Sussex Research Online

DNA entropy reveals a significant difference in complexity between housekeeping and tissue specific gene promoters

Article (Accepted Version)

Thomas, David, Finan, Christopher, Newport, Melanie and Jones, Susan (2015) DNA entropy reveals a significant difference in complexity between housekeeping and tissue specific gene promoters. Computational Biology and Chemistry, 58. pp. 19-24. ISSN 1476-928X

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/59371/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

DNA entropy reveals a significant difference in complexity between housekeeping and tissue specific gene promoters

David Thomas[1], Chris Finan[1] and Melanie J Newport[1] and Susan Jones[2]

[1]Brighton and Sussex Medical School, University of Sussex, Brighton, BN1 9PX, UK

[2]The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK.

Abstract

Background: The complexity of DNA can be quantified using estimates of entropy. Variation in DNA complexity is expected between the promoters of genes with different transcriptional mechanisms; namely housekeeping (HK) and tissue specific (TS). The former are transcribed constitutively to maintain general cellular functions, and the latter are transcribed in restricted tissue and cells types for specific molecular events. It is known that promoter features in the human genome are related to tissue specificity, but this has been difficult to quantify on a genomic scale. If entropy effectively quantifies DNA complexity, calculating the entropies of HK and TS gene promoters as profiles may reveal significant differences.

Results: Entropy profiles were calculated for a total dataset of 12,003 human gene promoters and for 501 housekeeping (HK) and 587 tissue specific (TS) human gene promoters. The mean profiles show the TS promoters have a significantly lower entropy ($p < 2.2e\text{-}16$) than HK gene promoters. The entropy distributions for the 3 datasets show that promoter entropies could be used to identify novel HK genes.

Conclusion: Functional features comprise DNA sequence patterns that are non-random and hence they have lower entropies. The lower entropy of TS gene promoters can be explained by a higher density of positive and negative regulatory elements, required for genes with complex spatial and temporary expression.

1. Introduction

In the human genome 5% of the DNA is estimated to be under selection pressure (Waterston et al. 2002), but only 1.5% is estimated to be coding (Lander et al. 2001). This indicates that elements of non-coding DNA are under selection pressure, and by implication have functional roles (Mu et al. 2011). Gene promoters comprise non-coding DNA but include large numbers of sequence features, including binding sites for transcription factors (TFs) that contribute to the regulation of gene expression. An increasing understanding of the importance of non-coding DNA has led to many methods being applied to the problem of differentiating functional and non-functional sites within them. Estimating the entropy of DNA, using concepts from the field of

information theory (Schneider 2010), is one way in which genomic elements have been analysed.

Definitions of entropy, including topological, Shannon, linguistic complexity and lossless compression have been applied to genomic regions from diverse genomes with varying results (Table 1). Five studies, applying definitions of topological entropy (Karamanos et al. 2006), Shannon entropy (Mantegna et al. 1995; Stanley et al. 1999), linguistic complexity (Troyanskaya et al. 2002) and lossless compression (Liu et al. 2008), conclude that non-coding DNA has a lower entropy than coding DNA. In contrast, 3 studies, one applying Shannon entropy (Mazaheri et al. 2010) and a 2 applying topological entropy (Koslicki 2011; Jin et al. 2014) conclude that non-coding DNA has higher entropy than coding DNA. The variation likely results from the differing and often very small DNA datasets used (Table 1), which for some analyses reflects the emphasis on the theoretical entropy calculation rather than its biological application. In addition some studies have defined noncoding DNA as intergenic DNA only (e.g Mazaheri et al. 2010) or intronic DNA only (Koslicki 2011; Jin et al. 2014), whilst others have included both types of DNA as non-coding (Karamanos et al. 2006). If differences in entropies between different types of DNA are small, then it is not surprising that studies using different datasets and definitions have reached different conclusions.

Two recent studies both apply definitions of topological entropy to systematic random samples of genes from all chromosomes in the human genome, and conclude that introns have a higher entropy than exons (Koslicki 2011; Jin et al. 2014). This can be explained by the fact that entropy is a measure of the randomness of a DNA sequence, and introns are expected to be more random as they have fewer functional signals and are less conserved than exons (Koslicki 2011; Jin et al. 2014). It has also been concluded that exons have a higher entropy than gene promoters (Jin et al. 2014), which could reflect the presence of multiple functional elements within the promoters which are under selection pressure. These include transcription factor binding sites (TFBSs) characterised by short sequence motifs that are highly degenerate. The TFBSs are bound cooperatively by TFs to form *cis*-regulatory modules (CRMs), which play a key role in gene regulation (Hardison & Taylor 2012). TFBSs represent DNA sequence patterns within promoters and hence have lower entropies, as observed in *E.coli* (Krishnamachari et al. 2004).

In addition to work comparing non-coding and coding DNA entropies, entropies have been used to create profiles for DNA sequence windows for complete chromosomes and genomes. Entropy profiles, based on linguistic complexity, created for 16 prokaryotic genomes, revealed differences in complexity between CG and AT rich genomes (Troyanskaya et al. 2002). Average mutual information (AMI) profiles created for chromosomes from eukaryotes, showed that such profiles are effective species signatures (Bauer et al. 2008). Topological entropy profiles calculated for *S.cerevisiae* also proved to be effective in quantifying the level of repetitive sequences in regions of DNA (Crochemore & Verin 1999). In additional work, entropy profiles based on the theory of chaotic dynamics (Jeffrey 1990) have been shown to quantify local DNA signatures (Dufraigne et al. 2005). This method was successfully applied to the identification of horizontal gene transfers between prokaryotic species (Dufraigne et al. 2005). In further work, profiles using Renyi entropies (a generalized of the Shannon entropy) were created and applied them to the identification of statistical significant of DNA sequence motifs, including TFBSs in prokaryotic gene promoters (Vinga & Almeida 2007).

As discussed, work has shown that entropy profiles can effectively measure both the complexity of local DNA sequences and act as global species signatures. In the current work the effectiveness of entropy profiles for measuring differences in DNA sequences at a level intermediate of the two is addressed. A topological definition of entropy is used to identify global sequence signatures in the promoters of genes with different transcriptional mechanisms in the human genome. Genes can be transcribed constitutively to maintain general cellular functions or be transcribed in restricted tissue and cells types for specific molecular events (Butte et al. 2001; Chang et al. 2011). The former are termed housekeeping genes (HK) and the latter tissue specific genes (TS). It is known that promoter features are related to tissue specificity (Schug et al. 2005; Farré et al. 2007), with TS genes having higher levels of nucleosome occupancy and higher densities of TFBSs (She et al. 2009). Such features facilitate the close transcriptional control required for expression in specific cell or tissue types. The hypothesis for the current work is that promoter features will give rise to different levels of DNA complexity that can be measured using entropy profiles. The identification of differences in DNA complexity of HK and TS gene promoters would be a first step in classifying additional genes to these two important transcriptional classes.

Study	Entropy Type	Dataset	Conclusions
(Colosimo & De Luca 2000)	Linguistic complexity	16 DNA sequences including eukaryotes (5 human) and prokaryotes (< 2650bp in length)	• native DNA < random DNA
(Troyanskaya et al. 2002)	Linguistic complexity	21 prokaryotic genomes	• C > NC
(Liu et al. 2008)	Lossless compression	human genome	• C > genomic
(Mantegna et al. 1995)	Shanon	2 phage genomes, 2 viral genomes *C.elegans* Chromosome III: Yeast Chromosomes III & XI 6 *E.coli*, 3 mouse & 9 human sequences	• C > NC
(Stanley et al. 1999)	Shanon	4 Yeast Chromosomes : Chr III, VI, IX, XI Primates in GenBank	• C > NC Intergenic only • C ==NC Introns only
(Mazaheri et al. 2010)	Shanon	*C.difficile* (G+C 29.1%) genome *B.bacteriovorus* (G+C 50.6%) genome	• C < NC Intergenic only
(Karamanos et al. 2006)	Topological	2 viral genomes and 4 human gene regions (max ~73K bp)	• C > NC
(Koslicki 2011)	Topological	human genome, 100 longest intron and exon sequences from 23 chromosomes	• C < NC Introns only
(Jin et al. 2014)	Topological	Human genome, random 100 introns + exons from each chromosome. 210K random gene promoters (-200bp upstream of TSS)	• C < NC Introns only

Table 1. Summary of previous publications in which entropies have been calculated for DNA sequences, and used to compare the relative entropy of coding and non-coding DNA.

2. Methods

2.1 Gene datasets: All genes with Human and Vertebrate Analysis and Annotation (HAVANA) annotations were extracted from the human GRCh38 genome assembly in Ensembl [release 77], using the application programming interface (Flicek et al. 2012). A gene subset, that comprised those with an intergenic region (defined as nucleotides between the transcription start site (TSS) of one gene and the 3' UTR of the proceeding gene) of ≥30K base pairs (bp), was then selected. This subset ensured that promoter regions did not include introns or exons of the proceeding gene when entropies of upstream regions >10K base pairs were calculated (see section 2.2.2). This dataset was denoted HAV1.

A subset of housekeeping (HK) and tissue specific (TS) genes were then extracted from the HAV1 dataset. This was based on a meta-analysis of 104 microarray datasets from 43 normal human tissues, that identified 2064 human HK and 2293 human TS genes (Chang et al. 2011). The HK and TS genes from this analysis were mapped to the HAV1 dataset, and the matched gene sets denoted HAV_HK and HAV_TS respectively. This extraction left 12003 genes remaining from the HAV1 dataset, and this was denoted HAV_12003.

2.2 Topological entropy calculations

2.2.1 Definition of topological entropy

The definition of topological entropy (H_{top}) as defined by Koslicki (Koslicki 2011) is as follows:

Let w is a finite sequence of length |w|, let n be the unique number such that

$$4^n + n - 1 \leq |w| < 4^{n+1} + (n+1) - 1$$

Then for $w_1^{4^n+n-1}$ *the first* 4^n*+n-1 letters of w*

$$H_{top}(w) := \frac{log_4(p_{w_1^{4^n+n-1}}(n))}{n}$$

where $p_w(n)$ represents the different number of n-length subwords that appear in *w*.

In this definition H_{top} is explained in terms of a DNA sequence having an alphabet of 4 bases. For this alphabet a string of length 1028 could contain every unique 5 base pair sequence (4^5 = 1024 (+ (5-1) for the rolling window)). H_{top} calculates the number of unique subsequences found by taking the log base 4 (alphabet size) of the count and

dividing by the length of the substring. For a 5bp subsequence within a 1028 base pair rolling window, the maximum possible number of unique subsequences is 1024 (4^5). If 50 unique subsequences are observed then the entropy is 0.56 ($\log_4(50)/5$), if 500 were observed the entropy would be 0.896 ($\log_4(500)/5$) etc. This means that systems with non-random sequences featuring functional patterns, have a low number of unique substrings, and hence a low entropy. Random sequences have higher entropies. Outline code of the Java module used to calculate the rolling window entropies is provided as supplementary data.

2.2.2 Entropy profiles of gene promoters

Entropy profiles were calculated for 4-mers in a forward rolling window of 259 base pairs for DNA regions surrounding the transcription start site (TSS). Profiles were initially calculated and tested on regions >10K base pairs upstream of the TSS for a small number of genes, but the compute time prohibited such large regions being analysed for the complete dataset. Hence, entropy profiles were calculated for a promoter region that extended from -2.5Kbp upstream to +0.5Kbp downstream of the TSS of each gene. For each profile a per base pair entropy was assigned by attributing the entropy of the forward rolling window to the first base pair in the window. A mean entropy profile was mapped for the complete HAV_12003 dataset and for the HAV_HK and HAV_TS datasets (Figure 1A). It should be noted that the forward rolling window of 259bp means that the calculation of entropy for base pairs from -258bp towards the TSS includes base pairs that are downstream of the TSS and potentially sample the first exons and/or introns of the gene. The resulting profiles are comparable to those presented in the entropy analysis of 21 prokaryotic genomes (Troyanskaya et al. 2002).

The significance of the differences observed in the profiles between the HAV_HK and HAV_TS datasets was quantified by calculating a P-value using the Wilcoxon signed rank non-parametric test of group differences, implemented by the Wilcox.test function in the MASS package (version 7.3-35) (Venables & Ripley 2002) in R (R Core Team 2014).

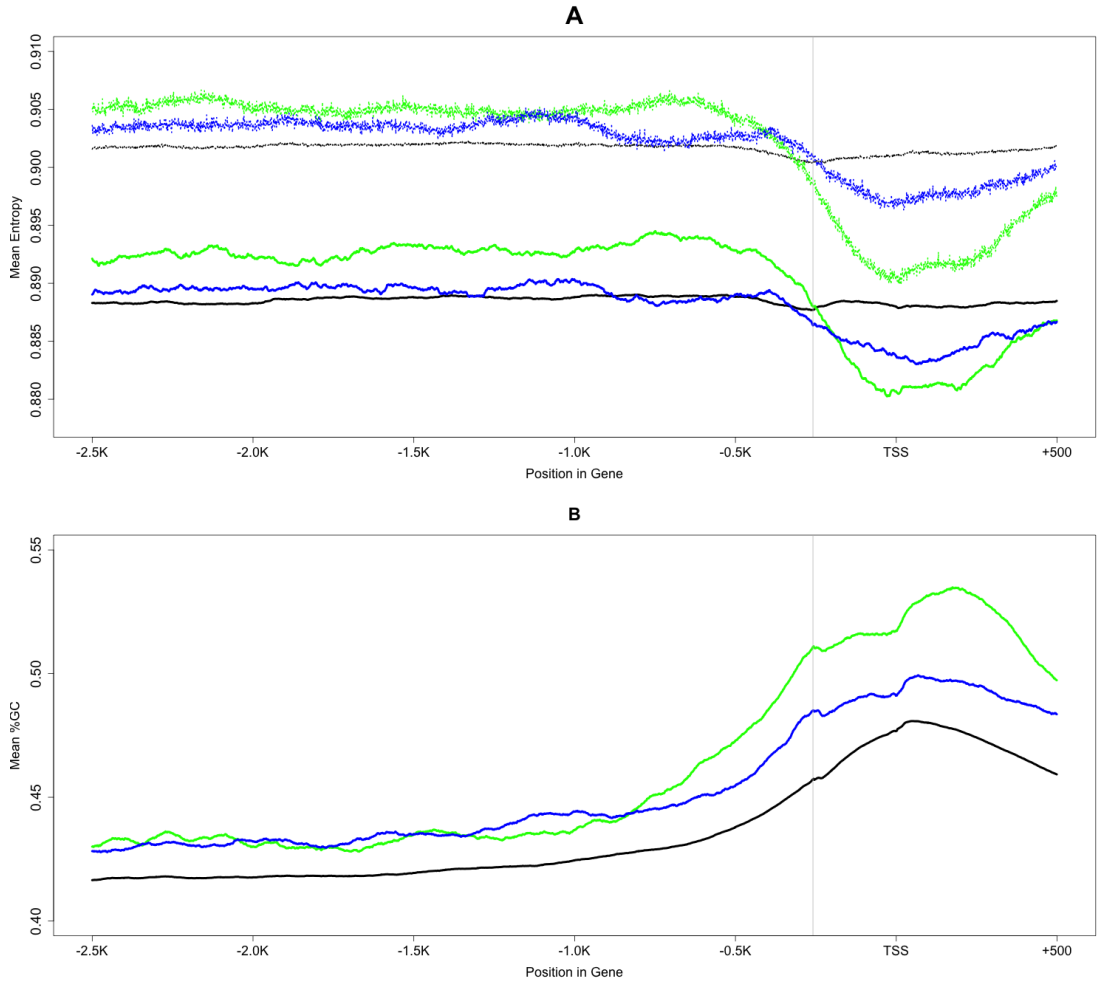

***Figure 1**. (A) Mean entropy profiles, based on a word size of 4 and a forward rolling window of 259 base pairs for the promoter regions of 3 gene datasets (solid lines) and for randomly shuffled DNA (dotted lines). (B) Mean %CG profiles calculated using a forward rolling window of 259 base a pairs. The pairwise differences in the GC profiles (HAV_HK vs HAV_TS: HAV_TS Vs HAV_12003: HAV_HK vs HAV_12003) are not significant when evaluated using a Wilcoxon signed rank test with continuity correction. Key to colours: HAV_TS = blue, HAV_HK=green, HAV_12003=black. The grey vertical line at position -259bp indicates the point at which the rolling window first samples beyond the TSS.*

2.2.3 Entropy profiles of random DNA

In order to measure the relative complexity of promoters, mean entropies were calculated for random DNA sequences. The entropies of random sequences were generated by permuting the DNA in a 259bp window repeatedly, as the window rolled across the promoter. For each random permutation of the 259bp window entropies were calculated based on a word size of 4. Random permutations of bases were generated using the Java random method. This gave random sequences that had a GC content that matched the base pairs of the forward rolling window in the real gene

promoters. A mean entropy profile was then calculated for -2.5Kbp to +0.5Kbp for the HAV_12003, HAV_HK and HAV_TS datasets (Figure 1A).

2.2.4 Distributions of entropy values of gene promoters

The profiles in Figure 1A represent mean entropies for each gene dataset. In order to assess the variation in entropies, the distribution of entropies for all genes in each dataset was plotted using the geom-density function of the ggplot2 (version 1.0) (Wickham 2009) package in R (R Core Team 2014) (Figure 2). The distributions were calculated for the region -2.5K to -259bp upstream of the TSS, so that the entropies were based on a forward rolling window that did not sample into the gene.

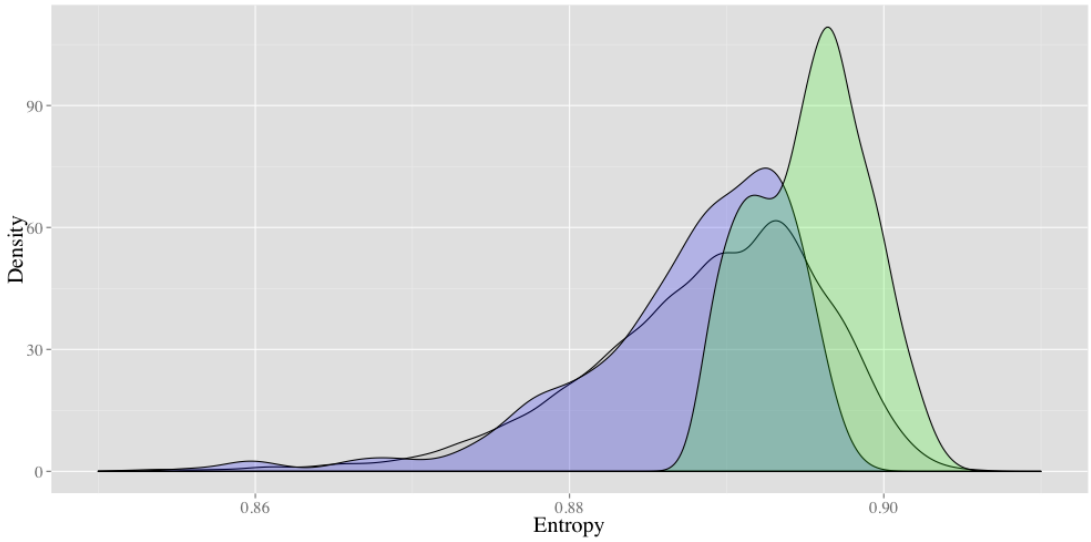

***Figure 2**: Distribution of entropy values calculated using a word size of 4 and a forward rolling window of 259bp, for promoter regions spanning -2.5k bp to -259bp upstream of the transcription start site. Key to colours: HAV_TS = blue, HAV_HK=green, HAV_12003=grey (the HAV_12003 density plot has the lowest density peak and lies behind those for HAV_HK and HAV_TS).*

2.3 CG content profile: DNA entropy is influenced by GC content (Troyanskaya et al. 2002). To assess the relationship between CG content and the entropy profiles, a percentage CG profile was calculated by attributing the CG content of a 259bp forward rolling window (to match the entropy profile rolling window size) to the first base pair in the window. Mean percentage CG profiles were calculated for -2.5Kbp to +0.5Kbp for the HAV_12003, HAV_HK and HAV_TS datasets (Figure 1B).

3. Results

3.1 Gene datasets

The HAV_12003 data set extracted from the GRCh38 human genome assembly, comprised 12003 genes. The HAV_HK dataset comprised 507 genes and the HAV_TS

dataset comprised 596 genes, which represent 4.1% and 4.9% of the HAV_12003 dataset respectively. The HAV_12003 gene dataset essentially comprises as yet unidentified TS genes (and to a much smaller extent HK genes), as well as genes with expression levels that do not fit the definition of HK and TS genes.

3.2 Entropy profiles

The mean entropy profiles for the promoter regions of all 3 datasets (HAV_12003, HAV_HK and HAV_TS) have lower entropies than random DNA, as would be expected (Figure 1). The real promoter regions are comprised of functional regions, such as TFBSs, which are under selection pressure and hence cannot evolve randomly. The key feature of the entropy profiles is that TS gene promoters have significantly lower entropies than HK gene promoters ($p < 2.2e\text{-}16$). This is likely to be reflective of an increased density of functional sequence features within the TS promoters.

The aligned %GC profiles (Figure 1B) show that entropy of the promoters is not simply explained by variations in GC content, as the %CG profiles do not mirror the variations in the entropy profiles. All the %CG profiles show an increase from -1.5K bp upstream to beyond the TSS, with the largest increases when the rolling window samples into the gene, and includes coding DNA, known to be more CG rich than non-coding DNA (Vinogradov 2003).

The definition of topological entropy used here, means that the size of the forward rolling window is related to the word sized (see section 2.2.1): the larger the word size the larger the rolling window. To test the effect of a different word size, profiles were also created for a word size of 5 that give a forward rolling window of 1059bp (Supplementary data: Figure S1). These profiles also showed TS gene promoters had significantly lower entropy than HK gene promoters. The same held true when a mean entropy profile for a combined word size of 4 and 5 was also calculated (as the average entropy in each window of the two word sizes) (Supplementary data: Figure S2). Hence, whilst using different word sizes moves the trough in the entropy profiles further upstream of the TSS (as the point at which the rolling window starts to sample into the gene moves further upstream), it does not change the overall conclusion; that TS gene promoters have lower entropy than HK gene promoters

3.3 Distributions of promoter entropies

The density plots of promoter entropies for the region between -2.5K bp and -259 bp shows that the HK gene promoters have an overlapping, yet distinct distribution (Figure 2). The density distribution for the TS genes is shifted towards lower entropies, whilst the HK genes are shifted towards higher entropies, as would be expected from the significantly different entropy profiles (Figure 1A). The distribution of entropies for genes in the HAV_12003 dataset is completely overlaid by the HK and TS distributions, but most resembles the TS density distribution with a long tail of lower entropies. This could be reflective of the HAV_12003 dataset comprising of a greater number of unidentified TS gene promoters and only a small number of HK genes (see discussion), but the densities are not definitive.

4. Discussion

By creating mean topological entropy profiles for gene datasets from the entire human genome, we effectively quantify variations in DNA complexity between the promoters of HK and TS genes. These variations can be attributed to functional features of the promoters, and not just compositional biases related to high GC content which is known to vary within promoters (Koudritsky & Domany 2008; Jaksik & Rzeszowska-Wolny 2012).

A gene promoter generically comprises a core promoter positioned +1bp to -120bp relative to the TSS, a proximal promoter -120bp to -1Kbp, and a distal promoter at an unknown distance from the TSS (Maston et al. 2006). In some cases TFs have >25% of their binding sites positioned >20Kb upstream of the TSS, indicating the importance of long-range gene expression regulation (Lee et al. 2012). Such data indicate that, in general, gene promoters do not have well defined boundaries, and architectures vary widely between genes (Hackanson et al. 2008; Vikman et al. 2009). However, even though promoters are ill-defined they are known to encode large numbers of sequence features, including GC rich regions, short sequence repeats, TFBS; as well as nucleosome occupancy and DNA curvature signatures. Our entropy profiles quantify the complexity of these specific features, as well as the background DNA within which they lie.

TS and HK genes can be considered as transcriptomic extremes (Chang et al. 2011), with TS genes being under complex regulatory control of multiple specific TFs, and HK genes having simpler regulatory mechanisms in which basal promoters predominate (Farré et al. 2007). In the current work promoters of TS genes were shown to have significantly lower entropies than the HK genes. The promoters of mammalian TS-

genes are more conserved than HK genes due to an increased density of functional sequence regions (Farré et al. 2007) and our TS promoter entropies reflect this. A high density of binding sites is likely to be required for the control of genes with complex spatial and temporary expression profiles, such as those with expression restricted to specific tissues. The promoters of HK genes also have lower levels of nucleosome occupancy, which is partly determined by sequence signals (Segal et al. 2006). Hence, the lack of such sequence signals could also contribute to the increased entropy of these promoters.

In this analysis we selected genes with >30K bp intergenic regions, to ensure the promoters did not include introns or exons of a proceeding gene, and increase the confidence that sequence signals in the promoter were linked to the selected gene. This selection does mean we have sampled genes from regions of lower gene density. Whilst there is some evidence that gene density is positively correlated with TFBS density the variance is large (Lee et al. 2012). Hence, whilst it is possible that the promoters we have analysed could have lower levels of TFBSs than if a smaller threshold had been used, this is not considered an important factor, and the same threshold has been applied to both the HK and the TS datasets.

In this work we show that TS gene promoters have significantly lower entropies than HK gene promoters, and it was initially hypothesized that entropies could be used to classify additional HK and TS genes from the HAV_12003 dataset. However, the densities of entropy values do overlap (Figure 2); and initial tests on developing a support vector machine to differentiate HK from TS genes gave a relatively low 63% accuracy in a 10-fold cross validation. Hence, it appears that the entropy values alone are not enough to differentiate HK from TS genes or alternatively that the overlap of entropy distributions could be reflective of the way in which HK and TS genes are defined.

Whilst it is known that the number of HK genes in the human genome will be relatively small, studies have identified significantly different numbers of such genes in the human genome (from 451 to 3,140) (Chang et al. 2011); and the definition of HK genes is currently being debated (Fantom Consortium, 2014). The mean number of HK genes defined in previous studies comprise approximately 10% of protein coding genes (Chang et al. 2011). A more recent study estimated the number to be even smaller at just 6%, when HK genes were defined as those showing ubiquitous and uniform expression (Fantom Consortium, 2014). The definition and hence numbers of HK

genes identified is dependent upon three factors: (a) the technology used to measure gene expression; these include microarrays (e.g. (Chang et al. 2011), RNA-sequencing (e.g. (Eisenberg & Levanon 2013) or single molecular cap analysis of gene expression (CAGE) (Fantom Consortium, 2014); (b) the analysis of expression by tissue or cell type (Shen-Orr et al. 2010) and (c) the statistical methods used to calculate expression thresholds (Dai et al. 2013).

The HK genes used in the current work were defined from 1431 samples from 42 normal human tissue types from 104 microarray data sets (Chang et al. 2011). Complex tissue types, such as the brain, have many different cell types expressed at different levels, and hence the expression measured will be strongly influenced by the sample variation on cell type frequencies (Shen-Orr et al. 2010). Hence, whilst the HK and TS genes datasets used in the current study meet one set of definitions, the use of expression data from other technologies and based on cell-type could re-classify some genes. The complexity of the transcriptional classification of genes (and whether such classifications are still valid in the light of new high-throughput gene expression data) is a key issue that needs to be considered when future models of transcription are developed.

As well as re-considering gene classification and functional DNA regions in promoter regions, future models of transcription also need to account for transcription factor specificity and affinity. These two parameters are complex; as specificity is difficult to quantify (Yan & Wang 2012), and affinity is difficult to measure *in-vitro*. Binding affinities have been measured for TFs in specific systems (Prouse & Campbell 2013; Wang et al. 2009), but methods present problems when scaling up to whole genomes. The affinity of TFs to bind specific TFBSs is also affected by flanking DNA sequences (Siggers & Gordân 2013) and by cooperative binding of additional TFs (He et al. 2009). The future development of more realistic models of transcription regulation requires a better understanding of the relationship between TFBS occupancy, TF binding affinities and their relationship to a revised transcriptional classification of genes. This will lead to dynamic and cooperative models of binding where TFs interact with both promoter DNA and multiple copies of the same or different TFs.

5. Acknowledgements

DT received 50% support from an MRC (UK) PhD studentship.

6. Data availability

The gene datasets and the entropy values used to create the profiles and density plots are available as supplementary data files (S1 and S2 respectively). An outline of the Java code used to calculate the entropy of the rolling windows of sequence is also included as supplementary material.

7. References

Bauer, M., Schuster, S.M. & Sayood, K., 2008. The average mutual information profile as a genomic signature. *BMC bioinformatics*, 9, p.48.

Butte, a J., Dzau, V.J. & Glueck, S.B., 2001. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *Physiological genomics*, 7(2), pp.95–6.

Chang, C.-W. et al., 2011. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PloS one*, 6(7), p.e22859.

Colosimo, a & De Luca, a, 2000. Special factors in biological strings. *Journal of theoretical biology*, 204(1), pp.29–46.

Crochemore, M. & Verin, R., 1999. Zones of low entropy in genomic sequences. *Computers & chemistry*, 23, pp.275–282.

Dai, H. et al., 2013. Mixed modeling and sample size calculations for identifying housekeepinggenes. *Statistics in Medicine*, 32(February), pp.3115–3125.

Dufraigne, C. et al., 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic acids research*, 33(1), p.e6.

Eisenberg, E. & Levanon, E.Y., 2013. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10), pp.569–574.

Fantom Consortium, 2014. A promoter-level mammalian expression atlas. *Nature*, 507(7493), pp.462–470.

Farré, D. et al., 2007. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome biology*, 8(7), p.R140.

Flicek, P. et al., 2012. Ensembl 2012. *Nucleic acids research*, 40(Database issue), pp.D84–90.

Hackanson, B. et al., 2008. Epigenetic modification of CCAAT/enhancer binding protein alpha expression in acute myeloid leukemia. *Cancer research*, 68(9), pp.3142–51.

Hardison, R.C. & Taylor, J., 2012. Genomic approaches towards finding cis-regulatory modules in animals. *Nature reviews. Genetics*, 13(7), pp.469–83.

He, X. et al., 2009. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PloS one*, 4(12), p.e8155.

Jaksik, R. & Rzeszowska-Wolny, J., 2012. The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences. *Gene*, 492(2), pp.375–81.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8), pp.2163–2170.

Jin, S. et al., 2014. A generalized topological entropy for analyzing the complexity of DNA sequences. *PloS one*, 9(2), p.e88519.

Karamanos, K. et al., 2006. Statistical compressibility analysis of DNA sequences by generalized entropy-like quantities: towards algorithmic laws for biology? In *Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications*. World Scientific and Engineering Academy and Society (WSEAS), pp. 481–491.

Koslicki, D., 2011. Topological Entropy of DNA Sequences. *Bioinformatics*, 27(8), pp.1061–1067.

Koudritsky, M. & Domany, E., 2008. Positional distribution of human transcription factor binding sites. *Nucleic acids research*, 36(21), pp.6795–805.

Krishnamachari, a, moy Mandal, V. & Karmeshu, 2004. Study of DNA binding sites using the Rényi parametric entropy measure. *Journal of theoretical biology*, 227(3), pp.429–36.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Lee, B. et al., 2012. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Research*, 22, pp.9–24.

Liu, Z., Venkatesh, S.S. & Maley, C.C., 2008. Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. *BMC genomics*, 9, p.509.

Mantegna, R.N. et al., 1995. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E Statistical Physics Plasmas Fluids And Related Interdisciplinary Topics*, 52(3), pp.2939–2950.

Maston, G. a, Evans, S.K. & Green, M.R., 2006. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics*, 7, pp.29–59.

Mazaheri, P. et al., 2010. Differentiating the protein coding and noncoding RNA segments of DNA using Shannon entropy. *International Journal of Modern Physics C*, 21, pp.1–9.

Mu, X.J. et al., 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic acids research*, 39(16), pp.7058–7076.

Prouse, M.B. & Campbell, M.M., 2013. Interactions between the R2R3-MYB transcription factor, AtMYB61, and target DNA binding sites. *PloS one*, 8(5), p.e65132.

R Core Team, 2014. R: A language and environment for statistical computing.

Schneider, T., 2010. A brief review of molecular information theory. *Nano Communications Networks*, 1(3), pp.173–180.

Schug, J. et al., 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology*, 6(4), p.R33.

Segal, E. et al., 2006. A genomic code for nucleosome positioning. *Nature*, 442(7104), pp.772–8.

She, X. et al., 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC genomics*, 10, p.269.

Shen-Orr, S.S. et al., 2010. Cell type-specific gene expression differences in complex tissues. *Nature methods*, 7(4), pp.287–289.

Siggers, T. & Gordân, R., 2013. Protein-DNA binding: complexities and multi-protein codes. *Nucleic acids research*, doi: 10.10, pp.1–13.

Stanley, H.E. et al., 1999. Scaling features of noncoding DNA. *Physica A*, 273(1-2), pp.1–18.

Troyanskaya, O.G. et al., 2002. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5), pp.679–688.

Venables, W. & Ripley, B., 2002. *Modern Applied Statitics with S* Fourth., New York: Springer.

Vikman, S. et al., 2009. Functional analysis of 5-lipoxygenase promoter repeat variants. *Human molecular genetics*, 18(23), pp.4521–9.

Vinga, S. & Almeida, J.S., 2007. Local Renyi entropic profiles of DNA sequences. *BMC bioinformatics*, 8, p.393.

Vinogradov, A.E., 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Research*, 31(7), pp.1838–1844.

Wang, Y. et al., 2009. Quantitative transcription factor binding kinetics at the single-molecule level. *Biophysical journal*, 96(2), pp.609–20.

Waterston, R.H. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–62.

Wickham, H., 2009. *ggplot2: elegant graphics for data analysis*, New York: Springer.

Yan, Z. & Wang, J., 2012. Specificity quantification of biomolecular recognition and its implication for drug discovery. *Scientific reports*, 2, p.309.