uOttawa

L'Université canadienne
Canada's university

t

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Gareth Palidwor
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)
GRADE / DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Tissue-Specific Codon Usage in Mouse

TITRE DE LA THÈSE / TITLE OF THESIS

Xuhua Xia
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

Stéphane Aris-Brosou

Guy Drouin

Michel Dumontier

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Tissue-specific codon usage in mouse

Gareth A. Palidwor

A thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in partial fulfilment of the requirements for the Master of Science Degree
In the Ottawa-Carleton Institue of Biology

Thèse soumise à
Faculté des études supérieures et postdoctorales
L'Université d'Ottawa
En vue de l'obtention de la maitrise en sciences
L'Institut de biologie d'Ottawa-Carleton

Your file Votre référence
ISBN: 978-0-494-79705-1
Our file Notre référence
ISBN: 978-0-494-79705-1

Abstract

Codon usage bias is due to a combination of biased mutation and selection effects. Isochore-related GC mutational bias has been shown to be the dominant cause of tissue-specific codon bias in adult human tissues and limited evidence for translational selection has been shown. This thesis is a comprehensive evaluation of the relative contribution of selection and mutation to variation in codon usage among different mouse embryonic tissues.

Through a detailed analysis of tissue-specific codon usage in relation to gene expression and gene location of mouse embryonic tissue-specific genes, I found that translational selection is partially responsible for some differences in codon usage among tissue-specific genes. This observation indicates profound impact of selection favouring codon-anticodon adaptation. The characteristic GC biases of tissue-specific gene sets are shown not to be caused by their clustering on the same isochore. Tissue-specific genes are no more clustered on the genome than randomly selected genes.

In mouse, the usage of some G-ending codons decreases with increasing GC bias, while all other G and C-ending codons increase. To understand this counterintuitive observation, we generate a continuous-time Markov chain model of GC-biased synonymous substitution which explains qualitative usage patterns of all codons, including non-linear and sometimes non-monotone codon usage in isoleucine, arginine and leucine. This effect is universal, extending to mouse and human genes as well as plant and prokaryotic genomes.

This work enriched our understanding of the codon-anticodon adaptation theory and extended it to the level of tissue-specific genes. The result suggests the possibility

of tissue-specific tRNA pools mediating tissue-specific codon-anticodon adaptation. The methods developed in the thesis can be easily extended to characterize this previously little explored facet of codon-anticodon adaptation.

Résumé

Le biais d'usage des codons est dû à une combinaison de mutation biaisée et d'effets de sélection. L'isochore lié au biais mutationnel GC a été démontré comme étant la cause principale de biais des codons spécifiques aux tissus dans les tissus humains adultes et peu de preuves d'appui pour la sélection de translation ont été observées. Cette thèse est une évaluation compréhensive de la contribution relative de la sélection et de la mutation à la variation de l'usage des codons chez les différents tissus de souris embryonnaires.

Grâce à une analyse détaillée de l'usage des codons spécifiques aux tissus par rapport à l'expression des gènes et à la localisation des gènes embryonnaires spécifiques aux tissus de la souris, j'ai trouvé que la sélection translationnelle est en partie responsable pour certaines différences dans l'usage des codons parmi les gènes spécifiques aux tissus.

Cette observation montre l'impact significatif de la sélection favorisant l'adaptation codon-anticodon. Les caractéristiques des biais GC des ensembles de gènes spécifiques aux tissus ne semblent pas être causés par leur regroupement sur le même isochore. Les gènes spécifiques aux tissus ne sont pas plus regroupés sur le génome que des gènes choisis au hasard.

Chez la souris, l'utilisation de certains codons se terminant en G diminue avec l'augmentation des biais GC, alors que tous les autres codons se terminant en G et en C augmentent. Pour comprendre cette observation paradoxale, nous générons un modèle de chaîne de Markov à temps continu de la substitution synonyme des CG biaisés qui explique les patrons d'utilisation qualitative de tous les codons, y compris

l'usage des codons non-linéaires et parfois non-monotones dans la isoleucine, la leucine et l'arginine. Cet effet est universel, s'étendant aux gènes de la souris et l'homme ainsi qu'à ceux des plantes et des génomes procaryotes.

Ce travail a enrichi notre compréhension de la théorie d'adaptation codon-anticodon et l'a étendue au niveau de gènes spécifiques aux tissus. Les résultats suggèrent la possibilité que les regroupements spécifiques aux tissus d'ARNt modèrent l'adaptation des codon-anticodon spécifiques aux tissus. Les méthodes développées dans la thèse peuvent facilement être amplifiées afin de caractériser cette facette jusque-là peu explorée de l'adaptation codon-anticodon.

Acknowledgements

I thank Dr. Xuhua Xia for accepting me as a graduate student and guiding my thesis work, Dr. Theodore Perkins for guidance in the development of the Continuous-Time Markov Chain models of codon bias, Dr. Stéphane Aris-Brosou for extraordinary support in the completion of this thesis, and my committee members Dr. Mads Kaern, Dr. Michel Dumontier and former co-supervisor Dr. Miguel Andradé for their ideas, advice and support.

My co-workers Christopher Porter, Paul Krzyzanowski, Pearl Campbell and Dr. Carolina Perez-Iratxeta have provided me with invaluable intellectual and moral support as I have pursued this research. Erinija Pranckeviciene, Anna van Weringh and Manon Ragonnet provided insightful comments and suggestions which significantly improved the manuscript.

I am grateful to my wife Michelle and children Enid and Amelia for providing me with the love, patience and time required to complete my studies.

It is easy, upon entering a field of study, to become overwhelmed and despair that everything of significance has already been found. The most gratifying result of this research for me was the realization that the careful investigation of apparently minor deviations from the expected can lead to novel discoveries, even in well studied areas. There is much still to be found.

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'

–Isaac Asimov

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Codon bias

In the genetic code most amino acids are encoded by multiple codons, and the use of synonymous codons with unequal frequencies is termed codon bias. Different species exhibit different codon usage bias with the degree and type of bias varying among species (Grantham et al., 1980, 1981; Moriyama & Hartl, 1993; Xia, 1996), among genes of a single species' genome (Gouy & Gautier, 1982; Ikemura, 1985; Sharp & Li, 1986) and even among sets of genes with tissue-specific expression in a single species' genome (Vinogradov, 2003; Plotkin et al., 2004; Semon et al., 2006).

Codon usage bias has been attributed to tRNA-mediated selection to maximize translational rate (Ikemura, 1981, 1982; Gouy & Gautier, 1982; Sharp et al., 1986; Bulmer, 1991; Xia, 1998) and accuracy (Akashi, 1994; Stoletzki & Eyre-Walker, 2007) as well as biased mutational effects (Muto & Osawa, 1987; Sueoka, 1988). These effects are not mutually exclusive, and observed codon bias is a balance between selection

and mutational effects as reviewed by Hershberg & Petrov (2008).

This study focuses on determining the causes of codon bias between tissue-specific sets of genes in *M. musculus* by testing for both selection and mutational effects and investigating apparent anomalies in codon usage.

1.1.1 Codon bias due to tRNA-mediated selection

Translation rate depends on the coding sequence of a protein and the cellular environment in which the translation occurs. The time required for any given codon being translated to receive a cognate tRNA is stochastic, though on average, codons corresponding to more abundant tRNA charged with an amino acid will be translated more rapidly (Garel, 1974). Thus proteins that use codons corresponding to more abundant charged tRNA species will generally be translated more quickly. There have been claims that codons cognate to rare tRNA may be selected to reduce translation rates (Zhang et al., 1991; Saier, 1995; Lavner & Kotlar, 2005), however these claims are poorly reasoned and have never received any substantial support (Bulmer, 1991).

Codon optimization for translational efficiency has been observed in many species and is most visible in highly expressed and ubiquitously expressed genes. The degree of translation-related codon bias tends to decrease as organismal complexity increases. More complex organisms generally have reduced population size and reproduce more slowly, subjecting genes to weaker selection effects (Subramanian, 2008). Strong translation-related codon bias effects have been identified in unicellular organisms such as prokaryotes (Ikemura, 1981; Gouy & Gautier, 1982) and yeast (Ikemura,

1982; Sharp et al., 1986) The effect has also been reported in simpler metazoans such as *C. elegans* (Duret, 2000), *D. melanogaster* (Moriyama & Powell, 1997), but is much weaker than that for unicellular organisms.

Just as codon bias may be selected to optimize for tRNA abundance, tRNA abundance may be selected to adapt to codon bias. This mutual selection is called *coadaptation* and has been reported in yeast (Percudani et al., 1997) and *C. elegans* (Duret, 2000).

In vertebrates the existence of tRNA-mediated selection has been reported for a number of organisms, though the effect is weak. tRNA-mediated selection has been claimed for *Xenopus laevis* (Musto et al., 2001) by the identification of the use of preferred codons in highly expressed genes. In *Homo sapiens* a weak positive correlation has been claimed between gene expression levels for broadly and ubiquitously expressed genes and cellular tRNA abundance as estimated from gene copy numbers (Comeron, 2004; Lavner & Kotlar, 2005). Yang & Nielsen (2000) analyzed several mammalian species and found that they could reject the null hypothesis of codon usage bias solely due to mutation for most genes. Despite these results, claims of tRNA-mediated selection in vertebrates are disputed due to the weakness of the effect and the difficulty in distinguishing between selection and other confounding effects such as biased gene conversion (Duret, 2002).

1.1.2 Codon bias due to GC mutational bias

GC mutational bias is a major influence on codon bias both in vertebrate (Bernardi et al., 1985) and prokaryotic genomes (Muto & Osawa, 1987), resulting in a close

association between GC% at the third codon position, also called GC3 (Sueoka, 1988) and GC-bias (as measured by GC% in prokaryotes or intergenic isochoric GC% in mammals). GC3 is most strongly affected by mutational bias as most mutations in this position are synonymous and will not affect the resulting coding sequence (Comeron, 2004).

As a result of this, there is a common belief that use of synonymous A/T-ending codons should decrease, and C/G-ending codons increase in frequency with increasing GC mutational bias. Despite widespread use of GC3 as a measure of the effect of GC mutational bias on coding sequences, an explicit model of the relationship between GC mutational bias and codon bias does not appear to have been published. Closely related work has been done; Knight et al. (2001) modelled codon and amino acid usage as a function of GC mutational bias in prokaryotes and eukaryotes showing that GC content drives codon and amino acid usage and provided a model of usage by compositional class, but did not directly address codon bias. In the absence of a null model for comparison, claims such as those of Kliman & Bernal (2005) that deviation from expected correlations between codon usage and GC-bias is evidence for selection are questionable. There is a clear relationship between the GC content of coding sequences and gene expression in mammals; highly expressed genes tend to be more common in GC rich isochores (Bernardi, 2000), thus genes with GC rich coding sequences tend to be more highly expressed.

1.1.2.1 Isochores as a form of mutational bias

Vertebrate genomes contain isochores: large regions (up to megabases in length) with relatively uniform GC content that can range from 30% to 60% in *Mus musculus*

(Cuny et al., 1981).

The GC3 of coding sequences is correlated to GC% of the isochore in which they are contained, as are intronic, 5' and 3' UTR GC values (Bernardi, 2000). Though individual isochores were initially believed to be of relatively uniform GC content (Cuny et al., 1981), analysis of full genomic sequences showed that the identification of a characteristic isochoric GC content can be problematic due to large local variations within isochores (Nekrutenko & Li, 2000).

Isochores of low GC content are more abundant than those of high GC content. The rarer GC rich regions contain the majority of genes while more common AT-rich regions contain few genes (Bernardi, 2000).

In mammals there is a strong relationship between breadth of gene expression and isochores; ubiquitously expressed so-called housekeeping genes are more likely to reside in the higher-GC isochores, while tissue-specific genes are more likely to be present in the lower GC range isochores (Vinogradov, 2003). Rodent genomes (*M. musculus* and *Rattus norvegicus*) have the weakest isochoric structure and narrowest GC range of the placental mammals (amniotes) for which full genomes are currently available; the human genome, in contrast, has one of the strongest isochoric structures and broadest GC ranges (Romiguier et al., 2010).

The mechanism behind the formation and maintenance of isochores has not been conclusively determined: it is not clear to what degree they are due to mutational or selection effects. It was initially hypothesized that isochores were due to regional mutational biases in affected genomes (Bernardi, 1993). A later study showed a general AT bias and no isochore specific mutational biases to more recent mutations (Alvarez-Valin et al., 2002). Though there is a possibility that isochores may be due

to a biased mutational effect that is not currently operating, it seems an unlikely explanation given the ubiquity of ischores in mammalian genomes (Romiguier et al., 2010).

Bernardi, who originally described isochores (Cuny et al., 1981), has proposed a so-called *"neo-selectionist"* model for isochore formation and maintenance (Bernardi, 2007) based on a combination of selection and biased mutational effects. In this model, regions of the genome stochastically accumulate AT-biased mutations, gradually becoming GC poor. Once a lower GC threshold is passed, the chromatin in that region will shift to a more closed conformation, inhibiting local gene expression. The changed chromatin conformation can spread and affect adjoining regions (Talbert & Henikoff, 2006) leading to inhibition of transcription in the affected regions. This effect will cause harmful suppression of gene expression and will be selected against, leading to conserved isochoric patterns in mammalian genomes. Bernardi (2007) claims this model explains the formation and maintenance of isochores and why isochoric GC ranges are more pronounced in mammals than in cold blooded vertebrates; the higher mammalian body temperature raises the lower GC threshold at which the deleterious chromatin effects occur, and raises the minimum GC for coding regions leading to more extreme isochoric ranges. They do not, however, provide any empirical evidence to support the claims of this model.

Biased gene conversion has been proposed as a mechanism to explain the formation and maintenance of isochores in mammalian genomes (Duret et al., 2006; Duret & Galtier, 2009). Biased gene conversion, reviewed in Marais (2003), is a consequence of the repair of double strand breaks by recombination where GC base pairs appear to be favoured. In mammalian cells the DNA repair mechanism for mismatches is

GC-biased (Brown & Jiricny, 1988); if this mechanism is part of the repair process during recombination, then the biased gene conversion resulting should itself be GC-biased. Supporting this hypothesis, it has been shown that recombination rates in the *H. sapiens* genome are positively correlated with GC content in the same region (Fullerton et al., 2001). Duret & Arndt (2008) used the biased genome conversion model to predict nucleotide substitution patterns in the *H. sapiens* genome and claim that their results allow them to reject selectionist models of isochore evolution. This study was limited to the human genome however, a recent study of 33 placental mammal genomes showed that, using the biased gene conversion model, species' genomic size and body mass can be used to predict the isochoric properties of their genome (Romiguier et al., 2010).

Isochores have, to date, not been definitively shown to be caused by mutation, selection or some combination thereof. It is outside of the scope of this work to determine the cause of this observed bias, though the balance of evidence seems to support the model of biased gene conversion (Duret & Galtier, 2009). For the purposes of this study we will treat this effect as mutational as the observed effect is very similar to that which is seen with GC mutational bias, with strong correlations between 5' and 3' UTRs as well as GC3, introns and flanking non-coding sequences (Bernardi, 2000).

Given the strong effect of isochoric bias on the content of codon sequence, it is possible that it may have an influence on tissue-specific codon bias if tissue-specific genes tend to be located on the same isochore or on isochores of characteristic GC content.

1.2 Tissue-specific codon bias

Two hypotheses have been put forward to explain observed tissue-specific codon bias.

The mutation hypothesis (Semon et al., 2006) argues that the large regional GC content variability (isochores) of vertebrate genomes affects the GC3 of the contained genes' coding regions. If genes specific to particular tissues tend to be from the same isochore class, this will lead to the observed tissue-specific codon bias. It predicts a correlation between the GC3 of coding sequences and the GC background content of neighbouring non-coding regions.

The selectionist hypothesis (Plotkin et al., 2004; Dittmar et al., 2006; Kotlar & Lavner, 2006) argues that codon usage in highly expressed tissue-specific genes has evolved to increase translation rates by using tissue-specific tRNA pools. This selection results in the observed differences in codon usage. This hypothesis predicts that highly expressed genes will show a higher codon usage bias than those with low expression levels. These more commonly used codons in the highly expressed genes will correspond to the most abundant charged tRNA in the corresponding tissue if they are subject to strong selection. The low expression genes, presumably less subject to selection, should show less difference in codon usage.

In the analyses of Lavner and Kotlar (Lavner & Kotlar, 2005; Kotlar & Lavner, 2006), the copy numbers of tRNA genes are used as an estimate of the average relative cellular abundance of tRNA in *H. sapiens*. This has been shown to be a reasonable assumption for unicellular species (Ikemura, 1981; Gouy & Gautier, 1982) and yeast (Ikemura, 1982; Sharp et al., 1986), and even simple multi-cellular eukaryotes such as *Caenorhabditis elegans* (Duret, 2000) and *Drosophila melanogaster* (Moriyama &

Powell, 1997). In more complex eukaryotes such as mammals tRNA levels can vary widely by tissue (Dittmar et al., 2006). Genomic tRNA gene counts may not therefore provide a reliable estimate of abundance under all conditions. It is notable that even with high tissue-specific variability of tRNA abundance in human, highly expressed genes have been shown to preferentially use codons that correspond to tRNA with the highest gene copy number in the genome (Comeron, 2004). Note that the analysis of Comeron (2004) was not restricted to tissue-specific genes; the genes analyzed were broadly or ubiquitously expressed. The results, therefore, describe the effect of selection on tRNA abundance averaged across many tissues.

There have been relatively few published studies of tissue-specific codon usage in mammals, and even fewer have analyzed differences in codon usage between tissue-specific genes. An early paper on this topic by Urrutia & Hurst (2001) noted that tissue-specific genes in *H. sapiens* showed weaker bias than ubiquitously expressed genes, but did not analyze differences between tissue-specific genes. Vinogradov (2003) demonstrated that the codon bias of tissue-specific genes in both *M. musculus* and *H. sapiens* is largely attributable to their isochoric association, and that tissue-specific genes tend to have a characteristic GC content for each tissue. This effect was later noted again by Semon et al. (2006), and remains unexplained, though it has been proposed that the isochoric association of tissue-specific genes may be a mechanism for controlling expression of those genes (Vinogradov, 2003, 2005). This would be consistent with tissue-specific co-adaptation of tRNA abundance and codon bias.

One study claimed to have observed tissue-specific codon bias not attributable to isochoric effects in *H. sapiens* (Plotkin et al., 2004). This result was convincingly

refuted as being an artifact of their analysis method by Semon et al. (2006) who used correspondence analysis to demonstrate that the majority of between-tissue codon bias is attributable to tissue-specific GC-bias. While the results of Semon et al. (2006) are credible, their analysis is a broad statistical description of all codons. Despite the majority of bias being explicable by isochoric bias in their analysis, it is possible that some codons may show differences in usage not correlated with isochoric bias which may be due to selection or other as-yet unidentified effects. Additionally, their analysis cannot distinguish between translational selection and mutational effects in cases where the mutational bias corresponds to translational selection; for example a G-ending codon experiencing translational selection in a tissue with a strong GC mutational bias. Due to the much weaker isochoric structure in *M. musculus* the relationship between the tissue specific bias and isochores may have differences from *H. sapiens*.

More recently Waldman et al. (2010) used the tRNA Adaptation Index (tAI) (Dos Reis et al., 2003) as an estimate of translation efficiency in *H. sapiens* and *M. musculus* and found it to be correlated with both tissue-specific and global expression levels, with high expression genes showing a high tAI. This is evidence of selection for translational efficiency. The effect in mouse was found to be weaker than in human; in addition adult tissues were found to show higher correlation than embryonic tissues.

1.3 Research Motivation

This thesis explores the relationship between codon usage and tissue-specific gene expression in *M. musculus* using genomic sequence and microarray data to test the

mutation and selection hypotheses.

This thesis identifies whether there are significant codon usage differences between tissue-specific gene sets, and whether those differences are attributable to selection or mutation effects. The genomic proximity of tissue-specific genes is investigated to determine to what extent it affects their GC-bias. Notable anomalies in codon usage with respect to GC-bias, particularly for the AGG and TTG codons, are identified and explained.

1.4 Overview of Subsequent Chapters

This chapter (Chapter 1) provides an overview of translation and mutation related codon bias, its relationship to tissue-specific codon bias, and explains the motivation for this study.

Chapter 2 is an analysis of between-tissue and between-lineage difference of codon usage in *M. musculus* based on microarray expression data from embryonic tissues.

Chapter 3 is an exploration of isochores and tissue-specific genes, testing the hypothesis that the characteristic GC3 observed in tissue-specific genes is due to those genes tending to be in close proximity to each other and thus being subject to similar isochoric GC-bias.

Chapter 4 provides an explanation of an apparent codon usage anomaly noted in *M. musculus* tissue-specific genes. To explain these observations, a continuous-time Markov chain model of GC-biased synonymous substitution is generated. This model correctly predicts the qualitative usage patterns of all codons, including non-linear codon usage in isoleucine, arginine and leucine. The model is shown to be applicable

to *M. musculus*, *H. sapiens* prokaryotic and plant sequences.

Chapter 5 contains a general discussion of the results contained in this thesis and some concluding remarks.

Chapter 2

Testing for tissue-specific codon usage

2.1 Abstract

In this chapter, significant differences in codon usage between sets of tissue-specific *M. musculus* genes are demonstrated. These differences are shown to be between tissues with genes having opposite isochoric GC mutational bias (GC rich versus AT rich) and the type of difference is found to match the GC-biases. Mutational effects on codon bias are distinguished from selection effects by splitting the genes into high and low expression sets and re-analyzing each individually; high expression genes exhibit significant differences in codon usage between tissues while low expression genes do not. This is evidence for translational selection to match tissue-specific tRNA pools, however in all cases codons showing differences between tissues correspond to the GC-biases of the tissues.

Tissue-specific genes clustered by major lineage groupings (mesoderm, external ectoderm and neuroectoderm) show significant differences between tissue-specific genes as a whole as well as for high and low expression subsets; as the high expression subset shows less difference than the low expression subset, there is no evidence for translational selection to match lineage tissue-specific tRNA pools.

2.2 Introduction

The mutational hypothesis (Semon et al., 2006) predicts that tissue-specific codon biases will be due to the GC mutational biases of the isochores in which the tissue-specific genes are resident. The observed codon bias is predicted to be correlated with the GC level of the host isochores as measured by GC3; in GC rich isochores G or C ending codons will be favoured and in AT rich isochores A or T ending codons will be favoured. Purely mutational bias will be unaffected by expression level and will be similar for both high and low expression genes.

The selection hypothesis predicts that tissue-specific biases are due to selection for translation efficiency and should be stronger for high expression genes than for low expression genes due to selective pressures (Plotkin et al., 2004).

Codon biases which are correlated with the GC mutational bias of the host isochores are suggestive of mutational effects but do not exclude the possibility that they are due to selection which corresponds to mutational bias. If, however, there are codon biases which are present in highly expressed genes and absent in low expression genes, the observation supports the hypothesis of codon selection for transcriptional efficiency.

Codons which appear to be subject to selection for transcriptional efficiency as described above may also be subject to matching mutational bias. The interaction between translational selection of codon bias, isochoric mutational bias and selection of tRNA specific pools may lead to the presence of tissue specific genes in certain isochoric ranges, as previously described in human tissues (Semon et al., 2006; Vinogradov, 2003).

			Interpretation	
Codon[1]	Isochore[2]	$tRNA_{AC}$[3]	Mutation[4]	tRNA sel.[5]
$P_G > P_A$	GC rich	$P_C > P_U$	X	X
		$P_C < P_U$	X	
	AT rich	$P_C > P_U$		X
		$P_C < P_U$		
$P_G < P_A$	GC rich	$P_C > P_U$	X	
		$P_C < P_U$	X	X
	AT rich	$P_C > P_U$		
		$P_C < P_U$		X

Table 2.1: Mutation and tRNA pool selection hypotheses predictions for NNR Codon families
(1)Proportion of G and A ending codons (P_G and P_A respectively)
(2)Isochore type
(3)Proportion of nucleotide in tRNA anticodon wobble position (mouse NNR codons have only G and A)
Hypotheses in reference to the empirical evidence. X indicates that the observations are consistent with the hypothesis.
(4) Mutational hypothesis that codon usage will match host isochore bias.
(5) tRNA selection hypothesis that tRNA abundance will be selected to match codon bias (and hence mutational bias).

Predictions of the mutational and tRNA abundance selection hypothesis are summarized in Table 2.1, taking into account potential non-Watson-Crick wobble pairing between the 3rd codon position and the first tRNA anticodon position. Consider a

tissue with a tRNA pool having more wobble U than wobble C tRNA for codons from an NNR family amino acid (2 codon families ending in A or G, e.g. phenylanyline (TTR)).

This would favour the usage of A-ending codons for that amino acid in that tissue. If the genes expressed in the tissue tend to be located in GC-rich regions with GC-biased mutations, then selection to increase A-ending codons and mutation to increase G-ending codons would work in opposition. If the genes expressed in this tissue tend to be located in AT-rich regions, then selection and mutation are in the same direction. This also implies a benefit for genes expressed in this tissue to be located in AT-rich regions and a cost for genes expressed in this tissue to be located in GC-regions. Analogous arguments can be made for codons from NNY family amino acids (2 codon families ending in T or C, e.g. asparagine (GAY)) and NNN family amino acids (those ending in A, C, G or T, e.g. proline (CCN)).

By this coadaptation mechanism, over time, tissue-specific tRNA pools would favour tissue-specific genes present in AT-rich or GC-rich regions with the combination of mutation and selection effects leading to isochore related tissue-specific codon usage. Thus, finding differential usage corresponding to isochoric bias is not against the hypothesis of selection mediated by tissue-specific tRNA pools, and if observed, provides some support for the existence of the above mechanism. Ideally, tissue specific tRNA abundance data would be used to test tRNA selection hypothesis as well, but no such data is available for these tissues.

To test the mutational and selection hypotheses, tissue gene expression data are required. StemBase (Sandie et al., 2009) is a source of curated data containing *M. musculus* Affymetrix GeneChip microarray data from samples with a high proportion

of stem cells, embryonic tissues and differentiation time series data. By using data from a single source it is hoped that lab specific variability will be reduced (Irizarry, et al., 2005). Though StemBase contains multiple chip types, the gene expression data used in the analysis is based solely on data from the Mouse Expression Set 430 (MOE430) GeneChip consisting of a matched pair of chips, MOE430A and B. Using data from only one chip set avoids complications associated with combining expression values across chip types (Hwang et al., 2004).

Gene expression microarrays provide measurements of the concentration of polyadenylated mRNA in samples. These do not directly describe protein translation rates or transcription levels due to the complexity of post-transcriptional regulatory mechanisms. GeneChip expression values may be considered as a proxy for protein expression levels in the absence of direct protein translation rate data. It has been estimated that a maximum of 40% of protein expression variation in mammalian cells is described by mRNA expression measurements with more highly expressed genes showing a better correlation between mRNA and protein abundance (Tian et al., 2004). As we are primarily concerned with the presence or absence of a given gene transcript, and whether it is of high or low expression, microarray data should be sufficiently accurate for the purposes of this analysis.

Ideally, tissue-specific analyses would use cell type specific gene expression data. In practice, however, it is usually impractical to extract and hybridize data from only a single cell type. Though some samples in StemBase are highly enriched in a single cell type (e.g. haematopoietic stem cells) most sample hybridizations were performed on the heterogenous cell populations which make up *M. musculus* tissues. Mammalian brain tissue, for example, consists primarily of neural and glial cells,

but also contains a small percentage of other cell types such as leukocytes and the endothelial cells of blood vessels. A microarray hybridization of mRNA extracted from such a tissue sample is therefore a measure of expression from an heterogenous populations of cell types, though is usually dominated by cells of a single or small set of lineage-related types.

Despite the limitations of microarray data in measuring protein expression levels and capturing cell type specific data, a number of tissue-specific gene expression analyses have been performed based on tissue samples hybridized to microarrays similar to those used in this study (Vinogradov, 2003; Lavner & Kotlar, 2005; Semon et al., 2006).

In this chapter we first determine whether differences in tissue-specific codon usage can be detected in our *M. musculus* data. The specific tissues and codons which exhibit differences are then examined in more detail. To distinguish between selection and mutational bias effects we split the tissue-specific data into high and low expression genes based on microarray expression data and determine whether it is possible to identify significant codon usage differences within each set.

The tissues being analyzed have developmental lineage relationships. If tRNAs show tissue-specific differences in abundance, it may also be the case that there are tRNA pools common to developmentally related tissues. To test for this, we group the experimentally determined tissue-specific genes by major lineage group (external ectoderm, mesoderm and neuroectoderm, see Figure 2.1 for hierarchic lineage relationships) and determine whether there are significant differences in codon usage between these.

The analysis shows significant differences in codon usage between tissue-specific *M.*

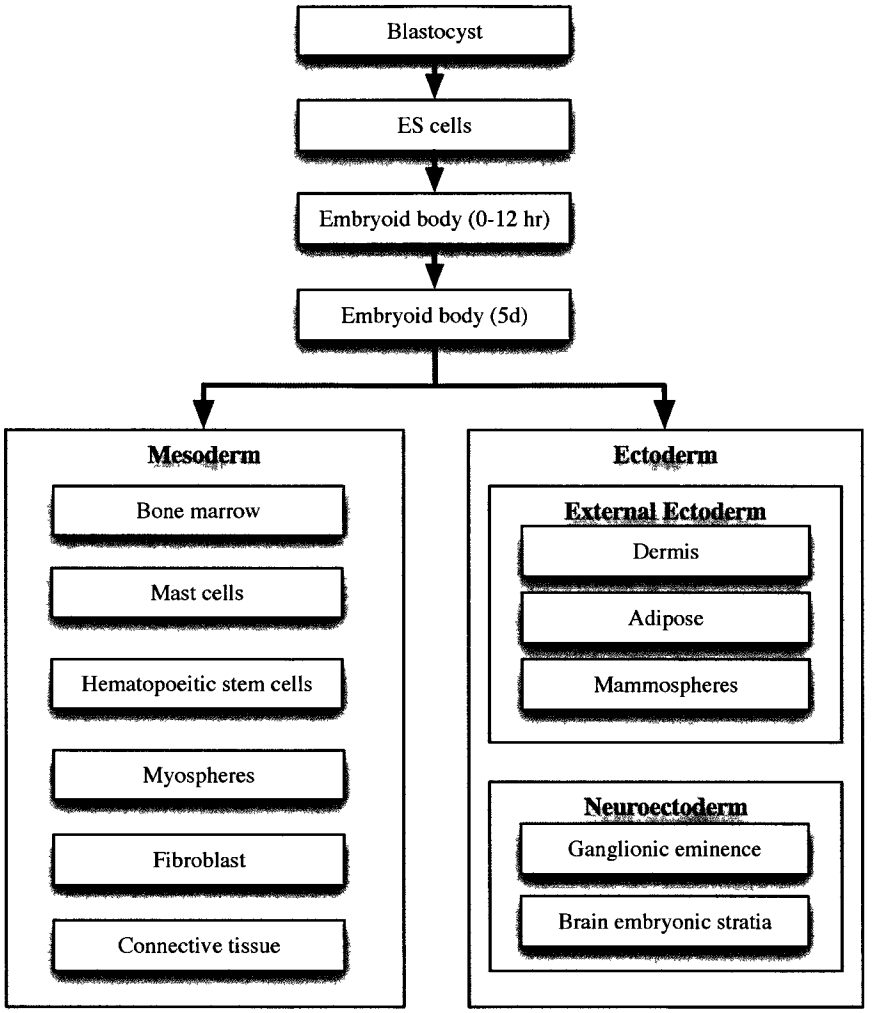

Figure 2.1: Lineage relationship of tissue samples used in analysis. Arrows indicate parent-child tissue relationship, boxes indicate major tissue class groupings (mesoderm, ectoderm, external ectoderm, neuroectoderm), nesting of boxes indicates parent-child relationship of lineage.

musculus gene sets, with these differences being between tissues with genes having opposite isochoric mutational bias. Though there is evidence for translational selection, in all cases this selection also matches the isochoric mutational bias. Tissue-specific genes grouped by lineage show significant differences in codon usage between lineages,

with most of these differences matching isochoric bias between gene sets. There is no evidence for translational selection by lineage.

2.3 Materials and Methods

2.3.1 Computing environments and data sources

The Ensembl (Hubbard et al., 2009) *M. musculus* genome version 52 based on the NCBI m37 assembly was used as the source of genomic sequences and annotation accessed via the Ensembl Perl API (Stabenau et al., 2004). Nucleotide and peptide sequences and genomic location of all introns and coding sequences for each of the 22,754 known protein-coding genes in Ensembl were extracted. The transcript with the longest coding sequence for each gene was selected when more than one splice variant was documented to maximize the number of codons available for analysis. Ideally the most frequently used splice variant would be selected but this information is not available. Nucleotide composition for background coding sequence was estimated by extracting sequence 5000 bases up and downstream from the start and end of the gene respectively. The start of the gene is defined as the beginning of the longest 5' UTR and the end of the gene is defined as the end of the longest 3' UTR associated with that gene. Other known gene exons identified within the background sequences were removed, truncating the sequence. The minimum sequence length used to calculate background resulting from this method is 2452 nucleotides. Codon usage, GC3 (GC content for the third codon position) intronic GC, GCg (GC content for the entire CDS) was calculated for each gene.

Manipulation of data and statistical calculations were performed in the R language Version 2.10.0 (R Development Core Team, 2010) using the Bioconductor library (Gentleman et al., 2004), and Perl using the Bioperl library version 1.5.1 (Stajich et al., 2002).

2.3.2 Codon Usage Measures

In this study, two types of codon usage measures are used: direct counts of codons usage and Relative Synonymous Codon Usage (RSCU). RSCU is a measure of non-uniform usage of synonymous codons (Sharp et al., 1986).

The RSCU value for the i^{th} codon of an amino acid with n codons is the ratio of observed codon count x_i to the expected codon count assuming no bias. More formally:

$$RSCU_i = \frac{x_i}{\frac{1}{n}\sum_{j=1}^{n} x_j} \tag{2.1}$$

The codons AUG and UGG corresponding to methionine and tryptophan respectively are excluded from this analysis as both amino acids have only one codon, thus their RSCU value will always be 1.

2.3.3 Tissues

The full set of samples in the StemBase database was reviewed and grouped based on tissue annotations. The samples were then organized into a developmental hierarchy based on the known embryonic developmental patterns in *M. musculus* (Gilbert,

2003); the developmental relationship between the samples used are shown in Figure 2.1. StemBase contains samples based on a variety of *M. musculus* embryonic stem cell lines; R1, J1 and V6.5 originally developed by the Jaenisch Laboratory at the Massachusetts Institute of Technology. These cell lines are commonly used for culturing ES derived tissues; they are biologically equivalent but genetically distinct as they are derived from different mouse strains. Though they can have differing expression levels for a given tissue or developmental stage it is expected that patterns of gene expression will be substantially similar for similar tissues (Sene et al., 2007).

2.3.3.1 Originating tissues

Blastocyst (BCYST) is the originating tissue, an early developmental structure of mammals which consists of an inner cell mass and an outer layer called the trophoblast which will form the placenta (Gilbert, 2003). The inner cell mass develops to form the embryo.

ES cells (ES2d) are extracted from the inner cell mass of the blastocyst in early development. When these ES cells are cultured they form aggregations of cells called embryoid bodies which are an *in vitro* model for embryo development (Kurosawa, 2007). There are two embryoid body tissue samples in the analyzed data, one at two days development (ES2d) and one at five days development (EB5d). Later in development, these embryonic cells differentiate into mesodermal, ectodermal and endodermal tissues. StemBase contains no endodermal tissues and is limited to originating tissues, ectodermal and mesodermal samples.

2.3.3.2 Mesoderm

The middle cell layer of the early embryo is the mesoderm which differentiates into tissues such as muscle, bone marrow and connective tissue. The bone marrow (BM) cells used in this analysis are highly enriched in stem cells. Mast cells are part of the immune system and found in various tissues. The mast cells in this analysis (MAST) are derived from bone marrow. Sorted from bone marrow, the population of cells in sample (SBM) is highly enriched in haematopoeitic stem cells. Connective tissue is collagen-rich fibrous tissue; the sample used in this analysis was extracted from day 15 embryos (ECT). Fibroblasts are the most common cells of connective tissue in animals, the fibroblasts in the samples used in this analysis (EFB) were grown in cell culture.

2.3.3.3 Ectoderm

The ectoderm is the outer layer of the early embryo which differentiates to tissues including the nervous system and the epidermis. The dermal (EDT) adipose (EAT) and brain embryonic tissues (BEe14s) were derived by extracting stem cells from skin, fat and brain stratia from 14 day old embryos and growing the cells in tissue culture. The mammosphere tissue (MMS) consists of mammary glands extracted from 5-8 week old mice and grown in culture. Tissue from the ganglionic eminence (GEE13.5), a transient structure of the mammalian fetal brain (Ulfig, 2002), was extracted from 13.5 day old embryos and cultured into neurospheres.

The tissues described can be divided into two subdivisions of the ectoderm (Gilbert, 2003). Brain embryonic and ganglionic eminence tissue are part of the neuroectoderm.

Dermal, adipose and mammosphere tissues are part of the external ectoderm.

ID	Description	Sample ID	Genes
BCYST*	Blastocyst	164,165,166, 167,169,219,220	7
BEe14S	Brain embryonic [E14] stratia	198	41
BM	Bone Marrow	294,295	36
EAT	Embryonic adipose tissue	199	22
EB5d	Embryoid body (5 days)	127	100
EBESv6.5*	Embryoid bodies and ES cells	153,154,155	15
ECT	Embryonic connective tissue	180	48
EDT	Embronic dermis tissue	200	23
EFB	Embryonic fibroblasts	286	52
ES2d	Embryonic stem cells (2 days)	128	41
GEE13.5	Ganglionic eminence E13.5	271	235
MAST	Mast cells	236,237	107
MMS	Mammospheres	255	173
MYS*	Myospheres	205	9
SBM	Hematopoietic stem cells (sorted)	233	96

Table 2.2: Details of tissue samples used in analysis

2.3.4 Microarray Analysis

M. musculus expression data sets were selected from the StemBase database (Sandie et al., 2009) and grouped into sets of tissues of relatively homogeneous cell type. All samples used (as identified by the StemBase Stem Cell Genomics Project Identifier or SCGP ID) consist of 3 biological replicates and were hybridized to both the MOE430A and MOE430B Affymetrix GeneChips, collectively referred to as the MOE430 chipset. The MOE430A design is based on genes that were well annotated in 2002 when the chipset was designed, while the MOE430B chip design is based on less reliable EST cluster gene predictions (Affymetrix, 2003).

Affymetrix gene chips use probesets to detect mRNA transcripts. For each target transcript there are multiple probes on the chip designed to hybridize with a portion of that transcript called *Perfect Match* (PM) probes. Each PM probe is paired with a *Mismatch* (MM) probe identical to the PM probe except for a single nucleotide. The MM probes are intended to detect background and non-specific hybridization. A probeset is the set of all PM and MM probes associated with a given target transcript. The optical intensity of each probe is proportional to the degree of hybridization, and these values are measured and used to calculate an estimated gene expression level.

For samples in time series, data from a point prior to the onset of differentiation (12h in the context of these data) were used to avoid complicating the analysis with gene expression values from samples containing multiple cell types. Genetically altered and cancer cell samples were excluded from the analysis due to the possibility of abnormal gene expression levels. Data for a total of 15 tissue types were extracted consisting of 25 sample of 3 replicates having 2 chips each, for a total of 150. Table 2.2 lists StemBase experiment identifiers, sample identifiers and descriptions for the tissue samples used in the analysis.

Gene expression values and present/absent calls were generated using the Affymetrix MAS5 gene expression calculation method (Affymetrix, 2001). For each probeset, MAS5 generates both a p-value to determine with what reliability a target transcript has been detected and an expression level for that transcript. The detection p-value is calculated based on the intensity differences of the PM and MM probes, and the expression level is calculated based on the weighted mean of the PM intensities with the MM intensities used to correct the PM values for non-specific hybridization and background.

MAS5 normalization was applied to the selected chips with the default parameters of $\alpha_1 = 0.04$, $\alpha_2 = 0.06$, $\tau = 0.015$. The alpha terms indicate p-values for probeset detection of the target mRNA: *Present* if p-value $< \alpha_1$, *Marginal* if $\alpha_1 < p$-value $< \alpha_2$ and *Absent* if p-value $> \alpha_2$). τ is a constant used in the Wilcoxon Signed Rank Test of probe pairs.

For each gene, Affymetrix probeset to gene mappings were determined using Ensembl provided annotations which directly map the individual probe sequences to the genome. The MOE430 chipsets used for this analysis were designed before a completed *M. musculus* genome was available and the reference genome information has changed considerably from that used to the design the chips; as a result, probes often do not map uniquely to coding sequences even when designed to do so. As the Affymetrix provided probeset to gene mappings can be inconsistent (Perez-Iratxeta & Andradé, 2005), mapping based on Ensembl annotations was chosen as more reliable. In this analysis, a probeset is considered to represent a given gene if all 11 PM probes of the probeset map to that gene and no probes map to any other genes of the genome. This is a stringent criterion which limits the usable probesets, however it reduces noise in the per-probeset signal due to cross hybridization between genes, providing higher quality, more precise expression values (Harrison et al., 2007). For each tissue type, expression values for each probeset were calculated and averaged for all samples and replicates, and the Affymetrix MAS5 *Present* call percentage for all replicates was calculated.

Genes were identified as being expressed in a given tissue if a probeset associated with a gene was called as *Present* by MAS5 in at least two out of three replicates for a given tissue. In the case of multiple probesets mapping to a single gene, any

probeset fulfilling the above criteria is sufficient to identify the gene as expressed in that tissue. Both *Marginal* and *Absent* MAS5 calls are treated as *Absent* for the purposes of this analysis. Genes are identified as tissue-specific if they are expressed in that tissue but not in any other tissue analyzed.

For each set of tissue-specific genes, codon counts and RSCU were calculated. Genes with representative transcripts containing less than 100 total codons were excluded from the analysis as short genes will have limited numbers of codons and the resultant RSCU values will be highly variable due to low codon counts. Tissues with 20 or fewer tissue-specific genes (BCYST, EBESv6.5, MYS) were excluded from this analysis as little selection pressure for biased codon usage in a tissue with so few uniquely expressed genes. As a result of this filtering 12 tissues remain containing a total of 974 tissue-specific genes.

Tissue-specific genes were grouped by developmental lineage into the main developmental lineage groupings (see Figure 2.1 for tissue / lineage relationships) of mesoderm (MD), external ectoderm (EED) and neuroectoderm (NED) to determine whether there are significant differences in lineage-specific gene expression.

Based on the gene expression levels for the tissue-specific genes, the data set used in the ANOVA analyses is divided into quartiles, those in the top 25% considered to be high expression and those in the bottom 25% considered to be low expression. This provides a data set which may be used to determine whether observed effects are due to selection. We expect high expression genes to be subject to higher levels of selection than low expression genes; they should therefore exhibit higher codon usage bias.

Multivariate statistical tests were performed to test the null hypothesis of no

difference in codon usage among the tissues and developmental groups of interest from the various data sets described above. The sum of RSCU values for any n codon amino acid will always add up to n; thus any given codon RSCU value may be derived from the $n-1$ others. To allow multivariate test such as MANOVA, variables must be independent so one codon was removed from each amino acid (NNT in NNY and NNN codon families and NNG in NNR families, TTG, AGG and TCG from leucine, arginine and serine respectively). In addition ATG and TGG were removed; as these are the only codons for methionine and tryptophan respectively, their RSCU value is unchanging. The resulting matrix of 41 codons with their RSCU values as variables and tissues or developmental groups as class variables are subject to a multivariate analysis of variance (MANOVA) using the R `manova` function (R Development Core Team, 2010). For this MANOVA analysis, tissue type is the independent variable and the the set of 41 codons above are the dependent variables.

MANOVA analyses identify whether independent variables have significant influence on dependent variables; in this case, tissues and codon RSCU values respectively. Three multivariate test statistics are be used to determine if there are significant differences in linear combinations of codons RSCU values between the different tissue types. Wilks' Lambda, Pillai's trace and the Lawley-Hotelling trace test statistics are applied to determine whether there are differences in average values and correlations among dependent variables between levels of independent variables. In the majority of cases these test statistics give similar results; Wilks' lambda is the most commonly used, while Pillai's trace is the most robust (Handy, 2004).

To identify which tissues have codon usage different from others and which codons exhibit tissue-specific or lineage specific codon usage differences we will use one way

ANOVA with Honest Significant Difference (HSD) as implemented in the R `aov` and `Tukey-HSD` functions respectively (R Development Core Team, 2010). HSD is used to control for Type 1 experiment-wise error rate, which we set at 0.05 for multiple comparisons in our analysis. ANOVA does not have the restrictions of MANOVA as to correlated variables; each codon is tested individually. We therefore perform the ANOVA analysis on a data set with only ATG, TGG and stop codons removed, with RSCU values as variables and 12 tissues as class variables. For this ANOVA analysis, tissue type is the independent variable and the codon RSCU value for the tissue-specific genes is the dependent variable, with the test applied sequentially to each codon.

2.4 Results and Discussion

2.4.1 Tissue-specific GC3, intronic GC and background GC

GC3, intronic GC and background GC are properties of genes which can give some insight into mutational or selection effects on those genes. Previous analyses of *H. sapiens* tissues have shown there can be significant differences in GC3 between tissue-specific gene sets (Semon et al., 2006; Vinogradov, 2003), though this phenomenon is unexplained. The observation of significant differences in GC3 between tissue-specific gene sets in the absence of significant differences in background and intronic GC is suggestive of translational selection acting on the coding sequences.

In the *M. musculus* data of this analysis, tissue-specific GC3 differs between tissues, with a broad distribution and large overlaps in values for all tissues (Table 2.3,

tissue	Mean GC3	Mean background GC	Mean intronic GC
EFB	0.562	0.448	0.435
ES2d	0.587	0.450	0.313
BM	0.605	0.460	0.470
SBM	0.611	0.458	0.427
EB5d	0.618	0.455	0.441
EAT	0.619	0.459	0.466
ECT	0.622	0.464	0.432
MMS	0.622	0.462	0.418
MAST	0.623	0.456	0.376
GEE13.5	0.624	0.461	0.406
BEe14S	0.638	0.448	0.451
EDT	0.650	0.455	0.261

Table 2.3: Mean values of GC3, background and intronic GC for *M. musculus* tissue-specific genes

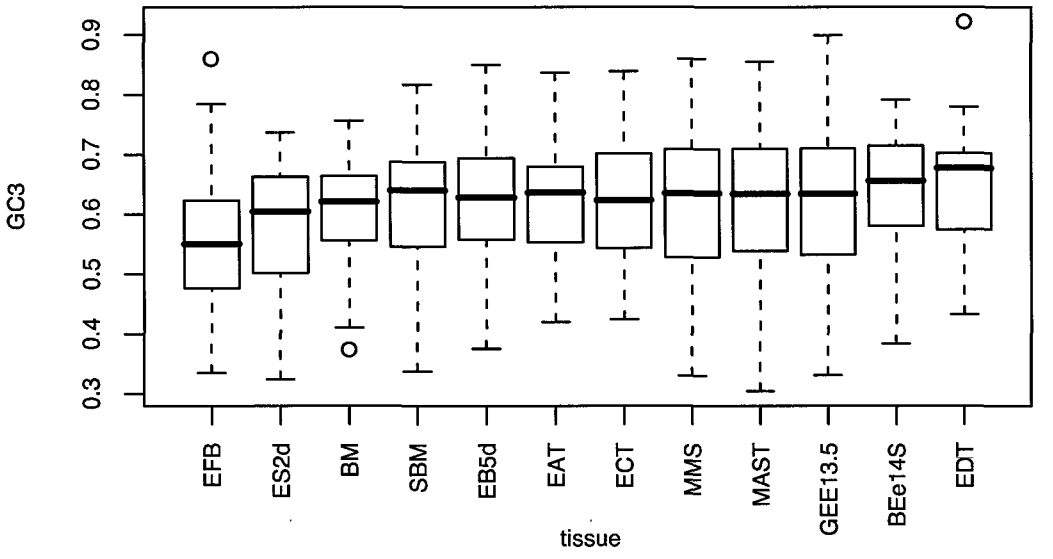

Figure 2.2: Boxplot of tissue-specific GC3. The tissues are ordered by increasing mean GC3, the box is the 1st to 3rd quartile, the line is the median, the whiskers are 1.5 times the inter-quartile range, and the circles are any outlier data points which lie beyond.

Figure 2.2). The means for tissues are significantly different from each other; ANOVA of the tissue GC3 values shows that we can reject at the 5% level the null hypothesis that the GC3 means for all tissues are equal with a p-value of 0.0352.

There are two pairs of tissues with significant differences between gene GC3 values (Table 2.4): MMS – EFB and GGE13.5 – EFB. EFB is the tissue with the lowest mean GC3 value, while MMS and GEE13.5 both have relatively high mean GC3 values (Table 2.3).

Tissues[1]	Diff [2]	Lower[3]	Upper[4]	p_{adj}[5]
MMS-EFB	0.06	0.00	0.12	0.03
GEE13.5-EFB	0.06	0.01	0.12	0.01

Table 2.4: Tissues with significant GC3 differences ($p < 0.05$)
[1]The tissues being compared, the first one having the highest GC3 value
[2](Tissue 1) - (Tissue 2) GC3 values
[3],[4]Lower and upper bounds for the 95% confidence interval
[5]HSD adjusted p-value
These descriptions apply to all subsequent ANOVA tables but are omitted for brevity.

There is a relationship between tissue-specific gene GC3, intronic GC and background GC (see Figure 2.3, Table 2.5). A linear fit of background GC to GC3 has a slope of 0.233 with $R^2 = 0.388$ while intronic GC (excluding the 31 tissue-specific genes without an intron) has a slope of 0.399 with $R^2 = 0.537$. Not all genes have introns and the linear fit for intronic GC to GC3 was determined using the subset of genes which do have them (see the ANOVA degrees of freedom column in Table 2.5).

Despite the correlation between GC3, intronic and background GC and the fact that some tissues have significant GC3 differences, intronic and background GC do not show significant differences between tissues; an ANOVA test of background and

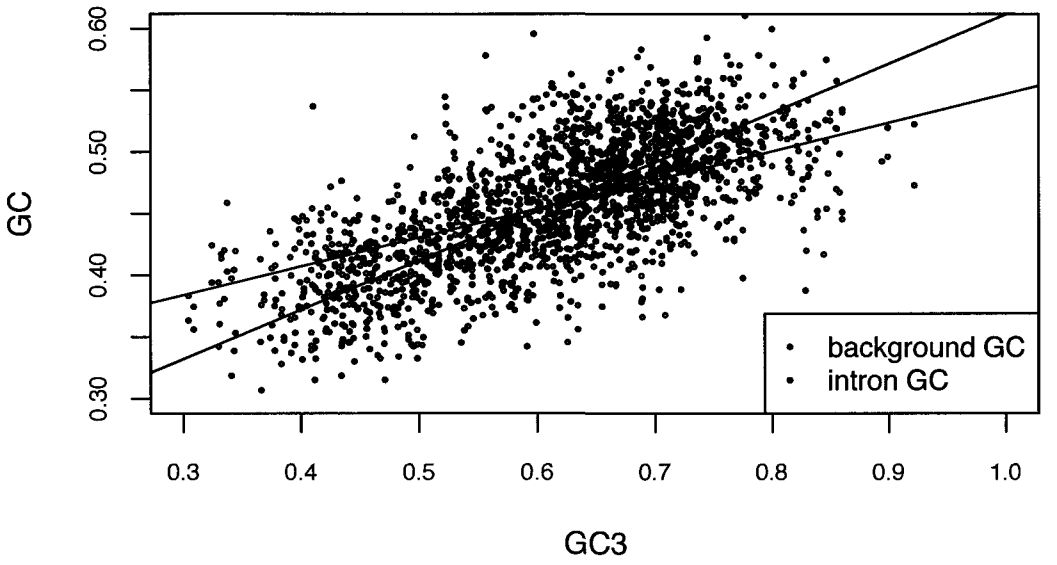

Figure 2.3: Relationship of intron and background GC to GC3 for tissue-specific genes. Lines indicate linear fits to each data set.

	Linear model slope [1]	R^2 [2]	ANOVA $p-value^{(3)}$	$F^{(4)}$	$dF^{(5)}$
intronic GC	0.399	0.537	$< 2.2\times10^{-16}$	1091.5	941
background GC	0.233	0.388	$< 2.2\times10^{-16}$	616.94	972

Table 2.5: Correlation of intronic and background GC with GC3
[1]slope of the linear model
[2]square of the linear model correlation coefficient
[3]ANOVA $p-value$
[4]ANOVA F ratio
[4]ANOVA degrees of freedom

intronic GC show no significant differences between tissue distributions (p-values of 0.487 and 0.417 respectively).

This result differs from prior analysis of *H. sapiens* which found significant dif-

ferences in intronic GC content between tissue specific genes (Semon et al., 2006). This difference may be due to the narrower isochoric GC range of the *M. musculus* genome and the proportionately narrower range of GC3 and associated intronic and background GC (Romiguier et al., 2010), or differences in the tissues analyzed; Semon et al. (2006) used mostly mature tissues, with the exception of placenta.

In this analysis, GC3, a property of coding sequences, shows significant variation between some tissues. Intronic GC and background GC, properties of non-coding sequences do not show significant variation. These observations are suggestive of translational selection effects. This raises the question of whether there are significant differences in codon usage between tissues. If the codon usage differences correspond to the GC3 differences between tissues, this is supportive of the hypothesis that tissues specific codon usage is caused by isochoric bias.

2.4.2 Tests of tissue-specific codon bias

There are significant differences in tissue-specific codon usage for some tissues and codons. We can reject the null hypothesis of no difference in codon usage among tissues with the MANOVA test giving a significant result ($p < 0.05$) for all three test statistics (Table 2.6).

	Value	F	$DF_{numerator}$	$DF_{denominator}$	p-value
Wilks' Lambda	0.5584	1.23	451	9845	0.00084
Pillai's Trace	0.5626	1.23	451	10252	0.00097
Hotelling-Lawley Trace	0.6039	1.23	451	10122	0.00072

Table 2.6: MANOVA tests of null hypothesis of no difference in codon usage among tissues for all tissue-specific genes

The ANOVA with HSD test identifies those codons with significant usage differences between specific pairs of tissues (Table 2.7).

Tissues	$GC3_{T1}$[1]	$GC3_{T2}$[2]	AA[3]	Codon	Diff	Lower	Upper	p_{adj}
BEe14S-EFB	0.638	0.562	Gln	CAG	0.237	0.021	0.452	0.017
EB5d-EFB	0.618	0.562	Gln	CAG	0.186	0.010	0.362	0.028
EFB-BEe14S	0.562	0.638	Gln	CAA	0.237	0.021	0.452	0.017
EFB-BEe14S	0.562	0.638	Pro	CCT	0.360	0.031	0.689	0.018
EFB-EB5d	0.562	0.618	Gln	CAA	0.186	0.010	0.362	0.028
EFB-EB5d	0.562	0.618	Ala	GCA	0.244	0.003	0.485	0.044
EFB-EB5d	0.562	0.618	Gly	GGA	0.291	0.007	0.576	0.039
EFB-ECT	0.562	0.622	Ala	GCA	0.284	0.002	0.566	0.046
EFB-GEE13.5	0.562	0.624	Lys	AAA	0.175	0.004	0.345	0.039
EFB-GEE13.5	0.562	0.624	Gln	CAA	0.205	0.047	0.363	0.001
EFB-GEE13.5	0.562	0.624	Ala	GCA	0.219	0.004	0.435	0.042
EFB-GEE13.5	0.562	0.624	Gly	GGA	0.282	0.027	0.537	0.016
EFB-MAST	0.562	0.623	Ala	GCA	0.263	0.025	0.501	0.016
EFB-MAST	0.562	0.623	Gly	GGA	0.294	0.013	0.575	0.031
EFB-MMS	0.562	0.622	Lys	AAA	0.176	0.001	0.352	0.048
EFB-MMS	0.562	0.622	Ala	GCA	0.312	0.089	0.534	0.000
EFB-MMS	0.562	0.622	Gly	GGA	0.290	0.027	0.553	0.016
ES2d-GEE13.5	0.587	0.624	Arg	CGT	0.245	0.020	0.471	0.020
ES2d-MMS	0.587	0.622	Asn	AAT	0.212	0.008	0.416	0.033
ES2d-MMS	0.587	0.622	Arg	CGT	0.251	0.019	0.483	0.021
ES2d-MMS	0.587	0.622	Ala	GCA	0.293	0.048	0.537	0.005
ES2d-SBM	0.587	0.611	Arg	CGT	0.273	0.024	0.522	0.018
GEE13.5-EFB	0.624	0.562	Lys	AAG	0.175	0.004	0.345	0.039
GEE13.5-EFB	0.624	0.562	Gln	CAG	0.205	0.047	0.363	0.001
GEE13.5-EFB	0.624	0.562	Pro	CCC	0.284	0.017	0.551	0.026
GEE13.5-EFB	0.624	0.562	Val	GTG	0.320	0.018	0.622	0.027
MAST-BEe14S	0.623	0.638	Pro	CCT	0.297	0.008	0.586	0.037
MAST-EFB	0.623	0.562	Ala	GCC	0.281	0.002	0.561	0.047
MAST-EFB	0.623	0.562	Gly	GGG	0.218	0.000	0.436	0.049
MMS-EFB	0.622	0.562	Lys	AAG	0.176	0.001	0.352	0.048

Continued on next page

Continued from previous page

Tissues	$GC3_{T1}$[1]	$GC3_{T2}$[2]	AA[3]	Codon	Diff	Lower	Upper	p_{adj}
MMS-EFB	0.622	0.562	Ala	GCC	0.304	0.042	0.565	0.008
MMS-EFB	0.622	0.562	Gly	GGG	0.227	0.023	0.431	0.015
MMS-ES2d	0.622	0.587	Asn	AAC	0.212	0.008	0.416	0.033
MMS-ES2d	0.622	0.587	Ala	GCC	0.312	0.024	0.599	0.020
SBM-BEe14S	0.611	0.638	Pro	CCT	0.335	0.041	0.628	0.011

Table 2.7: Codons whose RSCU values differ significantly between tissues ($p < 0.05$).
[1]Average GC3 of the first listed tissue
[2]Average GC3 of the second listed tissue
[3]Amino acid of the codon
These annotations apply to subsequent ANOVA tables but are omitted for brevity.

There are seven codon families (Ala, Arg, Asn, Gln, Gly, Lys, Pro, Val) corresponding to 14 distinct codons (AAA, AAC, AAG, AAT, CAA, CAG, CCC, CCT, CGT, GCA, GCC, GGA, GGG, GTG) with significant difference between 9 different tissues (BEe15S, EB5d, ECT, EFB, ES2d, GEE13.5, MAST, MMS, SBM).

The two pairs of tissue-specific genes with significant differences between GC3 values identified in section 2.4.1 (MMS - EFB and GGE13.5 - EFB) account for 14 of the 35 tissue-pair/gene found sets in Table 2.7. The fact that tissue pairs found to have significant differences in GC3 are represented so disproportionately in the significantly different codon sets is suggestive that the codon differences may be due to GC-bias.

If the codons which show significant differences in usage between tissues (Table 2.7) are considered in conjunction with the average GC3 of the corresponding tissues (Figure 2.2, Table 2.3) a clear pattern emerges. In a tissue pair, where a G or C ending codon is found to be significantly more highly used in one tissue, that tissue also has a higher average GC3. The opposite is also true, where an A or T ending

codon is found to be significantly higher in one tissue, that tissue has a lower average GC3.

This result is supportive of the mutational hypothesis in that codons used preferentially in tissue-specific genes correspond to the GC mutational bias to which those genes are subject. It is also consistent with the analysis of *H. sapiens* tissue specific genes by Semon et al. (2006). This is notable as *M. musculus* isochoric GC-bias is much weaker than that of *H. sapiens*.

These results, though supportive of the mutational hypothesis, are not incompatible with the selection hypothesis if the tissue-specific tRNA pools to which the genes are adapted correspond to the GC mutational bias as discussed in the introduction. For example EFB and ES2d tissues may have abundant tRNA that efficiently translate G and C-ending codons, whereas EDT, BEe14s and MAST tissues have abundant tRNA that efficiently translate A and T-ending codons.

It is possible to test for selection by incorporating gene expression information into the analysis. If the observed bias is the result of selection for translational efficiency based on tissue-specific tRNA pools, we may expect the tissue-specific differences to be stronger for highly expressed genes as they are more subject to selection and weaker or non-existent for low expression genes.

To test for selection, we divided the 974 tissue-specific genes into quartiles based on expression levels, assigning the highest 25% to the *high expression* category and the lowest 25% to the *low expression* category. MANOVA analyses were performed separately on the RSCU values of both the high expression and low expression gene sets to test the null hypothesis of no difference in codon usage among tissues.

Only the tests of the high expression genes allow us to reject the null hypothesis

Statistical test	Value	F	$DF_{numerator}$	$DF_{denominator}$	p-value
Wilks' Lambda	0.0831	1.21	451	2079	0.00330
Pillai's Trace	2.1390	1.19	451	2222	0.00745
Hotelling-Lawley Trace	2.9446	1.24	451	2092	0.00121

Table 2.8: MANOVA tests of null hypothesis of no difference in codon usage among tissues for high expression genes

	Value	F	$DF_{numerator}$	$DF_{denominator}$	p-value
Wilks' Lambda	0.1026	1.10	451	2079	0.09232
Pillai's Trace	2.0008	1.10	451	2222	0.10100
Hotelling-Lawley Trace	2.6206	1.11	451	2092	0.08216

Table 2.9: MANOVA tests of null hypothesis of no difference in codon usage among tissues for low expression genes

of no differences in codon usage among tissues (Table 2.8), while the low expression genes show no significant differences in codon usage among tissues (Table 2.9) at the 5% level. This result is supportive of the selection hypothesis as highly expressed genes will be subject to selection pressures to increase translation rates which may lead to codon biases.

ANOVA with HSD test identifies the codons which are found to be differentially used between specific pairs of tissues for high expression genes (Table 2.10). Note that the high expression gene set has different per-tissue GC3 from the full tissue specific gene set as they are based on different sets of genes. This is also the case with the low expression genes set which we will see in later analysis.

These results are similar to those shown in the previous section for all tissue specific genes. where the significantly different codons match the mean GC3s of the high-expression tissue-specific gene subsets.

Tissue	$GC3_{T1}$	$GC3_{T2}$	AA	Codon	Diff	Lower	Upper	p_{adj}
BM-EFB	0.637	0.574	Lys	AAG	0.580	0.074	1.086	0.010
BM-EFB	0.637	0.574	Val	GTG	1.020	0.053	1.988	0.029
ECT-EFB	0.712	0.574	Lys	AAG	0.729	0.165	1.293	0.002
ECT-EFB	0.712	0.574	Val	GTG	1.106	0.028	2.184	0.039
ECT-ES2d	0.712	0.596	Lys	AAG	0.572	0.061	1.082	0.014
ECT-ES2d	0.712	0.596	Ser	TCG	0.525	0.006	1.044	0.045
EDT-BEe14S	0.715	0.647	Leu	CTG	1.670	0.226	3.114	0.009
EDT-EB5d	0.715	0.629	Leu	CTG	1.409	0.203	2.616	0.008
EDT-EFB	0.715	0.574	Lys	AAG	0.763	0.175	1.351	0.002
EDT-EFB	0.715	0.574	Pro	CCG	0.702	0.004	1.400	0.047
EDT-EFB	0.715	0.574	Leu	CTG	1.695	0.251	3.139	0.008
EDT-ES2d	0.715	0.596	Lys	AAG	0.606	0.068	1.143	0.013
EDT-ES2d	0.715	0.596	Leu	CTG	1.479	0.160	2.799	0.014
EDT-GEE13.5	0.715	0.620	Leu	CTG	1.492	0.342	2.642	0.002
EDT-MMS	0.715	0.648	Leu	CTG	1.272	0.085	2.458	0.024
EDT-SBM	0.715	0.628	Leu	CTG	1.482	0.314	2.651	0.002
EFB-BM	0.574	0.637	Lys	AAA	0.580	0.074	1.086	0.010
EFB-ECT	0.574	0.712	Lys	AAA	0.729	0.165	1.293	0.002
EFB-ECT	0.574	0.712	Leu	CTT	0.863	0.166	1.560	0.003
EFB-EDT	0.574	0.715	Lys	AAA	0.763	0.175	1.351	0.002
EFB-GEE13.5	0.574	0.620	Lys	AAA	0.416	0.004	0.829	0.046
EFB-MAST	0.574	0.668	Lys	AAA	0.443	0.013	0.872	0.037
EFB-SBM	0.574	0.628	Lys	AAA	0.449	0.028	0.870	0.025
ES2d-ECT	0.596	0.712	Lys	AAA	0.572	0.061	1.082	0.014
ES2d-EDT	0.596	0.715	Lys	AAA	0.606	0.068	1.143	0.013
ES2d-MMS	0.596	0.648	Asn	AAT	0.384	0.002	0.765	0.047
GEE13.5-EFB	0.620	0.574	Lys	AAG	0.416	0.004	0.829	0.046
MAST-EFB	0.668	0.574	Lys	AAG	0.443	0.013	0.872	0.037
MAST-EFB	0.668	0.574	Val	GTG	0.931	0.110	1.751	0.012
MMS-EFB	0.648	0.574	Val	GTG	1.017	0.196	1.837	0.003
MMS-ES2d	0.648	0.596	Asn	AAC	0.384	0.002	0.765	0.047
SBM-ECT	0.628	0.712	Leu	CTT	0.568	0.017	1.118	0.037
SBM-EFB	0.628	0.574	Lys	AAG	0.449	0.028	0.870	0.025
SBM-EFB	0.628	0.574	Val	GTG	0.840	0.035	1.645	0.032

Table 2.10: Codons whose RSCU values differ significantly between tissues for high expression genes ($p < 0.05$).

Significant differences in tissue-specific codon usage present only in the high expression gene subset are suggestive of selection effects. The fact that in every case, significant differences in codon usage also match the relative tissue GC3 biases is a novel result. This effect may be due to co-adaptation of tissue GC-bias and tissue-specific tRNA pools as previously described with the effect being most apparent in high expression genes.

Though this analysis was only done on *M. musculus* embryonic tissues, the tissues show very similar properties to other mammalian tissues analyzed, for example *H. sapiens* adult tissue (Semon et al., 2006). It is possible that the correspondence between tissue-specific tRNA and tissue-specific GC3 is a general property of mammalian genomes. Co-adaptation of tissue specific GC3 and tRNA abundance is a potential explanation for statistically significant differences in mammalian tissue-specific gene GC3 noted in this study and previously (Semon et al., 2006; Vinogradov, 2003).

tRNA abundance data would be helpful in testing whether differences in codon usage correspond to tRNA abundance differences between tissues. Currently however, very little such data is available and none for the tissues used in this analysis. New sequencing technologies have shown promise in directly determining tRNA abundance (Sittka et al., 2008), however the diversity of post-translational modifications possible for tRNA (Iwata-Reuyl, 2008), and the high sequence similarity of many tRNA gene copies (Goodenbour & Pan, 2006) creates formidable technical challenges to accurately measuring tRNA abundance using these technologies.

2.4.3 Tests of lineage specific codon bias

The tissue-specific samples we are analyzing are related by hierarchic developmental relationships; as such the tRNA pools in groups related by lineage may be similar and lead to lineage specific codon usage bias. To create lineage based tissue-specific gene sets, we grouped the tissue-specific genes into external ectoderm (EED), mesoderm (MD) and neuroectoderm (NED) (see Figure 2.1 for lineage groupings).

There are significant differences in tissue-specific codon usage for some lineages and codons. We can reject the null hypothesis of no difference in codon usage among tissues with the MANOVA test giving a significant result ($p < 0.05$) for all three test statistics (Table 2.11).

	Value	F	$DF_{numerator}$	$DF_{denominator}$	p-value
Wilks' Lambda	0.8538	1.59	82	1580	0.00087
Pillai's Trace	0.1515	1.58	82	1582	0.00092
Hotelling-Lawley Trace	0.1651	1.59	82	1578	0.00082

Table 2.11: MANOVA tests of null hypothesis of no difference in codon usage of tissue-specific genes grouped by lineage

ANOVA with HSD identifies the codons found to be differentially used between specific pairs of lineages for high expression (Table 2.12) genes.

As with the previous analysis of tissue-specific genes, all lineage specific genes that show significant differences between lineage match the relative mean GC3 of the lineages. The average lineage mean GC3 values have rank: mesoderm < external ectoderm < neuroectoderm. These results are therefore consistent with the mutational hypothesis that they are due to GC-biased mutational effects.

Lineage	$GC3_{L1}$	$GC3_{L2}$	AA	Codon	Diff	Lower	Upper	p_{adj}
EED-MD	0.625	0.608	Asn	AAC	0.084	0.010	0.158	0.021
EED-MD	0.625	0.608	His	CAC	0.085	0.007	0.163	0.028
MD-EED	0.608	0.625	Asn	AAT	0.084	0.010	0.158	0.021
MD-EED	0.608	0.625	His	CAT	0.085	0.007	0.163	0.028
MD-EED	0.608	0.625	Ala	GCA	0.091	0.002	0.179	0.044
MD-NED	0.608	0.626	Arg	AGA	0.243	0.064	0.421	0.004
MD-NED	0.608	0.626	Ser	AGT	0.140	0.048	0.231	0.001
MD-NED	0.608	0.626	Ile	ATA	0.078	0.008	0.148	0.025
MD-NED	0.608	0.626	Gln	CAA	0.080	0.019	0.140	0.006
MD-NED	0.608	0.626	Pro	CCT	0.114	0.023	0.205	0.010
MD-NED	0.608	0.626	Val	GTA	0.091	0.023	0.159	0.005
NED-MD	0.626	0.608	Thr	ACG	0.103	0.026	0.180	0.005
NED-MD	0.626	0.608	Gln	CAG	0.080	0.019	0.140	0.006
NED-MD	0.626	0.608	Pro	CCG	0.082	0.003	0.161	0.040
NED-MD	0.626	0.608	Arg	CGC	0.175	0.021	0.330	0.022
NED-MD	0.626	0.608	Ser	TCG	0.087	0.019	0.155	0.008

Table 2.12: Codons whose RSCU values differ significantly between lineages ($p < 0.05$)

To test the selection hypothesis for differences in codon usage in a manner similar to that done in the previous section, we split the lineage specific genes into high expression and low expression data sets (using the top and bottom 25% of genes by expression level respectively), separately testing for differences in codon usage in those groups. We can reject the null hypothesis of no difference in codon usage between lineages for both sets (Tables 2.13, 2.14).

	Value	F	$DF_{numerator}$	$DF_{denominator}$	p-value
Wilks' Lambda	0.5188	1.56	82	330	0.00345
Pillai's Trace	0.5580	1.57	82	332	0.00327
Hotelling-Lawley Trace	0.7794	1.56	82	328	0.00365

Table 2.13: MANOVA tests of null hypothesis of no difference in codon usage of tissue-specific genes grouped by lineage for low expression genes

	Value	F	$DF_{numerator}$	$DF_{denominator}$	p-value
Wilks' Lambda	0.5610	1.36	82	332	0.03322
Pillai's Trace	0.4996	1.36	82	334	0.03334
Hotelling-Lawley Trace	0.6746	1.36	82	330	0.03314

Table 2.14: MANOVA tests of null hypothesis of no difference in codon usage of tissue-specific genes grouped by lineage for high expression genes

ANOVA with HSD test identifies codons differentially used between specific pairs of lineages for high expression (Table 2.10) and low expression (Table 2.14) sets:

Lineage	$GC3_{L1}$	$GC3_{L2}$	AA	Codon	Diff	Lower	Upper	p_{adj}
EED-NED	0.653	0.623	Asn	AAC	0.213	0.045	0.381	0.009
EED-NED	0.653	0.623	Ala	GCC	0.311	0.068	0.553	0.008
NED-EED	0.623	0.653	Asn	AAT	0.213	0.045	0.381	0.009
NED-EED	0.623	0.653	Ala	GCT	0.233	0.033	0.433	0.018
NED-MD	0.623	0.640	Thr	ACG	0.169	0.002	0.335	0.046
NED-MD*	0.623	0.640	Leu	TTA	0.153	0.026	0.279	0.013

Table 2.15: Codons whose RSCU values differ significantly between lineages in high expression genes ($p < 0.05$)

In most cases, for both the high (Table 2.15) and low (Table 2.16) expression lineage specific sets the significantly different codons between tissues correspond to the relative tissue GC3 values. The high expression average lineage GC3 values have rank: neuroectoderm < mesoderm < external ectoderm, while the low expression lineage GC3 values have rank: mesoderm < external ectoderm < neuroectoderm. The two exceptions to this tendency (marked with * in Tables 2.15 and 2.16) are AGG in the low expression data set and TTA in the high expression data set.

AGG is particularly notable as the codon usage has a negative correlation with GC3 over some ranges of GC3, so it may be expected to show the observed relationship

Lineage	$GC3_{L1}$	$GC3_{L2}$	AA	Codon	Diff	Lower	Upper	p_{adj}
EED-MD	0.609	0.573	Arg	CGG	0.308	0.018	0.598	0.035
EED-MD	0.609	0.573	Ala	GCC	0.228	0.026	0.429	0.023
EED-MD	0.609	0.573	Gly	GGG	0.208	0.047	0.369	0.007
EED-NED	0.609	0.642	Pro	CCT	0.209	0.004	0.413	0.044
MD-NED	0.573	0.642	Asn	AAT	0.162	0.013	0.312	0.029
MD-NED	0.573	0.642	Thr	ACA	0.270	0.076	0.463	0.003
MD-NED*	0.573	0.642	Arg	AGG	0.275	0.016	0.535	0.035
MD-NED	0.573	0.642	Ser	AGT	0.267	0.070	0.465	0.005
MD-NED	0.573	0.642	Ile	ATA	0.175	0.042	0.308	0.006
MD-NED	0.573	0.642	Gln	CAA	0.187	0.053	0.320	0.003
MD-NED	0.573	0.642	Pro	CCA	0.247	0.035	0.459	0.018
MD-NED	0.573	0.642	Pro	CCT	0.290	0.088	0.493	0.002
MD-NED	0.573	0.642	Glu	GAA	0.156	0.011	0.302	0.032
MD-NED	0.573	0.642	Ala	GCA	0.194	0.013	0.374	0.032
MD-NED	0.573	0.642	Gly	GGA	0.304	0.114	0.495	0.001
MD-NED	0.573	0.642	Val	GTA	0.140	0.017	0.263	0.022
MD-NED	0.573	0.642	Ser	TCT	0.231	0.012	0.449	0.036
NED-MD	0.642	0.573	Asn	AAC	0.162	0.013	0.312	0.029
NED-MD	0.642	0.573	Thr	ACC	0.240	0.020	0.460	0.028
NED-MD	0.642	0.573	Ser	AGC	0.259	0.019	0.499	0.031
NED-MD	0.642	0.573	Gln	CAG	0.187	0.053	0.320	0.003
NED-MD	0.642	0.573	Pro	CCC	0.319	0.107	0.530	0.001
NED-MD	0.642	0.573	Pro	CCG	0.218	0.031	0.405	0.017
NED-MD	0.642	0.573	Arg	CGG	0.313	0.041	0.584	0.019
NED-MD	0.642	0.573	Glu	GAG	0.156	0.011	0.302	0.032
NED-MD	0.642	0.573	Ala	GCC	0.203	0.015	0.392	0.031
NED-MD	0.642	0.573	Gly	GGG	0.182	0.031	0.332	0.013
NED-MD	0.642	0.573	Val	GTG	0.279	0.050	0.509	0.013

Table 2.16: Codons whose RSCU values differ significantly between lineages in low expression genes ($p < 0.05$)

to lineage GC3; this effect is explored in Chapter 4.

This analysis does not support the hypothesis that there is selection on codons to match lineage specific tRNA pools as both high and low expression sets show significant differences in codon usage between lineages. Selection related bias is expected to be stronger for the high expression gene set, however we observe that the bias for

high expression genes is weaker. This may be due to the absence of lineage-specific tRNA pools, or insufficient selection on those pools to significantly alter codon usage; the tests done here cannot distinguish between these alternatives. Given that the tissues used are all embryonic, there may be insufficient developmental divergence in the tissue groupings to detect translational selection on tRNA abundance; a similar analysis of adult differentiated tissues may be more likely to show such an effect.

The hypothesis of tissue-specific translational selection is supported by the analysis of this chapter while the hypothesis of selection on lineage specific pools is not. This is suggestive that there are tissue-specific tRNA pools but not lineage specific ones. Given the close lineage relationship between the tissues of the lineage grouping used here, this suggests that tRNA pool composition may change rapidly in differentiating tissues, a hypothesis that could be tested experimentally.

The isochoric structure of *M. musculus* is weak compared to that of *H. sapiens* and there are a range of isochoric intensities across mammalian genomes (Romiguier et al., 2010). A direct comparison of analogous tissue between species may provide insight into the contribution of isochores to the observed codon bias. This is explored in detail in Chapter 4 where the contribution of GC mutational bias to codon bias is quantified for mouse and human genomes.

2.5 Conclusion

We have demonstrated that there are significant tissue-specific differences in codon usage between the *M. musculus* tissues. In all cases, the differences between tissue-specific gene sets correspond to the average GC3 of the tissue-specific gene sets. Those

tissues which show the lowest GC3 also tend to use A or T ending codons; conversely, the tissues which show the highest GC3 tend to use G or C ending codons. The fact that this effect is present in the highly expressed gene sets while being absent in the low expression gene set is evidence that it may be considered both a mutational and a selection effect. It is possible that tRNA pools in different tissues correspond to GC-bias of those tissues, and high expression genes are particularly adapted to those pools. Testing this hypothesis will require measurements of tissue-specific tRNA abundance.

Tissue-specific genes grouped by lineage were found to have significant differences in codon usage between lineages which correspond to the average GC3 of the gene sets used, supporting the mutational hypothesis. The high and low expression subsets were also found to have significant differences in codon usage between lineage groupings, most of which corresponded to average lineage GC3 values. As both high and low expression sets show significant differences in codon usage between tissues, this analysis does not support the selection hypothesis that there is selection on codons to match lineage specific tRNA specific pools.

This analysis methodology is a general framework which may be applied to the analysis of other tissues and species to investigate the relationship of translational selection and codon bias in a broader context. Application of this method to analogous adult tissues in mouse may identify whether the tRNA pool changes significantly between the embryonic and adult states. A cross species analysis may be able to determine whether divergent species show common tissue-specific codon bias as a result of evolutionarily conserved tissue-specific tRNA pools.

Chapter 3

Are Tissue-Specific Genes Clustered in the Same Isochores?

3.1 Abstract

Tissue-specific *M. musculus* genes show statistically significant differences in GC3 between some tissues as demonstrated in chapter 2. In this chapter we test the hypothesis that these observed differences in tissue-specific GC3 are the result of genes clustering on the same isochore leading to common mutational biases. The results of the analysis show that for most tissues, the tissue-specific genes are no more clustered than would be expected from a random selection of protein coding genes of the *M. musculus* genome.

Significant clustering is noted in embryonic dermal tissue (EDT) and the host lineage external ectoderm. In embryonic dermal tissue we note that of the three pairs of genes found to be closer than 2MB, two of the genes Krt6b and Krt5 are members

of a keratin gene cluster; in the associated lineage grouping of external ectoderm which contains embryonic dermal tissue, we find four members; Krt6, Krt6b, Krt73 and Tencl.

While we observe some isolated tissue-specific and lineage-specific gene clustering, it is insufficient to explain observed tissue and lineage specific GC3 differences. In addition, the tissues shown to exhibit significant differences in tissue-specific gene GC3 do not show significant levels of clustering.

The evidence does not support the hypothesis that issue-specific GC3 biases are caused by the clustering of such genes in the same isochore.

3.2 Introduction

In Chapter 2 we have seen that tissue-specific gene sets exhibit characteristic statistically significant differences in codon usage between some tissues. We found that these codon usage differences are strongly associated with the GC3 content for each tissue and therefore may be attributable to isochoric mutational biases.

As the observed biases are strongly associated with GC3, a hypothesis that may be tested is that the common GC-bias observed in some tissue-specific genes is attributable to the proximity of such genes to each other on the genome. If large numbers of tissue-specific genes are clustered together on the same isochore, and therefore subject to the same mutational GC-bias, they may have similar GC3 and affect the tissue-specific GC3. In the *M. musculus* genome, isochores are usually 0.30 to 1.40 megabases wide, depending on isochore type, with few isochores larger than 2 megabases in size (Costantini et al., 2009); genes further apart than 2 megabases

are therefore unlikely to be on the same isochore.

There has been prior analysis of tissue-specific gene clustering in mammalian tissues, specifically in the human genome. A study using SAGE data for 14 human tissues found that tissue-specific genes did not cluster, though statistically significant clustering of broadly expressed housekeeping genes was found (Lercher et al., 2002). Semon et al. (2006) claimed that there is no significant clustering of tissue-specific genes in the human genome based on EST data from 18 tissues, though they did not provide a description of their methods or data in the publication.

Clusters of closely spaced genes with associated functions exist in mammalian genomes; some of these may be expected to show tissue-specific or lineage specific expression; for example the keratin gene clusters (Hesse et al., 2004), hox gene clusters (Krumlauf, 1994), histone gene clusters (Doenecke et al., 1997), and the osteocalcin gene cluster (Desbois et al., 1994). Co-expression of these closely clustered genes in a tissue-specific manner may have an influence on the GC-bias of tissue-specific genes.

Some clusters of genes are known to show lineage specific expression. For example the *M. musculus* and *H. sapiens* globin gene clusters show developmentally sequential expression (Li et al., 2002). Due to the development stage specific sequence of expression in such gene cluster members, the genes may not appear clustered in individual tissues, however such clustering may be visible by grouping tissue-specific genes within lineages.

Despite prior analysis having been published on tissue-specific and tissue lineage specific clustering in human tissues, no such analysis appears to have been published for *M. musculus* tissues and none on *H. sapiens* or *M. musculus* embryonic tissues. In addition no analysis of the effect of lineage on gene clustering appears to have been

published.

In this chapter we test the hypothesis that tissue-specific genes and lineage-associated tissue-specific genes are more clustered than would be expected from a random selection of protein coding genes on the *M. musculus* genome and investigate those genes which do exhibit such clustering.

3.3 Materials and Methods

We begin with a set of tissue-specific genes (see Materials and Methods section of Chapter 2) and associated chromosomal location, consisting of 974 tissue-specific genes from 12 tissues of the *M. musculus* genome. For each set of genes we will test the hypothesis that the genes are closer to each other on the genome than would be expected by chance.

Genes are distributed unevenly within chromosomes with a high gene density in the GC-rich isochores and low density in the more common AT-rich "gene deserts" isochores (Bernardi, 2007). Due to the complexity of gene distribution, a non-parametric sampling approach will be used to determine whether tissue-specific and lineage grouped tissue-specific genes are more clustered than would be expected by chance.

First, we must determine the adjacent gene distances for every tissue-specific gene set. For each tissue-specific gene set, we measure the distances between every pair of adjacent tissue-specific genes. The distances are measured between the closest Ensembl annotated portion of each gene per chromosome. For each chromosome with n genes specific to a given tissue where $n > 1$, there will be $n - 1$ distances. The number of inter-gene distances is therefore dependent on the distribution of tissue-

specific genes across chromosomes. Chromosomes with zero or one tissue-specific genes can yield no distance measures. The observed distribution of tissue-specific gene properties is shown in Table 3.1 and visualized as histograms in Figure 3.1. Tissue-specific gene pairs closer than 2MB are listed in Supplemental Table A.6.

The fact that the tissue-specific genes could be located in multiple clusters on a single chromosome complicates the analysis. Consider a hypothetical chromosome with four genes located at 3, 4, 50 and 51 megabases respectively. This distribution has genes in two clusters and three adjacent distance measures; 1, 46 and 1 megabases, with an average distance of 16 megabases. The pairwise adjacent distances of four other hypothetical genes at 1, 17, 33, 49 megabases would give the same mean distance despite their lack of clustering; a simple statistical measure such as mean or median will not, therefore, capture the degree of gene clustering. To identify gene clustering, we will determine if the tissue-specific genes are potentially on the same isochore (nearer than 2 megabases) (Costantini et al., 2009) and whether there are more such close gene pairs than are expected by chance.

Our null hypothesis is that the proportion of tissue-specific genes closer than 2 megabases should be no greater than that for the same number of genes randomly selected from the *M. musculus* genome. To test this hypothesis we will compare our observations of tissue-specific genes to simulated distributions of genes based on the set of all known protein coding genes in the *M. musculus* genome.

For each tissue we select n genes from the set of known protein coding genes in the *M. musculus* genome, where n is the number of tissue-specific genes, and determine the number that are closer together than 2 megabases. This is repeated 10,000 times, and the resulting distribution of numbers of gene pairs closer than 2 megabases will be

compared with the observed number for each tissue. The proportion of the sampled distribution which is *higher* than the observed count gives a sense of how likely the degree of observed clustering is, and provides an estimate of the p-value for our observed set. Though we have chosen $p < 0.05$ as the significance threshold for this analysis, as these tests are being repeated for each of 12 tissues, we use the Bonferroni method to adjust the significance threshold to $p < 0.0042$.

In addition, we will further test the hypothesis that lineage specific genes are more likely to be on the same isochore than would be expected by chance. To do this, we will group the tissue-specific genes based on their lineage relationship (Figure 2.1) into mesoderm, external ectoderm and neuroectoderm, and perform a clustering analysis, as above. Lineage grouped tissue-specific genes-pairs closer than 2 megabases are listed in Supplemental Table A.7. As these tests are being repeated for each of 3 lineage groupings, we use the Bonferroni method to adjust the significance threshold from $p < 0.05$ to $p < 0.0167$.

3.4 Results and Discussion

3.4.1 Tissue-specific gene proximity analysis

For the tissue-specific gene sets, with increasing numbers of genes there are decreasing means and medians for adjacent gene distances and an increasing number of genes closer than 2 megabases (Table 3.1), as would be expected for a random selection of genes. The observed distribution of tissue-specific protein coding genes over chromosomes is shown in Table 3.2. The genes are not confined to tight clusters on each

chromosome as can by seen in Figure 3.1.

Tissue	Genes	Pairs	Median	Average	Close pairs
BEe14S	41	24	30617006	40307725	0
BM	36	20	14438986	25704825	3
EAT	22	10	33360944	34479726	1
EB5d	100	80	13530056	20965454	8
ECT	48	34	17759813	25964866	3
EDT	23	10	16423902	20719352	3
EFB	52	34	21209874	32011040	3
ES2d	41	25	13846753	24841325	4
GEE13.5	235	215	6022264	9801016	54
MAST	107	87	12007895	16372104	15
MMS	173	153	6771963	11033274	41
SBM	96	76	9806101	17595161	15

Table 3.1: Statistical properties of pairwise gene distances (in base pairs)for each tissue type: genes indicates the tissue-specific genes, pairs indicates the number of adjacent, pairwise distances, median and average are for the adjacent pairwise distances, and close pairs indicates the number of gene pairs which are <2 megabases apart.

Of the tissue-specific gene sets only EDT tissue exhibits more gene pairs closer than 2 megabases than would be expected by chance (p-value 0.0025) (Table 3.3, Figure 3.2).

If the 23 tissue-specific EDT genes were selected randomly from protein coding genes of the *M. musculus* genome, few of these would be expected to be closer than 2 megabases to each other. There are only 3 observed gene pairs closer than 2 megabases (the full list of all tissue-specific gene pairs closer than 2 megabases can be found in Supplemental Table A.6). Two of the pairs are on chromosome 15 which contains 3 genes clustered within 2 megabases of each other: Krt6b, Krt5 and Gpr84. Gpr84 is putative G-coupled receptor protein which is not well characterized (Maglott et al., 2006). Krt6b and Krt5 are keratin genes, part of the type II keratin gene cluster

Chr	Genome	BEe14S	BM	EAT	EB5d	ECT	EDT	EFB	ES2d	GEE13.5	MAST	MMS	SBM
1	1263	5	2	1	5	1	1	1	1	14	6	11	7
2	1932	1	0	0	5	6	2	6	3	24	10	15	5
3	1068	3	4	3	5	0	3	2	4	11	7	4	7
4	1406	1	2	1	7	8	0	3	2	11	8	12	5
5	1334	4	3	0	7	5	0	1	7	16	3	14	8
6	1186	3	2	3	10	1	1	5	1	14	9	10	3
7	2017	2	3	0	6	0	2	2	3	16	10	14	9
8	1118	4	1	2	9	0	1	2	2	11	1	8	7
9	1285	0	0	3	5	4	2	0	2	17	7	13	5
10	1040	3	1	0	4	3	3	4	3	8	6	4	3
11	1738	2	4	1	5	3	0	1	0	28	10	14	13
12	730	1	0	0	3	1	1	4	0	9	1	5	5
13	881	2	1	1	3	2	0	4	1	6	5	10	1
14	873	0	2	2	8	0	1	2	1	4	4	5	3
15	831	1	2	1	0	2	3	0	2	9	2	8	4
16	703	1	3	0	3	6	1	3	4	8	4	4	1
17	1093	2	2	2	3	2	2	1	3	6	5	9	3
18	527	0	1	0	2	0	0	3	0	8	1	2	1
19	757	1	3	2	6	0	0	2	2	8	5	6	3
X	960	5	0	0	3	4	0	6	0	7	3	5	3
Y	12	0	0	0	1	0	0	0	0	0	0	0	0
total	22754	41	36	22	100	48	23	52	41	235	107	173	96

Table 3.2: tissue-specific gene chromosomal locations: The distribution of tissue-specific protein coding genes for the 12 tissues with 20 or more such genes whose sequences are $\geq$ 100 amino acids. The genome column indicates the total number of protein coding genes on that chromosome for comparison. The totals row indicates the total for each tissue-specific set and the *M. musculus* genome as a whole.

consisting of 26 closely spaced keratin genes having strong synteny with the equivalent human cluster (Hesse et al., 2004). It is not surprising that embryonic dermal tissue (EDT) expresses keratin genes, known to be highly expressed in dermal tissue (Roop et al., 1983). The clustering of keratin genes and likely tissue and lineage related expression of keratin genes explains the high number of closely spaced genes for this tissue type.

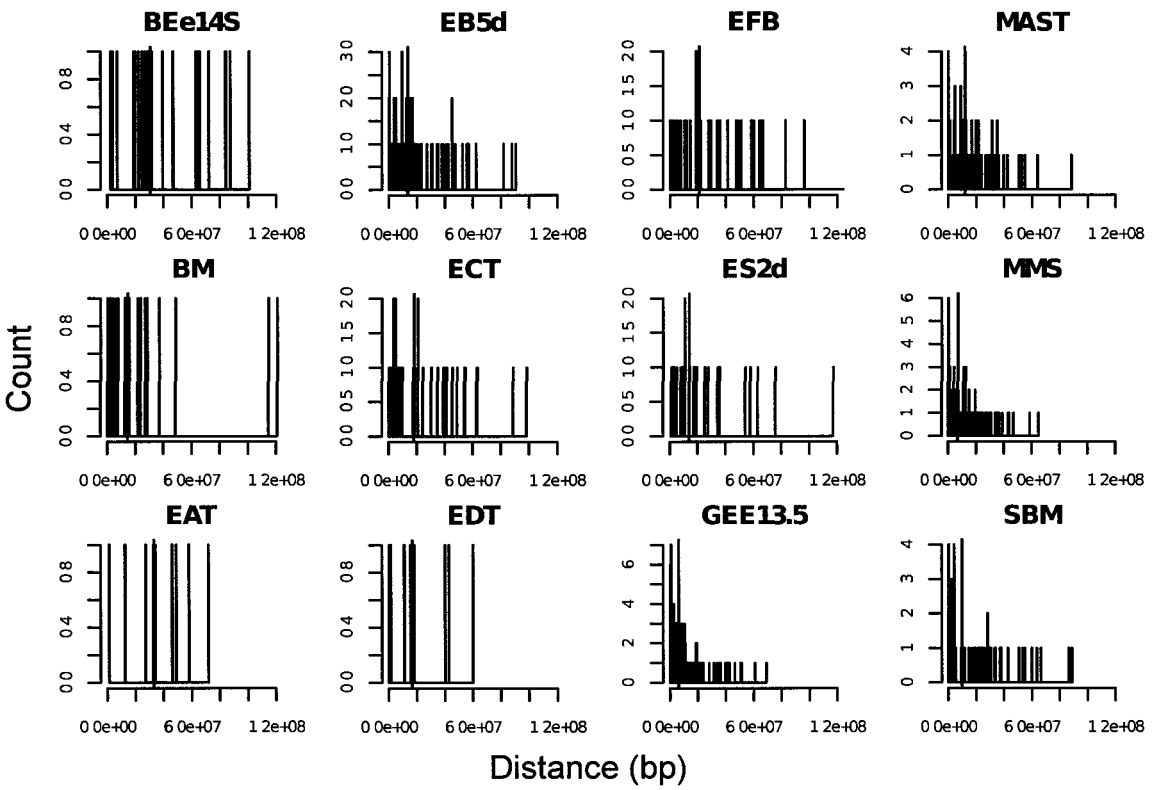

Figure 3.1: Histograms of adjacent tissue-specific gene distances, per tissue. X-axis indicates the distance between genes, Y-axis indicates the bin count. X-axis scales are identical, the red line indicates the median distance for each tissue.

It is notable that the tissues observed in chapter 2 to have significant differences in tissue-specific gene GC3 (GEE13.5, EFB, MMS) do not exhibit significant clustering. If genes clustered on the same isochores were a cause of tissue-specific GC3 differences these tissues would expected to show such clustering.

Within the set of all tissue-specific genes closer than 2 megabases, there are few clusters containing more than two genes. In general, we see very little clustering of genes on the same isochore. The absence of multiple genes clustering closely enough to inhabit the same isochore, and thus subject to the same isochoric GC-bias, does not support the hypothesis that clustering is the cause of the observed tissue-specific

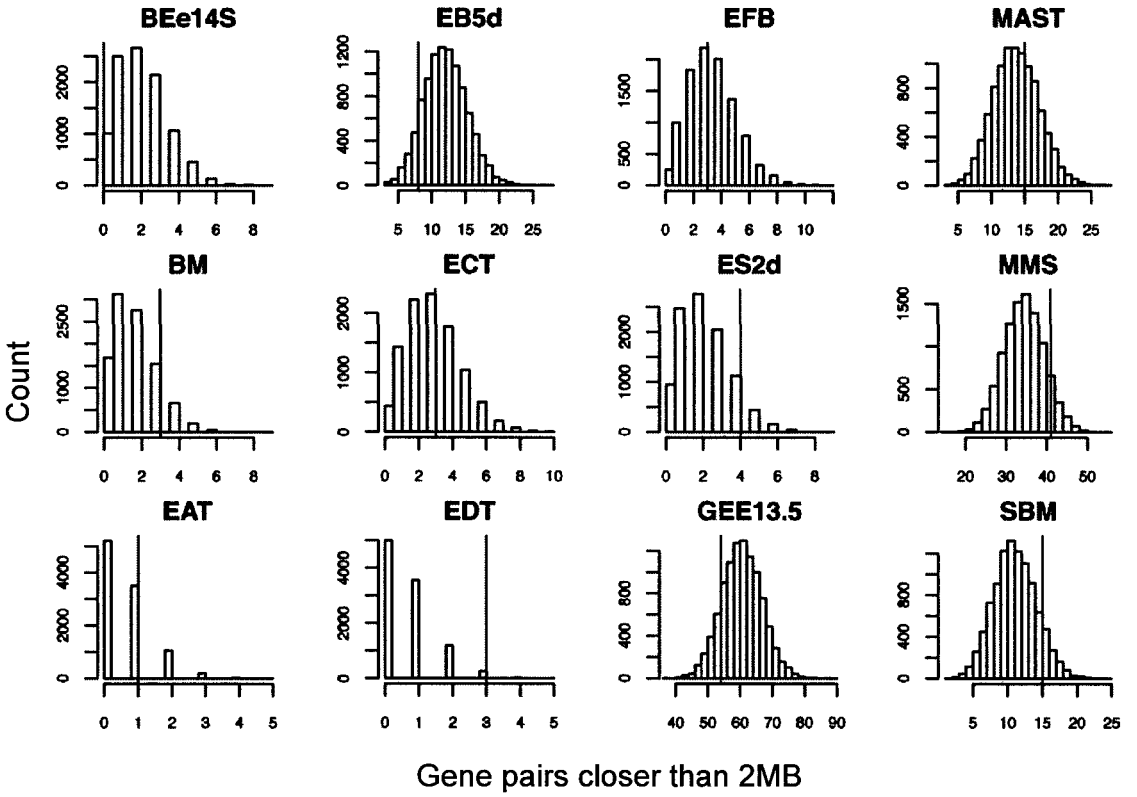

Figure 3.2: Simulated tissue-specific genes <2 megabases apart, resampled 10,000 times. Histograms of adjacent tissue-specific gene distances, per tissue. The X-axis indicates the number of gene pairs closer than 2 megabases for a given sampling, Y-axis indicates the bin count over all samples. The red line indicates the number of gene pairs closer than 2 megabases observed for each tissue. The proportion of bootstrap counts to the right of the observed number of gene pairs is an estimate of the p-value of the observed distribution as compared to a random selection of genes from the genome

GC-biases.

This result is consistent with prior analysis of *H. sapiens* adult tissues which found no significant clustering of tissue-specific genes (Semon et al., 2006).

Tissue	p-value
BEe14S	0.8990
BM	0.0895
EAT	0.1286
EB5d	0.9011
ECT	0.3588
EDT	0.0025
EFB	0.4726
ES2d	0.0649
GEE13.5	0.8566
MAST	0.3519
MMS	0.0905
SBM	0.1060

Table 3.3: p-value of median distance between tissue-specific genes as estimated by random sampling

3.4.2 Lineage specific gene proximity analysis

In the lineage grouped tissue-specific gene sets for increasing numbers of genes there are generally decreasing means and medians for adjacent gene distances and an increasing number of genes closer than 2 megabases (Table 3.4), as would be expected from a random selection of genes.

No lineage grouped tissue-specific gene set exhibits more gene pairs closer than 2 megabases than would be expected by chance (Table 3.5). The full list of all tissue-specific gene pairs closer than 2 megabases may be found in in Supplemental Table A.7.

Figure 3.3 shows the resampled distribution relative to the observed median and table 3.5 lists the estimated p-values for the tissue-specific genes grouped by lineage.

External ectoderm shows the lowest p-value of all the tested lineage groupings. In the 63 external ectoderm gene pairs closer than 2 megabases there are some gene

Tissue	Genes	Pairs	Median	Average	Close pairs
External Ectoderm	218	198	5347666	9606246	63
Mesoderm	339	319	3250547	6924525	120
Neuroectoderm	276	256	5063645	8379111	73

Table 3.4: Statistical properties of pairwise gene distances for major lineage groups: genes indicates the tissue-specific genes, pairs indicates the number of adjacent, pairwise distances, median and average are for the adjacent pairwise distances, and close pairs indicates the number of gene pairs which are <2 megabases apart.

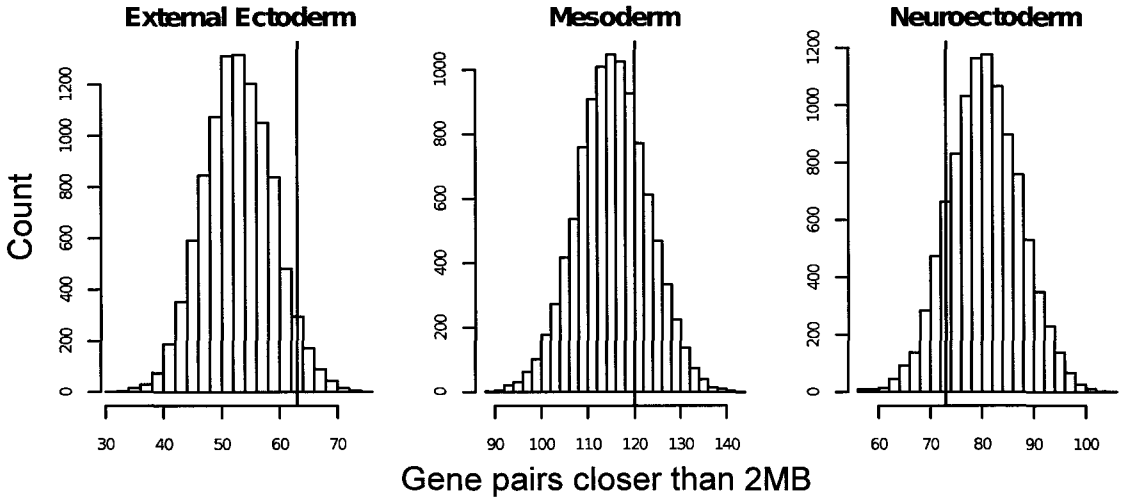

Figure 3.3: Simulated lineage specific genes <2 megabases apart, resampled 10,000 times. Histograms of adjacent tissue-specific gene distances, per tissue. The X-axis indicates the number of gene pairs closer than 2 megabases for a given sampling, Y-axis indicates the bin count over all samples, the red line indicates the number of gene pairs closer than 2 megabases observed for each lineage. The proportion of bootstrap counts to the right of the observed number of pairs is an estimate of the p-value of the observed distribution as compared to a random selection of genes from the genome

clusters represented. For example Hox gene cluster members Hoxb6 and Hoxb2 are within 2 megabases, and Krt6, Krt6b, Krt73 and Tencl are within 2 megabases; The Krt genes are members of the keratin gene cluster and Tencl is related in being associated with the Gene Ontology term "*collagen metabolic process [GO:0032963]*"

Tissue	p-value
External Ectoderm	0.0453
Mesoderm	0.2694
Neuroectoderm	0.8623

Table 3.5: p-value of median distance between lineage grouped tissue-specific genes as estimated by random sampling.

(Blake et al., 2006).

Within the set of all tissue-specific genes grouped by lineage, there are few clusters containing more than 2 genes within 2 megabases of each other, therefore there is little clustering of genes on the same isochore. The rarity of multiple genes clustering closely enough to inhabit the same isochore indicates that this is not the cause of lineage specific GC-biases.

Of the analyzed gene sets, only EDT (embryonic dermal tissue) and the parent external ectoderm lineage has more closely spaced genes than would be expected by chance. This appears to be caused in part by expression of genes in clusters known to have tissue and lineage specific expression. There are insufficient genes in these clusters to significantly affect GC3 of the tissue and lineage specific gene sets. If genes clustered on the same isochore were a major cause of tissue-specific GC3 differences in general, we would expect more than one tissue-specific gene set to show significant clustering.

This method has applications in the identification gene clusters in tissue and lineage specific gene sets. All genomes contain many genes with unknown function and few genes have been fully characterized. Identifying co-expressed gene clusters in tissue and lineage specific gene sets is a useful technique for finding novel functionally related gene. Prior work in this area has focused on co-expression (Zhang et al.,

2004); considering the proximity of co-expressed genes may lead to the discovery of novel functionally related gene clusters.

Though it has been shown in this chapter that tissue-specific genes are not more clustered along chromosomes than a random selection of genes, there are other measures of proximity which may be applied. Using Chromatin Conformation Capture (3C) techniques it has recently been established that mammalian chromosomes form a densely packed three dimensional structure in the nucleus that can place linearly distant portions of the genome in close proximity (Lieberman-Aiden et al., 2009). Open and closed chromatin were found to be partitioned into separate compartments in this structure. As open chromatin is more likely to be transcriptionally active, we would expect the tissue-specific genes to be in the open chromatin compartment. As they are potentially co-regulated they may also show spatial proximity in the three dimensional structure of the chromosomes despite large linear linear separations along chromosomes or even being on different chromosomes. The availability of 3C data for various cell types will allow this hypothesis to be tested.

These results for *M. musculus* of this chapter are consistent with prior analyses of adult *H. sapiens* tissue-specific genes (Lercher et al., 2002; Semon et al., 2006) which also showed no significant clustering. Given that *M. musculus* and *H. sapiens* have the weakest and strongest isochoric structure respectively, (Romiguier et al., 2010) it seems likely that this may be a general property of mammalian genomes.

3.5 Conclusions

The analysis in this chapter shows little indication of clustering either on tissue-specific level or when the tissue-specific genes are grouped by developmental hierarchy. The notable exception is embryonic dermal tissue (EDT) which has significantly more tissue-specific genes pairs closer together than 2 megabases than would be expected by chance. This result appears to have been influenced by the presence of genes of the keratin gene cluster. No significant clustering is seen in lineage grouped tissue specific genes. If genes clustered on the same isochore were a major cause of tissue-specific GC3 differences in general, we would expect more than one tissue-specific gene set to show significant clustering.

In conclusion we cannot reject our null hypotheses that tissue-specific genes and lineage specific genes are no more clustered than those randomly selected from genes of the genome, and there is insufficient isochore related gene clustering to cause significant differences in tissue-specific gene GC3.

Chapter 4

A General Model of Codon Bias Due to GC Mutational Bias

4.1 Abstract

Because most amino acids allow synonymous GC content-changing substitutions in the third codon position, the overall GC-bias of a genome or genomic region is highly correlated with GC3, a measure of third codon position GC content. However, in analyzing codon usage bias in *M. musculus* and *H. sapiens* genes as well as 897 prokaryotic and 197 plant genomes we find that two G-ending codons, AGG (arginine) and TTG (leucine), unlike all other G/C-ending codons, show overall usage that decreases with increasing GC-bias. Moreover, the usage of some codons appears non-linear, even non-monotone, as a function of GC-bias. To explain these observations, we propose a continuous-time Markov chain model of GC-biased synonymous substitution and apply the model to 2 and 4 codon families as well as isoleucine, arginine

and leucine. The resultant model predicts the decreasing usage of AGG and TTG with increasing GC-bias, the nonlinearity of arginine, leucine and isoleucine codon usage as function of GC-bias, and the fact that per-amino acid GC3 should linear for all amino acids with the exception of isoleucine. The models accounts for 72%, 64%, 52% and 36% of the observed variability of codon usage in prokaryotes, plants, *H. sapiens* and *M. musculus* respectively. When codons are grouped based on common GC content, 87%, 80%, 68% and 51% of the variation in usage is explained for prokaryotes, plants, *H. sapiens* and *M. musculus* respectively. This model establishes that GC-bias is the dominant factor in determining codon bias across a broad variety of life and that the form of the influence admits a particularly simple explanation. This model provides a natural null model for codon bias subject to GC mutational bias, relative to which further studies of codon usage may be measured.

4.2 Introduction

Codon bias, the unequal usage of synonymous codons, varies widely between species and, in some cases, between different regions of a genome in a single species(Bernardi, 1993). Factors influencing codon bias include selection for translational accuracy and efficiency (Bulmer, 1991; Gouy & Gautier, 1982; Ikemura, 1981, 1982; Xia, 1998) and GC-bias, on a genomic level in prokaryotes (Muto & Osawa, 1987) and on a regional or isochoric level in vertebrates (Bernardi, 2000).

The influence of GC-bias on codon bias results in a close association between GC% at the third codon position, also called GC3 (Sueoka, 1988), and GC-bias (genomic GC% in prokaryotes or isochoric GC% in mammals). This has led to a common

belief that the use of synonymous G/C-ending codons should increase in frequency with increasing GC-bias, while usage of A/T-ending codons should decrease (Kliman & Bernal, 2005). Though this is a reasonable assumption for most amino acids, those that allow GC-changing synonymous substitutions in the first as well as the third codon positions may be expected to act differently.

The recent growth in the availability of both partial and full genomic sequences has allowed for broad studies of codon usage subject to GC-bias across large numbers of species. For example Hershberg & Petrov (2009) identified favoured codons in prokaryotes and fungi and showed that they are strongly related to the species' intergenic GC content. Knight et al. (2001) modelled codon and amino acid usage as a function of GC mutational bias in bacteria, prokaryotes and eukaryotes showing that GC content drives codon and amino acid usage; they provided a model of usage by compositional class, but did not directly address codon bias. Despite the significant prior work in this area, a general predictive model of codon bias subject to GC mutational bias not been published.

4.3 Materials and Methods

Codon usage frequencies for prokaryotic and plant species were downloaded from the CUTG database (Nakamura et al., 2000) based on the NCBI GenBank Flat File Release 160.0 (Benson et al., 2008). The CUTG gbbct.spsum *prokaryotic* data set containing both bacterial and archaean species codon data and the gbpln.spsum *plant* codon usage data were used. These records sum the codon usage for all nuclear coding sequences in GenBank per species. Though species records can contain duplicated

genes, they provide a reasonable estimate of the genomic codon usage for each species, including those for which full, annotated genomic sequences are unavailable. For the 196 plant species and 897 prokaryotic species using the standard genetic code (NCBI genetic code 11) whose entries were based on 50 or more coding sequences (CDS), codon usage frequency per amino acid was calculated.

The Ensembl (Hubbard et al., 2009) v54_36p *H. sapiens* and v54_37g *M. musculus* genome databases were used as the source of genomic sequences accessed via the Ensembl Perl API (Stabenau et al., 2004). For each amino acid, the longest CDS for the 20884 *H. sapiens* and 22729 *M. musculus* protein-coding genes annotated as known containing at least ten codons for that amino acid were chosen for the respective analyses.

LOESS (locally estimated scatter plot smoothing) fits were used on the graphs to show general trends in scattered data without imposing a global model. The R 2.10.0 (R Development Core Team, 2010) `loess` function was used to generate the fit to the data (coloured lines on graphs) with default parameters (span=1.0, degree=2, least-squares fitting).

Percent variance from the model was calculated on a per-codon or per model-class basis by calculating one minus the variance of the difference between the model and the observations divided by the variance of the observations. Averages were then generated for all codons and model classes, each weighted by the number of codons used in the calculation.

The Harvey-Collier test for functional misspecification (Harvey & Collier, 1977) as implemented in the R `lmtest` package (Zeileis & Hothorn, 2002) was used to estimate the degree of non-linearity of codon usage as a function of GC3 in prokaryotes. The

Harvey-Collier test provides a test statistic which indicates the degree of non-linearity of the data; the higher the statistic the more non-linear the data, with perfectly linear data having a value of zero.

4.4 Results

4.4.1 General patterns of codon usage

In the tissue-specific analysis of the prior chapters, a general correspondence between imposed GC-bias, overall GC3 and codon usage is noted; G and C ending codons are more commonly used in isochores with high GC-bias, A and T ending codons are more commonly used in isochores with high AT bias. This trend in mammals has been previously noted and accounts for a large proportion of the codon usage bias observed in *H. sapiens* tissue-specific genes (Semon et al., 2006). Close inspection of the relationship between the usage bias of individual codon and GC3, however, reveals some notable deviations from the expected trend.

A principal component analysis on the first RSCU values for tissue-specific genes (see Materials and Methods, Chapter 2) demonstrates that the first principal component (PC1) accounts for 32% of the observed codon usage variance between genes, with the second principal (PC2) component accounting for only 4%. PC1 is strongly linearly correlated with GC3 ($R^2 = 0.970$). A plot of codon vectors projected onto the first two principal components (Figure 4.1) shows most A and T ending codons on the left side of the graph (negative PC1 values) and most G and C ending codons on the right side of the graph (positive PC1 values). The only clear exception to this

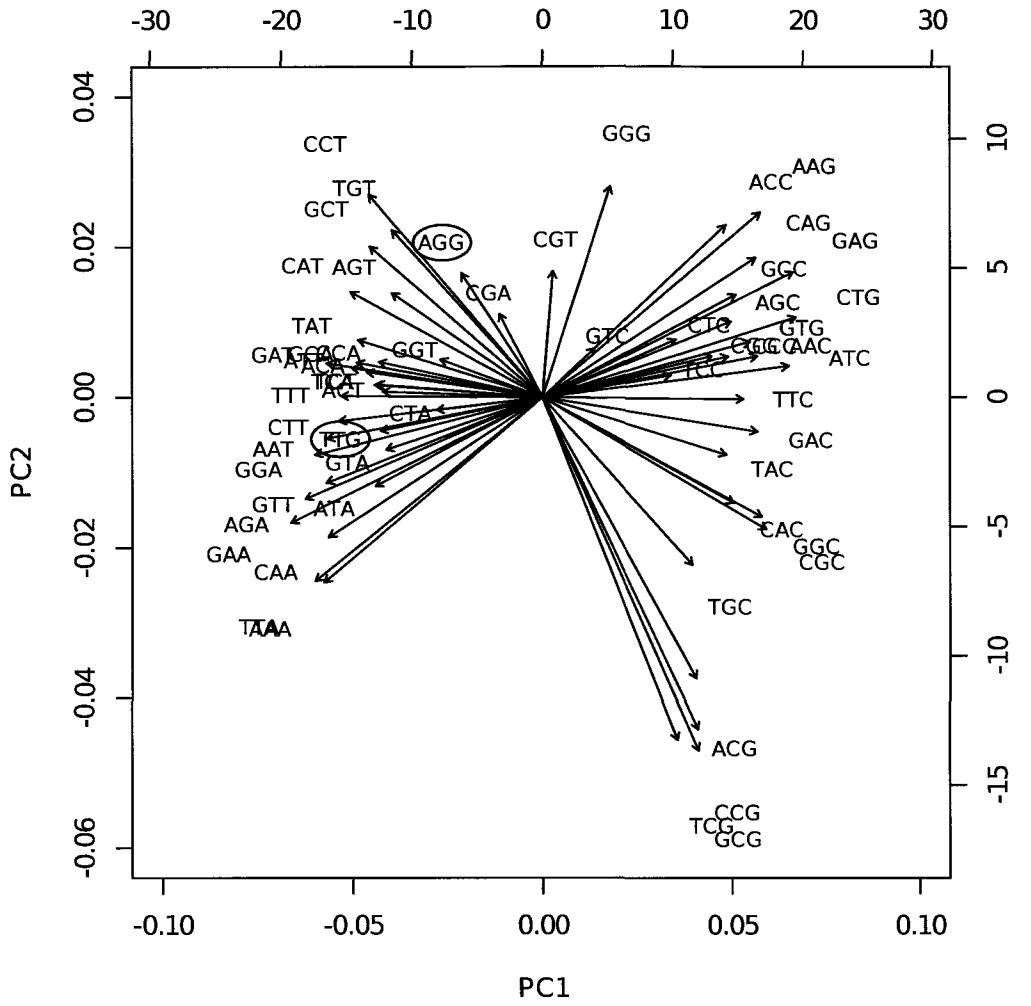

Figure 4.1: Codons projected onto the first two principal components of a principal component analysis of RSCU values for tissue-specific genes

trend are the codons TTG and AGG (red, circled codons in Figure 4.1); despite being G-ending they cluster with the A and T ending codons.

This pattern of usage was previously noted in the *H. sapiens* genome by Kliman & Bernal (2005) and attributed to species-specific selection effects. The analysis of this chapter shows that, rather than being an species specific anomaly, the negative correlation between AGG and TTG codon usage and GC-bias (as estimated by GC3)

appears to be ubiquitous and can be observed in the full set of protein coding genes in the *M. musculus* genome and *H. sapiens* genome as well as plant and prokaryotic genomes.

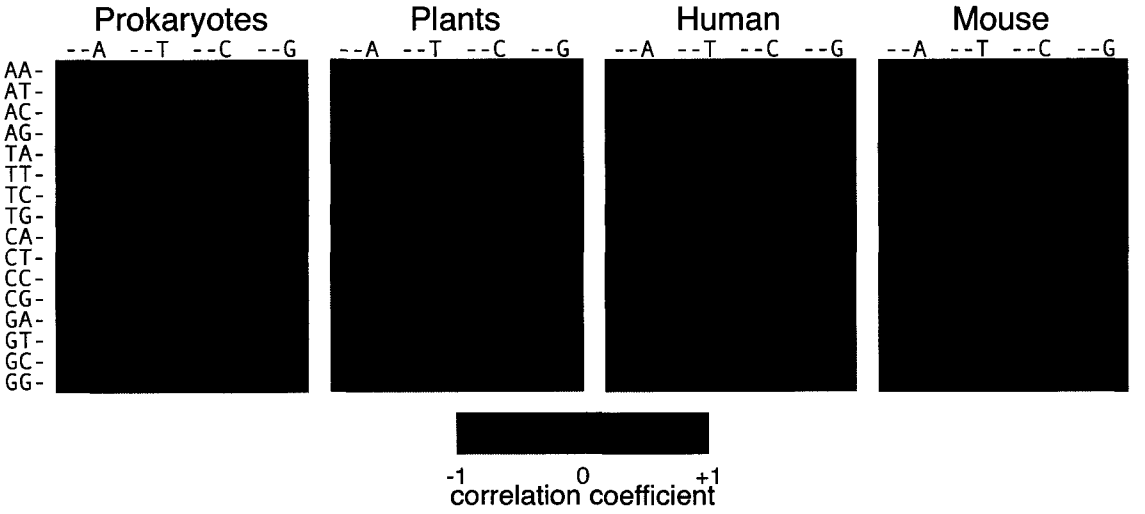

Figure 4.2: Heatmaps of GC3 vs codon frequency for prokaryotes, plants, *H. sapiens* and *M. musculus*. Green indicates a negative correlation with frequency of that codon *decreasing* with increasing GC3, red indicates a positive correlation with the frequency of that codon *increasing* with increasing GC3. Black indicates tryptophan and methionine (which each have only one codon) and stop codons.

Figure 4.2 displays heatmaps of the correlation coefficients between codon usage and GC3 (across species for prokaryotes and plants, and across genes for *H. sapiens* and *M. musculus*, see (see Supplemental Tables A.2, A.3, A.4 and A.5 for respective per-codon correlation values). In all three analyses, the usage of nearly all G- and C-ending codons is strongly positively correlated with GC-bias, and conversely for A-

and T-ending codons. However AGG (arginine) and TTG (leucine), defy this trend consistently across the four different groups of data by showing a negative correlation between usage and GC-bias. Two other arginine codons, CGA and CGT, display the expected negative correlation, but the correlation is very weak.

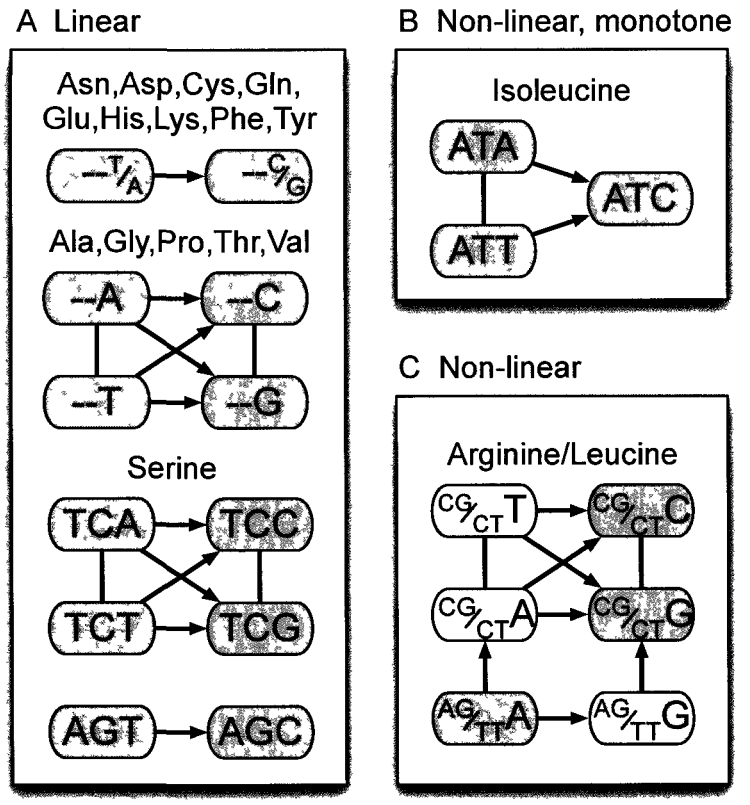

Figure 4.3: Networks of synonymous single site substitutions for all amino acids. GC increasing substitution directions are indicated with an arrow. GC preserving substitutions are represented by line without arrows. For amino acids with 2 and 4 codons as well as serine, G/C-ending codons are pink, and A/T-ending codons green. For arginine and leucine, codons with G/C in both synonymous positions (2xGC codons) are pink, those with only one G/C in a synonymous position (1xGC) are grey, and those with A/T in both synonymous positions (0xGC) are green. (A) Amino acids whose codons are predicted to have a linear response to codon bias. (B) Isoleucine, whose codons are predicted to have a non-linear but monotone response to codon bias. (C) Arginine and leucine, whose codons are predicted to have non-linear, and in some cases non-monotone, responses to codon bias.

Though the anomalous correlation of AGG and TTG with GC3 can clearly be seen in *H. sapiens*, *M. musculus*, plants and prokaryotes, the following analysis will focus primarily on prokaryotic data. The prokaryotic data consists of fewer data points (genomes) than the *M. musculus* or *H. sapiens* data (genes), however each prokaryotic genome value is based on at least 50 coding sequences and is therefore less subject to stochastic effects than the *H. sapiens* and *M. musculus* genes both of which provide much smaller sample of codon usage. This is particularly important as the RSCU values used to represent codon bias in this analysis are ratios. In addition, the prokaryotic genomes cover a broader range of GC3 than the *M. musculus* or *H. sapiens* genes and plant genomes.

The Harvey-Collier test shows that a large number of codons exhibit significantly non-linear usage (p-value < 0.01) in prokaryotes as a function of GC-bias (see Figure 4.4), though this may be influenced by violation of the constant-variance assumption made by the test or by the large number of data points involved. However, a number of codons, particularly those of arginine, leucine and isoleucine show very strong deviations from linearity.

We hypothesized that the unusual responses of these isoleucine, arginine and leucine codons may result from the structure of possible synonymous single site substitutions within these amino acids (Figure 4.3). Notably, arginine and leucine are the only amino acids that allow GC-changing synonymous substitutions in the first as well as the third position (Figure 4.3C). Isoleucine is the only amino acid with three codons and unequal numbers of G/C- and A/T-ending codons (Figure 4.3B). The remainder of the amino acids have equal numbers of A/T and G/C-ending codons and only allow GC-changing synonymous substitutions in the 3rd codon position. The

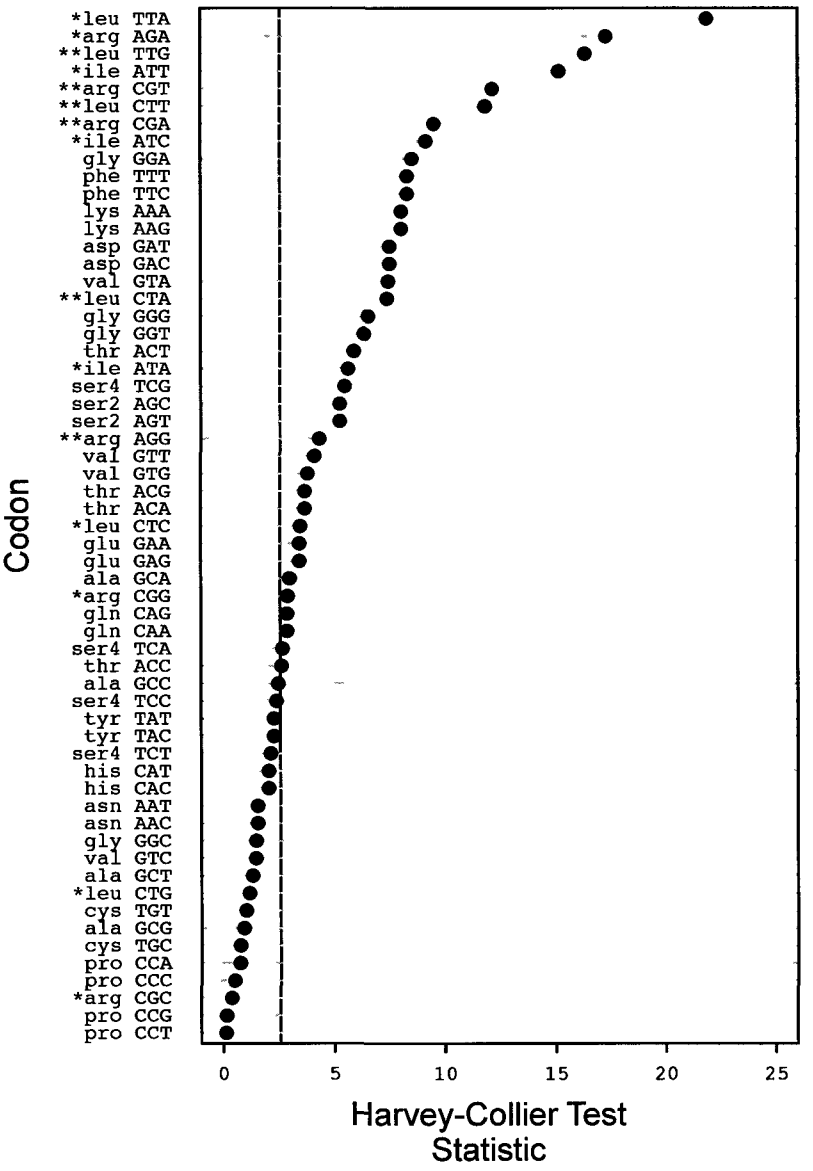

Figure 4.4: Harvey-Collier test statistic for codon bias as a function of GC3 for prokaryotes, ordered by decreasing magnitude. x-axis is the test statistic. The higher the test statistic value, the more non-linear the data. * and ** indicate codons that our model predicts to be non-linear and non-linear, non-monotone respectively. The vertical line indicates the $p = 0.01$ significance threshold.

usage frequencies of these codons is in many cases linear, or at least much closer to linear than shown by many of the leucine, isoleucine and arginine codons.

4.4.2 A model for codon usage based on synonymous mutations

We propose a continuous-time Markov chain model of codon evolution under point mutations that explains the observations above. We assume no non-synonymous mutations, with all variation in usage due to synonymous mutations. Though there are mechanisms by which synonymous mutation may affect protein function, they are more likely to be effectively neutral than non-synonymous mutations are (Chamary et al., 2006). Using this simplifying assumption, we are able to explain the major patterns of codon used without recourse to the added complexity of including non-synonymous mutations.

Consider the codons relating to any particular amino acid. For any two codons X and Y that differ by a single nucleotide, we assume that the rates of mutation from X to Y and from Y to X are the same if both codons have the same total number of G's and C's ($R_{XY} = R_{YX}$). If, however, Y has one more G or C than X, then we assume that mutation from X to Y happens at K times rate of mutation from Y to X ($R_{XY} = KR_{YX}$). Here, K is related to the GC-bias (B) as $K = B/(1-B)$. B ranges from total A/T bias to total G/C bias ($0 < B < 1$). Finally, we assume that the usages of the codons of each amino acid, in a set of genes or a genome, are equal to the equilibrium frequencies of those codons under the model. Letting, $[X]$ and $[Y]$ denote the usages of codons X and Y, then $[X]R_{XY} = [Y]R_{YX}$.

From the continuous time Markov-chain models for every amino acid except serine, one can solve for the equilibrium usage of all codons, as a function of GC-bias B, based on the set of possible synonymous single site mutations for each amino acid (Figure 4.3). The equilibrium solutions are summarized in Table 4.1. Importantly, the predicted equilibrium usages of every codon depend only on the GC-bias. They are not affected, for example, by differing transition and transversion mutation rates; as long as the rates are non-zero, the equilibrium solutions are the same.

Serine raises a small problem for our model in that it is the only amino acid that consists of disconnected blocks of codons—one set of four codons and another of two codons—which are not mutually accessible by any combination of synonymous point mutations. As a consequence, the model can only explain the usages of the TCA, TCC, TCG and TCT codons with respect to each other, and of the AGC and AGT codons with respect to each other. The relative usages of the first four codons compared to the last two is beyond the scope of our model. For the remainder of this chapter, we will treat the four TC-beginning codons as if they belong to a single amino acid, which we call serine4, and the two AG-beginning codons as if they belong to a different amino acid, which we denote serine2. Implicitly, this means that we redefine the usage of each TC-beginning codon as its number of occurrences divided by the total number of occurrences of TC-beginning codons, and likewise for the AG-beginning codons.

4.4.2.1 Linear usage predicted for all 2-codon and 4-codon amino acids.

For all the two-codon and four-codon amino acids, including the serine2 and serine4 groups, the model predicts codon usage that is linearly increasing or decreasing in

Table 4.1: Equilibrium solutions for codon frequencies

Amino acid / class	Codon	Frequency
Two-codon	$--$A or $--$T	$1-B$
	$--$C or $--$G	B
Four-codon	$--$A and $--$T	$\frac{1-B}{2}$
	$--$C and $--$G	$\frac{B}{2}$
Isoleucine	ATA and ATT	$\frac{1-B}{2-B}$
	ATC	$\frac{B}{2-B}$
Arginine / Leucine	AGA and TTA	$\frac{(1-B)^2}{1+B}$
	CGA, CGT, AGG, CTA, CTT and TTG	$\frac{B(1-B)}{1+B}$
	CGG, CGC, CTG and CTC	$\frac{B^2}{1+B}$

GC-bias, B, depending on whether the codon ends with an A/T or with a G/C. Consider first the two-codon amino acids. Each has one A/T-ending codon, which we will denote by X, and one G/C-ending codon, which we will denote by Y. Because Y has one more G/C than X, the model asserts that the mutation rates satisfy $R_{XY} = KR_{YX}$. At equilibrium, we have flux from X to Y must match the flux from Y to X:

$$[X]R_{XY} = [Y]R_{YX} = [Y]K^{-1}R_{XY} \Rightarrow K[X] = [Y]$$

The equilibrium frequencies must also sum to one, so we have:

$$[X] + [Y] = 1$$

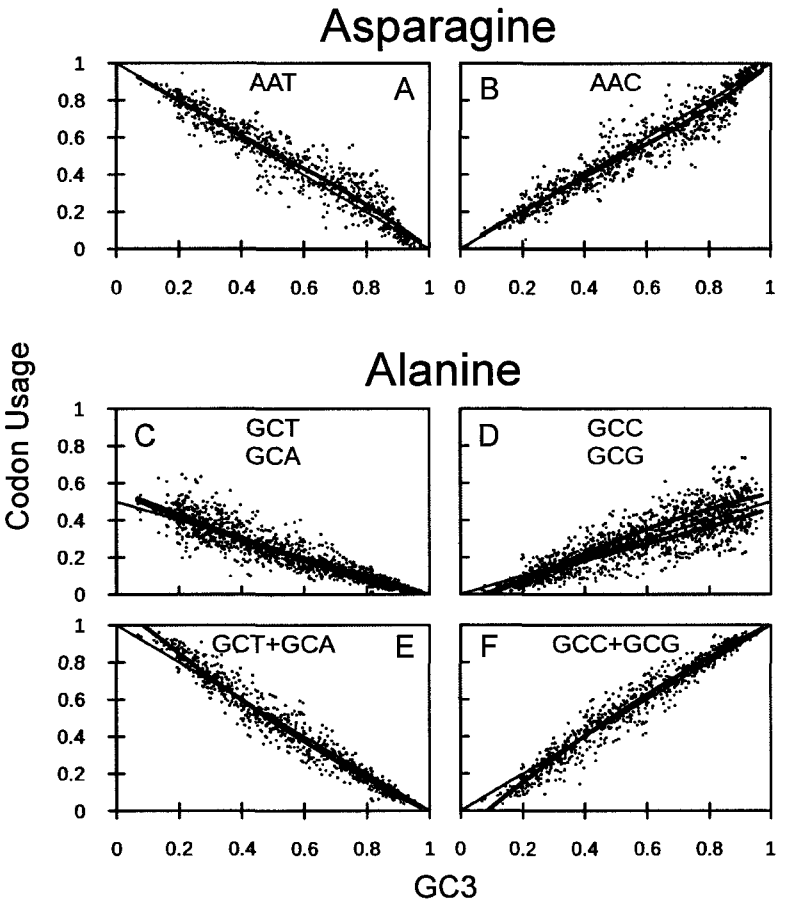

Figure 4.5: Codon usage frequency for asparagine and alanine based on prokaryotic data. The x-axis is genomic GC3 as an estimate of GC-bias, The y-axis is usage frequency. The coloured points and lines correspond to the observed codon usage frequencies for a given GC3 and the corresponding loess fit respectively. The black line is the model prediction. (A) Asparagine AAT codon. (B) Asparagine AAC codon. (C) Alanine GCT and GCA codons. (D) Alanine GCC and GCG codons. (E) The sum of the GCT + GCA alanine codons. (F) The sum of the GCC + GCG codons.

Combining these two equalities, and recalling that $K = B/(1-B)$ we can solve for the frequencies of X and Y:

$$[X] = \frac{1}{1+K} = 1 - B \qquad [Y] = B$$

As one would expect, usage of the A/T-ending codon decreases linearly with increasing GC-bias, with 0% usage in completely GC-biased situations ($B = 1$) and 100% usage in completely AT-biased situations ($B = 0$). The opposite holds for the G/C-ending codon.

For a four-codon amino acid, the derivation is similar. The model's assertion about mutation rates implies six different equalities, corresponding to the six possible pairs among the four codons. Let us focus on three:

$$R_{X_1X_2} = R_{X_2X_1} \qquad R_{Y_1Y_2} = R_{Y_2Y_1} \qquad R_{X_1Y_1} = KR_{Y_1X_1}$$

The remaining three relationships turn out to be redundant with these three. At equilibrium, the balance of fluxes imply that:

$$[X_1]R_{X_1X_2} = [X_2]R_{X_2X_1} \Rightarrow [X_1] = [X_2]$$

$$[Y_1]R_{Y_1Y_2} = [Y_2]R_{Y_2Y_1} \Rightarrow [Y_1] = [Y_2]$$

$$[X_1]R_{X_1Y_1} = [Y_1]R_{Y_1X_1} = [Y_1]K^{-1}R_{X_1Y_1} \Rightarrow K[X_1] = [Y_2]$$

Combining these with the fact that the frequencies sum to one, $[X_1] + [X_2] + [Y_1] + [Y_2] = 1$, we can solve to obtain:

$$[X_1] = [X_2] = \frac{1}{2} - \frac{1}{2}B \qquad [Y_1] = [Y_2] = \frac{1}{2}B$$

The usages of the A/T-ending codons are predicted to be exactly half of the usage of a single A/T-ending codon in a two-codon amino acid, and likewise for the G/C-ending

codons.

Figure 4.5 shows example codon usages for asparagine, a two-codon amino acid, alanine, a four-codon amino acid in the prokaryotic data. For asparagine, the model matches the observed usages with great accuracy. For alanine the overall trends of the A/T-ending and G/C-ending codons are well-captured, including the fact that their usages are approximately linear and have half the slope of the two-codon usages. However, one also observes that in the most GC-biased situations ($B = 1$), the two G/C-ending codons do not receive precisely one half of the usage. In some species, the G-ending codon receives more of the usage, while in other species the C-ending codon receives more usage. This inequality in usage between the two G/C-ending codons is correlated to their overall usage. A similar phenomenon occurs with the A/T-ending codons. Interestingly, for a four-codon amino acid, if one sums the usages of the two A/T-ending codons and of the two G/C-ending codons, one obtains usages that very accurately match what one would expect for a two-codon amino acid. This is most notable in glycine (Figure 4.7), where GGG shows low, non-linear usage while GGC shows balancing high usage with the summed GGC and GGG usage being quite linear.

4.4.2.2 Non-linear monotone usage predicted for isoleucine codons

For isoleucine, the derivation is similar, with the notable difference being the unequal numbers of G/C- and A/T-ending codons. The synonymous mutation relationships are shown in figure 2B. Let $[X], [Y], [Z]$ represent the equilibrium frequencies for the codons ATA, ATT and ATC respectively. Our assumptions regarding mutation rates imply that $R_{XY} = R_{YX}$ and $R_{XZ} = KR_{ZX}$. The flux balance at equilibrium then

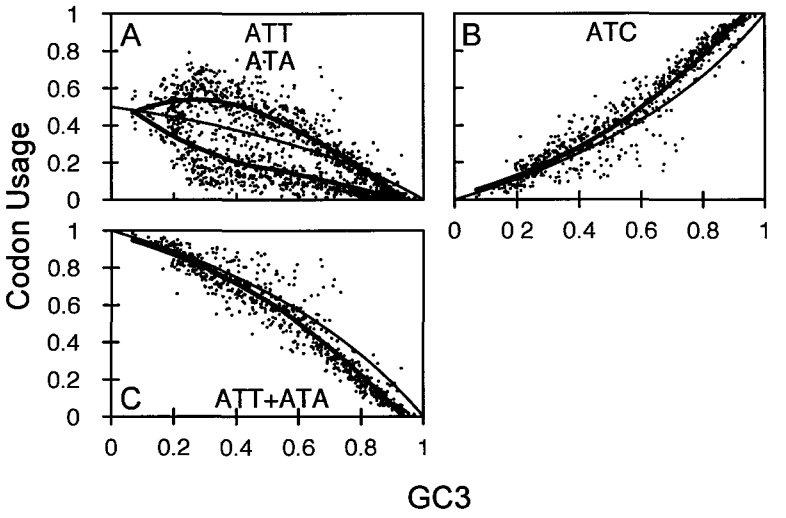

Figure 4.6: Codon usage frequency for isoleucine based on prokaryotic data. The x-axis is genomic GC3 as an estimate of GC-bias, The y-axis is usage frequency. The coloured points and lines correspond to the observed codon usage frequencies for a given GC3 and the corresponding loess fit respectively. The black line is the model prediction. (A) ATT and ATA codons. (B) ATC codon. (C) The sum of the ATA + ATT codon usages.

implies:

$$[X]R_{XY} = [Y]R_{YX} \Rightarrow [X] = [Y]$$

$$[X]R_{XZ} = [Z]R_{ZX} = [Z]K^{-1}R_{XZ} \Rightarrow K[X] = [Z]$$

Using $[X] + [Y] + [Z] = 1$ and $K = B/(B-1)$, we obtain

$$[X] = [Y] = \frac{1}{2+K} = \frac{1-B}{2-B} \qquad [Z] = \frac{B}{2-B}$$

The usage curves are shown in Figure 4.6, along with the prokaryote data. The

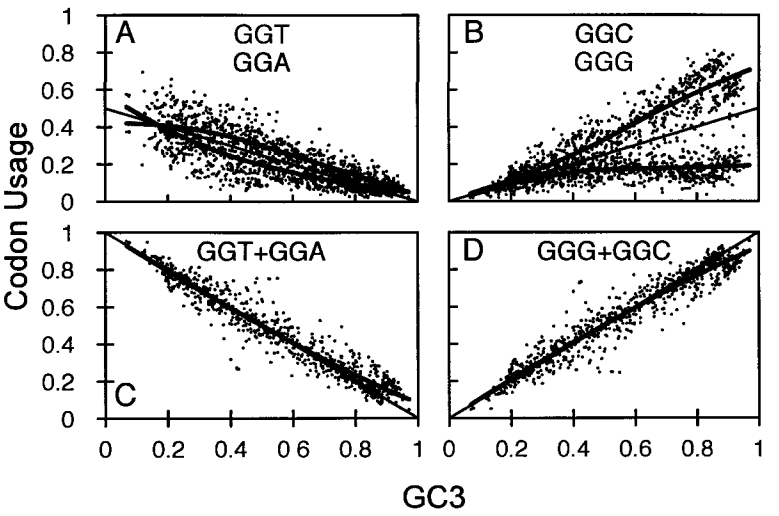

Figure 4.7: Codon usage frequency for glycine based on prokaryotic data. The x-axis is genomic GC3 as an estimate of GC-bias, the y-axis is usage frequency. The coloured points and lines correspond to the observed codon usage frequencies for a given GC3 and the corresponding loess fit respectively. The black line is the model prediction. (A) Glycine A/T-ending codon usage. (B) Glycine G/C-ending codon usage. (C) Glycine A/T-ending codon usage summed. (D) Glycine G/C-ending codon usage summed.

predicted and observed usages of ATC match very well showing positive slope and mild convexity. For ATA and ATT, the observed usages are decreasing in GC-bias B, particularly for $B \geq 0.3$. For that range of B, the usage curves also show the concavity predicted by the model, particularly ATT. However, the usage of ATT is approximately double that of ATA, in contradiction to the model prediction. In the vicinity of $B = 0.2$, there is a sudden "correction" of the relative usages, so that for the most A/T-biased species for which we have data, the usages of both ATT and ATA are approximately as predicted by the model. Despite these complicated patterns in ATA and ATT usage, as shown in Figure 4.6 panel C, the summed usages

of ATA and ATT correspond well to the model's prediction for their sum, over the whole range of GC-bias and is notably non-linear, with the model explaining 92% of observed variance.

4.4.2.3 Non-linear, and in some cases non-monotone, usage predicted for arginine and leucine codons

Arginine and leucine are notable as the only amino acids which allow synonymous, GC-changing point mutations in both the first and third codon positions as shown in Figure 4.3C.

Let $[U], [V], [W], [X], [Y]$ and $[Z]$ represent the equilibrium frequencies of the arginine codons AGA, CGA, CGT, AGG, CGG and CGC respectively or, equivalently, the equilibrium frequencies of the leucine codons TTA, CTA, CTT, TTG, CTG and CTC respectively. Our assumptions regarding mutation rates imply nine different equalities, corresponding to the nine possible pairs among the six codons. Only five non-redundant equalities must be specified: $R_{WV} = R_{VW}$, $R_{YZ} = R_{ZY}$, $R_{UX} = KR_{XU}$, $R_{XY} = KR_{YX}$, $R_{UV} = KR_{VU}$ The flux balance at equilibrium then implies:

$$[V]R_{VW} = [W]R_{WV} \Rightarrow [V] = [W]$$

$$[Y]R_{YZ} = [Z]R_{ZY} \Rightarrow [Y] = [Z]$$

$$[U]R_{UX} = [X]R_{XU} = [X]K^{-1}R_{XU} \Rightarrow K[U] = [X]$$

$$[X]R_{XY} = [Y]R_{YX} = [Y]K^{-1}R_{YX} \Rightarrow K[X] = [Y]$$

$$[U]R_{UV} = [V]R_{VU} = [V]K^{-1}R_{VU} \Rightarrow K[U] = [V]$$

Substituting the above into $[X] + [Y] + [Z] + [W] + [U] + [V] = 1$ and solving for U gives:

$$[U] + 3K[U] + 2K^2[U] = 1 \Rightarrow [U] = \frac{1}{1 + 3K + 2K^2}$$

Substituting $K = B/(1 - B)$ and solving for codon frequencies gives:

$$[U] = \frac{(1 - B)^2}{(1 + B)}$$

$$[V] = [W] = [X] = \frac{B(1 - B)}{(1 + B)}$$

$$[Y] = [Z] = \frac{B^2}{(1 + B)}$$

For both arginine and leucine, three distinct usage patterns are predicted, depending on whether a codon has total of zero, one or two G's and C's in the first and third positions, those that can vary synonymously. We will refer to these codon classes as 0xGC, 1xGC and 2xGC respectively, corresponding to the three solutions above.

0xGC (arginine: Figure 4.8C, leucine: figure 4.8I) and 2xGC (arginine: Figure 4.8A, leucine: Figure 4.8G) usage curves are monotonic and non-linear decreasing and increasing respectively with increasing GC3 described by $(1 - B)^2/(1 + B)$ and $B^2/(1 + B)$ respectively. 1xGC codons (arginine: Figure 4.8B, leucine: figure 4.8H) show a non-monotone, asymmetric, concave distribution described by $B(1 - B)/(1 + B)$ with a peak at $B = \sqrt{2} - 1 \approx 0.41$, below neutral bias ($B = 0.5$).

Intuitively, the 1xGC codons show peaked usage because when GC content is low the 0xGC codons are strongly favored, and when GC content is high the 2xGC codons are strongly favored. The asymmetry of the peak position is due to the fact

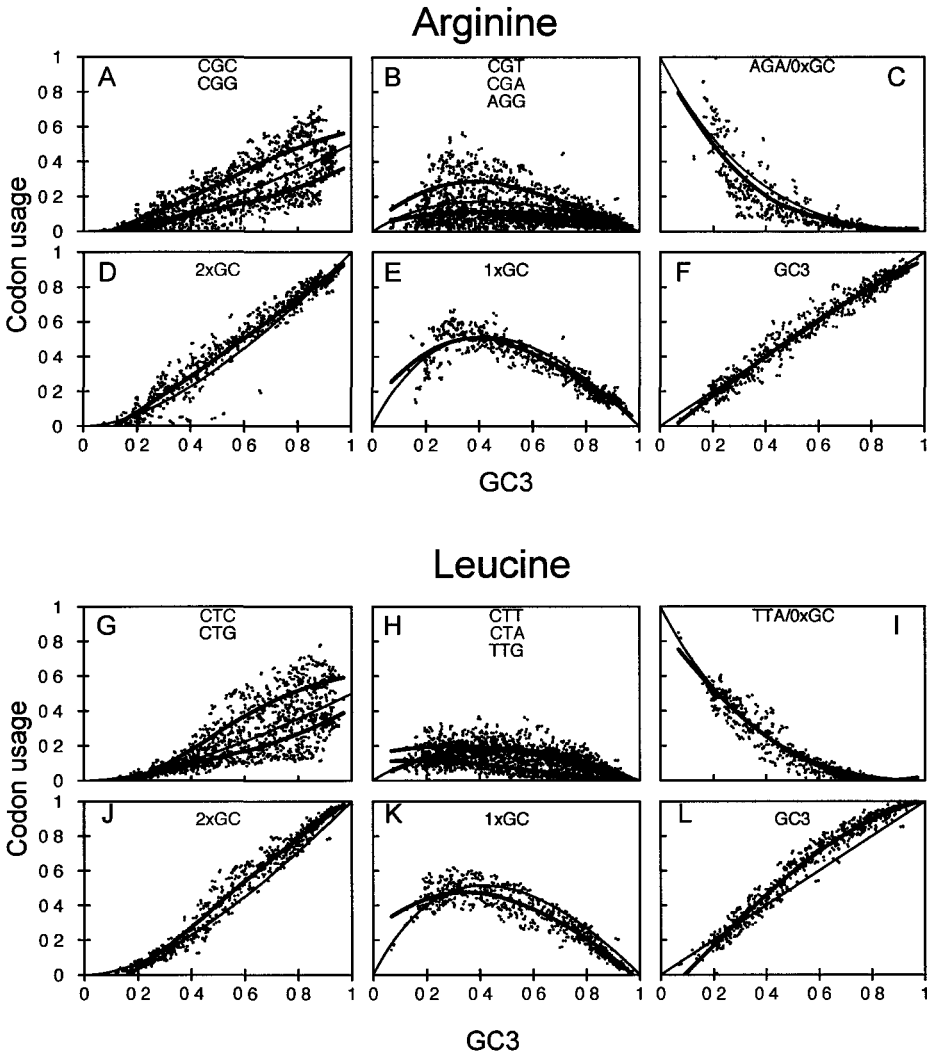

Figure 4 8 Codon usage frequency for arginine and leucine based on prokaryotic data The x-axis is genomic GC3 as an estimate of GC-bias, the y-axis is usage frequency The coloured points and lines correspond to the observed codon usage frequencies for a given GC3 and the corresponding loess fit respectively The black line is the model prediction (A) 2xGC arginine codons (B) 1xGC arginine codons (C) The 0xGC arginine codon (D) The sum of 2xGC arginine codons (E) The sum of 1xGC arginine codons (F) GC3 for arginine codons (G) 2xGC leucine (H) 1xGC leucine codons (I) 0xGC leucine (J) The sum of 2xGC leucine codons (K) The sum of 1xGC leucine codons (L) GC3 for leucine codons

that there are two 2xGC codons in each of arginine and leucine but only one 0xGC codon. The peak position is notable as the correlation of such a curve over the range of $B = [0, 1]$ will be negative, as shown for AGG (Figure 4.8B) and TTG (Figure 4.8H). The negative correlation is stronger when the data are richer in GC sequences, as is the case with the *H. sapiens* and *M. musculus* data. This result provides an alternative interpretation of the results of Kliman & Bernal (2005) who reported that AGG and TTG codons in *H. sapiens* are negatively correlated with GC3, intronic GC and expression and postulated that this may be the result of selection on these codons. AGG and TTG are the two G-ending 1xGC codons; their observed usage frequency patterns are predicted by our model for *H. sapiens* and *M. musculus* as well as prokaryote and plant genomes.

Our model accounts for 72% of the variation in codon usage in prokaryotic genomes 64% in plant genomes, 52% and 36% in *H. sapiens* and *M. musculus* genes respectively and predicts the observed decrease in usage of AGG and TTG with increasing GC-bias (see Supplemental Tables A.2, A.3, A.4 and A.5 for prokaryotic, plant, mouse and human codon data respectively) When individual codons are summed together for a class of solutions (0xGC, 1xGC and 2xGC for arginine and leucine (Figures 4.8D,E,J,K) and G/C- and A/T-ending for the other amino acids (Figures 4.5E,F, 4.6C, 4.7C,D) our model accounts for 87% of variability in class usage across prokaryotic species, 80% across plant species, 68% and 51% across *H. sapiens* and *M. musculus* genes respectively. Details of the models' fit to observations are provided in Supplemental Table A.1.

4.5 Discussion

The model presented in this chapter makes a number of predictions. The most important and notable prediction is that two G-ending codons (AGG, TTG) will show decreasing usage with increasing GC-bias as indicated by GC3. The second prediction of our model is that some codon frequencies will be non-linear and others linear as a function of GC-bias. The third prediction of our model is that the per-amino acid GC3 should be linear for all amino acids with the exception of isoleucine.

Our model predictions are most accurate in the prokaryotic data due to the large number of species in the data set, the fact that each species data point has a high codon count (at least 50 coding sequences and usually much more) and they cover a broad range of GC3 values. The plant data is much more limited, consisting of fewer genomes covering a narrower range of GC3 values. The *H. sapiens* and *M. musculus* data also has a narrower range of GC3 values than the prokaryotic data. The *M. musculus* genome, in particular, has less extreme isochores than the *H. sapiens* genome (Bernardi, 2000) and has a narrower range of GC3 values.

Though the *H. sapiens* and *M. musculus* data consists of many more data points, each data point is based on the relatively small number of codons in a single gene. As a result of this there are strong stochastic effects and the data are quite noisy resulting in lower prediction accuracy. Despite the limitations of the data, the key prediction of our model that AGG and TTG should have negative correlation with respect to GC3 holds true for all the data sets as clearly shown for prokaryotic, plant, *H. sapiens* and *M. musculus* data in Figure 4.2A.

Figure 4.4, which shows the Harvey-Collier test for linearity applied to codon

frequencies in the prokaryotic data set, shows that the 8 most non-linear codons as a function of GC3 are in the list of 12 that our model predicts to be non-linear. While the results for plant, *H. sapiens* and *M. musculus* data are less definitive, we attribute this to the properties of those data sets, as described above.

For all of the model solutions except isoleucine, the usages of G/C-ending codons sum to B, whereas the usages of A/T-ending codons sum to $1 - B$. Thus, the observed overall GC bias in the third codon position is predicted to be followed by each amino acid individually. Indeed, this is true for arginine and leucine as well, though somewhat surprising due to the nonlinearity of the individual codon class solutions. Isoleucine, which has an odd number of codons, is the sole exception, the nonlinear ATC codon usage $B/(2 - B)$ somewhat lower than B (Figure 4.6C). This property vindicates the use of $K = B/(1 - B)$ in our model to associate the mutational rate K with the overall GC bias as estimate by GC3.

There are a number of deviations from our model, where codons predicted to be linear show significant non-linearity (Figure 4.4). In some cases, deviations from the model may be indicative of biologically significant effects; for example GGC and GGG (glycine) show usage frequencies well above and below that of the model respectively in *H. sapiens*, *M. musculus*, prokaryotic and plant data (see Figure 4.7 for prokaryotic results). Low usage of GGG has been previously noted in *D. melanogaster* and attributed to the avoidance of guanine runs that may participate in the formation of G-quadruplexes, stable mRNA structures which can impede translation (Kreitman & Antezana, 1999). The presence of G-quadruplexes in *H. sapiens* 5' UTRs has been shown to repress translation (Beaudoin & Perreault, 2010). It appears from the results of this study that such an effect may have consequences to bacterial, *H.*

sapiens, *M. musculus* and plant codon usage as well.

Many other known influences on codon bias have not been included in our model. Selection for increased transcriptional efficiency can be particularly strong in prokaryotes (Xia, 1998) affects codon usage, and can vary between species. GC skew which is present in the human genome (Comeron, 2004), vertebrate mitochondria in general (Faith & Pollock, 2003), prokaryotes (Marín & Xia, 2008) and plants (Tatarinova et al., 2003) can effect codon bias. Context-specific mutations, where mutations are correlated with neighbouring bases (Jia & Higgs, 2008) may play a role in some of the observed deviations. Nearly-neutral non-synonymous mutations may also play a role: it has been shown that first and second position base content can vary with third position base content in metazoan mitochondria (Urbina et al., 2006). In addition, even if the model describes equilibrium usage correctly, coding sequences may not all be at or near the equilibrium state described by the model; genes having recently been subject to lateral gene transfer may not yet be in equilibrium with the host genome (Xie et al., 2003), and recently duplicated genes may not yet be in equilibrium with the GC-bias of their host isochore. It may be possible to incorporate some of these effects into future refinements of our model: in particular GC skew may be modelled using an adaptation of our method, however its position dependency poses challenges.

Despite the simplicity of this model and lack of fitted parameters, it broadly captures codon usage trends across multiple branches of life and for a wide range of GC-bias. The influence of GC-bias on codon usage appears to be felt most strongly at the level of codon classes rather than the level of individual codons, and a more detailed model will be required to explain the many deviations of individual codons from the model.

4.6 Conclusion

Motivated by unexpected codon usage patterns in tissue-specific *M. musculus* genes we have developed a general model of codon usage under GC-biased synonymous point mutations. The model is broadly applicable to prokaryotic and plant genomes *M. musculus* and *H. sapiens* genes.

Our model establishes that GC-bias is the dominant factor in determining codon bias across a broad variety of life and that the form of the influence admits a particularly simple explanation. This model provides a natural null model for codon bias subject to GC mutational bias, relative to which further studies of codon usage may be measured.

Chapter 5

Conclusions

Codon bias is a universal feature of protein coding genes. The specific nature of the codon bias observed at the level of a single gene, a set of tissue-specific genes or an entire genome can be indicative of the selection and mutation pressures to which the coding sequences have been subjected.

This thesis explored the relationship between codon usage and tissue-specific gene expression in *M. musculus* using microarray data, testing the mutation and selection hypotheses as well as investigating and shedding light on previously unexplained patterns of codon usage in the codon usage of tissue-specific genes that prove to be a general property of the genetic code subject to GC-bias.

Chapter 2 tested whether there were significant codon usage differences between tissue-specific gene sets. It was found that there are significant differences in codon usage but in every case the observed codon bias matched the differences in average GC3 between tissues; this is in keeping with the mutational hypothesis but does not exclude the possibility that at least some of the observed differences are due to

selection effects.

To test for selection effects, the tissue-specific genes were split into high and low expression sets. Significant differences in codon usage were found between high-expression tissue-specific genes for some tissues but not in the low expression category; this supports the selection hypothesis for some tissues, however the tissue-specific codon differences in the high expression tissue-specific gene set were found to correlate with the GC3 of the tissue-specific genes used. This effect may be due to co-adaptation of tissue GC-bias and tissue-specific tRNA pools with the effect being most apparent in high expression genes.

Finally, tissue-specific genes were grouped by lineage to test for lineage specific selection effects; though general differences in lineage specific codon usage were found, they too correspond to the GC3 of the lineage gene groupings. The fact that high and low expression genes show significant differences in codon usage between tissues does not support the hypothesis that there is adaptation to lineage specific tRNA pools.

Chapter 3 studied the proximity of tissue-specific and lineage specific genes and investigated whether groupings of such genes on the same isochore leading to similar GC-bias and hence GC3 may explain significant differences in tissue-specific GC3. Though some clustering of tissue-specific genes was found to be associated with gene clusters in embryonic dermal tissue, the tissues found to show significant gene GC3 differences were not found to have significant levels of gene clustering. Coexistence of multiple genes on the same isochore was therefore rejected as a cause of difference in GC3 between tissue-specific genes.

The analysis of Chapter 4 was based on an unexpected feature found in tissue-specific mouse genes. AGG and TTG were found to be the only G or C ending

codons whose usage *decreases* with increasing GC3; all others *increase*. This effect has previously been reported in the human genome by Kliman & Bernal (2005) who attributed it to species specific selection, however further investigation revealed that this property of AGG and TTG is ubiquitous, present in *H. sapiens* and *M. musculus* genes as well as bacterial and plant genomes. We hypothesized that the unusual responses of these codons may result from the structure of possible synonymous single site substitutions within arginine and leucine, and built a continuous-time Markov chain based model of codon bias subject to GC-bias for all amino acids. This model was shown to explain the observed AGG and TTG usage for mouse, human, plant and prokaryotes as well provide novel predictions of nonlinear and non-monotone codon usage under GC-bias for some codons. The model accounts for 72%, 64%, 52% and 36% of the observed variability of codon usage in prokaryotes, plants, human and mouse respectively. When codons are grouped based on common GC content, 87%, 80%, 68% and 51% of the variation in usage is explained for prokaryotes, plants, human and mouse respectively.

There are two key findings to this work. The first key finding is that tissue-specific codon bias due to translational selection appears to be present in mouse and corresponds to the GC3 biases of the tissue. This leads to the hypothesis that tRNA concentration itself corresponds to the tissue-specific GC3; this is evidence for co-adaptation of tissue GC-bias and tissue-specific tRNA pools. Validation of this hypothesis and further work on tissue-specific selection effects in mouse would be aided by tRNA abundance measures as it would indicate the specific tRNA that are being adapted to; currently however, very limited data of this type is available.

The second key finding of this work is that a simple continuous time Markov chain

model of codon usage subject to GC mutational bias can explain a large proportion of codon usage across all branches of life. The predictions made by this model for 2 and 4 codon family amino acids are in keeping with general expectations of codon usage under GC-bias; linear and corresponding to the 3^{rd} codon position. Predictions for arginine, leucine and isoleucine, however, are novel, predicting nonlinear and in some cases non-monotone usage patterns. The predicted negative correlation of AGG and TTG with GC-bias, in particular, is observable across all branches of life. This model establishes that GC-bias is the dominant factor in determining codon bias across a broad variety of life and that the form of the influence admits a particularly simple explanation. It provides a natural null model for codon bias subject to GC mutational bias, relative to which further studies of codon usage may be measured.

Bibliography

Affymetrix 2001. Affymetrix Microarray Suite User Guide, version 5.

Affymetrix 2003. Technical Note, Array Design and Performance of the GeneChip Mouse Expression Set 430.

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**: 927.

Alvarez-Valin, F., Lamolle, G. & Bernardi, G. 2002. Isochores, GC3 and mutation biases in the human genome. *Gene* **300**: 161–168.

Beaudoin, J. & Perreault, J. 2010. 5'-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Research* **38**: 7022–7036.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. 2008. GenBank. *Nucleic Acids Research* **36**: D25–30.

Bernardi, G. 1993. The vertebrate genome: isochores and evolution. *Molecular Biology and Evolution* **10**: 186–204.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3–17.

Bernardi, G. 2007. The neoselectionist theory of genome evolution. *Proceedings of the National Academy of Sciences* **104**: 8385–8390.

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

Blake, J., Eppig, J., Bult, C., Kadin, J. & Richardson, J. 2006. The mouse genome database (MGD): updates and enhancements. *Nucleic Acids Research* **34**: D562–567.

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

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**: 897–907.

Chamary, J., Parmley, J. & Hurst, L. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics* **7**: 98–108.

Comeron, J.M. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**: 1293–1304.

Costantini, M., Cammarano, R. & Bernardi, G. 2009. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics* **10**: 146.

Cuny, G., Soriano, P., Macaya, G. & Bernardi, G. 1981. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *European Journal of Biochemistry* **115**: 227–233.

Desbois, C., Hogue, D. & Karsenty, G. 1994. The mouse osteocalcin gene cluster contains three genes with two separate spatial and temporal patterns of expression. *Journal of Biological Chemistry* **269**: 1183–1190.

Dittmar, K.A., Goodenbour, J.M. & Pan, T. 2006. Tissue-specific differences in human transfer RNA expression. *PLoS Genetics* **2**: e221.

Doenecke, D., Albig, W., Bode, C., Drabent, B., Franke, K., Gavenis, K. & Witt, O. 1997. Histones: genetic diversity and tissue-specific gene expression. *Histochemistry and Cell Biology* **107**: 1–10.

Dos Reis, M., Wernisch, L. & Savva, R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli K-12* genome. *Nucleic Acids Research* **31**: 6976.

Duret, L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics* **16**: 287–289.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Current Opinions in Genetic Develpment* **12**: 640–649.

Duret, L. & Arndt, P. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics* **4**.

Duret, L., Eyre-Walker, A. & Galtier, N. 2006. A new perspective on isochore evolution. *Gene* **385**: 71–74.

Duret, L. & Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* **10**: 285–311.

Faith, J.J. & Pollock, D.D. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* **165**: 735–745.

Fullerton, S., Carvalho, B. et al. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Molecular Biology and Evolution* **18**: 1139–1142.

Garel, J. 1974. Functional adaptation of tRNA population. *Journal of Theoretical Biology* **43**: 211–225.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**: R80.

Gilbert, S.F. 2003. *Developmental Biology.* Sinauer Associates.

Goodenbour, J. & Pan, T. 2006. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Research* **34**: 6137–6146.

Gouy, M. & Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research* **10**: 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research* **9**: 213.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research* **8**: 197.

Handy, T.C. 2004. *Event-related potentials: a methods handbook.* The MIT Press.

Harrison, A.P., Johnston, C.E. & Orengo, C.A. 2007. Establishing a major cause of

discrepancy in the calibration of Affymetrix GeneChips. *BMC Bioinformatics* **8**: 195.

Harvey, A. & Collier, P. 1977. Testing for functional misspecification in regression analysis. *Journal of Econometrics* **6**: 103–119.

Hershberg, R. & Petrov, D. 2008. Selection on codon bias. *Annual Review of Genetics* **42**: 287–299.

Hershberg, R. & Petrov, D.A. 2009. General rules for optimal codon choice. *PLoS Genetics* **5**: e1000556+.

Hesse, M., Zimek, A., Weber, K. & Magin, T. 2004. Comprehensive analysis of keratin gene clusters in humans and rodents. *European Journal of Cell Biology* **83**: 19–26.

Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. & Flicek, P. 2009. Ensembl 2009. *Nucleic Acids Research* **37**: D690–697.

Hwang, K.B., Kong, S.W., Greenberg, S.A. & Park, P.J. 2004. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics* **5**: 159.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology* **151**: 389–409.

Ikemura, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Molecular Biology* **158**: 573–597.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**: 13.

Iwata-Reuyl, D. 2008. An embarrassment of riches: the enzymology of RNA modification. *Current Opinions in Chemical Biology* **12**: 126–133.

Jia, W. & Higgs, P. 2008. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Molecular Biology and Evolution* **25**: 339–351.

Kliman, R.M. & Bernal, C.A. 2005. Unusual usage of AGG and TTG codons in humans and their viruses. *Gene* **352**: 92–99.

Knight, R., Freeland, S. & Landweber, L. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology* **2**: 1–13.

Kotlar, D. & Lavner, Y. 2006. The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics* **7**: 67.

Kreitman, M. & Antezana, M. 1999. The population and evolutionary genetics of codon bias. *Evolutionary genetics: from molecules to morphology* **1**: 82–101.

Krumlauf, R. 1994. Hox genes in vertebrate development. *Cell* **78**: 191–201.

Kurosawa, H. 2007. Methods for inducing embryoid body formation: *in vitro* differentiation system of embryonic stem cells. *Journal of Bioscience and Bioengineering* **103**: 389–398.

Lavner, Y. & Kotlar, D. 2005. Codon bias as a factor in regulating expression *via* translation rate in the human genome. *Gene* **345**: 127–138.

Lercher, M., Urrutia, A. & Hurst, L. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics* **31**: 180–183.

Li, Q., Peterson, K., Fang, X. & Stamatoyannopoulos, G. 2002. Locus control regions. *Blood* **100**: 3077–3086.

Lieberman-Aiden, E., van Berkum, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B., Sabo, P., Dorschner, M. et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Maglott, D., Ostell, J., Pruitt, K. & Tatusova, T. 2006. Entrez Gene: gene-centered

information at NCBI. *Nucleic Acids Research* pp. D26–31.

Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genetics* **19**: 330–338.

Marín, A. & Xia, X. 2008. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: New substitution models incorporating strand bias. *Journal of Theoretical Biology* **253**: 508–513.

Moriyama, E. & Hartl, D. 1993. Codon usage bias and base composition of nuclear genes in Drosophila. *Genetics* **134**: 847–858.

Moriyama, E.N. & Powell, J.R. 1997. Codon usage bias and tRNA abundance in *Drosophila*. *Journal of Molecular Evolution* **45**: 514–523.

Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H. & Bernardi, G. 2001. Translational selection on codon usage in *Xenopus laevis*. *Molecular Biology and Evolution* **18**: 1703–1707.

Muto, A. & Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences* **84**: 166–169.

Nakamura, Y., Gojobori, T. & Ikemura, T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research* **28**: 292.

Nekrutenko, A. & Li, W. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research* **10**: 1986.

Percudani, R., Pavesi, A. & Ottonello, S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **268**: 322–330.

Perez-Iratxeta, C. & Andradé, M.A. 2005. Inconsistencies over time in 5% of NetAffx probe-to-gene annotations. *BMC Bioinformatics* **6**: 183.

Plotkin, J.B., Robins, H. & Levine, A.J. 2004. Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences* **101**: 12588–12591.

R Development Core Team 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL

http://www.R-project.org. ISBN 3-900051-07-0.

Romiguier, J., Ranwez, V., Douzery, E. & Galtier, N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research* .

Roop, D., Hawley-Nelson, P., Cheng, C. & Yuspa, S. 1983. Keratin gene expression in mouse epidermis and cultured epidermal cells. *Proceedings of the National Academy of Sciences* **80**: 716–720.

Saier, J., M. H. 1995. Differential codon usage: a safeguard against inappropriate expression of specialized genes? *FEBS Letters* **362**: 1–4.

Sandie, R., Palidwor, G.A., Huska, M.R., Porter, C.J., Krzyzanowski, P.M., Muro, E.M., Perez-Iratxeta, C. & Andradé-Navarro, M.A. 2009. Recent developments in StemBase: a tool to study gene expression in human and murine stem cells. *BMC Research Notes* **2**: 39.

Semon, M., Lobry, J. & Duret, L. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in Humans. *Molecular Biology and Evolution* **23**: 523–529.

Sene, K., Porter, C., Palidwor, G., Perez-Iratxeta, C., Muro, E., Campbell, P., Rudnicki, M. & Andradé-Navarro, M. 2007. Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics* **8**: 85.

Sharp, P. & Li, W. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **24**: 28–38.

Sharp, P.M., Tuohy, T.M. & Mosurski, K.R. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research* **14**: 5125–5143.

Sittka, A., Lucchini, S., Papenfort, K., Sharma, C.M., Rolle, K., Binnewies, T.T., Hinton, J.C. & Vogel, J. 2008. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genetics* **4**: e1000163.

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M. & Birney, E. 2004. The Ensembl core software libraries. *Genome Research* **14**: 929–933.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fu-

ellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D. & Birney, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* **12**: 1611–1618.

Stoletzki, N. & Eyre-Walker, A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Molecular Biology and Evolution* **24**: 374.

Subramanian, S. 2008. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics* **178**: 2429–2432.

Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences* **85**: 2653–2657.

Talbert, P.B. & Henikoff, S. 2006. Spreading of silent chromatin: inaction at a distance. *Nature Reviews Genetics* **7**: 793–803.

Tatarinova, T., Brover, V., Troukhan, M. & Alexandrov, N. 2003. Skew in CG content near the transcription start site in *Arabidopsis thaliana*. *Bioinformatics* **19**: 313–314.

Tian, Q., Stepaniants, S.B., Mao, M., Weng, L., Feetham, M.C., Doyle, M.J., Yi, E.C., Dai, H., Thorsson, V., Eng, J., Goodlett, D., Berger, J.P., Gunter, B., Linseley, P.S., Stoughton, R.B., Aebersold, R., Collins, S.J., Hanlon, W.A. & Hood, L.E. 2004. Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Molecular Cell Proteomics* **3**: 960–969.

Ulfig, N. 2002. Ganglionic eminence of the human fetal brain-new vistas. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology* **267**: 191–195.

Urbina, D., Tang, B. & Higgs, P. 2006. The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. *Journal of Molecular Evolution* **62**: 340–361.

Urrutia, A.O. & Hurst, L.D. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**: 1191–1199.

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

Vinogradov, A.E. 2003. Isochores and tissue-specificity. *Nucleic Acids Research* **31**: 5212–5220.

Waldman, Y., Tuller, T., Shlomi, T., Sharan, R. & Ruppin, E. 2010. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Research* **38**: 2964–2974.

Xia, X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* **144**: 1309–20.

Xia, X. 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics* **149**: 37–44.

Xie, G., Bonner, C., Brettin, T., Gottardo, R., Keyhani, N. & Jensen, R. 2003. Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in Xylella species and in heterocystous cyanobacteria. *Genome Biology* **4**: R14.

Yang, Z. & Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**: 32–43.

Zeileis, A. & Hothorn, T. 2002. Diagnostic Checking in Regression Relationships. *R News* **2**: 7–10.

Zhang, S.P., Zubay, G. & Goldman, E. 1991. Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. *Gene* **105**: 61–72.

Zhang, W., Morris, Q., Chang, R., Shai, O., Bakowski, M., Mitsakakis, N., Mohammad, N., Robinson, M., Zirngibl, R., Somogyi, E. et al. 2004. The functional landscape of mouse gene expression. *Journal of Biology* **3**: 21.

Appendix A

Supplemental Tables

Table A.1: Model correspondence with observed codon usage variance as a function of GC3

		prokaryotic		plant		human		mouse	
AA	model	codons	variance explained	codons	variance explained	codons	variance explained	codons	variance explained
phe	AT	10988263	0.898	2594593	0.823	173436	0.679	175479	0.528
phe	GC	12013093	0.898	2905632	0.823	192286	0.679	211917	0.528
tyr	AT	9201253	0.742	1870051	0.807	113688	0.652	113750	0.440
tyr	GC	12013093	0.898	2905632	0.823	166622	0.679	184539	0.520
his	AT	6178184	0.706	1784746	0.818	103321	0.688	101502	0.495
his	GC	6316853	0.706	1509844	0.818	136463	0.688	136838	0.495
gln	AT	9051291	0.861	2540964	0.698	130992	0.601	127313	0.448
gln	GC	12741003	0.861	2537456	0.698	355169	0.601	356614	0.448
asn	AT	11057828	0.911	2842844	0.814	172239	0.690	159909	0.487
asn	GC	10469487	0.911	2818743	0.814	178538	0.690	196444	0.487

Continued on next page

Continued from previous page

		prokaryotic		plant		human		mouse	
AA	model	codons	variance explained	codons	variance explained	codons	variance explained	codons	variance explained
lys	AT	15857845	0.839	3353480	0.779	267460	0.700	246815	0.545
lus	GC	12101933	0.839	4528104	0.779	330208	0.700	356853	0.545
asp	AT	16215879	0.887	4314822	0.875	229567	0.716	223640	0.473
asp	GC	15956872	0.887	3185110	0.875	254208	0.716	267643	0.473
glu	AT	19825452	0.680	4111542	0.832	323804	0.782	306083	0.613
glu	GC	15923811	0.680	4611173	0.832	425224	0.782	427625	0.613
cys	AT	2211343	0.890	1115436	0.724	94049	0.610	101605	0.414
cys	GC	3447583	0.900	1193908	0.710	104877	0.610	103613	0.414
ile	ile_AT	18891320	0.922	4318096	0.818	242612	0.753	240336	0.611
ile	ile_GC	16372164	0.879	2697546	0.689	197793	0.383	221570	0.199
val	AT	14834280	0.956	4046369	0.924	193136	0.770	199274	0.616
val	GC	27876359	0.956	4925598	0.924	437393	0.770	460214	0.616
pro	AT	9187802	0.922	4313479	0.874	375783	0.692	391518	0.471
pro	GC	17904668	0.922	2944543	0.874	293551	0.692	262431	0.471
thr	AT	10969085	0.943	4021832	0.861	300077	0.748	319255	0.570
thr	GC	20690712	0.943	3240293	0.861	252253	0.748	252340	0.570
ala	AT	18984535	0.953	5615206	0.884	369384	0.737	393130	0.555
ala	GC	40427712	0.953	4917463	0.884	380772	0.737	350838	0.555
gly	AT	17368030	0.917	5150990	0.820	290784	0.748	304092	0.565
gly	GC	27604982	0.917	4186564	0.820	412699	0.748	387849	0.565
arg	0xGC	3344449	0.720	1941778	0.624	128098	0.586	132926	0.444
arg	1xGC	12287867	0.395	4008656	0.436	292692	0.466	300988	0.324
arg	2xGC	20631556	0.892	2351844	0.826	231782	0.702	210099	0.566
leu	0xGC	9114156	0.920	1490706	0.476	85743	0.366	78921	0.221
leu	1xGC	18968003	0.762	6548104	0.709	364180	0.613	392758	0.460

Continued on next page

Continued from previous page

		prokaryotic		plant		human		mouse	
AA	model	codons	variance explained	codons	variance explained	codons	variance explained	codons	variance explained
leu	2xGC	32576166	0.949	4715407	0.927	630429	0.824	648195	0.677
ser	AT	14068344	0.501	6278147	0.480	441343	0.414	463796	0.330
ser	GC	20994445	0.501	5188879	0.480	453625	0.414	460583	0.330
ser4	AT	9301232	0.639	4678203	0.626	304890	0.478	319399	0.350
ser4	GC	12729104	0.639	3410580	0.626	240043	0.478	243426	0.350
ser2	AT	4767112	0.916	1599944	0.851	136453	0.523	144397	0.353
ser2	GC	8265341	0.916	1778299	0.851	213582	0.523	217157	0.353

Table A.2: Prokaryote codon usage correspondence with model

codon, AA and model	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
phe TTT (AT)	10988263	0.898	-0.951	8.298	3.87E-16
phe TTC (GC)	12013093	0.898	0.951	8.298	3.87E-16
tyr TAT (AT)	9201253	0.742	-0.880	2.278	0.023
tyr TAC (GC)	7867529	0.742	0.880	2.278	0.023
his CAT (AT)	6178184	0.706	-0.881	2.053	0.0404
his CAC (GC)	6316853	0.706	0.881	2.053	0.0404
gln CAA (AT)	9051291	0.861	-0.929	2.888	0.00397
gln CAG (GC)	12741003	0.861	0.929	2.888	0.00397
asn AAT (AT)	11057828	0.911	-0.956	1.544	0.123
asn AAC (GC)	10469487	0.911	0.956	1.544	0.123
lys AAA (AT)	15857845	0.839	-0.916	8.031	3.04E-15
lys AAG (GC)	12101933	0.839	0.916	8.031	3.04E-15
asp GAT (AT)	16215879	0.887	-0.945	7.500	1.54E-13
asp GAC (GC)	15956872	0.887	0.945	7.500	1.54E-13
glu GAA (AT)	19825452	0.680	-0.878	3.435	0.00062
glu GAG (GC)	15923811	0.680	0.878	3.435	0.00062
cys TGT (AT)	2211343	0.890	-0.944	1.025	0.306
cys TGC (GC)	3447583	0.900	0.950	0.757	0.449
ile ATT (AT)	13168072	0.715	-0.867	15.149	2.59E-46
ile ATA (AT)	5723248	0.408	-0.667	5.630	2.41E-08
ile ATC (GC)	16372164	0.922	0.962	9.145	3.94E-19
val GTT (AT)	8744470	0.868	-0.946	4.128	4E-05
val GTA (AT)	6089810	0.832	-0.912	7.445	2.28E-13
val GTC (GC)	12471508	0.811	0.904	1.462	0.144
val GTG (GC)	15404851	0.764	0.877	3.810	0.000148
pro CCT (AT)	4590218	0.854	-0.931	0.114	0.909
pro CCA (AT)	4597584	0.771	-0.892	0.751	0.453
pro CCC (GC)	6619106	0.670	0.828	0.499	0.618
pro CCG (GC)	11285562	0.705	0.888	0.151	0.88
thr ACT (AT)	5227150	0.852	-0.927	5.888	5.54E-09
thr ACA (AT)	5741935	0.842	-0.924	3.687	0.000241
thr ACC (GC)	12335617	0.736	0.902	2.637	0.00852
thr ACG (GC)	8355095	0.536	0.760	3.692	0.000236
ala GCT (AT)	9025169	0.858	-0.934	1.310	0.191
ala GCA (AT)	9959366	0.849	-0.922	2.999	0.00278
ala GCC (GC)	21341987	0.812	0.919	2.487	0.0131
ala GCG (GC)	19085725	0.711	0.845	0.917	0.359

Continued on next page

Continued from previous page

codon, AA and model	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
gly GGT (AT)	10125504	0.707	-0.843	6.352	3.38E-10
gly GGA (AT)	7242526	0.614	-0.789	8.519	6.77E-17
gly GGC (GC)	20415518	0.738	0.926	1.473	0.141
gly GGG (GC)	7189464	-2.022	0.455	6.542	1.02E-10
arg AGA (0xGC)	3344449	0.720	-0.786	17.291	5.21E-58
arg CGT (1xGC)	6181904	0.295	-0.515	12.138	1.71E-31
arg CGA (1xGC)	2530646	0.141	-0.443	9.502	1.82E-20
arg AGG (1xGC)	2328501	0.043	-0.208	4.328	1.68E-05
arg CGC (2xGC)	14023053	0.734	0.877	0.377	0.707
arg CGG (2xGC)	6608503	0.580	0.827	2.905	0.00377
leu TTA (0xGC)	9114156	0.920	-0.915	21.850	3.88E-85
leu CTT (1xGC)	7454352	0.444	-0.691	11.810	5.19E-30
leu CTA (1xGC)	3264509	0.375	-0.766	7.379	3.64E-13
leu TTG (1xGC)	8249142	0.435	-0.409	16.328	1.19E-52
leu CTC (2xGC)	11017176	0.601	0.818	3.486	0.000514
leu CTG (2xGC)	21558990	0.737	0.913	1.166	0.244
ser TCC (GC)	5747338	0.657	0.811	0.686	0.493
ser TCG (GC)	6981766	0.718	0.878	6.632	5.73E-11
ser TCT (AT)	4722792	0.839	-0.928	2.023	0.0433
ser TCA (AT)	4578440	0.824	-0.911	0.660	0.51
ser AGC (GC)	8265341	0.655	0.810	6.025	2.47E-09
ser AGT (AT)	4767112	0.856	-0.925	3.650	0.000277
ser4 TCC (GC)	5747338	0.702	0.838	2.398	0.0167
ser4 TCG (GC)	6981766	0.732	0.891	5.477	5.64E-08
ser4 TCT (AT)	4722792	0.861	-0.946	2.142	0.0325
ser4 TCA (AT)	4578440	0.820	-0.913	2.669	0.00774
ser2 AGC (GC)	8265341	0.916	0.960	5.257	1.83E-07
ser2 AGT (AT)	4767112	0.916	-0.960	5.257	1.83E-07

Table A.3: Plant codon usage correspondence with model

codon, AA and model	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
phe TTT (AT)	2594593	0.823	-0.908	3.594	0.000413
phe TTC (GC)	2905632	0.823	0.908	3.594	0.000413
tyr TAT (AT)	1870051	0.807	-0.902	1.366	0.173
tyr TAC (GC)	2040582	0.807	0.902	1.366	0.173
his CAT (AT)	1784746	0.818	-0.905	0.753	0.452

Continued on next page

Continued from previous page

codon, AA and model	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
his CAC (GC)	1509844	0.818	0.905	0.753	0.452
gln CAA (AT)	2540964	0.698	-0.836	1.410	0.16
gln CAG (GC)	2537456	0.698	0.836	1.410	0.16
asn AAT (AT)	2842844	0.814	-0.910	3.040	0.0027
asn AAC (GC)	2818743	0.814	0.910	3.040	0.0027
lys AAA (AT)	3353480	0.779	-0.884	0.970	0.333
lys AAG (GC)	4528104	0.779	0.884	0.970	0.333
asp GAT (AT)	4314822	0.875	-0.945	1.042	0.299
asp GAC (GC)	3185110	0.875	0.945	1.042	0.299
glu GAA (AT)	4111542	0.832	-0.912	1.070	0.286
glu GAG (GC)	4611173	0.832	0.912	1.070	0.286
cys TGT (AT)	1115436	0.724	-0.853	1.941	0.0537
cys TGC (GC)	1193908	0.710	0.845	1.428	0.155
ile ATT (AT)	2819055	0.666	-0.845	0.098	0.922
ile ATA (AT)	1499041	0.561	-0.766	0.856	0.393
ile ATC (GC)	2697546	0.818	0.921	0.627	0.531
val GTT (AT)	2854297	0.737	-0.918	0.982	0.327
val GTA (AT)	1192072	0.374	-0.776	2.162	0.0318
val GTC (GC)	2247969	0.542	0.755	0.802	0.423
val GTG (GC)	2677629	0.319	0.570	0.661	0.509
pro CCT (AT)	2211535	0.538	-0.737	0.635	0.526
pro CCA (AT)	2101944	0.542	-0.751	2.351	0.0197
pro CCC (GC)	1405766	0.594	0.780	0.243	0.808
pro CCG (GC)	1538777	0.554	0.753	2.096	0.0374
thr ACT (AT)	2102544	0.695	-0.840	4.373	2.01E-05
thr ACA (AT)	1919288	0.725	-0.859	2.031	0.0436
thr ACC (GC)	1926505	0.663	0.830	1.979	0.0493
thr ACG (GC)	1313788	0.475	0.690	3.092	0.00228
ala GCT (AT)	3224056	0.654	-0.826	6.337	1.61E-09
ala GCA (AT)	2391150	0.703	-0.843	3.266	0.00129
ala GCC (GC)	2773712	0.721	0.870	1.401	0.163
ala GCG (GC)	2143751	0.580	0.764	4.145	5.09E-05
gly GGT (AT)	2638534	0.393	-0.627	1.019	0.31
gly GGA (AT)	2512456	0.551	-0.742	0.222	0.824
gly GGC (GC)	2494847	0.670	0.939	4.163	4.73E-05
gly GGG (GC)	1691717	-1.161	0.040	1.613	0.108
arg AGA (0xGC)	1941778	0.624	-0.803	1.230	0.22
arg CGT (1xGC)	1107243	0.001	-0.159	0.430	0.667

Continued on next page

Continued from previous page

codon, AA and model	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
arg CGA (1xGC)	960894	-0.062	0.014	0.177	0.859
arg AGG (1xGC)	1617506	0.104	-0.350	2.639	0.00898
arg CGC (2xGC)	1297614	0.706	0.875	3.532	0.000517
arg CGG (2xGC)	1054230	-0.266	0.662	1.399	0.163
leu TTA (0xGC)	1490706	0.476	-0.779	5.638	6.04E-08
leu CTT (1xGC)	2579575	0.410	-0.713	3.318	0.00109
leu CTA (1xGC)	1264980	0.331	-0.594	3.207	0.00157
leu TTG (1xGC)	2703549	0.381	-0.769	0.930	0.353
leu CTC (2xGC)	2546375	0.602	0.792	1.315	0.19
leu CTG (2xGC)	2169032	0.670	0.822	2.087	0.0382
ser TCC (GC)	1959155	0.448	0.670	2.854	0.00479
ser TCG (GC)	1451425	0.587	0.774	2.640	0.00896
ser TCT (AT)	2603305	0.682	-0.840	0.952	0.342
ser TCA (AT)	2074898	0.732	-0.874	1.252	0.212
ser AGC (GC)	1778299	0.664	0.830	0.138	0.89
ser AGT (AT)	1599944	0.654	-0.818	0.402	0.688
ser4 TCC (GC)	1959155	0.632	0.796	3.027	0.00281
ser4 TCG (GC)	1451425	0.628	0.806	3.028	0.0028
ser4 TCT (AT)	2603305	0.723	-0.855	1.333	0.184
ser4 TCA (AT)	2074898	0.730	-0.862	0.868	0.387
ser2 AGC (GC)	1778299	0.851	0.924	1.724	0.0864
ser2 AGT (AT)	1599944	0.851	-0.924	1.724	0.0864

Table A.4: Human codon usage correspondence with model

codon, AA and model	genes	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
phe TTT (AT)	14030	173436	0.679	-0.824	4.803	1.58E-06
phe TTC (GC)	14030	192286	0.679	0.824	4.803	1.58E-06
tyr TAT (AT)	11445	113688	0.652	-0.807	8.966	3.55E-19
tyr TAC (GC)	11445	133052	0.652	0.807	8.966	3.55E-19
his CAT (AT)	10656	103321	0.688	-0.831	4.026	5.71E-05
his CAC (GC)	10656	136463	0.688	0.831	4.026	5.71E-05
gln CAA (AT)	15160	130992	0.601	-0.795	8.000	1.33E-15
gln CAG (GC)	15160	355169	0.601	0.795	8.000	1.33E-15
asn AAT (AT)	12803	172239	0.690	-0.831	6.833	8.67E-12
asn AAC (GC)	12803	178538	0.690	0.831	6.833	8.67E-12
lys AAA (AT)	16082	267460	0.700	-0.836	5.753	8.94E-09

Continued on next page

Continued from previous page

codon, AA and model	genes	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
lys AAG (GC)	16082	330208	0.700	0.836	5.753	8.94E-09
asp GAT (AT)	14945	229567	0.716	-0.846	7.267	3.87E-13
asp GAC (GC)	14945	254208	0.716	0.846	7.267	3.87E-13
glu GAA (AT)	16923	323804	0.782	-0.893	5.800	6.73E-09
glu GAG (GC)	16923	425224	0.782	0.893	5.800	6.73E-09
cys TGT (AT)	9247	94049	0.610	-0.781	3.025	0.0025
cys TGC (GC)	9247	104877	0.610	0.781	3.025	0.0025
ile ATT (AT)	14420	164370	0.512	-0.748	15.424	3.01E-53
ile ATA (AT)	14420	78242	0.478	-0.703	8.382	5.68E-17
ile ATC (GC)	14420	197793	0.753	0.887	8.384	5.59E-17
val GTT (AT)	17223	117047	0.593	-0.784	4.195	2.74E-05
val GTA (AT)	17223	76089	0.486	-0.709	19.607	1.13E-84
val GTC (GC)	17223	147911	0.091	0.438	16.664	7.27E-62
val GTG (GC)	17223	289482	0.502	0.751	0.567	0.571
pro CCT (AT)	17122	191056	0.403	-0.635	13.046	1.02E-38
pro CCA (AT)	17122	184727	0.437	-0.662	3.638	0.000275
pro CCC (GC)	17122	215796	0.484	0.718	2.320	0.0203
pro CCG (GC)	17122	77755	0.277	0.564	15.993	3.68E-57
thr ACT (AT)	16399	140788	0.461	-0.682	6.609	3.99E-11
thr ACA (AT)	16399	159289	0.431	-0.659	5.761	8.5E-09
thr ACC (GC)	16399	190611	0.503	0.746	2.539	0.0111
thr ACG (GC)	16399	61642	0.317	0.597	18.307	3.97E-74
ala GCT (AT)	18263	196875	0.442	-0.665	13.045	1.01E-38
ala GCA (AT)	18263	172509	0.486	-0.698	2.508	0.0121
ala GCC (GC)	18263	299152	0.505	0.746	1.420	0.156
ala GCG (GC)	18263	81620	0.197	0.517	12.858	1.13E-37
gly GGT (AT)	17891	113971	0.217	-0.556	4.274	1.93E-05
gly GGA (AT)	17891	176813	0.556	-0.779	3.588	0.000334
gly GGC (GC)	17891	237875	0.508	0.750	16.087	8.02E-58
gly GGG (GC)	17891	174824	0.108	0.442	17.050	1.14E-64
arg AGA (0xGC)	17232	128098	0.586	-0.796	8.542	1.43E-17
arg CGT (1xGC)	17232	47554	-0.070	-0.214	1.047	0.295
arg CGA (1xGC)	17232	64797	0.107	-0.335	10.183	2.76E-24
arg AGG (1xGC)	17232	125902	0.034	-0.113	26.824	2.62E-155
arg CGC (2xGC)	17232	109870	0.529	0.718	25.134	6.19E-137
arg CGG (2xGC)	17232	121912	0.392	0.646	3.188	0.00144
leu TTA (0xGC)	19432	85743	0.366	-0.759	35.358	1.86E-265
leu CTT (1xGC)	19432	145855	0.346	-0.715	1.029	0.303

Continued on next page

Continued from previous page

codon, AA and model	genes	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
leu CTA (1xGC)	19432	77598	0.209	-0.513	3.581	0.000343
leu TTG (1xGC)	19432	140727	0.299	-0.560	18.149	5.22E-73
leu CTC (2xGC)	19432	207573	-0.012	0.519	16.304	2.28E-59
leu CTG (2xGC)	19432	422856	0.635	0.848	8.463	2.78E-17
ser TCC (GC)	18651	191285	0.307	0.555	9.142	6.72E-20
ser TCG (GC)	18651	48758	0.121	0.493	15.613	1.31E-54
ser TCT (AT)	18651	167659	0.371	-0.613	9.239	2.75E-20
ser TCA (AT)	18651	137231	0.404	-0.637	4.151	3.33E-05
ser AGC (GC)	18651	213582	0.399	0.669	5.664	1.5E-08
ser AGT (AT)	18651	136453	0.354	-0.596	2.162	0.0306
ser4 TCC (GC)	18649	191285	0.416	0.672	4.502	6.78E-06
ser4 TCG (GC)	18649	48758	0.234	0.529	18.592	1.84E-76
ser4 TCT (AT)	18649	167659	0.355	-0.597	11.431	3.7E-30
ser4 TCA (AT)	18649	137231	0.371	-0.610	1.662	0.0965
ser2 AGC (GC)	18633	213582	0.523	0.724	2.230	0.0258
ser2 AGT (AT)	18633	136453	0.523	-0.724	2.230	0.0258

Table A.5: Mouse codon usage correspondence with model

codon, AA and model	genes	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
phe TTT (AT)	15071	175479	0.528	-0.728	2.382	0.0172
phe TTC (GC)	15071	211917	0.528	0.728	2.382	0.0172
tyr TAT (AT)	12370	113750	0.440	-0.664	3.462	0.000538
tyr TAC (GC)	12370	144220	0.440	0.664	3.462	0.000538
his CAT (AT)	10879	101502	0.495	-0.704	9.087	1.19E-19
his CAC (GC)	10879	136838	0.495	0.704	9.087	1.19E-19
gln CAA (AT)	15377	127313	0.448	-0.679	17.169	1.86E-65
gln CAG (GC)	15377	356614	0.448	0.679	17.169	1.86E-65
asn AAT (AT)	14002	159909	0.487	-0.698	10.347	5.3E-25
asn AAC (GC)	14002	196444	0.487	0.698	10.347	5.3E-25
lys AAA (AT)	17426	246815	0.545	-0.739	3.992	6.57E-05
lys AAG (GC)	17426	356853	0.545	0.739	3.992	6.57E-05
asp GAT (AT)	15506	223640	0.473	-0.691	11.141	1.02E-28
asp GAC (GC)	15506	267643	0.473	0.691	11.141	1.02E-28
glu GAA (AT)	17305	306083	0.613	-0.788	15.135	2.03E-51
glu GAG (GC)	17305	427625	0.613	0.788	15.135	2.03E-51
cys TGT (AT)	9816	101605	0.414	-0.644	4.657	3.25E-06

Continued on next page

Continued from previous page

codon, AA and model	genes	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
cys TGC (GC)	9816	103613	0.414	0.644	4.657	3.25E-06
ile ATT (AT)	15792	161555	0.326	-0.599	14.405	9.55E-47
ile ATA (AT)	15792	78781	0.353	-0.611	5.863	4.63E-09
ile ATC (GC)	15792	221570	0.611	0.809	9.548	1.51E-21
val GTT (AT)	18399	118177	0.429	-0.672	10.683	1.46E-26
val GTA (AT)	18399	81097	0.272	-0.538	18.920	4.37E-79
val GTC (GC)	18399	161901	0.017	0.300	12.312	1.06E-34
val GTG (GC)	18399	298313	0.355	0.638	6.498	8.35E-11
pro CCT (AT)	17864	201542	0.199	-0.448	14.936	3.84E-50
pro CCA (AT)	17864	189976	0.235	-0.485	20.332	7.15E-91
pro CCC (GC)	17864	194858	0.267	0.523	24.379	3.62E-129
pro CCG (GC)	17864	67573	0.178	0.446	11.900	1.58E-32
thr ACT (AT)	17609	146815	0.265	-0.515	7.597	3.18E-14
thr ACA (AT)	17609	172440	0.278	-0.531	1.857	0.0633
thr ACC (GC)	17609	193904	0.356	0.636	2.752	0.00593
thr ACG (GC)	17609	58436	0.132	0.432	8.358	6.86E-17
ala GCT (AT)	19101	217708	0.221	-0.472	22.327	5.05E-109
ala GCA (AT)	19101	175422	0.336	-0.584	3.250	0.00116
ala GCC (GC)	19101	279002	0.328	0.593	12.976	2.42E-38
ala GCG (GC)	19101	71836	0.151	0.431	13.155	2.35E-39
gly GGT (AT)	18700	122576	0.066	-0.372	10.790	4.63E-27
gly GGA (AT)	18700	181516	0.385	-0.652	2.149	0.0316
gly GGC (GC)	18700	225304	0.375	0.654	9.533	1.72E-21
gly GGG (GC)	18700	162545	-0.008	0.283	17.654	3.44E-69
arg AGA (0xGC)	18001	132926	0.444	-0.721	0.007	0.994
arg CGT (1xGC)	18001	49789	-0.106	-0.029	0.225	0.822
arg CGA (1xGC)	18001	71342	-0.023	-0.105	2.383	0.0172
arg AGG (1xGC)	18001	132693	0.046	-0.133	25.956	7.53E-146
arg CGC (2xGC)	18001	99327	0.417	0.645	24.260	5.83E-128
arg CGG (2xGC)	18001	110772	0.251	0.517	3.723	0.000197
leu TTA (0xGC)	20983	78921	0.221	-0.652	2.255	0.0242
leu CTT (1xGC)	20983	152022	0.233	-0.595	3.785	0.000154
leu CTA (1xGC)	20983	90311	0.082	-0.322	6.848	7.7E-12
leu TTG (1xGC)	20983	150425	0.148	-0.383	20.290	1.17E-90
leu CTC (2xGC)	20983	218998	-0.070	0.344	16.982	3.02E-64
leu CTG (2xGC)	20983	429197	0.496	0.753	4.377	1.21E-05
ser TCC (GC)	20080	196513	0.177	0.421	9.349	9.74E-21
ser TCG (GC)	20080	46913	0.067	0.391	10.816	3.43E-27

Continued on next page

Continued from previous page

codon, AA and model	genes	codon count	variance explained	correlation coefficient	Harvey-Collier Statistic	p-value
ser TCT (AT)	20080	183526	0.221	-0.474	16.122	4.12E-58
ser TCA (AT)	20080	135873	0.239	-0.489	11.003	4.43E-28
ser AGC (GC)	20080	217157	0.274	0.563	7.976	1.59E-15
ser AGT (AT)	20080	144397	0.208	-0.457	21.902	4.18E-105
ser4 TCC (GC)	20079	196513	0.274	0.544	1.076	0.282
ser4 TCG (GC)	20079	46913	0.142	0.423	13.055	8.6E-39
ser4 TCT (AT)	20079	183526	0.212	-0.462	1.855	0.0637
ser4 TCA (AT)	20079	135873	0.204	-0.452	6.809	1.01E-11
ser2 AGC (GC)	20067	217157	0.353	0.594	5.408	6.43E-08
ser2 AGT (AT)	20067	144397	0.353	-0.594	5.408	6.43E-08

Table A.6: Genes closer than 2MB by tissue

Tissue	Gene1	Gene2	Chrom	Distance
BM	Il6ra	Tpm3	3	159490
BM	Tpm3	S100a7a	3	591730
BM	Bex6	Muc13	16	1607088
EAT	Fgf23	AC123060.4	6	1357647
EB5d	AI894139	Zfp777	6	58900
EB5d	Cbln1	Zfp423	8	189216
EB5d	Zfp423	Nkd1	8	561799
EB5d	Punc	Zfp609	9	509605
EB5d	Myt1l	Dus4l	12	1719213
EB5d	Skiv2l2	Ndufs4	13	1361271
EB5d	Clybl	Tmtc4	14	516747
EB5d	Htr7	Tnks2	19	777178
ECT	Jph2	Wisp2	2	422906
ECT	Cgref1	Fosl2	5	1190442
ECT	Snx29	C530044N13Rik	16	48255
EDT	BC030307	Pah	10	745088
EDT	Krt6b	Krt5	15	26781
EDT	Krt5	Gpr84	15	1595348
EFB	Hoxa1	Hoxa2	6	4198
EFB	Foxf2	Serpinb1b	13	1452767
EFB	Gpr64	Scml2	X	619123
ES2d	Tmem79	Pklr	3	801720
ES2d	Mnx1	Slc5a6	5	1533399
ES2d	Susd2	Slc19a1	10	1389279

Continued on next page

Continued from previous page

Tissue	Gene1	Gene2	Chrom	Distance
ES2d	Kank3	Psors1c2	17	1710287
GEE13.5	Gad2	Cacna1b	2	1910017
GEE13.5	Cacna1b	Bmyc	2	943678
GEE13.5	Bmyc	Med22	2	1197597
GEE13.5	Abl1	Slc25a25	2	610261
GEE13.5	Stk39	Klhl23	2	1350205
GEE13.5	Klhl23	Gad1	2	725516
GEE13.5	9830001H06Rik	Vstm2l	2	835399
GEE13.5	Vstm2l	Slc32a1	2	666039
GEE13.5	Glrb	Mtap9	3	1444455
GEE13.5	Tdrkh	Sv2a	3	1749696
GEE13.5	Magi3	6530418L21Rik	3	1483434
GEE13.5	Srd5a2l2	Epha5	5	702596
GEE13.5	Rph3a	A930024E05Rik	5	1979836
GEE13.5	Slc29a4	Rnf216	5	554129
GEE13.5	Rnf216	Zfp12	5	122214
GEE13.5	Zfp12	Zfp316	5	3779
GEE13.5	Zfp316	Tmem130	5	1465166
GEE13.5	Asb4	Dlx5	6	1444783
GEE13.5	Hdac11	Wnt7a	6	189301
GEE13.5	Atp2b2	Slc6a11	6	89215
GEE13.5	Zfp61	Grik5	7	710533
GEE13.5	Ntrk3	Polg	7	871408
GEE13.5	Sez6l2	Zfp629	7	636425
GEE13.5	Deaf1	Tmem80	7	552
GEE13.5	Tmem80	Brsk2	7	612602
GEE13.5	Lig4	Ing1	8	1579766
GEE13.5	Nr2c2ap	Rab3a	8	620972
GEE13.5	BC060632	Terf2ip	8	1270007
GEE13.5	Sorl1	Tecta	9	205334
GEE13.5	Slc38a3	Celsr3	9	1158921
GEE13.5	Ascc2	Zmiz2	11	1703770
GEE13.5	Zmiz2	Adcy1	11	657327
GEE13.5	Nt5m	Mapk7	11	1612282
GEE13.5	Mapk7	3110043A19Rik	11	10190
GEE13.5	3110043A19Rik	Zfp286	11	1273332
GEE13.5	Efnb3	0610025P10Rik	11	341713
GEE13.5	0610025P10Rik	Wscd1	11	1836883
GEE13.5	Myo18a	Sez6	11	64963

Continued on next page

Continued from previous page

Tissue	Gene1	Gene2	Chrom	Distance
GEE13.5	D030028A08Rik	Stard3	11	1398322
GEE13.5	Stard3	Thra	11	360802
GEE13.5	Grin2c	2310004N24Rik	11	931646
GEE13.5	D11Bwg0517e	2900052L18Rik	11	1318282
GEE13.5	Galntl1	Smoc1	12	422935
GEE13.5	Gprin1	Sncb	13	9192
GEE13.5	Wdr41	Crhbp	13	408057
GEE13.5	4931414P19Rik	Thtpa	14	488927
GEE13.5	Triobp	Sox10	15	149214
GEE13.5	Sox10	AW544981	15	1506867
GEE13.5	Ccnt1	C230021P08Rik	15	851112
GEE13.5	C230021P08Rik	Accn2	15	212970
GEE13.5	Sema5b	Slc15a2	16	1085575
GEE13.5	Dscam	Prdm15	16	621244
GEE13.5	A630033E08Rik	Caskin1	17	1621657
GEE13.5	4930506M07Rik	Rab11fip2	19	829009
MAST	Med12l	Sucnr1	3	767115
MAST	Tnfaip8l2	Pex11b	3	1493076
MAST	Pex11b	EG433632	3	1284807
MAST	Slc30a2	Npal3	4	1092705
MAST	Npal3	Cnr2	4	400872
MAST	Adam8	Pkp3	7	1085789
MAST	Cx3cr1	Entpd3	9	471563
MAST	Oit3	Ccdc109a	10	6853
MAST	Pctk2	Lta4h	10	212570
MAST	Phospho1	Abi3	11	1751
MAST	Mboat1	1700018A04Rik	13	1331464
MAST	Ltb4r1	Cma1	14	172961
MAST	6330416L07Rik	Zfp40	17	812481
MAST	Gna14	Gcnt1	19	717855
MAST	Cd274	Pdcd1lg2	19	22825
MMS	C030014K22Rik	Lrrc52	1	354670
MMS	Tmem87b	Prosapip1	2	1778511
MMS	Prosapip1	Adam33	2	408265
MMS	Adam33	Spef1	2	106447
MMS	Spef1	2310035K24Rik	2	23142
MMS	2310035K24Rik	Prokr2	2	1127325
MMS	Hdhd3	Col27a1	4	713198
MMS	Hspb7	Zbtb17	4	19404

Continued on next page

Continued from previous page

Tissue	Gene1	Gene2	Chrom	Distance
MMS	Cmklr1	Usp30	5	449943
MMS	Gpr133	Vkorc1l1	5	707509
MMS	Chst12	Iqce	5	138269
MMS	C630007B19Rik	Mitf	6	1149860
MMS	Tspan11	Clec1a	6	1473657
MMS	Clec1a	Olr1	6	33425
MMS	5430432N15Rik	Hps5	7	1114698
MMS	Fbrs	Orai3	7	278931
MMS	Pstk	Gpr26	7	626793
MMS	Kcnq1	Osbpl5	7	261712
MMS	C230081A13Rik	Stra6	9	1711021
MMS	Glyctk	Dusp7	9	210494
MMS	Cacna1g	Hoxb6	11	1824974
MMS	Hoxb6	Hoxb2	11	50066
MMS	Hoxb2	Stac2	11	1682608
MMS	Tmem106a	Meox1	11	285736
MMS	Trim47	Mrpl38	11	21580
MMS	Mrpl38	Cygb	11	507282
MMS	Scfd1	Strn3	12	159549
MMS	Phactr1	Gfod1	13	57035
MMS	Gfod1	Rnf182	13	311809
MMS	Tspan17	Unc5a	13	152664
MMS	Nkd2	Cast	13	846674
MMS	Fam167a	Hmbox1	14	1351549
MMS	Hmbox1	Scara5	14	716556
MMS	Scara5	Kctd9	14	1950662
MMS	Mapk11	Abcd2	15	1996279
MMS	Krt73	Tenc1	15	300603
MMS	Hs3st6	C230078M08Rik	17	1684861
MMS	C230078M08Rik	Phf1	17	486814
MMS	Phf1	Lemd2	17	251868
MMS	Lemd2	Slc26a8	17	1433361
MMS	Entpd1	Ubtd1	19	1240161
SBM	Ermap	Scmh1	4	1215433
SBM	Tmem120b	Rsrc2	5	611106
SBM	Rsrc2	Setd8	5	690518
SBM	Wnk1	Mug1	6	1800762
SBM	Cd79a	Zfp526	7	319594
SBM	Zfp689	1110007A13Rik	7	1248609

Continued on next page

Continued from previous page

Tissue	Gene1	Gene2	Chrom	Distance
SBM	Klf1	Mylk3	8	419162
SBM	Cdyl2	Pkd1l2	8	262688
SBM	Ankrd43	Slc22a4	11	502852
SBM	Slc22a4	Anxa6	11	951018
SBM	Zfp39	A230051G13Rik	11	251144
SBM	Fntb	Plek2	12	1968052
SBM	Ccdc88c	Smek1	12	10540
SBM	Ctsb	Gata4	14	55781
SBM	Xrcc6	C920005C14Rik	15	301123

Table A.7: Genes closer than 2MB by lineage

Lineage	Gene1	Gene2	Chrom	Distance
External Ectoderm	AC119809.10	Il1r2	1	363745
External Ectoderm	C030014K22Rik	Lrrc52	1	354670
External Ectoderm	Disp2	Dll4	2	514523
External Ectoderm	Tmem87b	Prosapip1	2	1778511
External Ectoderm	Prosapip1	Adam33	2	408265
External Ectoderm	Adam33	Spef1	2	106447
External Ectoderm	Spef1	2310035K24Rik	2	23142
External Ectoderm	2310035K24Rik	Prokr2	2	1127325
External Ectoderm	Fcgr1	Itga10	3	351328
External Ectoderm	Casq2	Syt6	3	1428721
External Ectoderm	Hdhd3	Col27a1	4	713198
External Ectoderm	Rspo1	Zc3h12a	4	109370
External Ectoderm	Hspb7	Zbtb17	4	19404
External Ectoderm	Cmklr1	Usp30	5	449943
External Ectoderm	Gpr133	Vkorc1l1	5	707509
External Ectoderm	Chst12	Iqce	5	138269
External Ectoderm	C630007B19Rik	Mitf	6	1149860
External Ectoderm	Atg7	Plxnd1	6	1094216
External Ectoderm	Fgf23	Tspan11	6	806214
External Ectoderm	Tspan11	AC123060.4	6	486024
External Ectoderm	AC123060.4	Clec1a	6	950841
External Ectoderm	Clec1a	Olr1	6	33425
External Ectoderm	5430432N15Rik	Hps5	7	1114698
External Ectoderm	Fbrs	Orai3	7	278931
External Ectoderm	Pstk	Gpr26	7	626793
External Ectoderm	Kcnq1	Osbpl5	7	261712

Continued on next page

Continued from previous page

Lineage	Gene1	Gene2	Chrom	Distance
External Ectoderm	Rab20	Gas6	8	1986738
External Ectoderm	Nat2	Slc18a1	8	1535132
External Ectoderm	C230081A13Rik	Stra6	9	1711021
External Ectoderm	Clstn2	Mrps22	9	555760
External Ectoderm	Acpp	Glyctk	9	1815139
External Ectoderm	Glyctk	Dusp7	9	210494
External Ectoderm	BC030307	Pah	10	745088
External Ectoderm	Cacna1g	Hoxb6	11	1824974
External Ectoderm	Hoxb6	Hoxb2	11	50066
External Ectoderm	Hoxb2	Stac2	11	1682608
External Ectoderm	Aoc3	Tmem106a	11	240304
External Ectoderm	Tmem106a	Meox1	11	285736
External Ectoderm	Trim47	Mrpl38	11	21580
External Ectoderm	Mrpl38	Cygb	11	507282
External Ectoderm	Scfd1	Strn3	12	159549
External Ectoderm	Phactr1	Gfod1	13	57035
External Ectoderm	Gfod1	Rnf182	13	311809
External Ectoderm	Tspan17	Unc5a	13	152664
External Ectoderm	Nkd2	Cast	13	846674
External Ectoderm	Fam167a	Hmbox1	14	1351549
External Ectoderm	Hmbox1	Scara5	14	716556
External Ectoderm	Scara5	Kctd9	14	1950662
External Ectoderm	Scel	Pou4f1	14	848987
External Ectoderm	C1qtnf6	Sstr3	15	7397
External Ectoderm	Mapk11	Abcd2	15	1996279
External Ectoderm	Krt6b	Krt5	15	26781
External Ectoderm	Krt5	Krt73	15	80417
External Ectoderm	Krt73	Tenc1	15	300603
External Ectoderm	Tenc1	Gpr84	15	1191848
External Ectoderm	Stfa1	Pla1a	16	1940727
External Ectoderm	Hs3st6	C230078M08Rik	17	1684861
External Ectoderm	C230078M08Rik	Nkx2-5	17	392352
External Ectoderm	Nkx2-5	Phf1	17	91562
External Ectoderm	Phf1	Lemd2	17	251868
External Ectoderm	Lemd2	Slc26a8	17	1433361
External Ectoderm	Fosl1	0610038D11Rik	19	1528709
External Ectoderm	Entpd1	Ubtd1	19	1240161
Mesoderm	Lincr	Zap70	1	488436
Mesoderm	Dyrk3	Slc45a3	1	824733

Continued on next page

Continued from previous page

Lineage	Gene1	Gene2	Chrom	Distance
Mesoderm	Slc45a3	Atp2b4	1	1723683
Mesoderm	Oit3	Ccdc109a	10	6853
Mesoderm	Fgf22	Zbtb7a	10	1377675
Mesoderm	Zbtb7a	Glt8d2	10	1497438
Mesoderm	Pctk2	Lta4h	10	212570
Mesoderm	Lgr5	Tspan8	10	229491
Mesoderm	Rel	5730522E02Rik	11	1848117
Mesoderm	Ankrd43	Slc22a4	11	502852
Mesoderm	Slc22a4	Anxa6	11	951018
Mesoderm	4930438A08Rik	Zfp39	11	593864
Mesoderm	Zfp39	A230051G13Rik	11	251144
Mesoderm	A230051G13Rik	Nlrp3	11	377826
Mesoderm	Dnahc2	Slc2a4	11	393178
Mesoderm	1300007F04Rik	Nek8	11	627501
Mesoderm	Tbx4	Mpo	11	1877486
Mesoderm	Phospho1	Abi3	11	1751
Mesoderm	Wdr68	Gm885	11	696034
Mesoderm	Fn3k	Metrnl	11	251055
Mesoderm	Fntb	Plek2	12	1968052
Mesoderm	Plek2	Wdr22	12	1429739
Mesoderm	9030617O03Rik	Ccdc88c	12	15719
Mesoderm	Ccdc88c	Smek1	12	10540
Mesoderm	Mboat1	1700018A04Rik	13	1331464
Mesoderm	1700018A04Rik	Foxf2	13	270
Mesoderm	Foxf2	Mylk4	13	1077477
Mesoderm	Mylk4	Serpinb9b	13	245698
Mesoderm	Serpinb9b	Serpinb1b	13	43593
Mesoderm	Ltb4r1	Cma1	14	172961
Mesoderm	Ctsb	Gata4	14	55781
Mesoderm	C1qtnf3	Slc45a2	15	20585
Mesoderm	Dgat1	Apol8	15	1235981
Mesoderm	Apol8	Mfng	15	1000634
Mesoderm	Xrcc6	Naga	15	289448
Mesoderm	Naga	C920005C14Rik	15	2394
Mesoderm	Socs1	Snx29	16	620112
Mesoderm	Snx29	C530044N13Rik	16	48255
Mesoderm	Rtn4r	Tbx1	16	429307
Mesoderm	Thpo	Ephb3	16	470269
Mesoderm	Gp5	Bex6	16	1869021

Continued on next page

Continued from previous page

Lineage	Gene1	Gene2	Chrom	Distance
Mesoderm	Bex6	Muc13	16	1607088
Mesoderm	Popdc2	Upk1b	16	395001
Mesoderm	ORF63	Bach1	16	103619
Mesoderm	6330416L07Rik	Zfp40	17	812481
Mesoderm	Prss29	Msln	17	425934
Mesoderm	C920016K16Rik	H2-Ea	17	1273182
Mesoderm	H2-Ea	Notch4	17	276495
Mesoderm	2900055J20Rik	Sh3rf2	18	1795983
Mesoderm	Ptprcap	Ovol1	19	1392426
Mesoderm	Ovol1	Ehbp1l1	19	147316
Mesoderm	Gna14	Gcnt1	19	717855
Mesoderm	AC113961.14-203	Cd274	19	202377
Mesoderm	Cd274	Pdcd1lg2	19	22825
Mesoderm	Btrc	Poll	19	22262
Neuroectoderm	Armc9	Efhd1	1	986079
Neuroectoderm	Efhd1	Sh3bp4	1	1759577
Neuroectoderm	Sh3bp4	Rab17	1	1804345
Neuroectoderm	Rab17	Hdac4	1	963293
Neuroectoderm	Prox1	Slc30a1	1	1736087
Neuroectoderm	A130092J06Rik	9130404D14Rik	2	792531
Neuroectoderm	9130404D14Rik	C130021I20Rik	2	716078
Neuroectoderm	Rasgrp1	Fsip1	2	861891
Neuroectoderm	Jph2	Wisp2	2	422906
Neuroectoderm	Mc3r	2010011I20Rik	2	94523
Neuroectoderm	Gad2	Cacna1b	2	1910017
Neuroectoderm	Cacna1b	Bmyc	2	943678
Neuroectoderm	Bmyc	Med22	2	1197597
Neuroectoderm	Abl1	Slc25a25	2	610261
Neuroectoderm	Stk39	Klhl23	2	1350205
Neuroectoderm	Klhl23	Gad1	2	725516
Neuroectoderm	9830001H06Rik	Vstm2l	2	835399
Neuroectoderm	Vstm2l	Slc32a1	2	666039
Neuroectoderm	Med12l	Aadac	3	717060
Neuroectoderm	Aadac	Sucnr1	3	41717
Neuroectoderm	Il6ra	Tpm3	3	159490
Neuroectoderm	Tpm3	Dennd4b	3	165631
Neuroectoderm	Dennd4b	S100a7a	3	411982
Neuroectoderm	Tnfaip8l2	Hist2h2bb	3	1127367
Neuroectoderm	Hist2h2bb	Pex11b	3	356429

Continued on next page

Continued from previous page

Lineage	Gene1	Gene2	Chrom	Distance
Neuroectoderm	Pex11b	EG433632	3	1284807
Neuroectoderm	EG433632	Notch2	3	80709
Neuroectoderm	Dennd2d	Amigo1	3	1683305
Neuroectoderm	Dapp1	Adh7	3	236306
Neuroectoderm	Glrb	Mtap9	3	1444455
Neuroectoderm	Tdrkh	Sv2a	3	1749696
Neuroectoderm	Magi3	6530418L21Rik	3	1483434
Neuroectoderm	S1pr1	A930005H10Rik	3	166534
Neuroectoderm	Fkbp15	Zfp618	4	605025
Neuroectoderm	4922503N01Rik	4930522H14Rik	4	60624
Neuroectoderm	Ccdc24	Mpl	4	559260
Neuroectoderm	Mpl	Ermap	4	718107
Neuroectoderm	Ermap	Scmh1	4	1215433
Neuroectoderm	Scmh1	Cited4	4	136388
Neuroectoderm	Csf3r	Col8a2	4	242354
Neuroectoderm	Col8a2	Gjb4	4	1036756
Neuroectoderm	Gjb4	C77080	4	1865497
Neuroectoderm	2300002D11Rik	Slc30a2	4	844675
Neuroectoderm	Slc30a2	Npal3	4	1092705
Neuroectoderm	Npal3	Cnr2	4	400872
Neuroectoderm	Dnajb5	4930500O05Rik	4	709546
Neuroectoderm	Cgref1	Fosl2	5	1190442
Neuroectoderm	Fosl2	4933407H18Rik	5	1265261
Neuroectoderm	Aldh2	Tmem120b	5	1482717
Neuroectoderm	Tmem120b	Rsrc2	5	611106
Neuroectoderm	Rsrc2	Hip1r	5	224227
Neuroectoderm	Hip1r	Setd8	5	436731
Neuroectoderm	Gtf2ird2	Upk3b	5	1820357
Neuroectoderm	Flt3	Slc46a3	5	477948
Neuroectoderm	Srd5a2l2	Epha5	5	702596
Neuroectoderm	Rph3a	A930024E05Rik	5	1979836
Neuroectoderm	Fzd9	Vgf	5	1779065
Neuroectoderm	Slc29a4	Rnf216	5	554129
Neuroectoderm	Rnf216	Zfp12	5	122214
Neuroectoderm	Zfp12	Zfp316	5	3779
Neuroectoderm	Zfp316	Tmem130	5	1465166
Neuroectoderm	Hoxa1	Hoxa2	6	4198
Neuroectoderm	Hoxa2	Hoxa11	6	77277
Neuroectoderm	Figla	Mxd1	6	627986

Continued on next page

Continued from previous page

Lineage	Gene1	Gene2	Chrom	Distance
Neuroectoderm	Mxd1	Gata2	6	1529178
Neuroectoderm	Alox5	Rassf4	6	171830
Neuroectoderm	Wnk1	Iqsec3	6	1335154
Neuroectoderm	Iqsec3	Mug1	6	364863
Neuroectoderm	Asb4	Dlx5	6	1444783
Neuroectoderm	Hdac11	Wnt7a	6	189301
Neuroectoderm	Atp2b2	Slc6a11	6	89215
Neuroectoderm	Rasgef1a	Wnt5b	6	1340985
Neuroectoderm	Usp29	5730403M16Rik	7	147473
Neuroectoderm	Cd79a	Zfp526	7	319594
Neuroectoderm	AC164564.3	AI428936	7	1157898
Neuroectoderm	Kcne3	Dnajb13	7	316851
Neuroectoderm	Aqp8	4933440M02Rik	7	1861640
Neuroectoderm	4933440M02Rik	Itgal	7	1946488
Neuroectoderm	Itgal	Zfp689	7	108617
Neuroectoderm	Zfp689	Ctf1	7	263586
Neuroectoderm	Ctf1	1110007A13Rik	7	979578
Neuroectoderm	Adam8	Pkp3	7	1085789
Neuroectoderm	Zfp61	Grik5	7	710533
Neuroectoderm	Ntrk3	Polg	7	871408
Neuroectoderm	Sez6l2	Zfp629	7	636425
Neuroectoderm	Deaf1	Tmem80	7	552
Neuroectoderm	Tmem80	Brsk2	7	612602
Neuroectoderm	Gypa	Inpp4b	8	1204415
Neuroectoderm	Klf1	Mylk3	8	419162
Neuroectoderm	Cdyl2	Pkd1l2	8	262688
Neuroectoderm	Pcp2	BC068157	8	584013
Neuroectoderm	Lig4	Ing1	8	1579766
Neuroectoderm	Nr2c2ap	Rab3a	8	620972
Neuroectoderm	Pllp	Zfp319	8	629900
Neuroectoderm	BC060632	Terf2ip	8	1270007
Neuroectoderm	Dnm2	AB124611	9	19067
Neuroectoderm	AB124611	Cnn1	9	553950
Neuroectoderm	Trex1	Ccrl2	9	1980728
Neuroectoderm	Cx3cr1	Slc25a38	9	42117
Neuroectoderm	Slc25a38	Entpd3	9	415528
Neuroectoderm	Slc6a20a	Ccr2	9	432439
Neuroectoderm	Sorl1	Tecta	9	205334
Neuroectoderm	Slc38a3	Celsr3	9	1158921

Continued on next page

Continued from previous page

Lineage	Gene1	Gene2	Chrom	Distance
Neuroectoderm	Kif5a	Itga7	10	1670465
Neuroectoderm	Ascc2	Zmiz2	11	1703770
Neuroectoderm	Zmiz2	Adcy1	11	657327
Neuroectoderm	Nt5m	Mapk7	11	1612282
Neuroectoderm	Mapk7	3110043A19Rik	11	10190
Neuroectoderm	3110043A19Rik	Zfp286	11	1273332
Neuroectoderm	Efnb3	0610025P10Rik	11	341713
Neuroectoderm	0610025P10Rik	Wscd1	11	1836883
Neuroectoderm	Myo18a	Sez6	11	64963
Neuroectoderm	D030028A08Rik	Stard3	11	1398322
Neuroectoderm	Stard3	Thra	11	360802
Neuroectoderm	1500005I02Rik	Grin2c	11	453272
Neuroectoderm	Grin2c	2310004N24Rik	11	931646
Neuroectoderm	D11Bwg0517e	2900052L18Rik	11	1318282
Neuroectoderm	Galntl1	Smoc1	12	422935
Neuroectoderm	5033430I15Rik	Rnf144b	13	1156220
Neuroectoderm	Gprin1	Sncb	13	9192
Neuroectoderm	Jmy	Wdr41	13	1476659
Neuroectoderm	Wdr41	Crhbp	13	408057
Neuroectoderm	4931414P19Rik	Thtpa	14	488927
Neuroectoderm	Triobp	Sox10	15	149214
Neuroectoderm	Sox10	AW544981	15	1506867
Neuroectoderm	Mlc1	Shank3	15	520617
Neuroectoderm	Ccnt1	C230021P08Rik	15	851112
Neuroectoderm	C230021P08Rik	Accn2	15	212970
Neuroectoderm	Sema5b	Slc15a2	16	1085575
Neuroectoderm	Dscam	Prdm15	16	621244
Neuroectoderm	A630033E08Rik	Caskin1	17	1621657
Neuroectoderm	Ppp1r3c	Lgi1	19	1528129
Neuroectoderm	4930506M07Rik	Rab11fip2	19	829009
Neuroectoderm	Slc38a5	Otc	X	1972180
Neuroectoderm	Elf4	Slc25a14	X	169409
Neuroectoderm	Gpr64	Scml2	X	619123
Neuroectoderm	B630019K06Rik	Lancl3	X	305010
Neuroectoderm	Vsig1	Col4a6	X	125935