peck. 1998. Pleiotropy and the preservation of perfection. Science **279**:1210–1213.

Wheeler, D., D. Church, R. Edgar, et al. 2004. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res **32 Database issue**:D35–40.

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

Wilson, A. C., S. S. Carlson, and T. J. White. 1977. Biochemical evolution. Annu Rev Biochem **46**:573–639.

Winter, E. E., L. Goodstadt, and C. P. Ponting. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res **14**:54–61.

Woese, C. R. 1965. On the evolution of the genetic code. Proc Natl Acad Sci U S A **54**:1546–1552.

Wolf, Y. I. 2006. Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. Trends Genet **22**:354–357.

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

Wyckoff, G. J., C. M. Malcom, E. J. Vallender, and B. T. Lahn. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. Trends Genet **21**:381–385.

Xing, Y. and C. Lee. 2004. Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy. Trends Genet **20**:472–5.

Xing, Y. and C. Lee. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Proc Natl Acad Sci U S A **102**:13526–13531.

Xing, Y. and C. Lee. 2006a. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. Nat Rev Genet **7**:499–509.

Xing, Y. and C. Lee. 2006b. Can RNA selection pressure distort the measurement of Ka/Ks? Gene **370**:1–5.

Xing, Y., A. Resch, and C. Lee. 2004. The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. Genome Res **14**:426–41.

Xing, Y., Q. Xu, and C. Lee. 2003. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett **555**:572–578.

Yan, Y.-L., J. Willoughby, D. Liu, J. G. Crump, C. Wilson, C. T. Miller, A. Singer, C. Kimmel, M. Westerfield, and J. H. Postlethwait. 2005. A pair of Sox: distinct and overlapping functions of zebrafish sox9 co-orthologs in craniofacial and pectoral fin development. Development **132**:1069–1083.

Yanai, I., H. Benjamin, M. Shmoish, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21**:650–659. Evaluation Studies.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555–6.

Yang, Z. and J. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends in ecology and evolution **15**:496–503.

Yeo, G. W., E. Van Nostrand, D. Holste, T. Poggio, and C. B. Burge. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A **102**:2850–2855.

Yu, W.-P., S. Brenner, and B. Venkatesh. 2003. Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. Trends Genet **19**:180–183.

Zhang, J. and X. He. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol **22**:1147–1155.

Zhang, L. and W. H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol **21**:236–239.

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

Zhang, Z. and H. Kishino. 2004. Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. Genetics **166**:1995–1999.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7**:203–214.

Zuckerkandl, E. and L. Pauling. 1965. Evolutionary Divergence and Convergence in Proteins. *In* V. Bryson and H. Vogel, eds., Evolving genes and proteins. New York Academic Press, New York, pp. 97–166.