Gene expression in chromosomal Ridge domains: influence on transcription, mRNA stability, codon usage, and evolution
Gierman, H.J.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
A Model to Explain Natural Selection for Extreme Levels of Protein Expression in the Human Genome
A Model to Explain Natural Selection for Extreme Levels of Protein Expression in the Human Genome

Jan Koster*, Hinco J. Gierman*, Mireille H.G. Indemans and Rogier Versteeg

Department of Human Genetics, Academic Medical Center, University of Amsterdam, P.O. Box 22700, 1100 DE Amsterdam, the Netherlands.

* Joint first authors.

Submitted.

ABSTRACT

While human protein expression levels range from a few to over 150 million molecules per gene per cell, mRNA levels range from one to only a few thousand copies. We investigated how the extreme span in protein expression can be achieved. Highly expressed genes cluster in Ridges, which are chromosomal domains that increase the expression of embedded genes up to 8-fold. Ridges are GC-rich and are mainly found in GC-rich isochores. Preferred synonymous codons increase translation of transgenes and are also GC-rich in humans. Ridges may therefore superimpose efficient transcription and translation. We developed the Relative Codon Index to quantify genomic differences in codon usage bias. Ridge genes are indeed enriched for preferred codons, but highly expressed genes outside Ridges not. Also, Ridges have more GC-rich efficient translation-initiation sites. Paradoxically, consensus exists that small population sizes of higher organisms prevented natural selection for preferred codons, questioning a role for codon usage in humans. Rather, neutral mechanisms maintaining the GC-richness of isochores would have increased the GC content of codons as well. We propose a model that reconciles these paradoxes and explains how Ridges and isochores together facilitated selection for optimal protein expression. Genes that during evolution translocate into a Ridge, immediately obtain an increased expression, thus enabling selection for such an event. Subsequently, neutral mechanisms would slowly increase GC content and improve translation efficiency. The resulting efficient transcription and translation would exponentially increase the protein expression range, and explain how natural selection can achieve gradual adaptations in higher organisms.
INTRODUCTION

Quantitative mRNA analyses as e.g. provided by Serial Analysis of Gene Expression (SAGE), have shown that extremely highly expressed genes such as ribosomal proteins can reach expression levels up to a few thousand mRNAs per cell. Given that individual cells express a few hundred thousand mRNAs in total (Lewin 1980), this implies that mRNA levels of individual genes range from one to maximal a few thousand copies per cell. However, protein levels range from a few copies to over 150 million copies per cell for e.g. beta-actin proteins (Kislauskis 1997). This raises the fundamental question how the human expression machinery can create such an amplification of the range in protein expression levels compared to mRNA levels. Here we combine two lines of genome research that converge on the question how the organization of the human genome in functional domains relates to gene expression levels. The first research line deals with the clustering of highly expressed genes in chromosomal domains called Ridges (Caron 2001; Lercher 2002; Versteeg 2003; Gierman 2007). The second line of research studies the relation between GC content of genes, codon usage bias and translation efficiency (Lercher 2003).

Ridges are chromosomal domains of up to a few hundred genes that have an overall high gene expression (Caron 2001; Versteeg 2003). Ridges form stable domains that are highly expressed in all analyzed tissues and the individual genes in Ridges are often broadly expressed throughout different tissues (Lercher 2002). Ridges display a series of physical characteristics such as an overall high GC content, high gene-density, and short introns (Versteeg 2003; Lercher 2003). Weakly expressed genes also cluster in domains, called anti-Ridges, and have opposite characteristics. Recently, we have found that Ridge domains affect the expression of their embedded genes, as identical reporter genes inserted in Ridges have on average 4-fold higher expression than reporter genes integrated in anti-Ridges (Gierman 2007). The maximal measured effect of Ridge domains on the expression level of their embedded genes was 8-fold. This effect is possibly caused by the domain-wide open chromatin structure of Ridges (Goetze 2007). This demonstrated that besides classical regulation by transcription factor complexes, a second level of regulation exists that affects domain-wide expression levels in the genome. Relevant to this paper is that Ridges are GC-rich, and therefore relate to so-called isochores (Bernardi 1985b). However, the two concepts are not synonymous, as Ridges are functionally defined by expression levels of clustered genes, while isochores are defined by chromosomal banding patterns with GC content as physical basis.

The second research line focuses on the effect of codon usage and GC content on translation efficiency in an evolutionary context. Protein expression levels are not only controlled at the level of mRNA transcription, but also by translation. Most amino acids are encoded by several different codons (synonymous codons). The genome does not use synonymous codons in equal amounts. Many codons are used either more frequent (preferred) or less frequent (rare) than expected by chance. There are many examples that show that the use of preferred codons strongly increases the translation efficiency. These studies are usually performed by the expression
of non-human genes in human cells using so-called codon optimization (i.e. the replacement of all codons by the most preferred codons). Many studies report strong increases of up to a factor 10,000 in protein expression levels (Zolotukhin 1996; Levy 1996; Kim 1997; Leder 2001; Cid-Arregui 2003; Gao 2003; Bradel-Tretheway 2003; Gustafsson 2004). Although codon optimization studies with negative results probably remained unpublished, the large body of current literature suggests that the use of preferred codons can greatly enhance protein expression levels. The effect of preferred codons on translation is thought to be caused by differences in tRNA concentrations, which tend to be higher for preferred codons (Sharp 1995; Moriyama 1997; Kanaya 1999; Lavner 2005; Parmley 2007 and reviewed in Akashi 2001).

In bacteria, yeast, flies and worms, highly expressed genes display an overrepresentation of preferred codons, resulting in a strong correlation between codon usage bias and expression (Ikemura 1981; Ikemura 1982; Bennetzen 1982; Shields 1988; Stenico 1994). Very high protein expression in these organisms therefore in part stems from high mRNA expression, and in part from extra efficient translation of these transcripts due to efficient codon use. It is generally agreed upon that the use of preferred codons by highly expressed genes provided a selective advantage in these organisms and that the existing codon usage bias resulted from natural selection (Sharp 1994; Duret 2002). In contrast, the situation in humans and other warm-blooded vertebrates is more complicated, and topic of discussion. Studies in these organisms report only a weak or no correlation between the use of preferred codons and mRNA expression levels (Urrutia 2001; Smith 2001; Urrutia 2003; Plotkin 2004; Semon 2005; Semon 2006). Optimization of codon usage requires codon-by-codon mutational improvements and thus consists of many steps which each minimally affect translation efficiency. The small effective population size of mammals in general is thought to limit the power of natural selection for each fractional improvement, and would thus preclude codon optimization for highly expressed genes (Sharp 1995; Parmley 2007; dos Reis 2009).

Unlike bacteria, yeast, flies and worms, the genomes of mammals and other warm-blooded vertebrates display a strong variation in the GC content of large chromosomal domains (i.e. isochores) (Bernardi 1985b). The isochores are typically several hundreds of kilobases long and vary from roughly 30% to 60% GC content in humans. It is thought that not selection, but neutral mechanisms such as mutational bias and biased gene conversion, are mainly responsible for these regional differences in genomic GC content, and have shaped isochores (Filipski 1987; Sueoka 1988; Wolfe 1989; Duret 2008) (reviewed in Duret 2009). For example, Francino and Ochman have compared the changes in GC content of human globin pseudogenes residing in GC-poor and GC-rich chromosomal domains (Francino 1999). An originally GC-poor globin pseudogene that translocated to the GC-rich domain became GC-rich. As the pseudogenes are silent, they are not under selective pressure and the change in GC content is therefore driven by genetic drift (i.e. a neutral mechanism). These mechanisms affect the GC content of all sequences in genes, including the codons (Bernardi 1985a). As preferred codons in humans tend to be GC-rich, codon usage in mammals is thought to be largely determined by the GC content of the
local genomic background (Duret 2002). Many studies for increasing numbers of
genomes have confirmed the relationship between local genomic GC content and
codon usage (D’Onofrio 1991; Urrutia 2001). These observations have raised the
question whether codon optimization in human genes was the result of the effect
of the GC-rich neighborhood of genes (the neutral hypothesis or neutralist view), or
whether codon optimization was driven by a selective effect for increased translation
efficiency (the selectionist view). To answer this question, various studies have
looked for evidence of selection on codon usage by correcting for the GC content of
the genomic background of genes (Novembre 2002). Various studies reported either
no evidence of selection on synonymous codon usage (Duret 2000; Urrutia 2001) or
a significant but weak effect (Iida 2000; Urrutia 2003).

A satisfactory explanation of the evolutionary mechanism underlying codon usage
bias in the human genome is therefore still lacking (Eyre-Walker 1991; Sharp 1995;
Karlin 1996; Iida 2000; Nielsen 2006). Currently, no model explains the paradox of
how codon usage bias appears to strongly affect expression, whilst natural selection
seems incapable of driving codon usage in humans. Given that neutral mechanisms
mostly determine codon usage, the implication appears to be that codon usage is
determined by chance, and therefore raises the question whether codon usage
plays a role in the regulation of gene expression during human genome evolution
(Chamary 2006).

Here we present a thorough analysis of codon usage bias in Ridges. Highly
expressed genes outside Ridges are not enriched for preferred codons. High gene
expression by itself therefore seems not to trigger codon optimization. In contrast,
genomes in Ridges have a strong codon optimization. We propose a two step model,
in which natural selection can fix the translocation of a gene to a Ridge, due to
the strong effect of Ridges on the expression of embedded genes. Subsequently,
neutral mechanisms increase the GC content stepwise and optimize the codon
usage of the translocated gene, which can further increase expression. We also
show that the same mechanism improved translation initiation sequences of Ridge
genomes. The resulting model predicts that genes in Ridges not only have a high mRNA
expression, but also an efficient translation initiation and elongation. These combined
mechanisms would lead to an exponential increase of protein expression levels of
Ridge genes, providing an answer to the fundamental question how the human
genome can generate such a wide range of magnitudes in protein expression levels.

RESULTS

The Coding Sequences of Genes in Ridges Have a High GC Content
We previously reported that the genomic domains of Ridges are overall more GC-
rich (46.5%) than those of anti-Ridges (39.8%, the genomic average is 40.9%)
(Versteeg 2003). It is well-known that the GC content of the coding sequence of genes
corresponds to the GC content of the surrounding non-coding sequence (Bernardi
1985a; D’Onofrio 1991). However, no genome-wide study has been performed to
analyze the difference in GC content of the coding sequence of genes in Ridges compared to anti-Ridges. To determine the GC content of all genes, we separated the human genome in Ridge, intermediate and anti-Ridge domains as described previously (see Methods). All RefSeq sequences as annotated by the UCSC (HG17, http://genome.ucsc.edu/) were mapped to one of these categories and merged if they represented the same gene (Maglott 2000; Pruitt 2003; Karolchik 2003). In this way, 4,122 genes (24%) were annotated as Ridge, 1,868 genes (11%) as anti-Ridge, and the remaining 10,944 genes (65%) as ‘intermediate’. Analysis of the GC

**Figure 1.** Ridges are biased for gc-rich coding sequences. (A) The GC percentage (in 1 percent intervals) of the coding sequence (CDS) of genes is plotted against the relative occurrence within a domain (percentage of domain): Ridges (R; black line), anti-Ridges (A; light gray) and intermediate domains (I; dark gray). (B, C) Chromosome maps of chromosome 1 for expression (as measured by SAGE) and GC content of the coding sequence. Every RefSeq gene is placed at their genomic position (according to HG17 annotation). Height of the lines indicates the moving median (MM49) value over 49 RefSeqs (-24 to +24). Cytoband annotation is depicted beneath the profiles. Ridges are represented by boxes in red, anti-Ridges are represented by boxes in blue. (D) Chromosome map of chromosome 1 for the genomic GC content in 100,000 bp windows.
content of the full mRNA sequences in these domains confirmed the higher average GC content of Ridge genes (56%) compared to anti-Ridge genes (47%) (see Figure S1). This also holds true for just the coding sequence of the mRNAs: in Ridges the average GC content of the coding sequence is 58%, while this is 50% for anti-Ridge genes (Figure S1). Also the 3' and 5' untranslated regions (UTRs) of genes display a higher GC content in Ridges than in anti-Ridges (Figure S1). These differences are substantial, given that the dynamic range for the GC content of the coding sequence does not include the whole spectrum from 0–100%. In our analyses, the coding sequence of 98% of all genes have a GC content between 35% and 70%. This can be seen in Figure 1A, where the distribution of GC content of genes is plotted for Ridges (black), intermediate domains (dark gray) and anti-Ridges (light gray). Figure 1B-C shows for chromosome 1 that there is a genome-wide correspondence between the expression levels of chromosomal domains (Figure 1B), the GC content of the coding sequence of genes (Figure 1C) and genomic GC content (Figure 1D).

Ridge Genes are Enriched for Preferred Codons

The high GC content of the coding sequences in Ridges implies an effect on the codon usage of Ridges, as preferred codons in humans tend to be GC-rich (Sharp 1987; D’Onofrio 1991) and previous studies reported a correlation between genomic GC content and codon usage (D’Onofrio 1991). Therefore, we analyzed the differences in codon usage between Ridges and anti-Ridges. We utilized the Relative Synonymous Codon Usage (RSCU) index developed by Sharp and Li, which shows whether a codon is more or less frequently used than expected by chance (Sharp 1987). The RSCU for all codons in the human genome was calculated from codon usage tables as provided by Nakamura et al., which are based on updated versions of all human genes (Nakamura 2000). Analysis of the codon usage of all 18 amino acids with synonymous codons, showed that the average GC content of the most infrequently used codons is 39%, while this is 65% for the most frequently used codons. Figure 2A-B compares the use of the most frequent codons in Ridges and anti-Ridges. For 17 out of 18 amino acids, the preferred codons are more utilized in Ridges than in anti-Ridges. Figures 2D-E shows a chromosomal profile of this codon usage bias in Ridges and anti-Ridges for two different synonymous codons of valine. Ridges are biased for the most preferred codon for valine (GTG), which closely follows the expression profile (shown as moving medians of 49 genes) (Figure 2C-D). Conversely, the profile for the most infrequently used codon for valine (GTT) peaks in anti-Ridges (Figure 2C,E).

While these analyses reveal preferences for the codon usage of individual amino acids, they do not provide an overview of the overall codon usage of genes. To compare the overall codon usage of genes in different chromosomal domains, we utilized the RSCU to develop a Relative Codon Index (RCI) as a measure for the codon usage bias of genes. If a gene is composed of only the most preferred codons for every amino acid (based on RSCU), then the RCI is 1.0 (i.e. 100%); in the opposite case, when only the rarest codons are used, the RCI is 0. The RCI thus represents the percentage to which a gene utilizes preferred codons. The RCI values are independent of the amino acid composition or size of a gene and
can thus be used to compare different genes (note that the RCI is similar to the previously described Codon Adaptation Index (CAI) (Sharp 1987), but is based on more complete genomic data and has a normalization from 0 to 1 (see also Methods). Figure 3 shows the distribution of the RCI values of all genes. Genes in Ridges have a higher average RCI (0.68) than genes in anti-Ridges (0.58) (P <2*10^-5 Mann Whitney U test). Analogous to coding sequence GC content, RCI has a limited dynamic range from approximately 0.40 to 0.85, emphasizing that the difference in RCI between Ridges and anti-Ridges is substantial. As with GC content, there is a genome-wide correspondence between RCI and expression levels of chromosomal
domains in a chromosome map (Figure 2C,F). This shows that overall, genes in Ridges use more preferred codons than genes in anti-Ridges.

**Highly Expressed Genes outside Ridges are not Enriched for Preferred Codons**

The increased presence of preferred codons in Ridge genes compared to anti-Ridge genes, might either result from a neutral effect of the surrounding GC-rich chromosomal domain, or from evolutionary selection to improve gene expression efficiency (see Figure 4A). To discriminate between both possibilities, we explored our observation that only one-third of the highly expressed genes in the genome cluster in Ridges (see below). The remaining highly expressed genes are scattered over the genome. In addition, Ridges also contain poorly expressed genes. This allowed us to analyze our dataset for codon usage bias of highly (Hi) versus poorly (Lo) expressed genes in Ridges and outside Ridges. Highly, intermediately, and poorly expressed genes were set to 24%, 65% and 11% of all genes (these percentages are the same as the percentages of genes in Ridge, Intermediate, and anti-Ridge domains). Figure 4A shows that Ridge genes have a higher RCI (0.68) than anti-Ridge genes (0.58) (P <1*10^-5 Mann Whitney U test). However, there was no significant difference in RCI between the poorly (0.62) and highly expressed genes (0.63) in the genome (P = 0.09 Mann Whitney U; Figure 4B), suggesting that high gene expression itself does not trigger codon usage bias. To test this further, we divided the highly expressed genes (Hi) into two groups: Ridge (R) genes (31%) and anti-Ridge plus intermediate (A+I) genes (69%) and calculated average RCI. Figure 4C shows that highly expressed genes in Ridge have a higher RCI than highly expressed genes outside Ridges (HiR versus Hi(A+I) P <1*10^-5 Mann Whitney U test). Figure 4D shows that also poorly expressed genes in Ridges have a higher RCI than poorly expressed genes outside Ridges (LoR versus Lo(A+I) P <1*10^-5 Mann Whitney U test). Variations in the cut-
off values for either high or low expression produced similar results (see Figure S3). These analyses demonstrate that Ridges form chromosomal domains wherein all genes display a bias for preferred codons irrespective of their expression level, whilst genes outside of Ridges do not. These findings strongly argue that natural selection does not significantly improve codon usage of highly expressed genes in humans (in contrast to bacteria). Apparently, only the genes in Ridges underwent codon improvement due to the neutral effects of the GC-rich surrounding domain. Codon usage bias in humans therefore corresponds to the overall expression levels of chromosomal domains, but not to the expression levels of individual genes.

Ridge Genes Have Improved Translation Initiation Sites
The observation that Ridge genes are enriched for preferred codons, while highly expressed genes outside Ridges are not, raised the question whether the postulated effect of the Ridge domains on GC content also affected other parameters relevant for gene expression levels. Translation initiation efficiency is another important parameter determining protein expression levels. The optimal sequence for translation initiation in humans is the so-called Kozak consensus sequence: A/GCCAUGG (Kozak 1986; Kozak 1987; Kozak 1997; Kozak 2005). The presence of one or more of the consensus bases around the AUG start codon can enhance translation up to tenfold (Kozak 1986; Kozak 1987; Kozak 1997). The Kozak sequence is very GC-rich and Pesole and co-workers have shown that Kozak sequences are more frequent in GC-rich isochores (Pesole 1999). We therefore analyzed whether this holds true for Ridges, which correspond to, but are not synonymous with GC-rich isochores. We scored the presence of each of the optimal bases of the consensus sequence for all genes with a known translation initiation site (see Methods). Figure 5 shows that genes in Ridges use each of the preferred guanines and cytosines around the start codon significantly more often than genes in anti-Ridges (Figure 5). As the presence of each of these bases contributes to translation, many genes in Ridges will have
A model to explain extreme protein levels

an enhanced translation initiation site. At position -3 of the translation initiation site, both a guanine and an adenosine give optimal translation initiation. Figure 5 shows that the adenosine at position -3 is more frequent in anti-Ridges. This is consistent with the idea that the biases in codon usage and translation initiation are largely determined by genomic GC content. Importantly, the increase of optimal bases in the translation initiation site strongly increases the percentage of Ridge genes possessing a complete Kozak sequence, leading to an increase of 48% relative to anti-Ridges ($P = 8.4 \times 10^{-7}$ chi-square test); Figure 5).

A Model to Reconcile the Neutralist View With the Selectionist View

Based on our previous functional analyses of Ridges and the data on GC content and codon optimization of Ridge genes as reassessed in this paper, we propose a novel model to integrate current knowledge (Figure 6). We have recently shown that Ridges form chromosomal domains that upregulate the transcription of inserted genes up to 8-fold compared to anti-Ridge domains (Gierman 2007). This implies that when during human evolution, a gene happens to relocate into a Ridge domain by e.g. transposition, gene duplication, inversion or translocation, expression of the gene is likely to increase. If this increased expression provides a selective advantage, the event can be fixed in the population by natural selection. After translocation, the gene would still have the GC content and codon usage corresponding to an anti-Ridge. Over time, mutational bias would slowly increase the GC content of the gene. This will optimize both codon use and the translation initiation sequence. At the same time, the gene in question would acquire a GC content conform its

![Figure 5. Ridge genes are biased for preferred codons. The relative frequency of each of the known optimal translation initiation sites surrounding the AUG codon are shown for Ridges (light gray), intermediate domains (dark gray) and anti-Ridges (black). The optimal bases are at various positions relative to the AUG startcodon: A or G at -3, G at +4 and CC at -2 and -1. To the right the frequency of the entire Kozak sequence (GCCAUGG or ACCAUGG) is shown. All frequencies are normalized to that of anti-Ridges set at 100%. All preferred cytosines and guanines around the start codon occur more often in Ridges ($P < 2.9 \times 10^{-3}$ chi-square test).](image-url)
genomic domain and show no evidence of selection on codon usage when corrected for background GC content. This occurs without the need of natural selection to drive each of individual mutations in the gene. Rather, selection would only act on the increase in transcription due to translocation of a gene, which then in turn invokes a gradual increase in GC content and thereby optimization in codon usage.

**Ridges: Transcriptional and Translational Highways?**

In our model, the increase in GC content driven by neutral mechanisms will slowly optimize codon usage. The current body of literature suggests that codon optimization in mammalian cells leads to increased translation and protein levels. Figure 6 illustrates how several mechanisms would then act on expression levels of genes in Ridges. Firstly, the average mRNA expression of Ridge genes is higher due to the domain-effect of Ridges. Secondly, translation initiation of the transcripts is higher due to an optimized Kozak sequence. Finally, translation elongation is more efficient due to optimized codon use. Since these effects would exponentially add-up, Ridge genes could achieve extreme high protein expression levels. This mechanism will further increase expression without the necessity to select for each minor improvement. When levels would become too high, the genome has ample possibilities to allow for negative selection, e.g. by mutational adaptation of promoter or enhancer sequences. The model would act in two directions, both when a gene is translocated from an anti-Ridge into a Ridge, or when a Ridge gene migrates to an anti-Ridge. This system would therefore enable fractional improvements of fitness that are individually too weak to be selected for, to become fixed in the genome. Such a mechanism would potentially offer a huge advantage to mammalian cells, as it would provide a principle to achieve the vast range in expression levels required for proteins. Ridges would thus provide a physical domain where both mRNA expression and translation initiation and elongation of embedded genes are maximized.

**DISCUSSION**

One of the main questions in human evolution is how natural selection can be effective for traits with weak fitness gains. Codon usage bias is perhaps one of the most-studied examples in this respect. In bacteria and other lower organisms with a large effective population size, codon usage bias of highly expressed genes is well established and thought to be the result of natural selection. The appearance of higher organisms with small effective population sizes would have brought this essential evolutionary force to a standstill and made flexible adaptation of expression levels by codon optimization impossible. This problem might have been alleviated by the emergence of isochores that introduced a dichotomy of the genome. Although it is still debated how and when isochores originated, it is well established that isochores determine the GC content of their embedded genes (Filipski 1987; Sueoka 1988; Wolfe 1989; Galtier 2007; Duret 2008). This property of isochores enables the genome to increase GC content and improve codon usage and translation initiation of translocated genes, without the requirement of natural selection to drive this mechanism. The increase in GC content can occur relatively fast in evolutionary
terms. Natural selection for such a translocation would instead act one step earlier, when a gene is translocated into a Ridge and as a result, instantaneously obtains an increased expression level. Positive or negative selection decides whether the translocation will be fixed in the population. Over time, the GC content of the translocated gene would automatically increase due to neutral mechanisms (e.g. biased gene conversion) acting upon the entire Ridge (right). As a result, GC dependent mechanisms such as codon usage and translation initiation would potentially increase the protein levels of the gene further. The effect of chromosomal domains on multiple levels thus helps the genome to achieve a wide range of expression levels.

Our model makes several predictions that can be tested. First of all, the model implies that the use of preferred codons increases protein expression. Codon optimization studies suggest that preferred codons indeed strongly improve translation. However, these studies were mostly based on the complete codon optimization of non-human genes with very poor codons, while codon differences between Ridges and anti-
Ridges are less extreme. A systematic analysis of larger sets of genes is therefore needed to estimate the general effect of codon usage in Ridges and anti-Ridges. The advent of high-throughput proteomics promises to realize this goal in the near future. In contrast to codon usage, the quantitative effect of variations in the translation initiation site has been investigated systematically. Kozak has shown that each of the bases surrounding the translation start site affects protein translation and together they can modulate protein levels more than tenfold (Kozak 1986; Kozak 1987; Kozak 1997).

Our model implies that a systematic analysis of evolutionary related species will identify examples of genes that translocated to Ridges, and as a consequence have increased their mRNA expression and improved their codon usage. Another prediction is that genes for which an expression increase was beneficial during evolution, preferred translocation to Ridges over other chromosomal domains. Comparative analyses between the genomes of humans and related species will be needed to test this prediction.

Finally, other mechanisms might exist that utilize GC content to increase expression in Ridges. Several computational and experimental studies have suggested that mRNAs with high GC content have increased mRNA stability (Duan 2003; Nguyen 2004; Chamary 2005). We have recently found that indeed, mRNAs from Ridges compared to anti-Ridges have increased half-lives of 1.5–2 hours (Gierman submitted). This implies that together with codon usage and translation initiation, increased mRNA stability might constitute a third GC content-driven mechanism to increase expression levels in Ridges. Experiments by Kudla have even suggested that high GC content of a gene can by itself increase transcription (Kudla 2006), but these results were not supported by another study (Arhondakis 2008).

The results and the model presented here provide a novel and testable hypothesis on expression regulation and genome evolution in humans. Moreover, the model explains how the genome can utilize an exponential system of gene expression regulation to achieve the magnitudes of protein expression levels required by a cell.

METHODS

Datasets
RefSeq annotation on the human genome was obtained from UCSC (build HG17). RefSeq annotation for the coding sequences (CDS) and UTR was obtained from NCBI (2006-11-21) (Maglott 2000; Pruitt 2003; Karolchik 2003). Ridge anti-Ridge boundaries were used for all window sizes of 19 to 59 genes from our previously published data based on the UCSC build HG15 (Versteeg 2003). Overlapping regions were merged, resulting in a set of 45 Ridges and 36 anti-Ridges spanning 277mb and 614mb respectively. To obtain genomic positions for the HG17 build, we used the lift over algorithm at UCSC (http://genome.ucsc.edu/). GC content was
determined for each of the next datasets as follows: For the reference sequence the gene boundaries for every RefSeq aligned to the genome (HG17) was used; for the mRNA set the reference sequence was used; for the CDS and UTR sets the CDS and UTR within the reference sequence was used.

**RSCU calculations**
Relative synonymous codon usage (RSCU) values are estimated as the ratio of the observed codon usage to that value expected if there is uniform usage within synonymous groups multiplied by the number of codons for the amino acid. Data for the observed codon usage are taken from the codon usage database (http://www.kazusa.or.jp/codon/) (Nakamura 2000). A value of 1.0 then represents no bias, while a value higher or lower represents preference or the opposite respectively.

**RCI calculations**
The Reference sequence dataset was first checked for consistency using the following criteria: The coding sequence within every reference sequence should be dividable by 3, had to start with an ATG or CTG, should end with a stopcodon, must not possess more than 1 stopcodon and may only contain defined nucleotides (i.e. C, A, T, or G). To assess whether a RefSeq was positioned within a Ridge or anti-Ridge, the genomic location for every sequence was queried in the UCSC refflat table (HG17). For each RefSeq that passed these criteria, the RCI was calculated as follows: For each codon (c) within a reference sequence, the frequency (f) was multiplied by the RSCU index value (Ri) for that codon. The sum of those values generates the codon index (CI) for a gene: CI\text{gene} = \sum_{c=1}^{C} (f*Ri). For each amino acid the frequency was then multiplied by the RSCU index value of the best or worst codon. The sum of those values generates the minimal codon index (CI\text{min}) and maximal codon index (CI\text{max}). The Relative CI is then calculated by: (CI-CI\text{min})/(CI\text{max}-CI\text{min}). The Relative CI is then calculated by: (CI-CI\text{min})/(CI\text{max}-CI\text{min}). This RCI value then represents how well the sequence is making use of its synonymous codons (1.0 indicating the best synonymous codon for every amino acid within the sequence; 0.0 indicating the worst codon for every amino acid within the sequence). The RCI is thus relative within the boundaries of a gene’s codon usage, whereas the CAI does not utilize a lower boundary (the situation where the worst possible codons have been chosen). Nevertheless, there is a direct linear correlation between the RCI value of any gene versus the CAI value as they show a correlation of 0.95 (Pearson R²; see Figure S2). For most of the analyses in this manuscript the RefSeqs were merged to their respective genesymbols. When more than 1 RefSeq was encountered for a genesymbol, the GC content and the RCI are averaged.

**Chromosome maps**
All genomic profile images were created by a CGI web-application called Genome Profiler, which is developed within the department of human genetics at the AMC, the Netherlands by Jan Koster. In short, the domain-wide abundance for a given parameter such as gene expression is determined by calculating the median value for that parameter over a window of 49 genes, and sliding the window along the chromosome. Next, these calculated values are plotted at the base pair position of
the center gene within the current window (gene 25 out of 49), thereby generating a chromosome map.

**Translation initiation site analysis**

All analyses were performed using the same reference sequence dataset assigned as described under the RCI calculations, excluding only those sequences (n = 12) that use an alternative start codon (CTG). Sequences assigned to Ridges, anti-Ridges or intermediate domains were aligned on the annotated start codon (ATG) and for each optimal base, the presence in every sequence was scored. The relative frequencies of each base and the Kozak sequence (AGCCAUGG) were set to 100% for anti-Ridges. The differences in frequencies between were tested using a chi-square test.

**Acknowledgments**

We gratefully acknowledge the advice of M. Hamdi, M. Kool, F. Haneveld, L.J. Valentijn, G. Huizenga, M.J. Noordam, F. Berends, D. Geerts, E.E. Santo, B. Hooijbrink and F.J. van Hemert of the Academic Medical Center Amsterdam. We thank T.J. Ettema of Uppsala University and W.G. Forssmann of the Hannover Medical School for support and we thank the reviewers for useful comments. This work was supported by grants from the European Commission for the FP6 3D-Genome project (contract LSHG-CT-2003-503441). JK, HJG, and RV conceived and designed study. JK, HJG, and MHGI performed experiments. JK contributed analysis tools and wrote Perl scripts. HJG conceived model. JK and HJG analyzed data. JK, HJG and RV wrote paper.

**Supplementary information**

Supplementary figures S1–S3.

**Conflict of interest**

The authors declare no conflict of interest.
A model to explain extreme protein levels

References

Bradel-Tretheway BG, Zhen Z, Dewhurst S. 2003. Effects of codon-optimization on protein expression by
secondary structure in mammals. Genome Biol 6(9):R75.
Chamary JV, Pamrley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in
Cid-Arregui A, Juarez V, zur Hausen H. 2003. A synthetic E7 gene of human papillomavirus type 16 that
yields enhanced expression of the protein in mammalian cells and is useful for DNA immunization
compositional properties of human genes, codon usage, and amino acid composition of proteins. J
Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA
Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern
Filipski J. 1987. Correlation between molecular clock ticking, codon usage fidelity of DNA repair,
Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of
Gao F, Li Y, Decker JM, Peyerl FW, Bibollet-Ruche F, Rodenburg CM, Chen Y, Shaw DR, Allen S,
Musonda R, et al. 2003. Codon usage optimization of HIV type 1 subtype C gag, pol, env, and
nef genes: in vitro expression and immune responses in DNA-vaccinated mice. AIDS Res Hum
Retroviruses 19(9):817-23.
Gierman HJ, Koster J, Indemans MH, Versteeg R. Genes in chromosomal Ridge domains have increased
mRNA folding stability and half-life, further contributing to their high expression. Submitted
Kozak M. 1997. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J 16(9):2482-92.
A model to explain extreme protein levels


SUPPLEMENTARY INFORMATION

Supplementary figures
Figure S1
Figure S2
Figure S3
**Figure S1.** Ridges display a high GC content over multiple mRNA properties. The average GC content was determined over all reference sequences for the complete messenger RNA (mRNA), the coding sequence (CDS), and the 5 prime and 3 prime untranslated region (UTR) sequences. For gene symbols containing multiple RefSeqs, the average was taken over the reference sequences. Refseqs were divided in 3 groups: Ridges (R), anti-Ridges (A) and Intermediate (I) (see Methods).

**Figure S2.** The RCI and CAI of a gene have a strong direct linear correlation. Correspondence between the Codon Adaptation Index (CAI) and the Relative Codon Index (RCI) for all genes in the human genome. Note that the RCI is relative, while the CAI is only bounded by a maximal value of 1. Both measures made use of the same Relative Synonymous Codon Usage (RSCU) table. RCI vs CAI show a correlation of 0.95 (Pearson R²).
Figure S3. High and low expression genes do not differ in RCI regardless of the cut-off value. Relative Codon Index (RCI) distributions on lowest (black line) versus highest (gray line) expression levels in the current dataset (HG17, 385 SAGE libraries). Genes were divided into Lower or Higher expression based on a large variety of Serial Analysis of Gene Expression (SAGE) expression cut-offs. For every cut-off value the average RCI was calculated for the genes in the Higher and Lower group. (A) The average RCI for different cut-off values. (B) the number of genes within a group for different cut-off values.