Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Codeness: a measurable index that separates self-templating epigenetic codes from sequence echoes across the tree of life

Davis W. Cheng

chengdw1980@gmail.com

Independent Researcher

Method Article

Keywords: epigenetic code, DNA methylation, mutual information, heritability, chromatin, transcriptional memory, codeness index

Posted Date: October 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11183912/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Whether the epigenome constitutes a "code" has been debated for a quarter-century, largely qualitatively. We make the question measurable by defining an epigenetic layer's *codeness* as the geometric mean of three axes — heritability, causality and sequence-predictability — with heritability estimated as the mutual information a mark's daughter-cell state carries about its mother-cell state conditional on DNA sequence, separating genuine self-templating from sequence-driven re-derivation.

Methods

We built a Monte-Carlo engine that parameterizes each axis from published perturbation and sequence-prediction data, and reanalyzed two primary datasets to ground the two load-bearing axes: Hammer-seq DNA-methylation inheritance and the genome-wide CRISPRoff transcriptional-memory screen.

Results

Across seven mammalian layers the index resolves two mechanistic tiers — self-templating codes (DNA methylation, repressive marks, CENP-A) and sequence echoes (chromatin accessibility, active marks, most non-coding RNA). DNA methylation is the most code-like layer, whereas the most sequence-predictable layer is among the least code-like, so codeness is not predictability. Genome-wide methylation-inheritance mutual information is 0.72, and readability predicts CRISPRoff memory with a cross-validated AUC of 0.71. The index extends across bacteria, fungi, plants, invertebrates and mammals, where it suggests a separable "universal epigenetic codeness law" — a conserved layer ranking rescaled by a per-kingdom gain, broken only by carrier switches — which we advance as a falsifiable hypothesis.

Conclusions

Codeness is a measurable, versioned and reproducible quantity that concentrates in DNA methylation in mammals and separates self-templating codes from sequence echoes; the self-templating memory function is conserved across the tree of life, but the molecule that carries it is lineage-specific.

Background

For a quarter-century, "is there an epigenetic code?" has been posed as a yes/no question and answered largely qualitatively. The genetic code is a fixed, near-universal cipher mapping 64 codons to 20 amino acids; by analogy, the strong "histone code" hypothesis proposed that combinations of chromatin marks constitute a comparably readable cipher [1]. Critics counter that most chromatin marks correlate with

transcriptional state rather than instruct it, follow rather than lead, and lack the fixed, position-wise readout that defines a code [2, 3].

The impasse is partly definitional. "Code" bundles at least three separable properties that are individually measurable: whether a mark's state is faithfully inherited through cell division independent of the underlying sequence; whether it causally instructs downstream state; and whether it is predictable from sequence. These need not co-occur. DNA methylation is faithfully maintained after replication and causal for silencing but only moderately sequence-predictable [4], whereas chromatin accessibility is the most sequence-predictable layer yet is largely re-derived each cycle rather than copied.

Here we replace the qualitative question with a quantitative one. We define the *codeness* of each epigenetic layer as the geometric mean of three axes scaled to [0,1], C = (H · K · P)^(1/3), where H is heritability, K is causality and P is sequence-predictability. The geometric mean enforces that a layer is code-like only if it scores on all three axes. Critically, we estimate H as the mutual information between a mark's mother- and daughter-cell states conditional on DNA sequence, separating self-templating from sequence-driven re-derivation. We compute the index for seven molecular layers in mammals (Table 1, Fig. 1), test five falsifiable predictions (Table 3), ground the heritability and readability axes by reanalysis of primary data — Hammer-seq DNA-methylation inheritance [4] and the genome-wide CRISPRoff transcriptional-memory screen [5] — and extend the index across bacteria, fungi, plants, invertebrates and mammals (Fig. 6, Table 2). Throughout, we are explicit about which numbers are measured and which are current best estimates.

Table 1
Mammalian codeness by layer.

Layer	H (cond.)	K	P	C	Rank	Tier
DNA methylation	0.38	0.65	0.72	0.56	1	self-templating code
Histone (repressive)	0.14	0.75	0.52	0.38	2	self-templating code
Histone variant (CENP-A)	0.28	0.30	0.42	0.33	3	self-templating code
Histone (activating)	0.03	0.62	0.75	0.24	4	sequence echo
3D genome architecture	0.07	0.20	0.62	0.21	5	sequence echo (potentiator)
Chromatin accessibility	0.01	0.30	0.85	0.13	6	sequence echo
ncRNA guidance	0.03	0.14	0.48	0.12	7	sequence echo

Table 2
Most code-like layer per kingdom.

Kingdom	Top layer	Codeness	Basis
Bacteria	DNA methylation	0.53	cell-cycle adenine methylation; restriction/regulon
Fungi	DNA methylation	0.53	RIP/DIM-2 heterochromatin
Plants	DNA methylation	0.68	MET1 maintenance + RdDM/paramutation
Invertebrates (nematode)	Repressive marks (+ piRNA 0.26)	0.45	5mC absent; Argonaute/HRDE-1 inheritance
Mammals	DNA methylation	0.56	DNMT1 maintenance

Table 3
Prediction scorecard.

#	Prediction	Verdict	Evidence
P1	DNA methylation ranks most code-like	Supported	Index rank 1 (Table 1)
P2	Readable written states persist	Supported (primary data)	CRISPRoff: OR 5.1; multivariate AUC 0.71
P3	3D reorganization aids inheritance of a written state	Supported for establishment	Fab-7 interchromosomal contacts; 3D term
P4	Histone marks score below methylation on all three axes	Partly supported	Repressive K high; H lower
P5	Combinatorial syntax, no per-position cipher	Supported	+ 0.4–0.6 bit syntax gain; variant MI ≈ 0

Methods

The codeness index and Monte-Carlo engine

For each layer, C = (H·K·P)^(1/3) with H, K, P ∈ [0,1]. The index is evaluated by Monte-Carlo simulation (N = 200,000 draws) over triangular priors on each axis, reported as a median with a 90% credible interval per layer. The engine is implemented in Python (numpy, pandas, scipy, matplotlib); all random seeds are fixed at 20260920. Engine versions v0–v7 differ only in which priors have been replaced by data (Additional file 1).

Conditional-mutual-information heritability

We use a three-variable generative model. A latent binary sequence class S (high- versus low-affinity, prior π) sets the mother-mark prevalence; the mother mark M is drawn from that prevalence; the daughter mark D equals a copy of M with weight c and a sequence re-draw with weight (1 − c), where the copy channel has maintenance a = p(D = 1|M = 1) and de-novo rate b = p(D = 1|M = 0). Heritability is the normalized conditional mutual information H = I(M;D|S)/H(M|S), computed analytically from the 2×2 joints within each sequence class and averaged over π. The unconditional comparator is I(M;D)/H(M) on the sequence-marginalized joint (Fig. 2).

Axis parameterization

Causality K is the fraction of targets with a causal effect in perturbation studies (enhancer perturbation [6]; hit-and-run epigenome editing [5, 7]; protein degrons); predictability P is the reported accuracy of sequence-to-epigenome models (DeepCpG for methylation [8]; Enformer for chromatin [9]); the copying weight c is anchored in persistence-after-withdrawal experiments [5, 7, 10]. Elicited components (the sequence-channel prevalences; the 3D contact-coverage and potentiation efficiency) are labeled as such and given triangular priors.

WP1 — methylation-inheritance mutual information

Processed per-CpG event files for GEO GSE131098 [4] — counts of the four mother→daughter methylation states on both replication arms — were downloaded per sample from GEO. For wild-type chase samples (0 min, 30 min, 2 h, 10 h, 24 h), counts were summed genome-wide to a 2×2 mother→daughter table, from which the normalized mutual information H = I(M;D)/H(M), the maintenance rate a and the de-novo rate b were computed. Density stratification binned CpGs by count in 1-kb windows (sparse ≤ 2, low 3–7, dense ≥ 8) and computed H within each; the density-conditional H is the coverage-weighted within-stratum average (Fig. 3).

WP2 — readability predicts persistence

From the genome-wide CRISPRoff hit-and-run screen [5], each gene carries a CpG-island flag and a memory (hit-and-run) call. The binary test compared persistence rates by CpG-island status (Fisher exact test). For the graded analysis, each target was anchored at its guide coordinate (screen sgIDs are hg19), the chromosome was obtained from UCSC refGene (hg19), and the coordinate was lifted to hg38 (pyliftover). Graded readability used (i) ENCODE HEK293 H3K9me3 fold-change signal (experiment ENCSR000FCJ, file ENCFF902RQI, GRCh38) averaged over the promoter (± 2 kb) via pyBigWig [12], and (ii) CpG observed/expected ratio, CpG density and GC content over the promoter (± 1 kb) from the hg19 two-bit sequence. AUC was computed as the Mann–Whitney statistic; the multivariate model was L2-regularized logistic regression (scikit-learn) on standardized features, evaluated by stratified 5-fold cross-validation. Initial-silencing strength was excluded after producing a perfect AUC, which a leakage check attributed to the memory call being thresholded on the silencing score (Fig. 4, Table 4).

Table 4
WP2 readability proxies (n = 18,984 targets).

Readability proxy	AUC (→ persistence)
ENCODE HEK293 H3K9me3 (repressive mark)	0.44
CpG island (binary flag)	0.60
CpG observed/expected (graded)	0.64
Multivariate, readability-only (5-fold CV)	0.71 ± 0.01

3D interaction term

A 3D contact raises the copying weight of the enclosed mark toward a contact-stabilized ceiling: $c_{eff}(L) = c(L) + \varphi \cdot (c_{ceiling} - c(L)) \cdot \theta(L)$, with $\varphi \sim$ triangular(0.5, 0.75, 0.9) (from the sufficiency of engineered Fab-7 contacts [11]), $c_{ceiling} = 0.9$, and θ the fraction of a layer's loci in stabilizing contacts. The 3D "systemic value" is the summed increase in codeness across layers (Fig. 5).

Head-to-head comparisons and cross-kingdom index

Both head-to-head comparisons used identical generative parameters to isolate the effect of the definition; ranking correlations are Spearman ρ (Fig. 2). For the cross-kingdom index (Fig. 6, Table 2), organism-specific causality, predictability and copying-weight anchors were applied where perturbation or prediction data exist (bacterial [16] and plant [17] methylation; nematode small-RNA/Argonaute inheritance [15]; fungal and plant repressive marks); the remaining cells inherit mammalian values and are reported as transferred. All analysis code includes a synthetic-data self-test with fixed seed (see the archived analysis code; Availability of data and materials).

Separable-law analysis

To test whether codeness is separable across layers and kingdoms, the cross-kingdom matrix (29 present layer×kingdom cells) was fit to a two-way multiplicative model, $\log C(L,K) = \log C0 + \log \alpha_L + \log \beta_K$, by least squares on the present cells; the layer constant α_L and per-kingdom constant β_K are the recovered effects and the interaction $\varepsilon(L,K)$ is the residual. Model fit is reported as the fraction of log-codeness variance explained (R^2). The whole-epigenome aggregate is $C_{system} = C_{dominant} + \rho \cdot [(1 - \prod_L (1 - C_L)) - C_{dominant}]$, where $\rho \in [0,1]$ interpolates between fully redundant layers ($\rho = 0$, dominant layer only) and fully independent layers ($\rho = 1$); its sensitivity was assessed by sweeping ρ. The full derivation, constants, redundancy sensitivity, falsification predictions and a data-collection plan are provided in Additional file 4.

Results

Mammalian layers resolve into two mechanistic tiers

Scored with data-derived causality for all seven layers, the mammalian ranking is DNA methylation (C = 0.56) > repressive histone marks (0.38) > the CENP-A histone variant (0.33) > activating histone marks (0.24) > 3D genome architecture (0.21) > chromatin accessibility (0.13) > non-coding-RNA guidance (0.12) (Table 1, Fig. 1). This resolves the epigenome into **self-templating codes** (DNA methylation, repressive Polycomb/heterochromatin marks, and the replication-independent CENP-A variant, whose centromeric chromatin is propagated by a dedicated licensing pathway [13]) and **sequence echoes** (chromatin accessibility, activating marks, most ncRNA, much of 3D architecture) that are readable but re-derived each cycle. DNA methylation is the most code-like layer; chromatin accessibility, the most sequence-predictable layer, is nearly the least code-like.

Codeness is not predictability, and conditioning on sequence creates the tiers

Two head-to-head comparisons, computed from identical generative parameters so they isolate the effect of the definition rather than the inputs, distinguish codeness from simpler measures (Fig. 2). First, ranking the layers by codeness versus by sequence-predictability alone yields essentially unrelated orders (Spearman $\rho = -0.07$): chromatin accessibility is the most sequence-predictable layer ($P = 0.85$, rank 1) yet ranks sixth of seven on codeness, whereas DNA methylation is predictability-rank 3 but codeness-rank 1 (Fig. 2A). A predictability-only score would mislabel the most readable layer as the most code-like. Second, replacing the conditional mutual information $I(M;D|S)$ with the naïve unconditional $I(M;D)$ inflates every layer's heritability, but most for the sequence-driven layers (accessibility + 0.27, activating marks + 0.25) and negligibly for self-templating ones (histone variants + 0.02); under the unconditional measure a sequence echo outranks genuinely self-templating layers and the two-tier structure collapses (Fig. 2B). Conditioning on sequence is what creates the tiers.

Primary-data heritability: DNA methylation is a high-fidelity self-templating register

Reprocessing Hammer-seq data, which measures methylation on the parent and daughter strands of the same CpG after replication (GSE131098 [4]), we summed the per-CpG event counts genome-wide over 10.1 million CpG sites in wild-type cells at the end of the cell cycle. The normalized mother–daughter mutual information is **$H = 0.72$**, with maintenance fidelity $a = 0.933$, de-novo rate $b = 0.007$ and prevalence $p(M = 1) = 0.785$ (Fig. 3). The maintenance kinetics reproduce the original study directly from the raw counts: fidelity rises 0.55 → 0.82 → 0.89 → 0.93 at 0 min, 30 min, 2 h and 10–24 h of chase, with H rising 0.24 → 0.72 in step (Fig. 3A). Maintenance is essentially uniform across local CpG density (dense 0.934, low 0.928, sparse 0.921), and conditioning H on local CpG density barely changes it (0.725 → 0.724), indicating that the mother–daughter signal is genuine copying rather than sequence re-derivation at this resolution (Fig. 3B).

Primary-data readability: readable states persist, and the reader is the DNA-methylation substrate

The framework predicts that written states which persist after the writer is removed sit in a more readable context (Prediction 2). On the genome-wide CRISPRoff screen [5], across 21,240 genes, CpG-island promoters (a readability proxy) maintain silencing far more often than non-island promoters: P(persist | CpG-island) = 11.5% versus 2.5%, odds ratio 5.1, AUC 0.62, Fisher one-sided p = 5 × 10^−19⁰ (Table 4). A graded-readability analysis localizes the operative reader (Fig. 4). Anchoring each of 18,984 targets at its guide coordinate, the repressive mark CRISPRoff itself deposits — ENCODE HEK293 H3K9me3, in the screen's own cell line [12] — does **not** predict persistence (AUC 0.44, slightly reversed), because CpG-island promoters are H3K9me3-poor. Replacing the binary flag with graded CpG-island strength (CpG observed/expected over the promoter) raises the AUC to 0.64 (Fig. 4A), and a readability-only logistic model (CpG O/E + CpG density + GC + H3K9me3), evaluated by 5-fold cross-validation, reaches **AUC = 0.71 ± 0.01** (Fig. 4B, Table 4). Initial-silencing strength was excluded: adding it drives the AUC to a spurious 1.0, because the screen's memory call is itself thresholded on the silencing score — a label-leakage trap. The readable feature that carries CRISPRoff memory is thus the DNA-methylation substrate (CpG-island character), not the repressive histone mark.

The 3D genome potentiates other layers' heritability

By a per-layer, steady-state-causality axis the 3D genome scores low (C = 0.21): cohesin/CTCF degrons abolish loops and TADs yet change little steady-state expression. But engineered interchromosomal Fab-7 contacts are sufficient to trigger stable transgenerational H3K27me3 inheritance in *Drosophila* [11], i.e., the role of 3D contacts is to potentiate other layers' heritability, which a per-layer axis cannot represent. Modeling a contact as raising the copying weight of the enclosed mark toward a stabilized ceiling, the summed potentiation ("3D systemic value") is ≈ 0.22 — about equal to 3D's standalone codeness, so counting only the standalone score roughly halves its true contribution (Fig. 5). Gains concentrate in the low-copying layers (activating marks, accessibility) and repressive Polycomb marks, while the self-templating codes barely change. This term is explicitly hypothesis-driven and parameterized from the Fab-7 result, not a new measurement.

The epigenome is combinatorial, not a per-position cipher

Converting explained variance to information ($I = -\frac{1}{2}\cdot\log_2(1 - r^2)$), single chromatin marks carry ~ 0.7–1.0 bit about expression, mark combinations carry ~ 1.5–2.0 bit, and the syntax gain from combining marks is + 0.4–0.6 bit (~ 80%), supporting a combinatorial "language" (Prediction 5, Table 3). Per-position variant-effect predictors carry only ~ 0.002–0.016 bit, i.e., there is no fixed per-position cipher.

The code exists across the tree of life, but its carrier is lineage-specific

Recomputing the index with organism-specific anchors resolves a cross-kingdom map (Fig. 6, Table 2). The most code-like layer is DNA methylation in bacteria (0.53; cell-cycle-regulated adenine methylation [16]), fungi (0.53), plants (0.68 — the highest cell anywhere, reflecting MET1-maintained CG methylation that is transgenerationally stable [17]) and mammals (0.56). In nematodes, which lack 5-methylcytosine, the top slot is filled instead by repressive histone marks (0.45) plus small-RNA (Argonaute/HRDE-1) inheritance (0.26) [15]. The functional register — a heritable, self-templating memory layer — is universal; the carrier molecule is lineage-specific. Roughly one-third of cross-kingdom cells are organism-anchored and two-thirds are transferred from mammalian data, labeled as such.

A separable "universal epigenetic codeness law" (hypothesis)

Box 1. The Universal Epigenetic Codeness Law, in words

Across the tree of life, how strongly any given molecular layer of the epigenome behaves as a heritable code is set by two largely independent things: an intrinsic property of the layer itself, which does not depend on the organism, and a single overall setting that belongs to the lineage. Every heritable layer carries a fixed intrinsic “code-ability” — a tendency to store information and transmit it faithfully through cell division — and the layers fall in the same rank order in every kingdom examined: DNA methylation is the most code-like, followed by repressive histone marks and specialized histone variants, then active histone marks, three-dimensional genome architecture, non-coding-RNA guidance, and finally chromatin accessibility, the least code-like. Each lineage then applies one overall “coding gain” that turns all of its layers up or down together, so that some lineages rely on epigenetic coding more than others while preserving the same internal ranking. The single systematic exception occurs where evolution has swapped the molecular carrier of memory — for instance, lineages that lack DNA methylation and shift the burden onto small RNAs or repressive marks. There, and essentially only there, the ranking is locally rearranged.

In everyday terms, each layer is a different writing surface ranked by how well it holds a permanent message — DNA methylation like ink on paper, chromatin accessibility like footprints in sand — and that ranking of surfaces is the same across bacteria, fungi, plants, animals, and humans; what differs between them is only how strongly each relies on epigenetic memory overall. The only real exceptions are lineages that switched to a different writing material altogether.

Box 2. The General Codeness Model, in brief

The claim, in one sentence. Across seven epigenetic layers and five kingdoms, a layer’s codeness is well described by a universal, layer-specific “code-ability” rescaled by a single per-kingdom gain, plus a small correction confined to lineages that switch the carrier molecule:

$C(L,K) = C0 \cdot \alpha_L \cdot \beta_K \cdot \varepsilon(L,K)$.

The terms. C0 is a baseline (the geometric grand mean of codeness). **α_L** — the universal layer constant, or “code-ability” — gives the same layer ranking in every kingdom (above 1, more code-like than average; below 1, less). **β_K** — the per-kingdom gain — is one scalar per lineage that raises or lowers all of its layers together. **$\varepsilon(L,K)$** — the interaction — is ≈ 1 wherever the law holds and departs from 1 only at carrier switches.

Fusing the code with the echoes. Treating each layer’s codeness as its chance of carrying heritable memory, the layers combine into a whole-epigenome codeness in which the dominant self-copying code sets a floor and the remaining layers (the “echoes”) add diminishing increments. Across kingdoms this ranking is robust (plants highest, bacteria lowest), while the absolute value depends on how independent the layers are assumed to be.

Fitting the separable model $C(L,K) = C0 \cdot \alpha_L \cdot \beta_K \cdot \varepsilon(L,K)$ on log-codeness (Fig. 7A) explains 95.7% of the variance in the separable part: the layer constant α_L ranks DNA methylation 2.34 > repressive 1.54 > CENP-A variant 1.23 > activating 0.88 > 3D 0.76 > ncRNA 0.68 > accessibility 0.49 (Fig. 7B), and the per-kingdom gain β_K ranges invertebrates 1.14 > plants 1.10 > fungi 1.02 > mammals 0.98 > bacteria 0.80 ($C0 = 0.252$). The residual interaction ε accounts for the remaining ~ 4% and is confined to carrier switches — small-RNA and repressive-mark codeness rise in the 5-methylcytosine-less invertebrate lineage, and ncRNA codeness falls in mammals; excluding invertebrates raises the separable fit to $R^2 = 0.972$. The whole-epigenome aggregate $C_{system} = C_{dominant} + \rho \cdot [(1 - \prod_L(1 - C_L)) - C_{dominant}]$, with a redundancy parameter ρ, ranks plants highest and bacteria lowest at every ρ, so the kingdom ranking is robust while its absolute value is not (Fig. 7C). We advance this separable relationship as a "universal epigenetic codeness law" only in the form of a falsifiable hypothesis: because two-thirds of cross-kingdom cells are transferred from mammalian data, the separability is partly built in, and its high R^2 reflects our construction as much as biology (Discussion). The informative part is the ~ 4% interaction, which comes from the independently anchored cells and lands precisely on the known carrier switches.

Predictions tested

Of five falsifiable predictions, four are supported and one is partly supported (Table 3): P1 (methylation ranks first) supported; P2 (readable states persist) supported on primary data; P3 (3D reorganization aids inheritance of a written state) supported for establishment by the Fab-7 result; P4 (histone marks score below methylation on all three axes) partly supported; P5 (combinatorial syntax, no per-position cipher) supported.

Discussion

The codeness index turns a 25-year qualitative debate into a measured quantity and yields three results a narrative review cannot: the epigenome resolves into self-templating codes and sequence echoes; codeness is not predictability (the most sequence-predictable layer is among the least code-like); and the same functional register is carried by different molecules in different kingdoms. The primary-data anchors matter because they convert the two most load-bearing axes from literature estimates to measurements — the methylation heritability from Hammer-seq ($H = 0.72$, Fig. 3) and the readability-to-persistence relationship from CRISPRoff (cross-validated AUC = 0.71, Fig. 4).

Three deeper caveats bound the interpretation, and we state them plainly. First, directionality: causality is scored for steady-state expression, and if it proves uniformly low across layers — marks following transcription rather than instructing it — the picture shifts from "instructive code" toward "heritable record"; the index confronts this by measuring causality rather than assuming it, so a low-causality result is an honest finding, not a failure. The 3D genome already exposes this, its standalone steady-state causality understating its role in establishing heritability [11], which also requires sequence-specific Polycomb recruitment [14]. Second, construct validity: codeness is measured exactly as defined, and the

geometric-mean aggregation is a reasoned modeling choice, not a derived law; a weighted arithmetic mean reorders the middle of the table, and the engine supports reporting both. Third, several parameters remain elicited and are labeled wherever they appear — the sequence-channel prevalences of the heritability model, the contact-coverage and potentiation efficiency of the 3D term, and the copying weight, which is anchored in persistence experiments but is cell-line-specific; the cross-kingdom map is two-thirds transferred from mammalian data.

These caveats bear directly on the separable "universal epigenetic codeness law" (Fig. 7). The law is appealing — a conserved layer ranking (α_L) rescaled by a lineage gain (β_K), broken only by carrier substitution (ε) — and, if real, would reframe the field's question into a single equation with interpretable constants. We are careful to present it as a hypothesis rather than a result, precisely because the transferred cross-kingdom cells make part of the separability automatic. The decisive test is an independently measured matrix, and the law makes sharp, pre-registrable predictions that such a matrix could falsify: (i) the layer ranking α_L should be conserved across kingdoms that share a carrier (rank correlation ≥ 0.7); (ii) each kingdom should reduce to a single scalar β_K, without strong layer-specific kingdom effects away from carrier switches; (iii) the interaction ε should remain confined to carrier-switch lineages; and (iv) the separable model should retain $R^2 \geq 0.8$ on independently measured data. A concrete plan to build that matrix — replacing the transferred cells with organism-specific measurements of causality, predictability and heritability, prioritizing the layers that currently force separability — is given in Additional file 4. Until then, the universal epigenetic codeness law is a testable conjecture about the index, not an established property of biology.

Two self-corrections from the primary-data reanalysis are worth recording, as both cut against tidier earlier expectations. A provisional "two-speed" methylation code (near-perfect at dense CpGs, weak at isolated sites) is not borne out by the raw Hammer-seq counts, where maintenance is uniform across CpG density; and in the readability analysis the obvious repressive-mark reader (H3K9me3) fails, the operative reader being the DNA-methylation substrate. That the readable feature is CpG-island character rather than the histone mark is consistent with methylation-versus-transcription-factor competition at CpG-rich promoters [18]. The readability AUCs, while highly significant given $n \approx 19{,}000$, are modest in effect size (0.60–0.71), because a single promoter feature only partly determines memory.

Relative to the strong histone-code hypothesis, these results favor a language, not a cipher: chromatin marks form a combinatorial, context-dependent, partly heritable regulatory language with a memory layer concentrated in DNA methylation, rather than a fixed position-wise lookup table. We state this lean explicitly so that a reader weighting the strong-code evidence more heavily can reach a different balance from the same numbers.

Conclusions

Codeness — heritability × causality × sequence-predictability, with heritability conditioned on sequence — is a measurable quantity that differs by layer, concentrates in DNA methylation in mammals, and

separates self-templating codes from sequence echoes; the self-templating memory function is conserved across the tree of life, but the molecule that carries it is lineage-specific. Two primary-data reanalyses ground its most consequential axes, and the cross-kingdom pattern suggests a separable universal epigenetic codeness law that we frame as a testable hypothesis. The question the field has argued about since 2000 becomes one that can be measured, versioned and reproduced.

Declarations

Ethics approval and consent to participate

Not applicable. This study reanalyzes publicly available, de-identified datasets and involved no human participants, human material or animals.

Consent for publication

Not applicable.

Competing interests

The author declares no competing interests.

Funding

No funding was received for this work.

Author Contribution

D.W.C. is the sole author. D.W.C. conceived and established the study's ideas and hypotheses, designed the analytical approach, identified and collected the primary datasets and the public bioinformatics tools and resources used, performed and directed the analyses, interpreted the results, and drafted, reviewed and approved the manuscript. D.W.C. takes full responsibility for all content.

Acknowledgements

The author conceived and established all ideas and study designs, identified and collected the data and the public bioinformatics tools and resources, and drafted the manuscript. Computational analyses, analysis code, figures, and manuscript drafting and editing were then carried out with the assistance of an AI system (Anthropic Claude) used as a tool under the author's direction and verification. The AI system does not meet authorship criteria and is not listed as an author. This use of AI is disclosed in

accordance with ICMJE, COPE, and publisher policy on the use of AI tools; the author takes full responsibility for all content, including any AI-assisted elements.

Data Availability

No new sequencing data were generated in this study; all primary datasets analysed are publicly available. DNA-methylation inheritance (Hammer-seq): Gene Expression Omnibus accession GSE131098. CRISPRoff transcriptional memory: GEO super-series GSE168012 and the published supplementary Table S3. H3K9me3 chromatin signal: ENCODE experiment ENCSR000FCJ, file ENCFF902RQI (HEK293). Genome sequence and annotation: UCSC refGene, the hg19 two-bit sequence, and the hg19-to-hg38 liftOver chain (GRCh37/hg19 and GRCh38/hg38).The complete analysis code — the codeness index engine (versions v0–v7), the work-package analyses, the separable-model fit, and the head-to-head comparison — together with the derived data tables and all result files, is openly available and permanently archived at Zenodo under DOI 10.5281/zenodo.22946911, released under the MIT License. The datasets and figures generated during this study are also included within this article and its additional files.

References

1. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293:1074–80.
2. Henikoff S, Greally JM. Epigenetics, cellular memory and gene regulation. Curr Biol. 2016;26:R644–8.
3. Ptashne M. Epigenetics: core misconcept. Proc Natl Acad Sci USA. 2013;110:7101–3.
4. Ming X, Zhang Z, Zou Z, Lv C, Dong Q, He Q, et al. Kinetics and mechanisms of mitotic inheritance of DNA methylation and their roles in aging-associated methylome deterioration. Cell Res. 2020;30:980–96.
5. Nuñez JK, Chen J, Pommier GC, Cogan JZ, Replogle JM, Adriaens C, et al. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. Cell. 2021;184:2503–19.
6. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.
7. O'Geen H, Tomkova M, Combs JA, Tilley EK, Segal DJ. Determinants of heritable gene silencing for KRAB-dCas9 + DNMT3 and Ezh2-dCas9 + DNMT3 hit-and-run epigenome editing. Nucleic Acids Res. 2022;50:3239–53.
8. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18:67.
9. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwińska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods.

2021;18:1196–203.
10. Bintu L, Yong J, Antebi YE, McCue K, Kazuki Y, Uno N, et al. Dynamics of epigenetic regulation at the single-cell level. Science. 2016;351:720–4.
11. Fitz-James MH, Sabarís G, Sarkies P, Bantignies F, Cavalli G. Interchromosomal contacts between regulatory regions trigger stable transgenerational epigenetic inheritance in Drosophila. Mol Cell. 2025;85:677–91.
12. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
13. Jansen LET, Black BE, Foltz DR, Cleveland DW. Propagation of centromeric chromatin requires exit from mitosis. J Cell Biol. 2007;176:795–805.
14. Coleman RT, Struhl G. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a Drosophila HOX gene. Science. 2017;356:eaai8236.
15. Buckley BA, Burkhart KB, Gu SG, Spracklin G, Kershner A, Fritz H, et al. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. Nature. 2012;489:447–51.
16. Kozdon JB, Melfi MD, Luong K, Clark TA, Boitano M, Wang S, et al. Global methylation state at base-pair resolution of the Caulobacter genome throughout the cell cycle. Proc Natl Acad Sci USA. 2013;110:E4658–67.
17. Mathieu O, Reinders J, Čaikovski M, Smathajja C, Paszkowski J. Transgenerational stability of the Arabidopsis epigenome is coordinated by CG methylation. Cell. 2007;130:851–62.
18. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015;528:575–9.

Figures

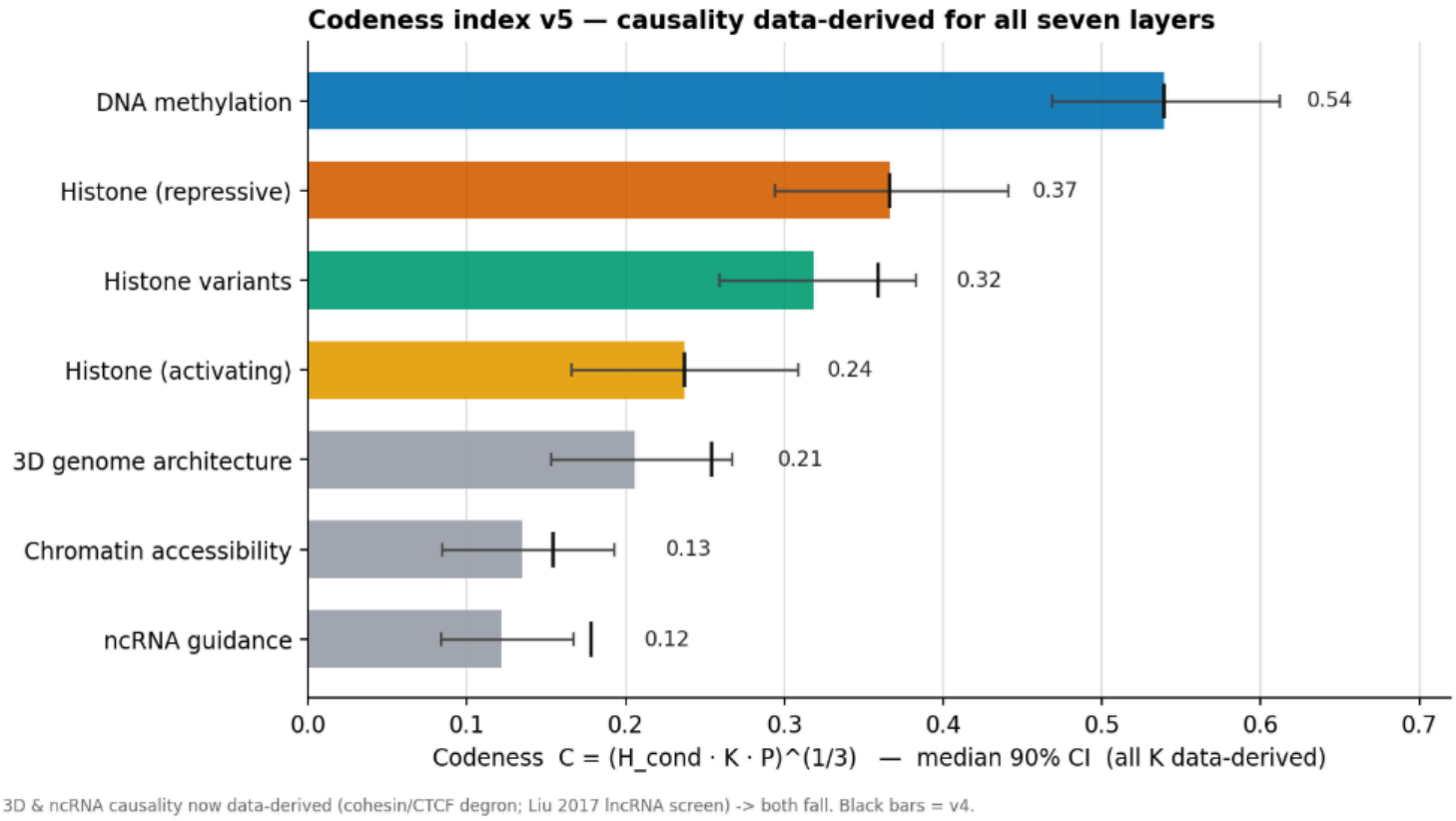


Figure 1

Mammalian codeness across seven molecular layers, showing the self-templating-code versus sequence-echo tiers.

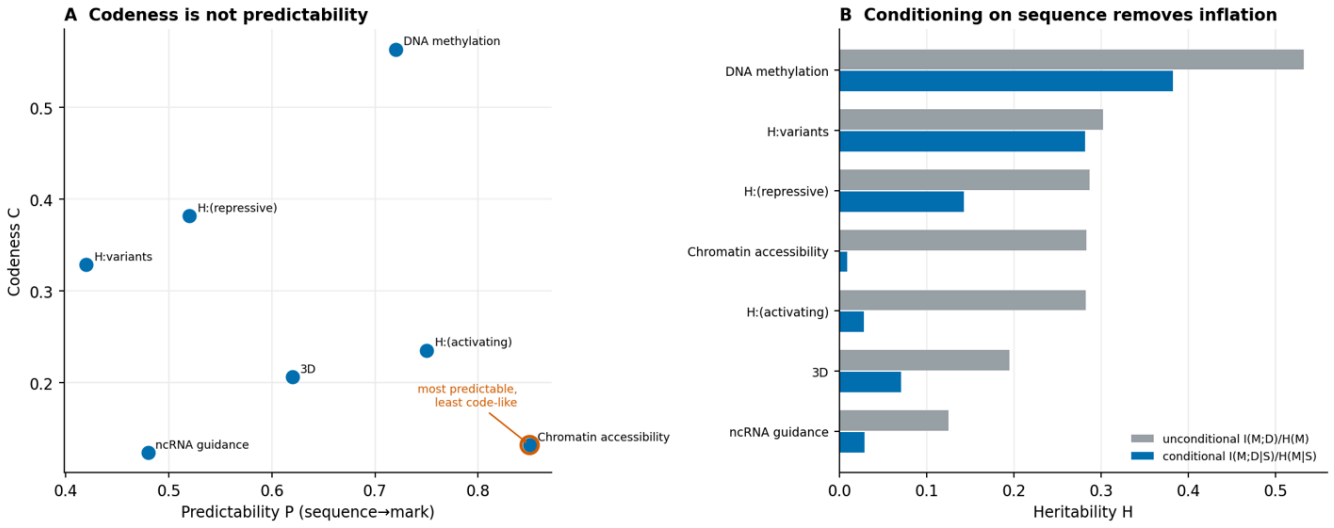
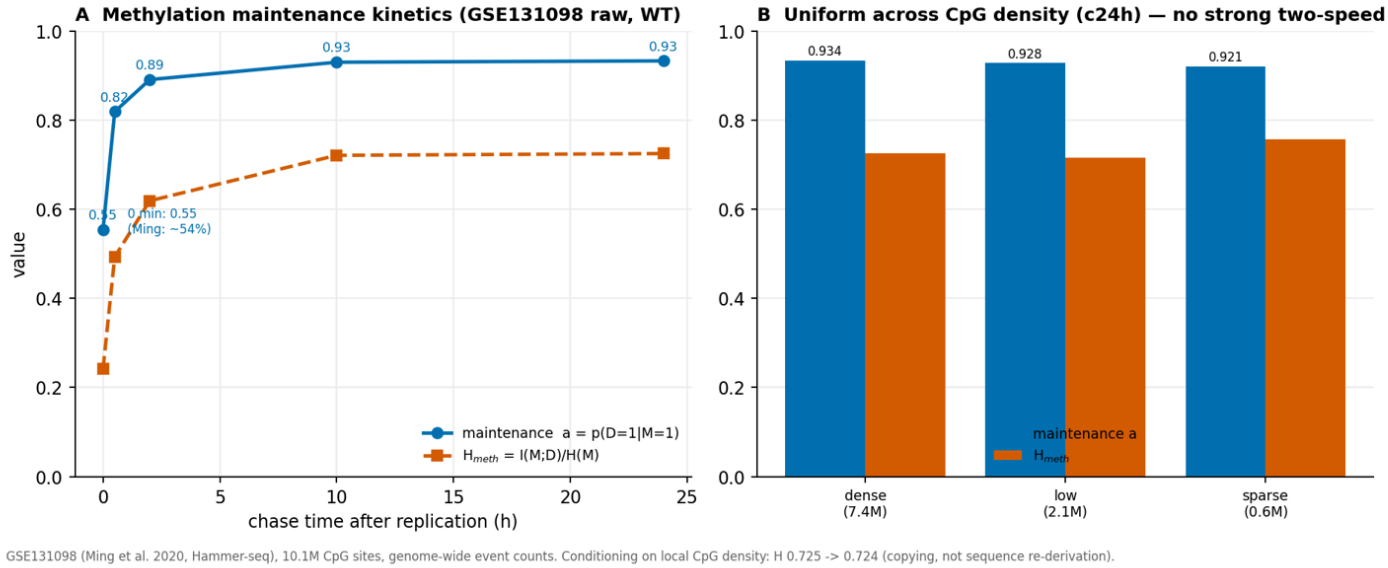


Figure 2

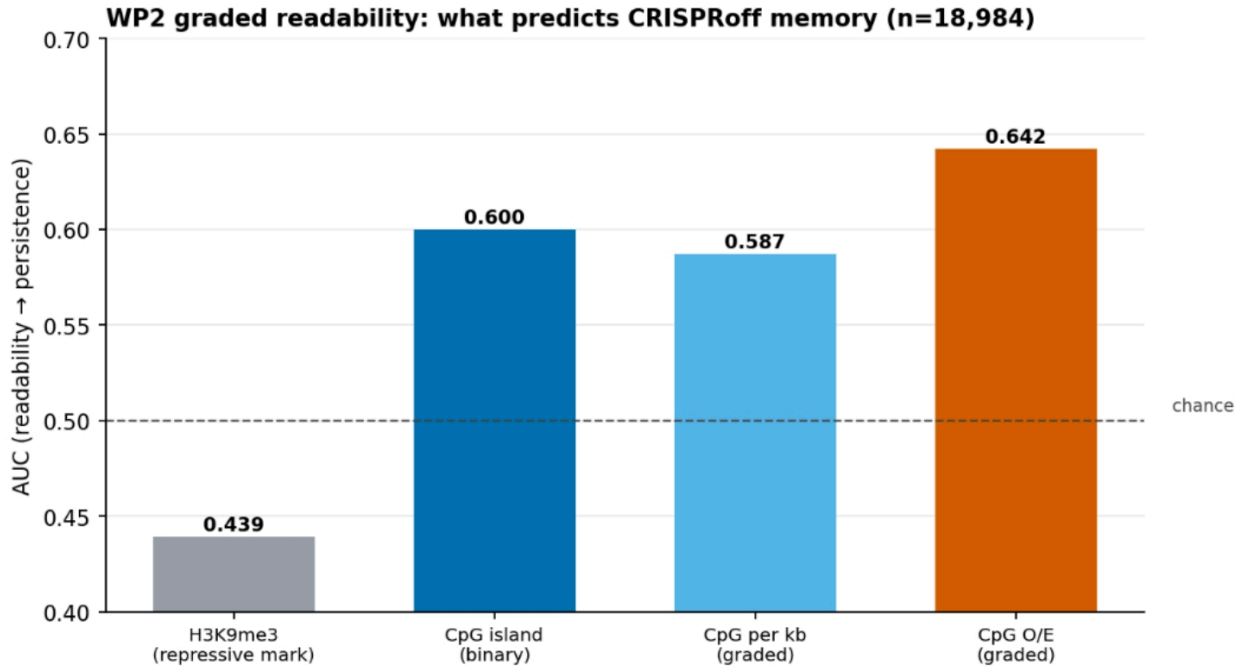
Codeness is not predictability. (A) Codeness versus sequence-predictability rankings (Spearman ρ = −0.07). (B) Conditional versus unconditional mutual information; conditioning on sequence creates the tiers.



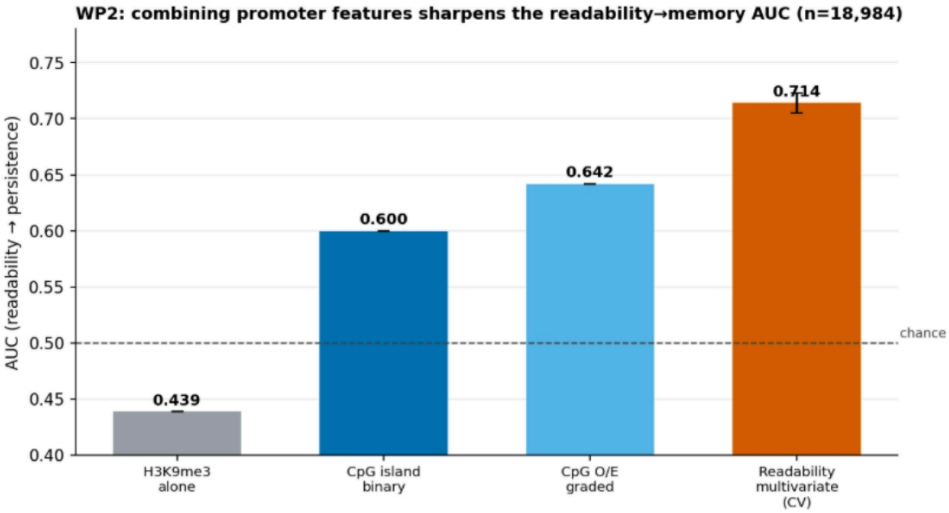
GSE131098 (Ming et al. 2020, Hammer-seq), 10.1M CpG sites, genome-wide event counts. Conditioning on local CpG density: H 0.725 -> 0.724 (copying, not sequence re-derivation).

Figure 3

DNA-methylation maintenance from raw Hammer-seq data (GSE131098). (A) Maintenance fidelity and normalized mutual information versus chase time, reproducing the published kinetics. (B) Uniformity across CpG density.



Repressive histone mark (ENCODE HEK293 H3K9me3) does NOT predict persistence (AUC<0.5): CpG-island promoters are H3K9me3-poor. The reader that matters is the DNA-methylation substrate — graded CpG O/E sharpens the AUC to 0.64 vs 0.60 for the binary flag.



Multivariate = 5-fold CV logistic on CpG O/E + CpG/kb + GC + H3K9me3 (readability only). Initial-silencing strength is EXCLUDED: it gives AUC 1.0 by leakage — the Nuñez hit-and-run 'memory' call is thresholded on that score.

Figure 4

Readability predicts CRISPRoff transcriptional memory. (A) AUC by readability proxy: H3K9me3 fails; CpG observed/expected sharpens over the binary flag. (B) A cross-validated multivariate readability model reaches AUC 0.71; initial-silencing strength is excluded as label leakage.

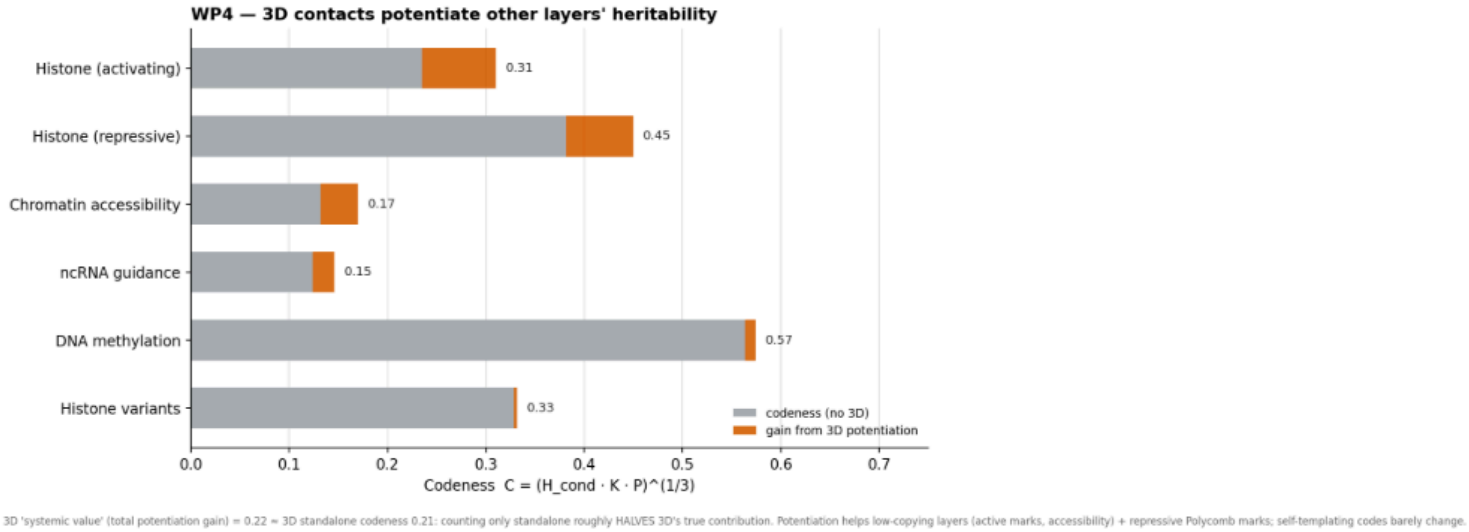


Figure 5

The 3D genome potentiates other layers' heritability; its "systemic value" roughly doubles its standalone codeness.

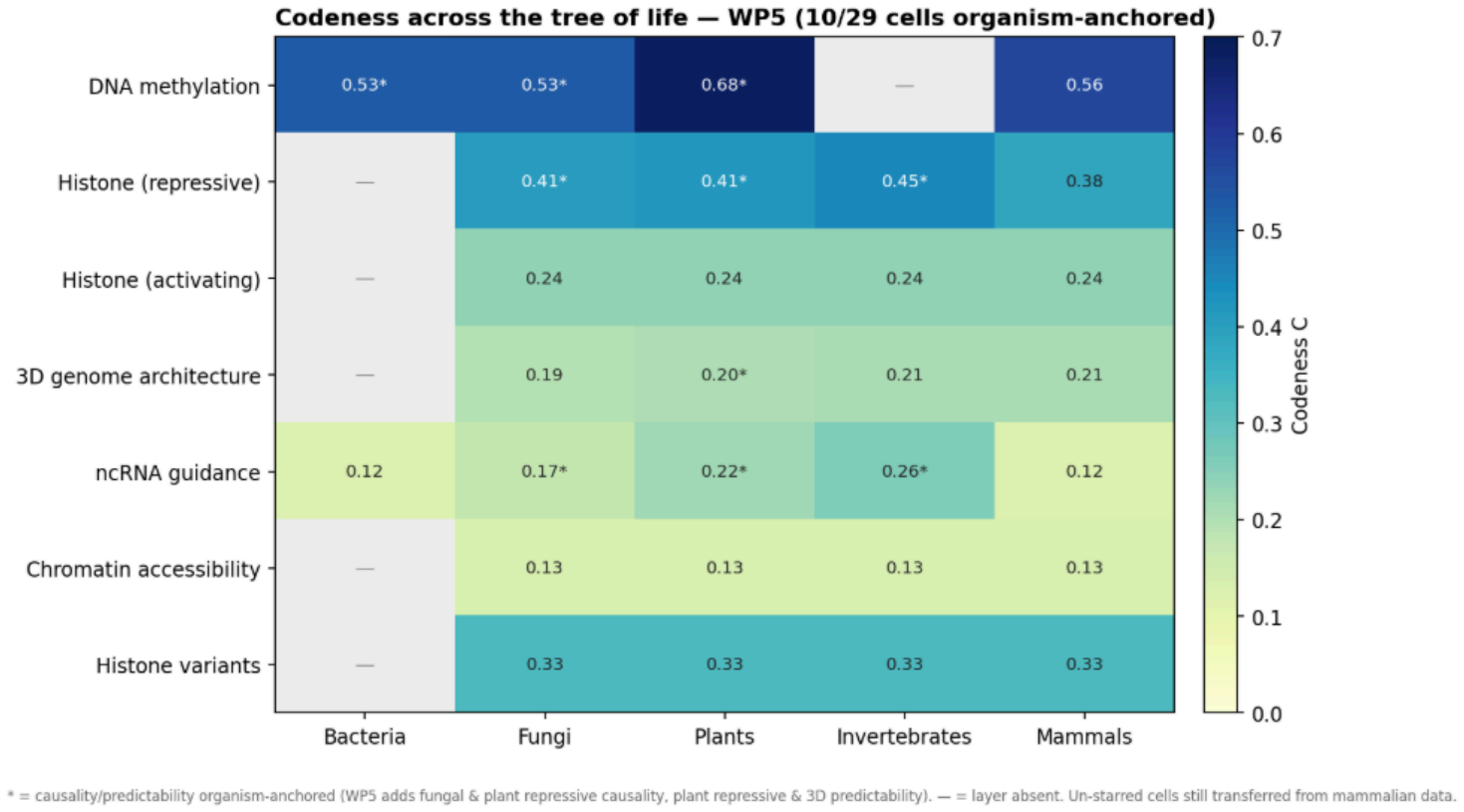


Figure 6

Codeness across the tree of life; the most code-like layer per kingdom, with a lineage-specific carrier molecule.

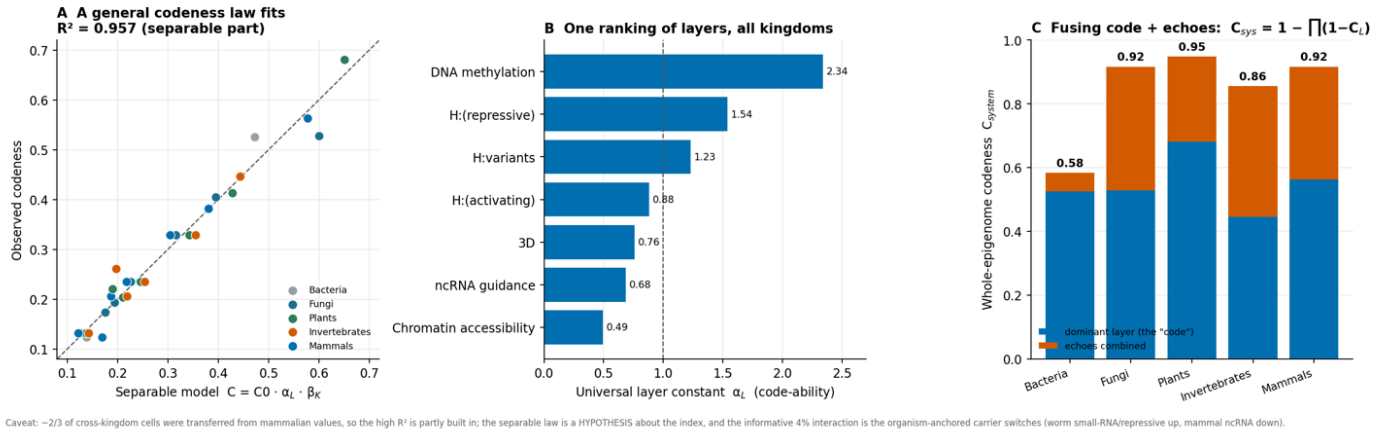


Figure 7

A candidate separable "universal epigenetic codeness law" (hypothesis). (A) Observed codeness versus the separable model C0·α_L·β_K (R^2 = 0.957). (B) The universal layer constant α_L (one ranking for all kingdoms). (C) Whole-epigenome codeness C_system, fusing the dominant self-copying code (blue) with the echoes (orange). The high separable R^2 is partly a construction artifact (transferred cross-kingdom cells); the law is advanced as a falsifiable hypothesis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1indexversionhistory.xlsx
- AdditionalFile2WP2pertargettable.csv
- AdditionalFile3crosskingdommatrix.xlsx
- AdditionalFile4generalcodenessmodelnote.docx