Human Accelerated Geometry

J. R. Dunkley

December 9, 2025

Abstract

Human genomes contain extensive noncoding sequence whose large scale organisation is usually modelled by neutral drift and simple stochastic processes on the line. Recent work in Universal Information Hydrodynamics has shown that a Bernoulli bounded entropy functional with Fisher metric gradient terms admits cored Yukawa type equilibrium profiles for a scalar field σ_{F}, interpreted as a logit of a local occupation probability. Here we ask whether this Fisher bounded entropy ansatz provides a useful description of noncoding sequence around human transcription start sites. We treat GC fraction in the UCSC `gc5Base` track as a one dimensional field around 142 585 human promoters in GRCh38 and fit three competing parametric profiles to ±5 kb windows: a Gaussian diffusion model, an exponential decay model, and a Fisher cored profile imported from the bounded entropy Fisher functional. Model selection by Akaike Information Criterion shows that Gaussian and exponential baselines are preferred for the vast majority of loci, but a small subset of promoters, around 0.4%, are best fit by a Fisher cored profile and require it at a strong threshold $\Delta\mathrm{AIC} > 10$ when compared with both baselines. We then intersect these Fisher core promoters with an extended catalogue of Human Accelerated Regions (HARs) in hg38 and measure enrichment as a function of linear distance. At a scale of 100 kb, Fisher core promoters show a non trivial enrichment in HAR neighbourhoods. The resulting overlap gene set includes several canonical developmental regulators, such as *PAX6*, *HAND2*, *SIM1*, *CDX2*, *WT1*, and *PRDM8*. We conclude that a small, non random subset of human promoters exhibits GC profiles consistent with a Fisher bounded entropy geometry and that these loci are more likely than average to lie in regulatory neighbourhoods that experienced lineage specific acceleration on the human branch.

Contents

1 Introduction

Large eukaryotic genomes devote most of their length to noncoding sequence. In the human genome, protein coding exons occupy only a few percent of the total, while introns, promoters, enhancers, untranslated regions, and a variety of repetitive elements fill the remaining ninety plus percent. Although many noncoding bases evolve under approximate neutral drift, it is now clear that the spatial organisation of noncoding DNA around genes encodes regulatory information and chromatin state.

Recent work on phase separated chromatin condensates shows that internucleosomal linker length and local interaction networks control the thermodynamic stability and material properties of chromatin droplets [1]. This provides an experimental substrate on which it is plausible to interpret our Fisher stiffness and screening length as coarse grained parameters for a nucleosome level energy landscape.

A simple and robust coarse grained descriptor of this organisation is GC fraction. At kilobase scales, GC rich regions correlate with CpG islands, open chromatin, promoter activity, and certain classes of enhancers, while GC poor regions correlate with more compact chromatin states. This has motivated a range of models that treat GC variation along the genome as the result of stochastic dynamics such as biased random walks, diffusion with local constraints, or context dependent mutation processes.

In parallel, the Universal Information Hydrodynamics (UIH) programme has developed a geometric framework in which both reversible and irreversible dynamics are generated by a unified operator $K = G_{\mathrm{F}} + iJ_{\mathrm{F}}$ acting on probability densities on an information manifold [2, 3]. In the scalar sector, one considers a field σ_{F} that is the logit of a local Bernoulli occupation probability and studies a bounded entropy Fisher functional of the form

$$\mathcal{F}[\sigma_{\mathrm{F}}] = \int \left\{ \alpha \, |\nabla \sigma_{\mathrm{F}}|^2 + V_{\mathrm{Bern}}(\sigma_{\mathrm{F}}) - J(x)\, \sigma_{\mathrm{F}}(x) \right\} \mathrm{d}x, \tag{1.1}$$

where α is a stiffness, V_{Bern} is a bounded entropy potential, and J is a source. Under mild conditions this functional admits Bogomolny type completions and cored Yukawa equilibrium profiles for σ_{F} [4]. These profiles, characterised by an amplitude, a core radius, and a screening length, provide an information geometric analogue of screened potentials in statistical field theory.

In previous work, this Fisher bounded entropy sector was applied to static galactic halos, with the scalar field interpreted as a logit of an effective occupation number in a vacuum sector [4]. Here we consider a different system where the natural degree of freedom is again a Bernoulli variable that takes values in a compact interval. For a fixed window around a transcription start site (TSS), the local GC fraction from a five base pair smoothed track can be viewed as such a field, with values in the unit interval at each position.

This suggests a simple question. If one regards the GC fraction profile around a promoter as the equilibrium configuration of an effective scalar field on a line, are the data better described by simple neutral baselines, such as Gaussian or exponential profiles, or by the cored profiles that arise from a bounded entropy Fisher functional When fitted to a large and homogeneous dataset, does the Fisher ansatz single out any special subset of promoters

Human accelerated noncoding elements provide an independent axis along which to assess evolutionary importance. Human Accelerated Regions (HARs) are short segments of noncoding DNA that are highly conserved across vertebrates yet show an unusual cluster of substitutions along the human lineage [5, 6]. Many HARs have been functionally characterised as developmental enhancers, particularly in the forebrain and limb. If Fisher cored promoter profiles do mark loci under atypical regulatory constraints, one might expect some overlap with the regulatory neighbourhoods of HARs.

In this note we address these questions using only public genome tracks and simple parametric models. We assemble a catalogue of human TSSs in GRCh38, extract GC fraction profiles from the UCSC `gc5Base` track in fixed windows around each TSS, and fit three models to each profile: a Gaussian diffusion profile, an exponential decay profile, and a Fisher cored profile imported directly from the bounded entropy Fisher functional. Model comparison by Akaike Information Criterion (AIC) identifies a small subset of promoters that require the Fisher cored profile at a strong threshold. We then intersect these Fisher core promoters with an extended Pollard and Capra HAR set in hg38 and quantify enrichment as a function of linear distance.

Our aim is deliberately modest. We do not attempt a full genomic or evolutionary analysis, and we do not optimise the Fisher ansatz beyond importing its functional form from earlier work. Rather, we document that a Fisher bounded entropy profile is empirically preferred for a small but non random subset of promoters and that these loci are modestly enriched in human accelerated regulatory neighbourhoods.

2 Methods

2.1 Data sources

All analyses are performed on the GRCh38 human reference genome.

GC fraction. We use the UCSC `gc5Base` bigWig track for hg38, which gives GC percentage in sliding windows of length five bases along each chromosome. Values are in the range 0 to 100. For modelling we convert these to GC fractions by dividing by 100.

Transcription start sites. Gene annotations are taken from GENCODE v44 basic gene models for GRCh38 [7]. From the GTF file we construct a TSS table by taking, for each annotated transcript, the first genomic coordinate on the corresponding strand. For most analyses we collapse to one TSS per gene symbol by selecting a representative transcript per gene; different choices have negligible impact on the statistics reported here. The final TSS catalogue contains 142 585 distinct promoter centred windows with sufficient GC data coverage.

Human Accelerated Regions. Human Accelerated Regions are taken from the Pollard and Capra extended HAR set for hg38, obtained from the UCSC Table Browser

[5, 6]. The table provides genomic intervals for each HAR on UCSC chromosome naming. We treat the midpoints of these intervals as reference positions when computing distances to TSSs.

All coordinates are handled consistently on the GRCh38 assembly. Where necessary, NCBI chromosome accessions for TSS entries are mapped to UCSC style chromosome names.

2.2 Extraction of GC profiles around TSSs

For each TSS we extract a symmetric window of GC values of radius R around the start site from the `gc5Base` bigWig. In the main analysis we take $R = 5\,000$ base pairs, giving windows of length $2R = 10\,000$ positions. For a TSS at coordinate x_{TSS} on chromosome c, the intended window runs from $x_{\mathrm{TSS}} - R$ to $x_{\mathrm{TSS}} + R$. If this window overlaps the start or end of the chromosome, we trim to the available range and pad missing positions with `NaN` values.

For each TSS we record:

- the gene identifier and, when available, the gene symbol,
- the chromosome name and strand,
- the TSS coordinate on the reference assembly,
- a vector $\{g_i\}_{i=-R}^{R-1}$ of GC fractions for positions at offset i from the TSS.

For genes on the negative strand, the extracted GC vector is reversed so that in all cases the index $i = 0$ corresponds to the TSS and positive offsets correspond to downstream positions in the sense of transcription. This produces a consistent coordinate system in which each promoter is represented by a one dimensional GC profile $g(i)$ indexed by integer offsets $i \in \{-R, \ldots, R-1\}$.

Windows with fewer than a minimum number of finite GC entries after padding are excluded from further analysis. In practice, with $R = 5\,000$ this removes only TSSs that are very close to chromosome ends.

2.3 Models for GC profiles

We fit three parametric models to each GC profile. In each model the predicted GC fraction at offset $s \in \mathbb{R}$ is described by a baseline level plus a local perturbation controlled by a small number of parameters. For fitting we treat the offsets $i \in \{-R, \ldots, R-1\}$ as positions s_i in base pairs.

These three forms have simple probabilistic interpretations. The Gaussian profile represents pure diffusion of a scalar field from a local constraint under neutral drift. The exponential profile represents the stationary covariance of a linear Ornstein Uhlenbeck process, that is, relaxation toward the background with a single screening length but no central plateau. The Fisher profile adds one additional parameter, r_{core}, which allows a finite width plateau near the TSS; this is the static equilibrium implied by a bounded entropy Fisher functional for a scalar field on a line. In parallel, super enhancer catalogues are beginning to use entropy based scores to classify regulatory domains [8]; our Fisher functional can be viewed as a continuous, geometry based refinement of this trend, replacing discrete sequence entropy by a field theoretic stiffness and screening length.

All profiles are fitted on a symmetric window of width $2R$ around the annotated TSS. In the main analysis we use $R = 5000\,\mathrm{bp}$, so the profiles are constrained by data on a 10 kb interval. The fit bounds on r_{core} and L are chosen to be of order the window size, so cases where the optimum sits on the upper bound for either parameter should be interpreted as lower bounds: the underlying plateau or screening length is at least of order 2 kb or 10 kb, but may in reality extend beyond the fitting window.

2.3.1 Gaussian diffusion profile

The first baseline model represents the GC profile as the equilibrium of a diffusion process from a point like constraint. The predicted GC fraction at offset s is

$$f_{\mathrm{gauss}}(s; A, \sigma, b) \;=\; b + A\exp\left(-\frac{s^2}{2\sigma^2}\right), \tag{2.1}$$

where b is the far field GC fraction, A is the amplitude of the perturbation, and σ is a characteristic diffusion length. For $A > 0$ this yields a symmetric GC peak around the TSS; for $A < 0$ it yields a symmetric GC dip.

2.3.2 Exponential decay profile

The second baseline model represents a profile with a central feature that decays exponentially with distance, as might arise from a simple screened process. The predicted GC fraction is

$$f_{\mathrm{exp}}(s; A, L, b) \;=\; b + A\exp\left(-\frac{|s|}{L}\right), \tag{2.2}$$

where L is a screening length. This form is also the zero temperature limit of a Fisher sector with vanishing core radius and is included here as a simple, widely used one dimensional profile.

2.3.3 Fisher cored profile

The Fisher bounded entropy sector considered in [4] starts from a functional for a scalar field σ_{F} that is the logit of a Bernoulli occupation probability and includes a bounded entropy potential. In one dimension, under a Bogomolny type completion, the equilibrium profiles of σ_{F} lead to cored Yukawa type profiles for observables that depend monotonically on σ_{F}. A convenient analytic approximation for the GC fraction in that case is

$$f_{\mathrm{F}}(s; A, r_{\mathrm{core}}, L, b) \;=\; b + A \exp\left(-\frac{\sqrt{s^2 + r_{\mathrm{core}}^2}}{L}\right), \tag{2.3}$$

where r_{core} is a core radius and L is a screening length inherited from the Fisher stiffness and effective mass in the underlying scalar field theory. For $|s| \ll r_{\mathrm{core}}$ the factor inside the exponential is approximately constant, giving a central plateau; for $|s| \gg r_{\mathrm{core}}$ the profile decays approximately exponentially with length scale L. The amplitude A and background b are defined as before.

We emphasise that (2.3) is not introduced as an arbitrary flexible curve. Its form and parameter interpretation are inherited from the Fisher bounded entropy functional developed in earlier work [2, 4]. In this paper we do not rederive that functional, but simply import the resulting profile as a candidate model for promoter GC geometry.

2.4 Nonlinear fitting and model selection

For each GC profile $g(i)$ associated with a TSS window, we fit the three models f_{gauss}, f_{exp}, and f_{F} by nonlinear least squares on the finite entries of the GC vector. We denote by g_k the GC fraction at position s_k in the window after masking missing values, with $k = 1, \ldots, n$.

Before fitting, GC percentages from `gc5Base` are converted to fractions in $[0, 1]$. Profiles that appear to be reported in percentage units are rescaled accordingly. For each model $f(\cdot; \theta)$ with parameter vector θ, we minimise the residual sum of squares

$$\mathrm{RSS}(\theta) \;=\; \sum_{k=1}^{n} \big(g_k - f(s_k; \theta)\big)^2 \tag{2.4}$$

using the `curve_fit` routine from SciPy. To aid stability and avoid unphysical fits, we impose simple box constraints on parameters:

- background $b \in [0, 1]$,
- amplitude $A \in [-1, 1]$,
- diffusion scale $\sigma \in [10, 50\,000]$ bp,
- screening length $L \in [10, 50\,000]$ bp,
- core radius $r_{\mathrm{core}} \in [0, 10\,000]$ bp.

Initial guesses for amplitudes are taken from the difference between the central value and the median of the outer part of the window. Initial guesses for b are taken as the median of the outermost few hundred positions. Initial scales for σ, L, and r_{core} are set to values of order 10^3 bp.

For each fitted model we record the residual sum of squares RSS and the number of free parameters k (three for the Gaussian and exponential, four for the Fisher cored model). To compare models we use the Akaike Information Criterion

$$\mathrm{AIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2k, \tag{2.5}$$

where n is the number of data points used in the fit. For each promoter we compute $\mathrm{AIC}_{\mathrm{gauss}}$, $\mathrm{AIC}_{\mathrm{exp}}$, and $\mathrm{AIC}_{\mathrm{F}}$.

We define the best model at a TSS as the model with the smallest AIC. To quantify the strength of evidence for the Fisher profile we consider the differences

$$\Delta_{\mathrm{F,gauss}} = \mathrm{AIC}_{\mathrm{F}} - \mathrm{AIC}_{\mathrm{gauss}}, \qquad \Delta_{\mathrm{F,exp}} = \mathrm{AIC}_{\mathrm{F}} - \mathrm{AIC}_{\mathrm{exp}}. \tag{2.6}$$

Negative values indicate that the Fisher cored profile is preferred over the corresponding baseline. In keeping with standard model selection conventions, we regard $\Delta < -10$ as strong evidence in favour of the Fisher model against that baseline.

In the results we distinguish:

- loci where the Gaussian is the best model,
- loci where the exponential is the best model,
- loci where the Fisher cored model is the best model,

and we define a promoter as *Fisher core required* if:

1. the Fisher cored model is the best by AIC, and
2. $\Delta_{\mathrm{F,gauss}} < -10$ and $\Delta_{\mathrm{F,exp}} < -10$.

This criterion demands that the Fisher profile not only fits better than each baseline, but does so by a margin that penalises the extra parameter in (2.3).

2.5 Distances to Human Accelerated Regions and enrichment analysis

To relate Fisher core promoters to Human Accelerated Regions we compute the linear distance from each TSS to its nearest HAR on the same chromosome. For each HAR interval with start a and end b we define the midpoint

$$x_{\mathrm{HAR}} = \frac{a+b}{2}. \tag{2.7}$$

For a TSS at coordinate x_{TSS} on chromosome c, the signed distance to that HAR is $d = x_{\mathrm{TSS}} - x_{\mathrm{HAR}}$; the absolute distance is $|d|$. For each TSS we compute the minimum absolute distance to any HAR on the same chromosome,

$$d_{\mathrm{min}} = \min_{\text{HAR on chromosome}} |x_{\mathrm{TSS}} - x_{\mathrm{HAR}}|. \tag{2.8}$$

We then define, for a given distance cutoff D, a binary indicator

$$I_{\text{near}}(D) = \begin{cases} 1, & \text{if } d_{\min} \le D, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

In the main analysis we consider $D = 10$ kb, 50 kb, and 100 kb, which span typical scales for cis regulatory interaction in one dimensional genomic coordinates.

For each cutoff D we tabulate counts in a 2 × 2 contingency table that cross Fisher core status with HAR proximity:

	HAR near ($d_{\min} \le D$)	HAR far ($d_{\min} > D$)
Fisher core required	$n_{\text{core,near}}(D)$	$n_{\text{core,far}}(D)$
Not Fisher core required	$n_{\text{noncore,near}}(D)$	$n_{\text{noncore,far}}(D)$

From this table we compute an odds ratio

$$\text{OR}(D) = \frac{n_{\text{core,near}}(D)\, n_{\text{noncore,far}}(D)}{n_{\text{core,far}}(D)\, n_{\text{noncore,near}}(D)}, \tag{2.10}$$

and a one sided Fisher exact test P value for enrichment of Fisher core promoters among HAR near TSSs. The odds ratio $\text{OR}(D)$ measures how much more likely a promoter is to be Fisher core required if it lies within distance D of a HAR, compared with promoters lying further than D from any HAR.

We also extract the list of genes whose promoters are Fisher core required and lie within a given cutoff, for example 50 kb, of a HAR. For these genes we inspect known functional annotations to provide qualitative context in the results. A more detailed Gene Ontology enrichment analysis for the full set of Fisher core required genes can be provided as a supplementary table but is not central to the present note.

3 Results

3.1 Model selection across 142 585 promoters

We first applied the three profile models from Section 2.3 to the GC fraction windows around all 142 585 TSSs in the GENCODE v44 catalogue. Fits converged and yielded finite AIC values for every promoter.

Table 1 summarises the best model by AIC. The Gaussian and exponential baselines account for the majority of promoters. The Fisher cored profile is selected as the best model at 4 978 promoters, corresponding to about 3.5% of the total.

To identify promoters where the Fisher sector is not only preferred, but required, we applied the strong AIC threshold from Section 2.4. A TSS was classified as *Fisher core required* if the Fisher cored model was the best by AIC and improved on *both* Gaussian and exponential baselines by at least $\Delta\text{AIC} < -10$. This yields 620 promoters, corresponding to 0.435% of the catalogue.

Table 1: Best fitting GC profile model by AIC across 142 585 human promoters.

Best model	Count	Fraction of promoters
Gaussian	105 359	73.9%
Exponential	32 248	22.6%
Fisher cored	4 978	3.5%

The full distribution of ΔAIC values shows that for most promoters the Fisher profile either does not improve over the baselines or does so by less than the penalty for the extra parameter. The Fisher core required subset occupies the extreme negative tail of this distribution. A histogram of ΔAIC values and a bar chart of best model counts provide a compact visual summary of this behaviour (Figure 1).

The small size of this subset is a statement of specificity. By construction, the three way model comparison rejects the Fisher profile for the majority of promoters where a simple Gaussian or exponential decay already provides an adequate description of the GC profile. The fact that only about 0.4% of loci require a cored Fisher profile to achieve a strong information theoretic improvement over both baselines therefore suggests that the Fisher geometry is acting as a high specificity filter for promoters with unusually extended, plateau like GC structure rather than as a generic description of promoter GC content.

3.2 Fisher parameter ranges for Fisher core required promoters

For the 620 Fisher core required promoters, the fitted Fisher parameters $(A, r_{\mathrm{core}}, L, b)$ exhibit a characteristic range. In a representative subset with additional quality cuts, the Fisher screening length L has median value of order 800 bp, with an interquartile range from a few hundred to just over 1 000 bp. The core radius r_{core} typically lies between tens and a few hundred base pairs, with a median around 100 bp, consistent with a smoothed central region on the scale of a nucleosome footprint.

The Fisher amplitudes A are centred on positive values around 0.4, indicating GC enrichment at the TSS relative to the flanking background for most Fisher core required promoters, but the distribution also includes negative values down to approximately -1, corresponding to local GC depletion relative to the surrounding sequence. Background levels b lie in the expected range for GC fraction in the local genomic context.

For a small number of Fisher core required promoters the fitted parameters reach the upper bounds imposed in the optimisation for r_{core} and L. In these cases the GC profile within the ±5 kb window appears comparatively broad and flat around the TSS, with only a slow decay toward the window edges. As a result, the data constrain only lower bounds on the core size and screening length and do not allow a sharp separation between these scales. We treat these bound hitting fits as indicating an extended plateau and long range decay within the chosen window, but we do not interpret the numerical values of r_{core} and L in such cases as precise estimates.

Distributions of r_{core}, L, and A for the Fisher core required promoters are shown as histograms in Figure 2. A scatter plot of core radius versus amplitude illustrates that large positive amplitudes are compatible with both narrow and moderately wide cores,

Table 2: Enrichment of Fisher core required promoters near Human Accelerated Regions. For each distance cutoff D we report counts of Fisher core required and non core promoters that lie within D of a HAR (HAR near) or further than D from any HAR (HAR far), together with the odds ratio $\mathrm{OR}(D)$ and one sided Fisher exact test P value for enrichment of Fisher core promoters among HAR near TSSs.

Cutoff D	$n_{\text{core,near}}$	$n_{\text{core,far}}$	$n_{\text{noncore,near}}$	$n_{\text{noncore,far}}$	$\mathrm{OR}(D)$
10 kb	7	613	1 285	140 680	1.25
50 kb	29	591	5 378	136 587	1.25
100 kb	62	558	10 668	131 297	1.37

One sided P values for enrichment: $P = 3.3 \times 10^{-1}$ at 10 kb, $P = 1.5 \times 10^{-1}$ at 50 kb, and $P = 1.4 \times 10^{-2}$ at 100 kb.

whereas negative amplitudes tend to coincide with wider cores and flatter profiles.

3.3 Enrichment of Fisher core promoters near Human Accelerated Regions

We next compared Fisher core status with distance to Human Accelerated Regions using the Pollard and Capra extended HAR set in hg38. For each promoter we computed the minimum absolute distance $d_{\min}$ to any HAR midpoint on the same chromosome, then classified promoters as HAR near or HAR far at three distance cutoffs $D = 10$ kb, 50 kb, and 100 kb. Table 2 reports the resulting contingency tables, odds ratios, and P values.

At 10 kb and 50 kb the odds ratios are modest and the P values are not significant at conventional thresholds. At 100 kb the odds ratio increases to $\mathrm{OR}(100\text{ kb}) \approx 1.37$, and the one sided P value falls to $P \approx 1.4 \times 10^{-2}$. This indicates that, at a scale of order 100 kb in linear genomic distance, Fisher core required promoters are more likely than average to lie within the regulatory neighbourhood of a HAR. The effect size is not large, but it is consistent with the idea that Fisher core promoters mark loci where the local noncoding environment has experienced atypical evolutionary constraints.

3.4 HAR proximal Fisher core promoters

Focusing on the 50 kb cutoff, we extracted the genes whose promoters are Fisher core required and lie within 50 kb of a HAR. This yields 29 promoter instances corresponding to a smaller number of unique gene loci once duplicates and pseudogenes are collapsed. Table 3 lists a subset of these genes together with their distances to the nearest HAR and fitted Fisher parameters.

Several of the genes in Table 3 are well studied developmental regulators. *PAX6* is a master regulator of eye and forebrain development. *HAND2* participates in limb and heart patterning. *SIM1* is involved in hypothalamic development and energy balance. *CDX2* and *WT1* regulate axial patterning and urogenital development respectively. The presence of these genes among HAR proximal Fisher core promoters is consistent

Table 3: Example Fisher core required promoters within 50 kb of a Human Accelerated Region. For each gene we report the distance to the nearest HAR midpoint, the Fisher amplitude A, core radius r_{core}, and screening length L. Parameters that reach the imposed upper bound are indicated.

Gene	Distance to HAR (bp)	A	r_{core} (bp)	L (bp)
HAND2	31 130	1.00	1 936	1 124
PAX6	8 732	−1.00	1 718	775
SIM1	8 187	0.23	1 999*	1 664
CDX2	3 186	0.66	2 000*	15 000*
WT1	41 967	−0.52	2 000*	15 000*
PRDM8	15 205	0.78	2 000*	1 146
CTBP2	36 756	1.00	1 742	600
PNOC	8 671	1.00	1 416	540

*Parameter at or very near the upper bound of the allowed fit range, indicating that only a lower bound on the scale is constrained by the data in the ±5 kb window.

with prior evidence that many HARs act as developmental enhancers for brain and limb [5, 6].

We stress that, although the fitted Fisher parameters for some of these loci reach high values for r_{core} and L, those numerical values are partly determined by parameter bounds and limited window size. The robust conclusion is that these promoters exhibit broad, slowly varying GC profiles around the TSS that are better captured by a cored Fisher profile than by the simpler baselines.

4 Discussion

The methodological proof-of-principle is significant. We have used a simple, large scale test to ask whether a bounded entropy Fisher geometry, originally introduced in the context of scalar halos in information hydrodynamics, is useful as an effective (high-specificity filter) of noncoding sequence around human promoters.

The fitted Fisher parameters admit a simple qualitative interpretation. For the bulk of Fisher preferred promoters we find core radii r_{core} of order 100 bp to 200 bp and screening lengths L of order 500 bp to a few kilobases. These scales are consistent with the size of nucleosome scale GC rich features and promoter associated CpG islands, and with the expectation that open, accessible promoter regions should extend over at least a few hundred base pairs around the TSS while relaxing toward the genomic background on kilobase scales. For the small number of promoters where r_{core} or L saturate the parameter bounds, the natural reading is that the underlying plateau or domain extends beyond the 10 kb window used here; a multi scale analysis with larger windows would be required to distinguish such extended cores from genuinely flat backgrounds.

The modest enrichment of Fisher core promoters near human accelerated regions suggests one possible evolutionary role for this geometry. Extended GC plateaux provide a stable, accessible chromatin environment in which lineage specific substitutions can accumulate in nearby regulatory elements without destroying the overall promoter architecture. Our data are consistent with the view that Fisher like cores may act as stable platforms that support accelerated regulatory evolution, but they do not in themselves establish a direct mechanistic link between the two.

The test uses only one dimensional GC fraction profiles, three parametric curves, and standard information criteria. Within that restricted setting, three main observations emerge.

First, when all three models are allowed to compete, Gaussian diffusion and exponential decay profiles account for the majority of promoter GC windows. This supports the general view that for most promoters, local GC structure is adequately described by relatively simple processes that do not require an explicit core scale beyond the smoothing already present in the data. The Fisher cored profile is selected as the best model for a few percent of promoters, but the strong AIC threshold used here identifies a much smaller subset in which the extra parameter is justified by the data.

Second, the Fisher core required promoters show a characteristic range of fitted parameters. Most have core radii on the order of a nucleosome footprint and screening lengths on the order of one kilobase. These scales are broadly compatible with the idea that Fisher cores correspond to local regions where GC content has been smoothed and stabilised over the footprint of chromatin and its immediate vicinity. The presence of both positive and negative amplitudes indicates that the Fisher geometry does not enforce a particular direction of GC change, but rather captures the presence of a plateau like region around the TSS relative to the flanking sequence.

Third, Fisher core required promoters show a modest enrichment in the neighbourhood of Human Accelerated Regions at scales of order 100 kb in linear distance. The effect size is not large and the enrichment is not visible at the smallest cutoffs, which is unsurprising given that enhancers often act at tens to hundreds of kilobases in one dimensional coordinates.

Nonetheless, the combination of a geometrically defined promoter subset and an independently defined catalogue of accelerated noncoding elements produces a coherent signal: promoters that require a Fisher core geometry are somewhat more likely than average to sit in regulatory neighbourhoods that experienced lineage specific acceleration on the human branch.

Independent evidence indicates that GC biased gene conversion can account for a substantial fraction of lineage specific acceleration in ultraconserved regulatory elements across mammals and birds [9]. The Fisher cores we observe may therefore be read, cautiously, as geometric equilibria of such long term GC biased processes, providing a stable plateau on which lineage specific substitutions can accumulate without immediately disrupting promoter function.

The overlap gene set in Table 3 provides a qualitative cross check. Several of the genes are canonical developmental regulators with well characterised roles in brain, limb, and axis patterning, and several have been linked directly or indirectly to HAR function in previous work [5, 6]. This alignment is not proof of mechanism, but it is consistent with the idea that the Fisher core geometry marks loci at which noncoding sequence has been shaped by evolutionary constraints beyond neutral drift.

GC rich promoter geometries are already implicated in human disease, for example through tandem repeat expansions in AFF3 and related loci that show strong associations with neurodevelopmental phenotypes [10]. This suggests that outlier Fisher cores, especially extreme plateaus or very long screening lengths, could offer a compact way to flag regulatory regions where GC geometry itself is part of the pathogenic mechanism.

We have used only GC fraction as an observable and only one dimensional windows of fixed radius. Extending the model to integrate CpG island annotations, chromatin accessibility data, and histone modification tracks would allow more direct biological interpretation of the fitted parameters.

Likewise, allowing the window radius to vary as a function of promoter class or incorporating multi scale fits could reduce parameter degeneracies for bound hitting cases. The exponential model considered here is also the zero core limit of a Fisher sector, so a fuller comparison of Fisher based models may be informative.

On the evolutionary side, our enrichment calculation uses linear genomic distance as a proxy for regulatory proximity and a single catalogue of HARs. Incorporating chromatin conformation data to use contact based distances and cross referencing multiple sets of accelerated elements would refine the picture. More sophisticated statistical models could also take into account the local density of genes and regulatory elements when assessing enrichment.

The present analysis provides a compact, initial result, from a quick and simple probe: when GC fraction around human promoters is modelled as the equilibrium of an effective scalar field, a small subset of promoters is better described by a cored Fisher profile imported from a bounded entropy Fisher functional than by two standard baselines.

These Fisher core promoters are enriched near Human Accelerated Regions, and the overlap genes include several key developmental regulators. This suggests that the Fisher bounded entropy geometry, motivated by information hydrodynamics, may have a role as an effective description of regulatory landscapes in the human genome.

Acknowledgements

We thank the maintainers of the UCSC Genome Browser and the GENCODE consortium for making the underlying data available in convenient machine readable form. Computational experiments for this work were performed using publicly available tools in the Python scientific ecosystem.

References

[1] H. Zhou, J. Huertas, M. J. Maristany, K. Russell, J. H. Hwang, et al. Multi-scale structure of chromatin condensates rationalizes phase separation and material properties. *bioRxiv*, 2025. Preprint 2025.01.17.633609.

[2] J. R. Dunkley. Universal Information Hydrodynamics Preprint, 2025.

[3] J. R. Dunkley. Hypocoercive renormalisation Preprint, 2025.

[4] J. R. Dunkley. Emergent Fisher Halos from Information Geometry Preprint, 2025.

[5] K. S. Pollard, S. R. Salama, N. Lambert, M. A. Lambot, S. Coppens, et al. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, 443:167–172, 2006.

[6] J. A. Capra, G. D. Erwin, G. McKinsey, J. L. R. Rubenstein, K. S. Pollard. Many human accelerated regions are developmental enhancers. *Philosophical Transactions of the Royal Society B*, 368:20130025, 2013.

[7] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. Harrow, R. Guigó, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2019.

[8] B. Shi, J. Zhao, Y. Li, Y. Gu, et al. SEA version 4.0: a major expansion and update of the Super-Enhancer Archive. *Nucleic Acids Research*, 2025.

[9] A. Liu, N. Wang, G. Xie, Y. Li, X. Yan, X. Li, et al. GC-biased gene conversion drives accelerated evolution of ultraconserved elements in mammalian and avian genomes. *Genome Research*, 33(10):1673–1689, 2023.

[10] B. S. Jadhav, A. Ameur, A. R. Quinlan, N. C. L. Ng, A. T. Nordin, et al. A phenome-wide association study of methylated GC-rich repeats identifies a GCC repeat expansion in AFF3 as a novel cause of intellectual disability. *Nature Genetics*, 56(11):2322–2332, 2024.