The Accumulation of Gene Regulation Through Time

Maria Warnefors* and Adam Eyre-Walker
School of Life Sciences, University of Sussex, Brighton, United Kingdom
*Corresponding author: E-mail: maria.warnefors@gmail.com.
Accepted: 28 February 2011

Abstract
Gene expression is governed by an intricate combination of transcription factors (TFs), microRNAs (miRNAs), splicing factors, and other regulators. Genes cannot support infinitely complex regulation due to sequence constraints and the increased likelihood of harmful errors. However, the upper limit of regulatory complexity in the genome is not known. Here, we provide evidence that human genes are currently not operating at their maximum capacity in terms of gene regulation. We analyze genes spanning the full spectrum of eukaryote evolution, from primate-specific genes to genes present in the eukaryote ancestor, and show that older genes tend to be bound by more TFs, have more conserved upstream sequences, generate more alternative isoforms, house more miRNA targets, and are more likely to be affected by nonsense-mediated decay and RNA editing. These results cannot be explained by overrepresentation of certain functional categories among younger or older genes. Furthermore, the increase in complexity is continuous over evolutionary time, without signs of saturation, leading to the conclusion that most genes, at least in the human genome, have the capacity to evolve even more complex gene regulation in the future.

Key words: human genome, gene regulation, gene age, saturation.

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

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

To what extent have human genes reached their maximum regulatory capacity? This question can be addressed by analyzing the level of regulation associated with genes that arose at different evolutionary times. Four potential scenarios are illustrated in figure 1. In the first (fig. 1A), genes are continuously acquiring regulatory features and have not yet reached their maximum capacity. In the second scenario (fig. 1B), older genes are saturated in terms of gene regulation and do not show a further increase in complexity. These two scenarios assume that gene regulatory features accumulate over time. It might, however, be that different forms of regulation dominate in genes of different age categories (fig. 1C) or that regulation and age are uncorrelated factors (fig. 1D). This last scenario does, however, appear unlikely as evolutionary age is known to correlate with aspects of gene architecture, including gene length and intron density (Wolf et al. 2009), as well as with gene expression, such that older genes tend to be expressed in more tissues (Milinkovitch et al. 2010) and at higher levels (Wolf et al. 2009) than younger genes.

To distinguish between these scenarios, we have collected information on a variety of regulatory mechanisms...
operating in the human genome and related this to the evolutionary age of the affected genes. We found that older genes tend to be bound by more TFs, have more conserved upstream sequences, use more alternative transcription start sites (TSSs), produce more alternative splicing isoforms, and use more alternative polyadenylation sites. Furthermore, older genes are more likely to be affected by miRNAs, nonsense-mediated decay (NMD), and RNA editing. Based on this and the lack of apparent saturation, we draw the conclusion that the majority of human genes could support higher levels of regulation than what we currently observe.

Materials and Methods
To group human genes according to time of origin, we used the phylostratigraphic classifications established by Domazet-Lošo and Tautz (2010), with the additional requirement that the genes should be represented in release 59 of the Ensembl database (Flicek et al. 2010). We excluded human genes shared by archaea and bacteria from our analysis as many of the regulatory mechanisms that we consider are specific to eukaryotes. The number of genes for each of the 18 age categories is shown in Table 1.

Next, we calculated eight measures of the regulatory complexity of human genes. First, we estimated the complexity of transcriptional regulation for each gene, by counting the number of TFs that bound within 10 kb upstream of the TSS in the human cell line GM12878. This data set came from ENCODE ChIP-seq experiments performed at the HudsonAlpha Institute (Birney et al. 2007) and was available through the HAIB TFBS track for the human genome (release hg18) in the UCSC Genome Browser (Kent et al. 2002). The following 20 TFs were analyzed: BATF, BCL3, BCL11, EBF, Egr-1, GABP, IRF4, NRSF, p300, PAX5c, PAX5n, Pbx3, POU2F, POU5F1, RARG, RARG2, RARG3, RARZ, RXRa, RXRα, RXRβ, RXRγ.
Sin3A, SP1, SRF, TAF1, TCF12, USF-1, and ZBT33. As a secondary measure of transcriptional regulation, we calculated the degree of conservation of sequences within 10 kb upstream of the TSS as the proportion of bases that were identified as conserved within primates by the phastCons program (Siepel et al. 2005). This information was taken from the Conservation track in the UCSC Genome Browser.

Our next three complexity measures were based on the number of transcripts that are generated due to alternative use of TSSs, alternative splicing, and alternative polyadenylation,
to make GOslim terms for “molecular function” and “biological process” (Ashburner et al. 2000). To this end, we downloaded genes into functional categories based on gene ontology even when we corrected for gene function, we first grouped the complexity values of each gene, not the averaged values, which are provided for overview in figure 2.

We calculated eight measures of regulatory complexity, by calculating the Pearson correlation. This analysis was based on the complexity values of each gene, not the averaged values, which are provided for overview in figure 2.

To examine whether the observed correlations persisted even when we corrected for gene function, we first grouped genes into functional categories based on gene ontology terms (Ashburner et al. 2000). To this end, we downloaded GOslim terms for “molecular function” and “biological process” from Ensembl release 59 (Flicek et al. 2010). To make sure we had sufficient power to detect any correlations, we restricted our analysis to terms that matched at least 1,000 genes in our data set. We also excluded terms that were children to any of the other included terms, with exception for the term “binding,” which due to its generality was further divided into “protein binding” and “nucleic acid binding.” We then repeated the analysis described above for each functional category.

### Results and Discussion

We have examined the accumulation of regulatory complexity in human genes, by analyzing several aspects of gene expression in genes of different evolutionary ages. To group genes according to time of origin, we used the classifications given by Domazet-Lošo and Tautz (2010). These age estimates rely on ortholog identification by BLAST (Altschul et al. 1997), which could mean that some faster-evolving genes escape detection. However, simulations indicate that overall this strategy is reliable (Albà and Castresana 2007).

In total, human genes were divided into 18 age categories, with the oldest category including human genes that were present in the eukaryote ancestor and the youngest category consisting of primate-specific genes (table 1). Divergence times for the different categories were taken from the TimeTree database (Hedges et al. 2006), except in cases of contradictory estimates, where instead we interpolated the divergence time from the surrounding categories by taking the average time (table 1). Qualitatively similar results were obtained when we excluded these categories, as well as when we performed the analysis using the category numbers rather than the time estimates.

We calculated eight measures of regulatory complexity, based on publicly available data (see Materials and Methods). To estimate the level of transcriptional regulation, we analyzed sequences within 10 kb upstream of the TSS. First, we counted the number of TFs that bind to this region in the human lymphoblastoid cell line GM12878. To exclude nonexpressed genes, only genes that were bound by at least one TF were included in the analysis. Figure 2A shows the average number of TFs that bind to genes of different ages, with a clear increase in TF binding for old relative to young genes. As the data are rather noisy and some of the age categories contain relatively few genes (table 1), differences between individual age categories should be interpreted with caution in this and the following graphs. A list of means and standard errors for all investigated regulatory mechanisms is provided as Supplementary Material online. Analysis confirmed that evolutionary age is significantly correlated with TF-binding diversity, such that older genes are typically associated with more types of TFs ($P = 2 \times 10^{-16}$, Pearson correlation, note all correlations are performed on the raw data, not the means shown in the figures). To estimate the magnitude of the increase in diversity, we fitted a linear model to the data, which showed that genes in the youngest category are typically bound by 4.1 TFs, whereas the oldest genes are bound by 5.4 TFs (table 2).

Second, we assessed the level of conservation of upstream sequences, by counting the number of bases within

### Table 2

Differences in Complexity between the Youngest and Oldest Age Categories

| Category                  | Youngest Genes (Primates) | Oldest Genes (Eukaryotes) | Ratio |
|--------------------------|---------------------------|---------------------------|-------|
| TF-binding sites         | 4.12                      | 5.38                      | 1.31  |
| Conserved bases upstream | 396                       | 547                       | 1.38  |
| TSSs                     | 2.35                      | 4.92                      | 2.09  |
| Splicing isoforms        | 2.76                      | 5.72                      | 2.07  |
| Polyadenylation sites    | 2.26                      | 4.80                      | 2.12  |
| miRNA sites              | 0.0017                    | 0.0573                    | 33.7  |
| NMD proportion           | 0.058                     | 0.168                     | 2.90  |
| RNA editing proportion   | 0.052                     | 0.161                     | 3.10  |

The estimates were obtained by fitting a linear model to the data.
10 kb of the TSS that were identified as conserved among primates by the phastCons program (Siepel et al. 2005). Again, we found a significant correlation with age, where older genes tend to have more conserved upstream sequences than younger genes \( (P = 1 \times 10^{-10}) \), such that the upstream regions of the oldest genes contain almost 40% more conserved bases compared with younger genes (table 2). Thus, both TF binding and upstream conservation show a highly significant correlation with evolutionary age.

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

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

Next, we investigated the distribution of verified miRNA-binding sites across the 18 categories (fig. 2F) and found that older genes are enriched in this type of regulation \( (P < 5 \times 10^{-11}) \), with the number of miRNA targets per gene increasing more than 30-fold from 0.0017 to 0.0573. We also found significant positive correlations between gene age and the likelihood for genes to be targeted by the less common regulatory mechanisms NMD \( (P < 2 \times 10^{-16}) \) and RNA editing \( (P < 2 \times 10^{-16}) \). For both of these mechanisms, around 5% of the youngest genes are affected, whereas the proportion among the oldest genes is three times larger.

In theory, the results described above could be influenced by an uneven distribution of gene functions among the age categories. If “early” genes predominantly are of a functional type that requires a certain level or mode of regulation, whereas “late” genes have other functions and therefore different regulatory needs, then we might see a superficial correlation between age and regulatory complexity. To test this possibility, we further divided our data set according to
gene ontology terms (Ashburner et al. 2000) and repeated the analysis for a number of functional categories (see Materials and Methods). In the vast majority of cases, the correlations between complexity and gene age remained positive even for functional subsets of genes (fig. 4), showing that the positive correlations that we obtained for the full data set are not due to functional bias.

Based on these results, we can exclude the last two possibilities shown in figure 1, (no increase in complexity with time and certain types of complexity being associated with particular time periods) as all forms of regulatory complexity investigated here show a significant increase over time. We are therefore left to determine whether the oldest human genes have reached regulatory saturation, that is, whether the pace at which genes accumulate new features has slowed down for older genes. To do this, we performed a regression analysis involving a quadratic term. However, in all eight cases, this term was either not significant or it indicated that the pace is higher for older genes. Thus, we have not found any evidence to suggest that human genes have reached regulatory saturation, that is, whether the pace at which genes accumulate new features has slowed down for older genes. To do this, we performed a regression analysis involving a quadratic term. However, in all eight cases, this term was either not significant or it indicated that the pace is higher for older genes. Thus, we have not found any evidence to suggest that human genes have reached regulatory saturation, that is, whether the pace at which genes accumulate new features has slowed down for older genes.

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

**Supplementary Material**

Supplementary material is available at Genome Biology and Evolution online (http://gbe.oxfordjournals.org/).

**Acknowledgments**

We thank Tomislav Domazet-Loso for assistance with the age categories. This work was supported by the University of Sussex.
Literature Cited
Albà MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 7:53.
Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.
Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.
Birney E, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.
Domazet-Lošo T, Tautz D. 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol. 8:66.
Flicek P, et al. 2010. Ensembl's 10th year. Nucleic Acids Res. 38:D557–D562.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 22:2971–2972.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.
Kiran A, Baranov PV. 2010. DARNED: a DAtabase of RNA EDiting in humans. Bioinformatics. 26:1772–1776.
Milinkovitch MC, Helaers R, Tzika AC. 2010. Historical constraints on vertebrate genome evolution. Genome Biol Evol. 2:13–18.
Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. 2009. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res. 37:D155–D158.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5:e14.
Roux J, Robinson-Rechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res. 21:357–363.
Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Tazi J, Bakkour N, Stamm S. 2009. Alternative splicing and disease. Biochim Biophys Acta. 1792:14–26.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. Inaugural article: the universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A. 106:7273–7280.

Associate editor: George Zhang