Potential energy landscapes identify the information-theoretic nature of the epigenome

Garrett Jenkinson[1,2], Elisabet Pujadas[1,3], John Goutsias[2] & Andrew P Feinberg[1,3,4]

Epigenetics is the study of biochemical modifications carrying information independent of DNA sequence, which are heritable through cell division. In 1940, Waddington coined the term "epigenetic landscape" as a metaphor for pluripotency and differentiation, but methylation landscapes have not yet been rigorously computed. Using principles from statistical physics and information theory, we derive epigenetic energy landscapes from whole-genome bisulfite sequencing (WGBS) data that enable us to quantify methylation stochasticity genome-wide using Shannon's entropy, associating it with chromatin structure. Moreover, we consider the Jensen–Shannon distance between sample-specific energy landscapes as a measure of epigenetic dissimilarity and demonstrate its effectiveness for discerning epigenetic differences. By viewing methylation maintenance as a communications system, we introduce methylation channels and show that higher-order chromatin organization can be predicted from their informational properties. Our results provide a fundamental understanding of the information-theoretic nature of the epigenome that leads to a powerful approach for studying its role in disease and aging.

In his seminal work, Conrad Waddington employed deterministic differential equations to define epigenetics as the emergence of a phenotype that can be perturbed by the environment but whose endpoints are predetermined by genes[1]. However, we have proposed a role for epigenetic stochasticity in development and disease[2,3], which has led to relatively simple probabilistic models of epigenetic landscapes that account for randomness in DNA methylation by adding a 'noise' term to deterministic models[4,5]. Some authors have also characterized methylation stochasticity using the notion of epipolymorphism[6,7], a form of nonadditive Tsallis entropy whose measurement is limited to a small portion of the genome and can underestimate heterogeneity in WGBS data[6,7] (**Supplementary Note**).

Here we take a foundational approach to understanding epigenetic information using principles from statistical physics and information theory that organically incorporate stochasticity into the mathematical framework and apply this approach on diverse WGBS data sets. In contrast to metaphorical 'Waddingtonian' landscapes, we present a rigorous derivation of epigenetic potential energy landscapes that encapsulate the higher-order statistical properties of methylation, fully capturing behavior that is opaque to customary mean-based summaries.

We quantify methylation stochasticity using Shannon's entropy and provide a powerful information-theoretic methodology for distinguishing epigenomes using the Jensen–Shannon distance between sample-specific energy landscapes associated with stem cells, tissue lineages, and cancer. Moreover, we establish a relationship between entropy and topologically associating domains (TADs) that allows prediction of their boundaries from WGBS samples. We also introduce methylation channels as models of DNA methylation maintenance and show that their informational properties can effectively predict higher-order chromatin organization using machine learning. Finally, we introduce a sensitivity index that quantifies the rate by which environmental perturbations influence methylation stochasticity along the genome.

This merger of epigenetic biology and statistical physics yields many fundamental insights into the relationship between information-theoretic properties of the epigenome and nuclear organization in normal development and disease. Moreover, it provides novel methods for evaluating the informational properties of individual samples and their chromatin structure and for quantifying differences between tissue lineages, aging, and cancer at high resolution and genome-wide.

RESULTS

Stochastic epigenetic variation and energy landscapes

Current methods for methylation analysis are limited predominantly to modeling stochastic variation at individual CpG sites while ignoring statistical dependence among neighboring sites[8]. However, fully characterizing the stochastic and polymorphic nature of epigenetic information requires knowledge of the probability distribution of methylation patterns (epialleles) formed by groups of CpG sites[6,7]. At present, this distribution is estimated empirically, requiring much higher coverage than WGBS data routinely provide (**Fig. 1** and **Supplementary Note**).

To remedy this problem and better understand the relationship between epigenetic stochasticity and phenotypic variability, we

[1]Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [2]Whitaker Biomedical Engineering Institute, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [4]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. Correspondence should be addressed to J.G. (goutsias@jhu.edu) or A.P.F. (afeinberg@jhu.edu).

Received 7 October 2016; accepted 13 February 2017; published online 27 March 2017; doi:10.1038/ng.3811

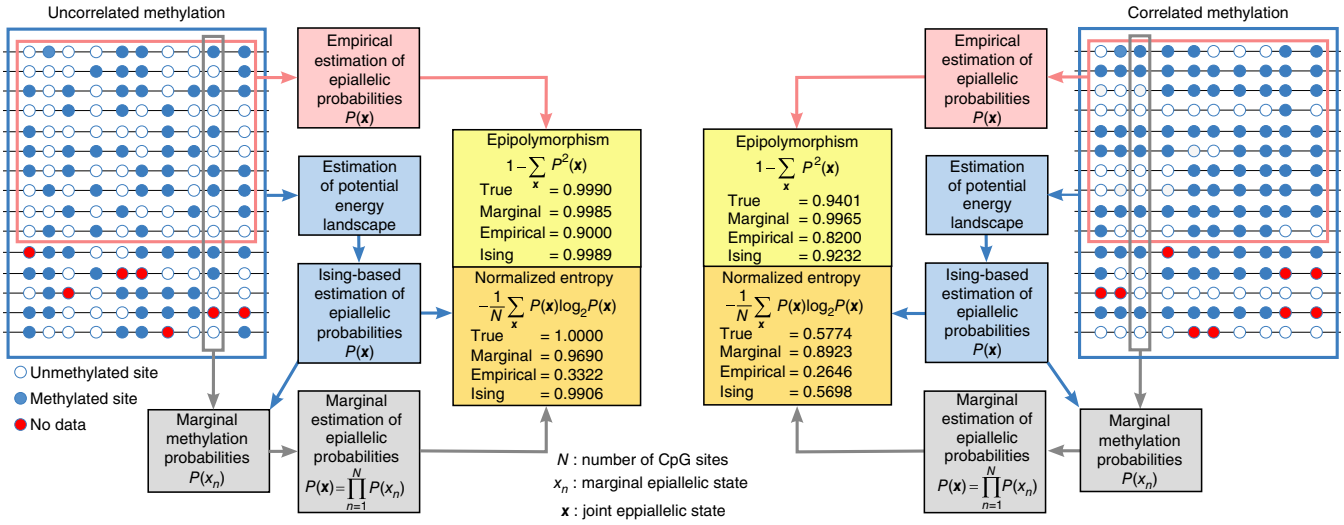

Figure 1 Estimation of epiallelic probabilities, epipolymorphisms, and normalized epiallelic entropies. Multiple WGBS reads within a genomic region are used to form a methylation matrix whose entries represent the methylation status of each CpG site (blue, methylated; white, unmethylated; red, no data). Most methods estimate marginal probabilities at individual CpG sites using only data within each column of the methylation matrix, which can then be employed to estimate epiallelic probabilities by assuming statistical independence. At low levels of correlation, this method may accurately estimate epipolymorphisms and entropies, but it will overestimate these quantities when high correlation is present. Empirical estimation of epiallelic probabilities uses only fully observed rows of the methylation matrix and can underestimate the epipolymorphisms and entropies regardless of correlation level. Estimation of epiallelic probabilities using an Ising potential energy landscape employs all data available in the methylation matrix and consistently provides accurate estimates.

employed an approach based on statistical physics and information theory. We represented methylation within a genomic region containing N CpG sites by a random vector $\mathbf{X} = [X_1, X_2, \ldots, X_N]$, where X_n takes value 1 or 0 depending on whether the nth CpG site is methylated or unmethylated, respectively. We then modeled $\mathbf{X}$ using the Boltzmann–Gibbs distribution

$$P(\mathbf{x}) = \frac{1}{Z}\exp\{-U(\mathbf{x})\}$$

where $U(\mathbf{x})$ is the energy of the methylation pattern $\mathbf{x}$ and

$$Z = \sum_{\mathbf{x}} \exp\{-U(\mathbf{x})\}$$

is the partition function.

The function $V(\mathbf{x}) = U(\mathbf{x}) - U(\mathbf{x}^*)$, where $\mathbf{x}^*$ is a methylation pattern with the least energy (ground state), defines a potential energy landscape whose elevation (potential) provides a measure of the improbability of finding the methylation pattern $\mathbf{x}$ relative to the most likely pattern $\mathbf{x}^*$. This landscape possesses one or more 'potential wells' corresponding to local minima of $U(\mathbf{x})$ with each well associated with an attractor representing the most probable methylation pattern to be found within a genomic region among all patterns associated with the well (**Fig. 2a**). Demethylation and *de novo* methylation allow methylation patterns to be modified, with higher probability for changes that move patterns toward lower potential energy. At steady state, a genomic region can be associated with a 'cloud' of methylation patterns that fluctuate around an attractor, resulting in pattern variability controlled by the width of the potential well. Notably, a potential energy landscape could be associated with two distinct attractors producing 'bistable' behavior (**Fig. 2a**). DNA methylation is subject to this type of behavior, which was found to be associated with gene imprinting (**Supplementary Note**).

Using the maximum-entropy principle[9], we determined an energy function that is consistent with methylation means and nearest-neighbor correlations, given by

$$U(\mathbf{x}) = -\sum_{n=1}^{N}(\alpha+\beta\rho_n)(2x_n-1) - \sum_{n=2}^{N}\frac{\gamma}{d_n}(2x_n-1)(2x_{n-1}-1)$$

where α, β, and γ are parameters characteristic to the genomic region, ρ_n is the CpG density, and d_n is the distance between CpG sites n and $n-1$, leading to the 1D Ising model of statistical physics that takes into account non-cooperative and cooperative factors in methylation (**Supplementary Note**). This choice encapsulates the notion that methylation depends on two distinct factors: the CpG architecture of the genome, quantified by CpG densities and distances, and the local biochemical environment provided by the methylation machinery, quantified by parameters α, β, and γ. Moreover, it allows computation of potential energy landscapes, joint probabilities of methylation patterns, marginal probabilities at individual CpG sites, and a number of novel measures for methylation analysis (**Supplementary Note**). Simulated data provided evidence that, in contrast to empirically estimating epiallelic probabilities, methylation pattern analysis using the Ising model can consistently produce accurate results using relatively low-coverage data (**Fig. 1** and **Supplementary Note**).

Using a maximum-likelihood approach (Online Methods), we estimated methylation potential energy landscapes from WGBS data corresponding to 35 diverse samples that enabled detailed local profiling throughout the methylome (**Supplementary Table 1**). This analysis showed, for example, that most methylation patterns associated with the CpG island (CGI) of *WNT1* in normal colon exhibit high potential values (**Fig. 2b**), implying little variability from the pattern with the lowest potential (attractor), which in this case coincides with the fully unmethylated pattern. The accompanying violin plot (**Fig. 2b**) shows that any deviation from the attractor toward a pattern of higher potential will be rapidly 'funneled' back toward the attractor, leading

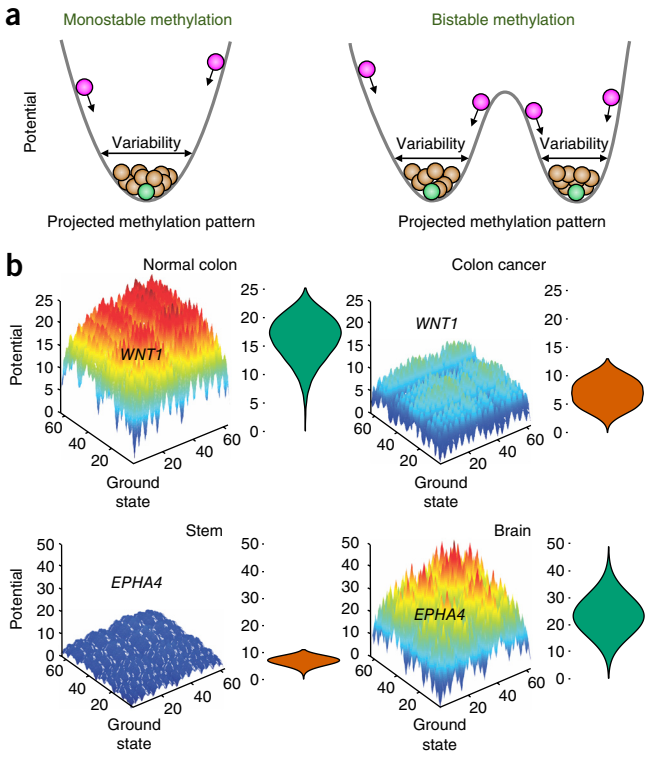

Figure 2 Potential energy landscapes. (**a**) Hypothetical monostable and bistable potential energy landscapes illustrating the presence of potential wells that correspond to attractors (green balls) and associated clouds of methylation patterns (brown balls). The magenta balls indicate unlikely methylation patterns drawn toward lower potential energies during maintenance. (**b**) Potential energy landscapes associated with 12 CpG sites within the CGIs of *WNT1* in normal colon and colon cancer and within the CGI of *EPHA4* in embryonic stem cells and brain. Each point in the domain of the potential energy landscape marks a methylation pattern, with the point at (0,0) indicating the fully unmethylated state, which is the ground state in both examples. The 2^{12} potential values are distributed over a 64 × 64 grid using a 2D version of the Gray code (Online Methods). Violin plots summarize distributions of potential values.

to low methylation stochasticity. However, most methylation patterns in colon cancer manifest much lower potential values than in normal colon (**Fig. 2b**), implying a substantial gain in pattern variability and increased methylation stochasticity in cancer. Similarly, most methylation patterns associated with the CGI of *EPHA4*, a key developmental gene, had low potential values in embryonic stem cells (**Fig. 2b**), implying substantial pattern variability from the attractor, which also coincides with the fully unmethylated state. In contrast, *EPHA4* showed higher potential values in the brain (prefrontal cortex) (**Fig. 2b**), yielding lower pattern variability and lower methylation stochasticity than in embryonic stem cells.

Epigenetic entropy quantifies methylation stochasticity

To facilitate genome-wide analysis of methylation information, we partitioned the genome into nonoverlapping units and performed methylation analysis at a resolution of 1 genomic unit. Consistent with the length of DNA within a nucleosome (~146 bp), we chose genomic units of 150 bp each to strike a balance between leveraging as much information as possible within a genomic unit and performing high-resolution methylation analysis. We then quantified methylation within each genomic unit using the methylation level (average methylation), whose probability distribution is calculated from the Ising model, and used its mean and normalized Shannon entropy to measure methylation stochasticity (Online Methods).

In agreement with the literature[10], the mean methylation level was globally higher in embryonic stem cells and brain tissues than in normal colon, lung, and liver. This was also true for $CD4^+$ lymphocytes and skin keratinocytes, but methylation levels were reduced in cancer and were progressively lost in cell culture (**Fig. 3a**,**b**). We also observed low normalized methylation entropy in embryonic stem cells, brain cells, $CD4^+$ lymphocytes, and skin keratinocytes associated with young subjects as well as a global increase of entropy in most cancers—but not in two liver cancer samples with profound hypomethylation accompanied by a less entropic methylation state (**Fig. 3a**). Differential entropy changes in cancer were often, but not always, associated with changes in mean level (**Supplementary Fig. 1a–c**), demonstrating that changes in stochasticity are not necessarily related to changes in mean methylation, and indicating that both must be assessed when interrogating biological samples (**Supplementary Fig. 2**). Furthermore, genome-wide mean level and entropy distributions over selected genomic features demonstrate lower and more variable values within CGIs and transcription start sites (TSSs) than in other genomic features, such as shores, exons, or introns (**Supplementary Fig. 3a**,**b**).

Consistent with a previous analysis comparing newborns to centenarians[11], our results showed global hypomethylation in all three $CD4^+$ samples from older people (ages 82–86 years), as compared to three samples from younger individuals (ages 18–25 years), and the same was true in skin keratinocytes (**Fig. 3a**,**c**). This was accompanied by gain in methylation entropy, with changes in entropy more pronounced than in mean (**Fig. 3a**,**c**). Although passage number in fibroblasts was also associated with global hypomethylation, it was more appreciable than in $CD4^+$ samples, whereas entropy loss was globally observed at later passages (**Fig. 3a**), in stark contrast to global gain in entropy observed in $CD4^+$ samples from older people. We also assessed changes in methylation stochasticity while accounting for biological, statistical, and technical variability, which confirmed that aged $CD4^+$ cells often show changes in entropy (**Supplementary Note**). Using the Jensen–Shannon informational distance (Online Methods), we investigated differences between the young and old in the $CD4^+$ samples, as well as dissimilarities with passage in cultured fibroblasts. Our results (**Supplementary Note**) suggest that fibroblast cultures with high passage numbers inaccurately model methylation stochasticity in aging, which is associated with global gain in informational dissimilarity driven by increased entropy.

Informational distances delineate lineages and identify key developmental genes

Previous studies indicate that epigenetic discordance within gene regulatory elements, such as enhancers, might fully account for observed epigenetic dissimilarities between two samples[12,13]. However, when we computed genome-wide distributions of Jensen–Shannon distances within several genomic features, we did not find consistent containment of epigenetic dissimilarity within a particular feature (**Supplementary Fig. 4**). Therefore, to understand the relationship between epigenetic information and phenotypic variation, we used the Jensen–Shannon informational distance to quantify epigenetic discordance between pairs of samples (Online Methods). We then asked whether we could distinguish colon, lung, and liver from one another and from matched cancers, as well as from embryonic stem cells, brain, and $CD4^+$ lymphocytes. For computational feasibility, we limited the analysis to 17 representative samples, and visualized the results using multidimensional scaling (Online Methods).

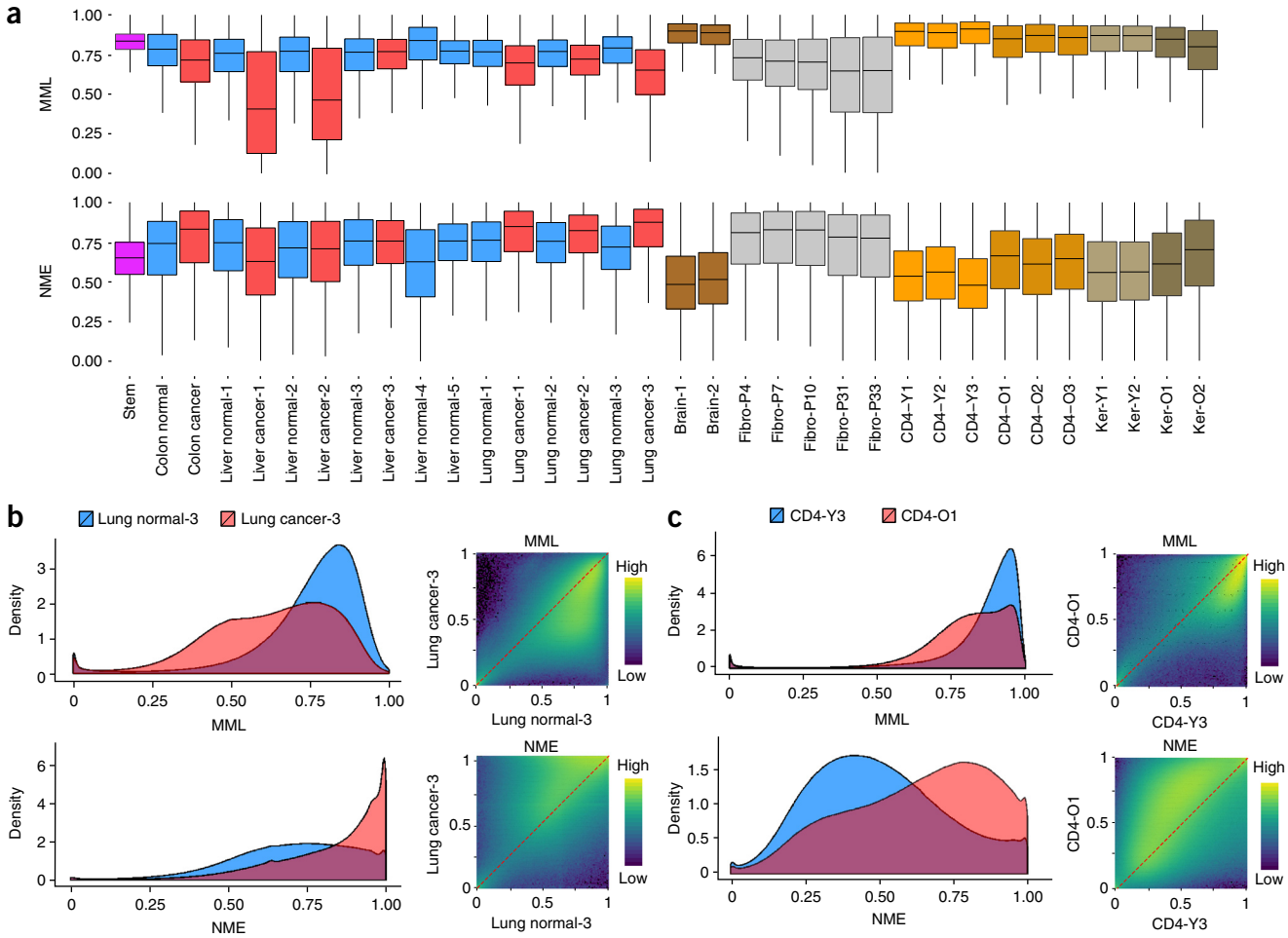

Figure 3 Mean methylation level and normalized entropy. (**a**) Genome-wide distributions of mean methylation level (MML) and normalized methylation entropy (NME) values in all samples used in this study. Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5 × IQR. (**b**,**c**) Genome-wide 1D and 2D $\log_{10}$-transformed MML and NME densities associated with normal lung and lung cancer (**b**) and CD4$^+$ lymphocytes from younger (CD4-Y3) and older (CD4-O1) subjects (**c**).

The samples fell into clear categories based on developmental germ layers, equidistant from embryonic stem cells, with cancers being well separated from normal samples (**Fig. 4**).

Given the relationship between the stem cell sample and the three germ layers, we examined genes that had differences in mean methylation level in embryonic stem cells as compared to differentiated tissues and genes that possessed epigenetic discordance quantified by the Jensen–Shannon distance. We ranked genes on the basis of absolute difference in mean methylation level within their promoters as well as using the Jensen–Shannon distance (**Supplementary Table 2** and Online Methods). Some genes involved in development and differentiation (for example, *FOXD3*, *SALL1*, *SOX2*, and *ZIC1* when comparing stem cells to lung) had relatively small changes in mean methylation but large Jensen–Shannon distances, affirming that the probability distributions of methylation levels within their promoters were different, despite little differences in mean. We further explored whether non-mean-related methylation differences could identify genes between sample groups that are occult to existing mean-based analyses by employing a relative ranking scheme that assigned higher scores to genes with larger Jensen–Shannon distances but smaller absolute changes in mean methylation level (Online Methods). In the stem-cell-to-brain comparison, for example, many key genes (*IGF2BP1*, *FOXD3*, *NKX6-2*, *SALL1*, *EPHA4*, *ASCL2*, and *OTX1*) topped the relative ranking list (**Supplementary Table 2a**), with gene ontology (GO) enrichment analysis[14] identifying key process categories associated with stem cell maintenance and brain development (**Supplementary Table 3a**). Moreover, 30 significant (FDR q value < 0.01) GO process categories showed tenfold or greater enrichment using the relative scheme, compared to five using the mean-based scheme. We obtained similar results when comparing stem cells to lung, with the relative scheme identifying key developmental processes and genes in both mesodermal and embryonic stem cell categories (**Supplementary Tables 2b** and **3b**). Comparison of embryonic stem cells to CD4$^+$ lymphocytes uncovered enrichment for immune-related functions, driven by differential mean level, and many developmental and morphogenesis process categories, predominantly driven by the Jensen–Shannon distance (**Supplementary Tables 2c** and **3c**).

When comparing differentiated tissues, mean-based GO analysis resulted in highly enriched process categories, related mostly to differentiated function, such as cellular regulation and signaling. However, the relative scheme resulted in highly enriched categories related largely to development and differentiation (**Supplementary Tables 2d–f** and **3d–f**), probably because relative rankings are low for genes that are methylated in only one cell type. When we compared normal lung to cancer, the relative scheme produced a larger number of highly enriched process categories than mean-based analysis, and

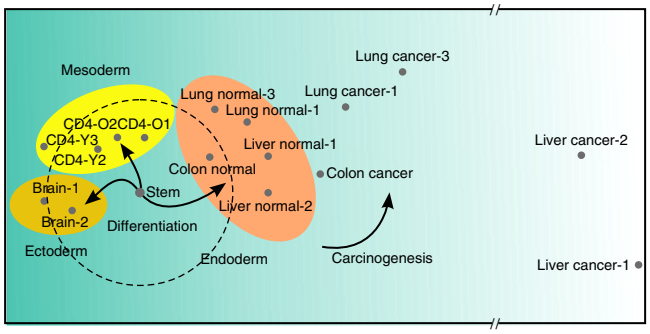

Figure 4 Informational distances and lineages. Visualization of genomic dissimilarity in 17 diverse cell and tissue samples using multidimensional scaling, evaluated by Jensen–Shannon distance. Tissues derived from endoderm (colon, lung, liver), mesoderm ($CD4^+$ lymphocytes), and ectoderm (brain) are located roughly equidistant from embryonic stem cells (dashed circle). Cancerous tissues are well separated from normal tissues and stem cells, and two liver cancers are far removed from their matched normal counterparts.

these were again related to developmental morphogenesis categories (**Supplementary Tables 2g** and **3g**). There were 40 significant (FDR q value < 0.01) GO process categories with tenfold or greater enrichment when using the relative scheme, compared to 7 GO categories when using the mean-based scheme. These results suggest that major changes may occur in the probability distributions of methylation levels associated with developmentally critical genes and that the shape of these distributions, rather than their means per se, may be related to pluripotency and fate lineage determination in development and cancer.

Finally, by assessing a relationship between transcription factor binding and Jensen–Shannon distances in development, we found a strong association between Jensen–Shannon distances and Polycomb repressive complex 2 (PRC2) binding within enhancer regions (**Supplementary Note**). This raises the possibility that PRC2 not only influences the mean behavior of DNA methylation, as has been established previously[15], but also may control the stochastic behavior of methylation in a developmentally relevant and targeted manner.

Entropy blocks predict TAD boundaries

TADs are highly conserved structural features of the genome across tissues and species[16–18]. Loci within these domains tend to interact frequently, with much less frequent interactions taking place between loci in adjacent domains. Although genome-wide detection of TAD boundaries is experimentally challenging, these boundaries can be reasonably predicted from ChIP–seq data (using CTCF transcription factor and monomethylation of histone H3 Lys4) using a computational approach[19]. We therefore examined the possibility of locating TAD boundaries genome-wide using WGBS data.

In many of our samples, known TAD boundary annotations were visually proximal to boundaries of entropy blocks, large genomic regions of consistently low or high normalized entropy values (**Fig. 5a**, **Supplementary Fig. 5**, and Online Methods). We thus hypothesized that TAD boundaries may be located within genomic regions that separate successive entropy blocks. As a first test, we computed entropy blocks in the embryonic stem cell data and identified 404 regions predictive of TAD boundaries (Online Methods). We found that these predictive regions significantly overlapped or were close to the 5,862 annotated TAD boundaries in H1 stem cells[16] (**Supplementary Note**). Using 90% of the computed predictive regions (Online Methods and **Supplementary Note**), we correctly identified 6% of the annotated TAD boundaries (362 out of 5,862).

Because TADs are thought to be cell-type invariant[16,18], we can predict the location of more TAD boundaries by combining information from entropy blocks derived from additional phenotypes (**Fig. 5b**). We therefore computed entropy blocks using WGBS data from 17 cell types, determined predictive regions for each cell type, and combined these regions into a list (6,687 predictive regions) that encompasses information from all cell types (Online Methods). Moreover, we combined the TAD boundary annotations for H1 stem cells with available annotations for IMR90 lung fibroblasts[16] to obtain a total of 10,276 'ground-truth' annotations. We then obtained results similar to the case of stem cells with TAD boundaries falling within or near identified predictive regions significantly more often than expected by chance (**Supplementary Note**). This resulted in 62% correct identification of the annotated TAD boundaries (6,369 out of 10,276) derived from 95% of computed predictive regions (Online Methods and **Supplementary Note**), which can be further improved by including additional phenotypes.

These predictions are further supported by the relatively small errors, as compared to TAD size, obtained when locating TAD boundaries at the centers of predictive regions. This is demonstrated by estimates of the probability density and the corresponding cumulative probability distribution showing that location errors are stochastically smaller than TAD sizes (**Fig. 5c**) and the fact that the median location error was an order of magnitude smaller than the median TAD size (92 kb versus 760 kb). Together, these results provide statistical evidence of an underlying relationship between TADs and entropy blocks, wherein a TAD is associated with consistently low or high normalized entropy values, which can be used to computationally identify TAD boundaries accurately genome-wide from WGBS data.

Methylation channels explain epigenetic memory maintenance

To investigate epigenetic memory maintenance from an information-theoretic perspective, we modeled this process using a methylation channel, which quantifies transitions of the binary methylation state at each CpG site of the genome using the local probabilities of demethylation and *de novo* methylation (**Fig. 6a** and **Supplementary Note**). This results in a model for methylation maintenance known in information theory as the asymmetric-noise binary communication channel[20].

A methylation channel is limited to transmitting a maximum amount of information, quantified by its information capacity[20]. Moreover, appreciable consumption of free energy, which must be dissipated to the surroundings in the form of heat, is required to achieve high transmission reliability essential for normal cellular function. We assessed methylation reliability using the notion of relative dissipated energy and identified approximate relationships between channel capacity, relative dissipated energy, and methylation entropy (**Supplementary Note**). We predicted that highly reliable methylation maintenance is achieved through high-capacity methylation channels that produce less entropic methylation at the expense of higher energy consumption (**Fig. 6b**, blue), whereas low-capacity methylation channels dissipate less energy, achieving lower reliability and a disordered methylation state (**Fig. 6b**, red). This establishes the influence of methylation channels on epigenetic memory by providing a fundamental link between their information-theoretic properties and the nature of epigenetic memory maintenance.

We computed capacities, relative dissipated energies, and entropies genome-wide in individual samples and comparative studies (**Fig. 6c**

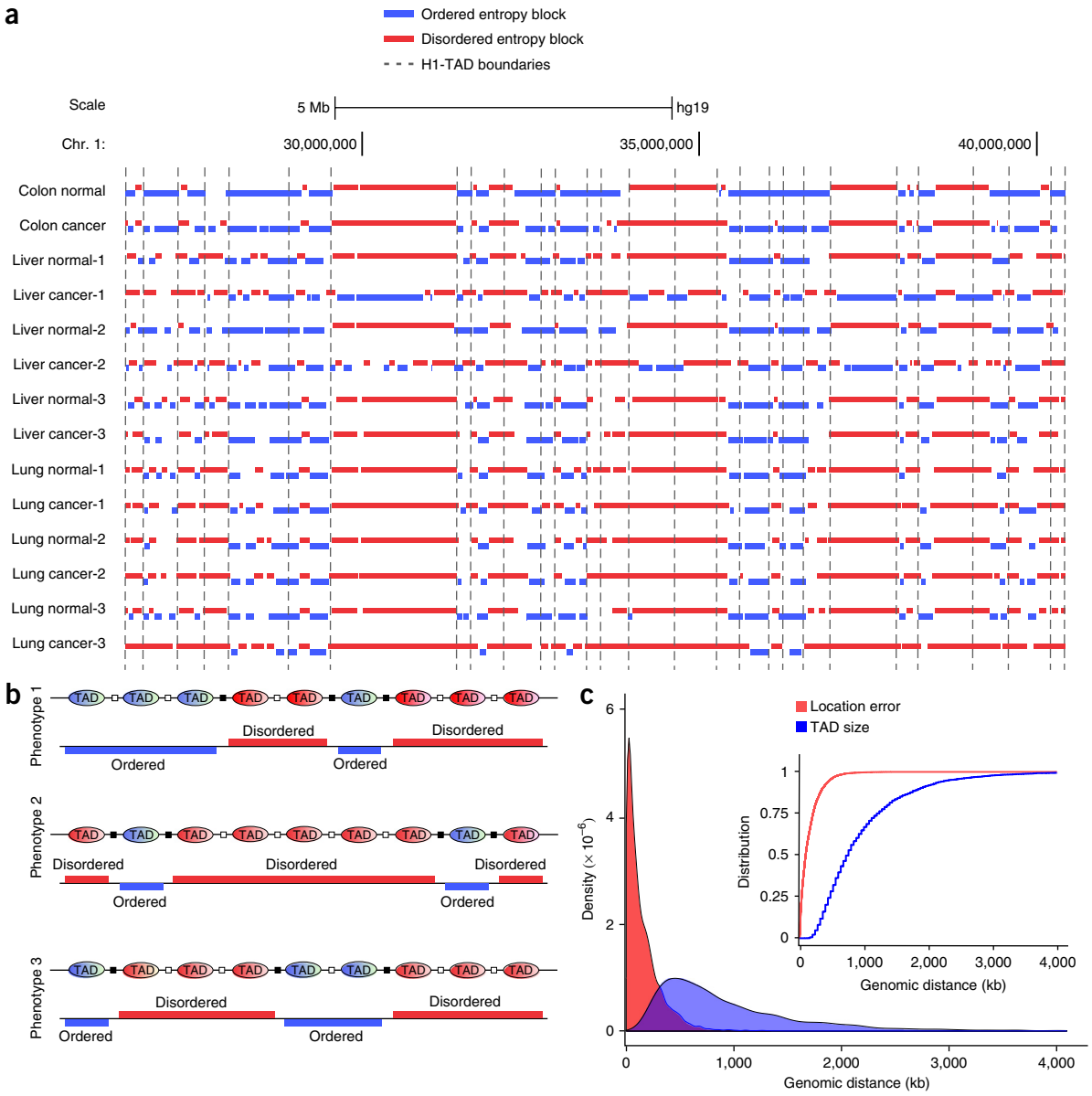

Figure 5 Entropy blocks and TAD boundaries. (**a**) In the normal–cancer panel, a subset of known TAD boundary annotations in H1 stem cells appear to be correlated with the boundaries of entropy blocks (blue, ordered; red, disordered), suggesting that TADs may maintain a consistent level of methylation entropy within themselves. (**b**) Regions of entropic transitions can be used to identify the location of some TAD boundaries (black squares). As TADs are cell-type invariant, the location of more TAD boundaries can be identified using additional WGBS data corresponding to distinct phenotypes. (**c**) Probability densities and cumulative probability distributions (inset) of the TAD boundary location error and TAD sizes.

and **Supplementary Fig. 3c**,**d**). We observed global loss of capacity and relative dissipated energy accompanied by global gain in CpG methylation entropy in colon and lung cancer (**Fig. 6c**,**d**), although this was not always true in liver cancer (**Fig. 6c**). Moreover, brain cells, $CD4^+$ lymphocytes, and skin keratinocytes showed high capacities and relative dissipated energies, which decreased with age, whereas stem cells had a narrow range of relatively high capacities and relative dissipated energies (**Fig. 6c**,**e**). Sorting genes by normalized methylation entropy within their promoters in stem cells (**Supplementary Table 4**), we discovered many genes characterized by high information capacity and relative dissipated energy, such as *FOXO3*, *TGIF1*, *SATB2*, *IGF2BP1*, *SMAD7*, *ZIC2*, *SALL1*, and *SOX2*, which are involved in stem cell regulation, pluripotency, and differentiation[21–28]. We also found that the methylation state within CGIs and TSSs is maintained by methylation channels whose capacities are higher overall than those within shores, shelves, open seas, exons, introns, and intergenic regions, and this is accomplished by higher energy consumption (**Supplementary Fig. 3c**,**d**). These results highlight an information-theoretic view of epigenetic organization that explains methylation stochasticity in a way that is consistent with the need of cells to manage limited energy resources in a strategic manner. According to this model, reliable transmission of methylation information within critical regions of the genome is facilitated by high-capacity methylation channels that result in low methylation stochasticity at the cost of high energy consumption. However, methylation transmission within other regions of the genome is transmitted by low-capacity methylation channels that consume less energy but produce higher methylation stochasticity.

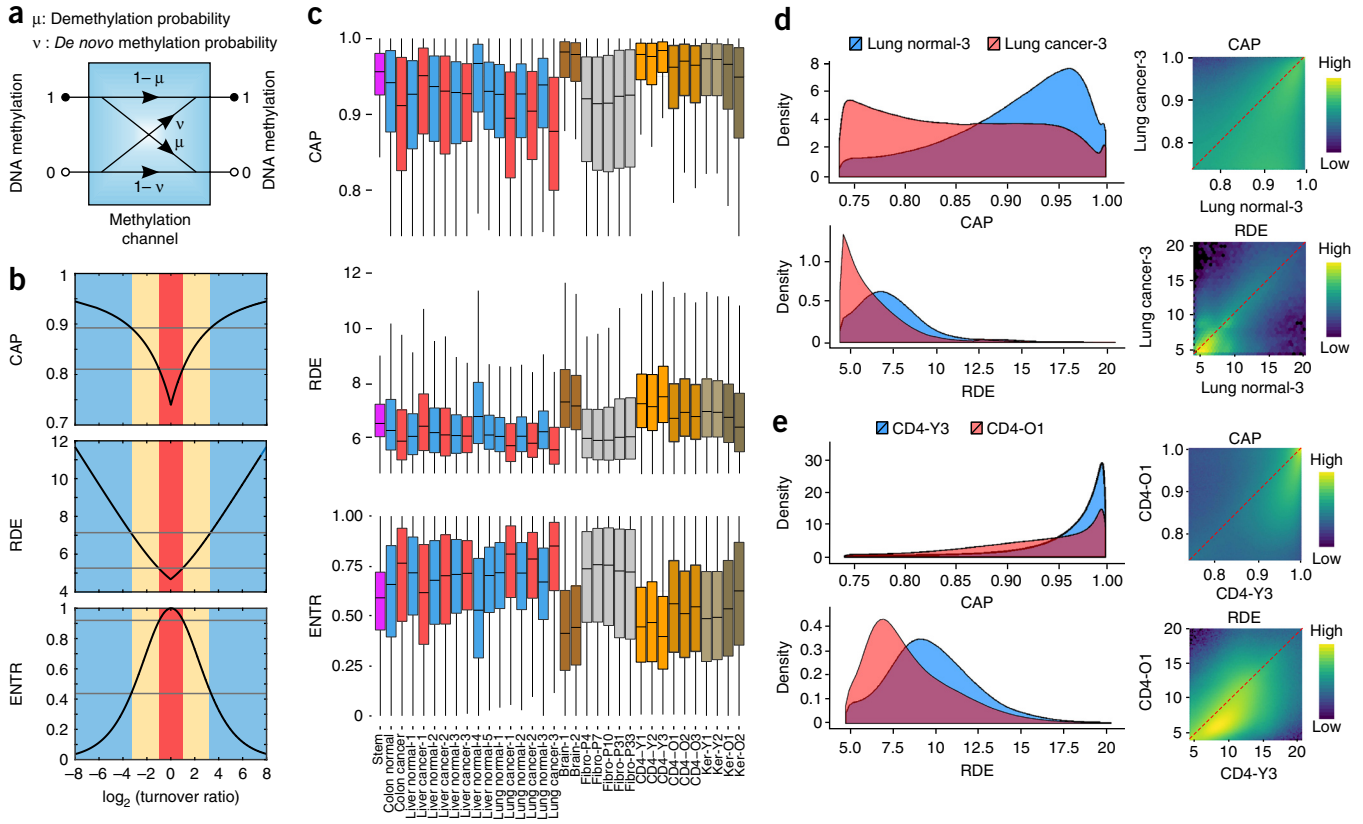

Figure 6 Information-theoretic properties of methylation channels. (**a**) A methylation channel maintains the methylation state at a CpG site (1, methylated; 0, unmethylated) using four conditional probabilities (μ, demethylation probability; ν, *de novo* methylation probability). (**b**) Theoretical curves of the capacity (CAP), relative dissipated energy (RDE), and input/output entropy (ENTR) of a methylation channel in terms of the $\log_2$ ratio of the probability of *de novo* methylation to the probability of demethylation (turnover ratio). These thresholds correspond to entropy levels used to identify ordered and disordered genomic units that build entropy blocks (Online Methods). (**c**) Genome-wide distributions of CAP, RDE, and ENTR at individual CpG sites. Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5 × IQR. (**d**,**e**) Genome-wide 1D and 2D $\log_{10}$-transformed CAP and RDE associated with normal lung and lung cancer (**d**) and younger (CD4-Y3) and older (CD4-O1) $CD4^+$ lymphocytes (**e**).

Prediction of chromatin changes in development and cancer

Recent work in chromatin organization has found the existence of cell-type-specific compartments A and B, known to be associated with gene-rich transcriptionally active open chromatin and gene-poor transcriptionally inactive closed chromatin, respectively[16,18,29]. Although identifying compartments A and B experimentally is challenging[29], it can be achieved computationally[30]. We therefore sought to identify compartments A and B in individual WGBS samples from local informational properties of the methylome.

When we compared Hi-C data from Epstein–Barr virus (EBV)-transformed cells to our methylation channel results, we observed enrichment of low information capacity, low relative dissipated energy, and high normalized methylation entropy in compartment B, and the opposite in compartment A (**Fig. 7a**,**b**). This suggested the possibility of predicting compartments A and B from informational properties of methylation maintenance. We therefore employed a random forest regression model to learn the informational structure of compartments A and B from available ground-truth data (Online Methods). We achieved reliable prediction, with cross-validated average correlation of 0.82 and average agreement of 91% between predicted and true A and B signals using a calling margin of 0.2 (**Supplementary Fig. 6a** and Online Methods), suggesting that a small number of local information-theoretic properties of methylation maintenance can be highly predictive of large-scale chromatin organization.

Consistent with the fact that compartments A and B are cell-type specific, and in agreement with the finding of extensive A and B compartment reorganization during early stages of development[31], we observed many differences in predicted compartments between tissues and in carcinogenesis (**Supplementary Fig. 6b–f**). Our methylation data also showed that our predicted compartment transitions corresponded often to TAD boundaries identified from Hi-C data[31] (**Supplementary Fig. 6b**). We also quantified observed differences in compartments A and B by computing percentages of switching in all sample pairs (**Supplementary Table 5** and Online Methods) and clustered the samples by using net percentage of A–B switching as a dissimilarity measure (**Fig. 7c** and Online Methods). The clusters had 31 out of 34 samples grouped in a biologically meaningful manner, providing evidence that A–B switching, as determined by methylation information, can accurately quantify phenotypic differences in the samples. Notably, stem cell differentiation is associated with high levels of chromatin reorganization (**Fig. 7c**), whereas differentiated lineages and cancer are clustered together but distinct from each other. Moreover, fibroblasts form one cluster, whereas young $CD4^+$ samples form their own, and the same is true for skin cells.

The chromatin organization of normal samples was markedly different from that of matched cancer samples (**Fig. 7c**). Previous studies found large hypomethylated blocks in cancer that are remarkably consistent across tumor types[32]. These blocks correspond closely to

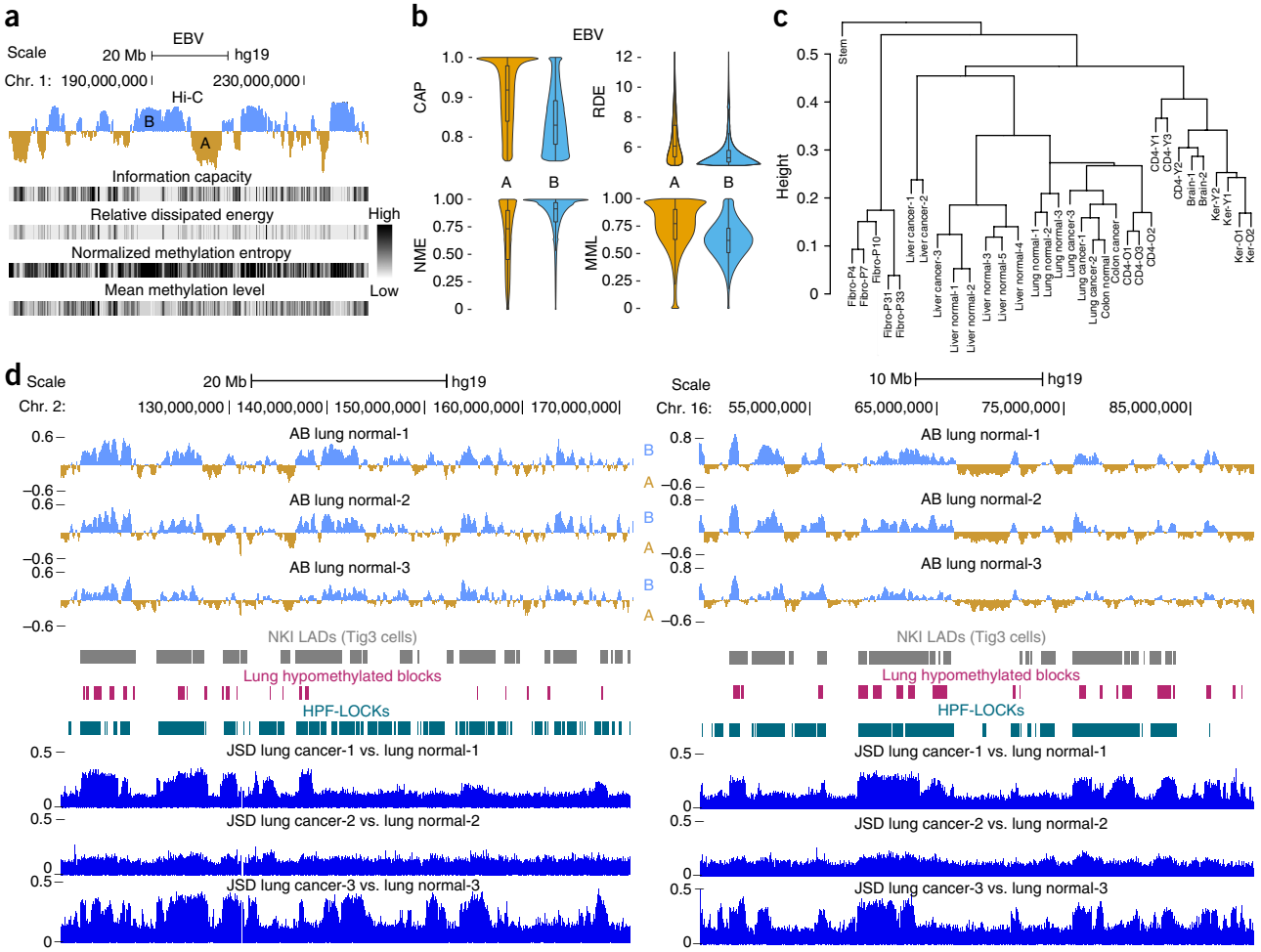

Figure 7 Information-theoretic prediction of large-scale chromatin organization. (**a**) Analysis of Hi-C and WGBS data shows that maintenance of the methylation state within compartment B (blue) in EBV-transformed cells is mainly performed by low-information-capacity methylation channels that dissipate low amounts of energy and result in a relatively disordered and less methylated state than in compartment A (brown). (**b**) Violin plots of genome-wide distributions of information capacity (CAP), relative dissipated energy (RDE), normalized methylation entropy (NME), and mean methylation level (MML). Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5 × IQR. (**c**) Hierarchical clustering of samples using the net percentage of A and B compartment switching as a dissimilarity measure. At a given height, a cluster is characterized by lower overall compartment switching than an alternative grouping of samples. (**d**) UCSC genome browser images of two chromosomal regions showing overlap of compartment B in normal lung (blue) with hypomethylated blocks, LADs, and LOCKs. Gain in Jensen–Shannon distance (JSD) is observed within compartment B (blue) in lung samples during carcinogenesis.

large-scale regions of chromatin organization, such as lamin-associated domains (LADs) and large organized chromatin histone H3 Lys9 (K9) modifications (LOCKs)[3,33]. Consistent with our observations on the information-theoretic properties of compartment B and of carcinogenesis (**Figs. 6c** and **7a**,**b**), we asked whether hypomethylated blocks are associated mainly with compartment B (Online Methods). We found significant overlap with compartment B in normal lung (**Fig. 7d**), and the same was true for LADs and LOCKs (**Supplementary Table 6**). Compartment B in normal tissue exhibited regions of large Jensen–Shannon distances (**Fig. 7d**), suggesting that considerable epigenetic changes may occur within this compartment during carcinogenesis, which is further supported by the genome-wide distributions of the Jensen–Shannon distance values between normal and cancer samples within compartments A and B in the normal samples (**Supplementary Fig. 7**). The observed association of compartment B in normal tissue with hypomethylated blocks and large Jensen–Shannon distances indicates that compartment B demarcates genomic regions with methylation information that is more likely to be degraded in cancer.

Entropic sensitivity to environmental variability

Epigenetic changes integrate environmental signals with genetic variation to modulate phenotype. We therefore sought to investigate the influence of environmental exposure on methylation stochasticity. We viewed environmental variability as a process that directly influences the parameters of the methylation potential energy landscape and employed a probabilistic approach that enabled us to compute from WGBS data a sensitivity index that quantifies the rate by which environmental perturbations influence methylation entropy along the genome (**Fig. 8**, **Supplementary Figs. 3e** and **8**, and **Supplementary Note**). For example, we observed entropic sensitivity within a CGI associated with *WNT1* in normal colon, with a portion gaining entropy and losing sensitivity in the matched cancer sample (**Fig. 8a**).

Globally, we observed differences in entropic sensitivity among tissues (**Fig. 8b** and **Supplementary Fig. 8**), with embryonic stem cells and brain showing higher levels of entropic sensitivity than the rest of the samples. As brain cells are highly methylated (**Fig. 3a**), high levels of entropic sensitivity would predict that the brain can show high

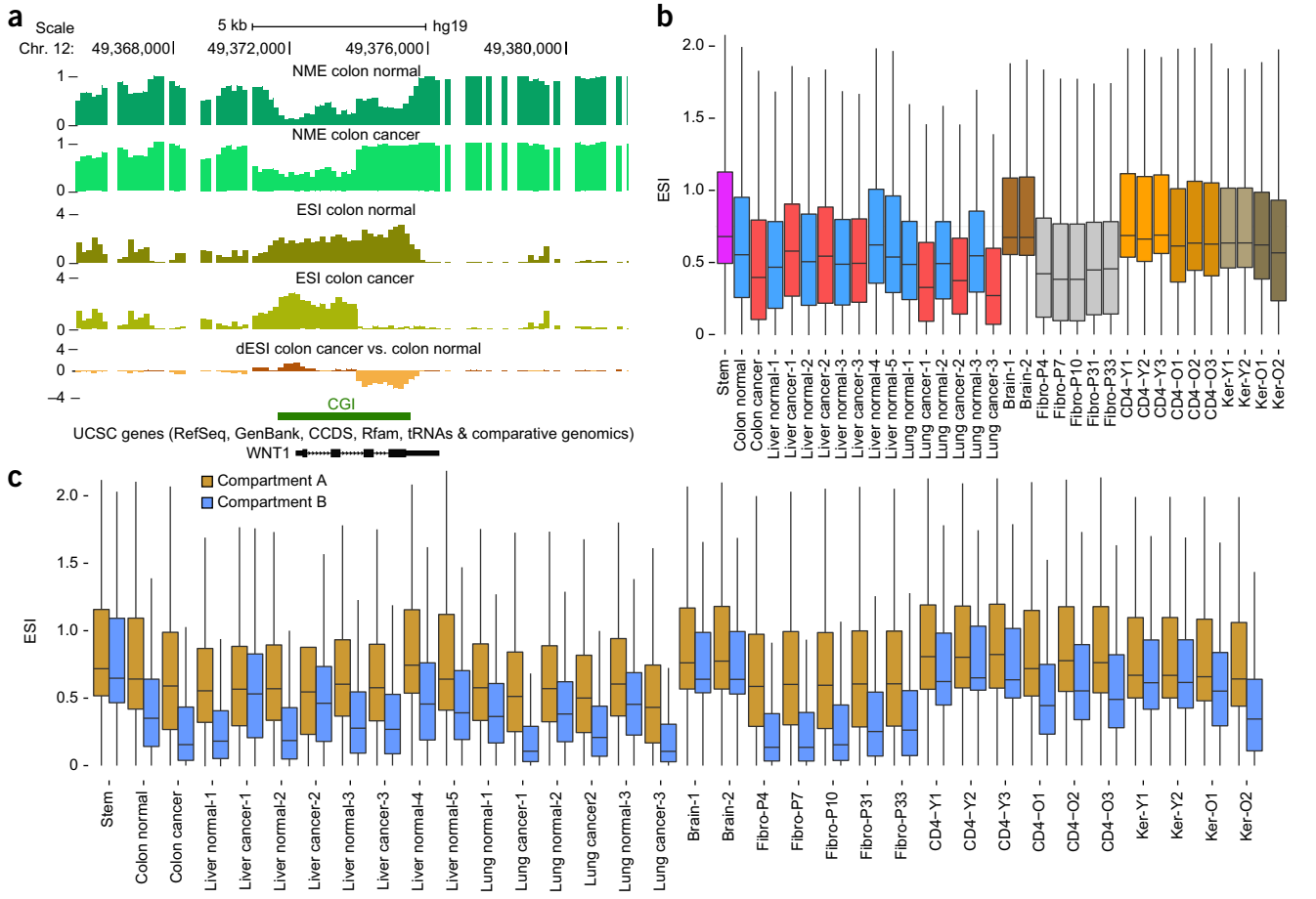

Figure 8 Entropic sensitivity distributions in single samples and comparative studies. (**a**) Gain in entropy and loss of entropic sensitivity are observed within a portion of the CGI associated with *WNT1*. NME, normalized methylation entropy; ESI, entropic sensitivity index; dESI, differential entropic sensitivity index. (**b**,**c**) Genome-wide distributions of entropic sensitivity across cell types (**b**) and within compartments A (brown) and B (blue) (**c**). Center lines, median; boxes, interquartile range (IQR); whiskers, 1.5 × IQR.

rates of demethylation in response to environmental stimuli, consistent with recent data showing that TET3 acts as a synaptic activity sensor that epigenetically regulates neural plasticity through active demethylation[34]. Colon and lung cancer showed global loss of entropic sensitivity (**Fig. 8b** and **Supplementary Fig. 8a**), whereas liver cancer showed a gain. Moreover, $CD4^+$ lymphocytes and skin keratinocytes exhibited global loss of entropic sensitivity in older individuals (**Fig. 8b** and **Supplementary Fig. 8b**), whereas cultured fibroblasts had lower sensitivity. Notably, we observed higher and more variable entropic sensitivity values within CGIs and at TSSs than in other genomic features, such as shores, exons, and introns (**Supplementary Fig. 3e**). However, some unmethylated CGIs showed low entropic sensitivity, whereas changes in entropic sensitivity within CGIs were observed between normal and cancer samples, as well as in older individuals (**Supplementary Fig. 8c–h**). Notably, differences in entropic sensitivity were not simply due to entropy itself, as many regions of low entropy had small sensitivity values, whereas other such regions displayed high values (**Supplementary Fig. 8c–e,g**). Last, we ordered genes in terms of entropic sensitivity within their promoters (Online Methods) and found many key developmental regulators or environmental sensors to be associated with high entropic sensitivity in stem cells and colon (**Supplementary Table 7**).

Entropic sensitivity within compartment A was higher than in compartment B in all samples (**Fig. 8c**), consistent with the idea that the transcriptionally active compartment A should be more responsive to stimuli. Moreover, differences between normal and cancer tissues were confined largely to compartment B (**Fig. 8c**). We observed substantial loss of entropic sensitivity in compartment B in older $CD4^+$ lymphocytes and skin keratinocytes, but not in compartment A. In contrast, cell culture led to gained sensitivity within compartment B (**Fig. 8c**).

To further investigate entropic sensitivity changes between tissues, we ranked genes according to differential entropic sensitivity within their promoters between colon normal and cancer samples (**Supplementary Table 8** and Online Methods). Several highly ranked genes were found to encode LIM-domain proteins, such as *LIMD2*, and are implicated in colon and other types of cancer, such as *QKI* and *HOXA9*, a canonical rearranged homeobox gene[35] that is dysregulated in cancer, and *FOXQ1*, which is overexpressed and enhances tumorigenicity in colorectal cancer[36]. Together, these results indicate that environmental exposure may influence epigenetic stochasticity in cells with sensitivity that varies along the genome and between compartments in a cell-type-specific manner. This presents the possibility that disease, environmental exposure, and aging are associated with substantial changes in entropic sensitivity, thus compromising integration of environmental cues regulating cell growth and function.

DISCUSSION

Our information-theoretic approach to epigenomics using the Ising model of statistical physics has shown that a formal approach

to methylation analysis can precisely extract and quantify the information content of experimental data to yield fundamental insights into epigenetic behavior. Here we have provided a formal definition of potential energy landscapes, characterized intrinsic epigenetic stochasticity, derived epigenetic entropy and methylation channels, associated chromatin organization with informational properties of methylation, and estimated entropic sensitivity to environmental conditions. We have also developed high-resolution computational tools for analyzing stochasticity in WGBS methylation data with low (10- to 20-fold) coverage, for quantifying epigenetic distances using normal–disease pairs that could be crucial in personalized medicine, and for predicting 3D chromatin structure from individual methylation samples in health and disease.

Shannon entropy varied markedly among tissues, across the genome, and among features of the genome. Entropy was increased with aging in skin and blood, but not in cell culture, suggesting a link between increased entropy and epigenetic aging. Jensen–Shannon distances precisely quantified epigenetic discordances between individual samples, demonstrating that cancer is informationally distant from both embryonic stem cells and normal tissues, thus providing a potential clinical advantage of identifying specific differences between two samples. Notably, epigenetic discordance was found to be associated with changes in entropy or large Jensen–Shannon distances and not necessarily with differences in mean methylation, and should thus be routinely used in epigenetic analysis.

We discovered that TAD boundaries are potential transition points between high- and low-entropy blocks and that information-theoretic properties of methylation channels could effectively predict chromatin organization in terms of compartments A and B. Computed compartments B demonstrated lower capacity, lower relative dissipated energy, higher Shannon entropy, lower entropic sensitivity, and larger Jensen–Shannon distance values in carcinogenesis, as well as significant overlap with hypomethylated blocks, LOCKs, and LADs. Moreover, A–B switching accurately quantified differences in phenotype, with marked switching in development and carcinogenesis. Finally, some cancers and aging were associated with global loss of entropic sensitivity that could be related to the autonomous nature of tumor cells and the well-known reduced physiological plasticity of aging.

This study demonstrates a relationship between chromatin structure, methylation channels, and entropic sensitivity that may maximize an organism's efficiency in storing epigenetic information and help explain developmental plasticity. In this model, pluripotent stem cells require relatively high energy to maintain high-capacity methylation channels within a portion of the genome, achieving reduced methylation stochasticity. Other regions characterized by increased entropic sensitivity are associated with highly deformable potential energy landscapes, which may correspond to differentiation branch points, as metaphorically suggested by Waddington. After differentiation, some large genomic domains, such as regions associated with pluripotency, need not maintain high channel capacities and energy consumption, with their sequestration providing increased energy efficiency but resulting in high epigenetic stochasticity and reduced responsiveness.

Furthering this model, our observation that compartment B shows reduced energy expenditure and channel capacity, and thus does not accurately maintain methylation information, explains the observed significant overlap of compartment B in normal tissue with hypomethylated blocks in cancer, suggesting that compartment B is more dysregulated than compartment A in carcinogenesis, in agreement with the higher Jensen–Shannon distance values observed. We therefore hypothesize that cancer cells gain a microevolutionary advantage upon reorganization of dysregulated B domains, thereby amplifying epigenetic stochasticity to increase plasticity and adaptability beyond that of the primary tissue.

The stochastic nature and properties of DNA methylation and their close relationship with chromatin structure raise the possibility that epigenetic information is carried by a population of cells as a whole and that this information helps not only to achieve and maintain a differentiated state but also to mediate developmental plasticity throughout the life of an organism.

URLs. UCSC genome browser, http://genome.ucsc.edu/; VISTA enhancer browser, http://enhancer.lbl.gov/; NCBI GEO, https://www.ncbi.nlm.nih.gov/geo/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank X. Li, A. Vandiver, and J. Walston (Johns Hopkins University) for cells and/or FASTQ files; R. Trygvadottir, B. Berndsen, A. Idrizi, and C. Callahan (Johns Hopkins University) for sequencing; J.-P. Fortin and K. Hansen for providing A and B compartment data; I. Morrison (University of Otago) and A. Gimelbrant (Dana-Farber Cancer Institute) for access to imprinted gene and MAE data sets; A. Meissner and M. Ziller (Broad Institute) for access to bisulfite sequencing data sets; and W. Timp and K. Hansen (Johns Hopkins University) for critical reading of the manuscript. This work was supported by US National Institutes of Health (NIH) grants R01CA054358 and DP1ES022579 to A.P.F., National Science Foundation grants CCF-1217213 and CCF-1656201 to J.G., and NIH grant AG021334 to J. Walston. E.P. was supported by the Medical Scientist Training Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

A.P.F., E.P., G.J., and J.G. designed the study. G.J. and J.G. developed the mathematical and computational methods. G.J. wrote the computer code and implemented the methods. A.P.F. and E.P. designed and led the experiments. E.P. procured outside data and performed quality control, preprocessing, and bisulfite alignment. G.J., E.P., A.P.F., and J.G. analyzed the data. A.P.F., G.J., and J.G. wrote the manuscript with the assistance of E.P.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Waddington, C.H. *The Strategy of the Genes* (Allen and Unwin, 1957).
2. Feinberg, A.P. & Irizarry, R.A. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. USA* **107** (Suppl. 1), 1757–1764 (2010).
3. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
4. Pujadas, E. & Feinberg, A.P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148**, 1123–1131 (2012).
5. Timp, W. & Feinberg, A.P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* **13**, 497–510 (2013).
6. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
7. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).
8. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–719 (2012).
9. Pressé, S., Ghosh, K., Lee, J. & Dill, K.A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **85**, 1115–1141 (2013).
10. Cedar, H. & Bergman, Y. Programming of DNA methylation patterns. *Annu. Rev. Biochem.* **81**, 97–117 (2012).

11. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. USA* **109**, 10522–10527 (2012).
12. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
13. Bell, R.E. *et al.* Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.* **26**, 601–611 (2016).
14. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
15. Mohn, F. *et al.* Lineage-specific Polycomb targets and *de novo* DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30**, 755–766 (2008).
16. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
17. Nora, E.P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
18. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49**, 773–782 (2013).
19. Huang, J., Marco, E., Pinello, L. & Yuan, G.C. Predicting chromatin organization using histone marks. *Genome Biol.* **16**, 162 (2015).
20. Cover, T.M. & Thomas, J.A. *Elements of Information Theory* (John Wiley & Sons, 1991).
21. Savarese, F. *et al.* Satb1 and Satb2 regulate embryonic stem cell differentiation and *Nanog* expression. *Genes Dev.* **23**, 2625–2638 (2009).
22. Karantzali, E. *et al.* Sall1 regulates embryonic stem cell differentiation in association with Nanog. *J. Biol. Chem.* **286**, 1037–1045 (2011).
23. Liu, K. *et al.* The multiple roles for Sox2 in stem cell maintenance and tumorigenesis. *Cell. Signal.* **25**, 1264–1271 (2013).
24. Ozair, M.Z., Noggle, S., Warmflash, A., Krzyspiak, J.E. & Brivanlou, A.H. SMAD7 directly converts human embryonic stem cells to telencephalic fate by a default mechanism. *Stem Cells* **31**, 35–47 (2013).
25. Gopinath, S.D., Webb, A.E., Brunet, A. & Rando, T.A. FOXO3 promotes quiescence in adult muscle stem cells during the process of self-renewal. *Stem Cell Rep.* **2**, 414–426 (2014).
26. Mahaira, L.G. *et al. IGF2BP1* expression in human mesenchymal stem cells significantly affects their proliferation and is under the epigenetic control of TET1/2 demethylases. *Stem Cells Dev.* **23**, 2501–2512 (2014).
27. Lee, B.K. *et al.* Tgif1 counterbalances the activity of core pluripotency factors in mouse embryonic stem cells. *Cell Rep.* **13**, 52–60 (2015).
28. Luo, Z. *et al.* Zic2 is an enhancer-binding factor required for embryonic stem cell specification. *Mol. Cell* **57**, 685–694 (2015).
29. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).
30. Fortin, J.P. & Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).
31. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
32. Timp, W. *et al.* Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.* **6**, 61 (2014).
33. Berman, B.P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. *Nat. Genet.* **44**, 40–46 (2011).
34. Yu, H. *et al.* Tet3 regulates synaptic transmission and homeostatic plasticity via DNA oxidation and repair. *Nat. Neurosci.* **18**, 836–843 (2015).
35. Nakamura, T. *et al.* Fusion of the nucleoporin gene *NUP98* to *HOXA9* by the chromosome translocation t(7;11)(p15;p15) in human myeloid leukaemia. *Nat. Genet.* **12**, 154–158 (1996).
36. Kaneda, H. *et al.* FOXQ1 is overexpressed in colorectal cancer and enhances tumorigenicity and tumor growth. *Cancer Res.* **70**, 2053–2063 (2010).

ONLINE METHODS

Samples for WGBS. We used previously published WGBS data corresponding to ten samples, which included H1 human embryonic stem cells[37], normal and matched cancer cells from colon and liver[12], keratinocytes from skin biopsies of sun-protected sites from younger and older individuals[38], and EBV-immortalized lymphoblasts[39]. We also generated WGBS data corresponding to 25 samples that included normal and matched cancer cells from liver and lung, prefrontal cortex (brain), cultured human neonatal fibroblasts at five passages, and sorted $CD4^+$ T cells from younger and older individuals, all with institutional review board (IRB) approval. We obtained prefrontal cortex samples from the University of Maryland Brain and Tissue Bank, which is a Brain and Tissue Repository of the NIH NeuroBioBank. Peripheral blood mononuclear cells (PBMCs) were isolated from peripheral blood collected from healthy subjects and separated using a Ficoll density-gradient separation method (Sigma-Aldrich). $CD4^+$ T cells were subsequently isolated from PBMCs by positive selection with MACS magnetic bead technology (Miltenyi). Post-separation flow cytometry assessed the purity of $CD4^+$ T cells to be 97%. Primary neonatal dermal fibroblasts (mycoplasma free) were acquired from Lonza and cultured in DMEM (Gibco) supplemented with 15% FBS (Gemini BioProducts).

DNA isolation. We extracted genomic DNA from samples using the Masterpure DNA Purification kit (Epicentre). High molecular weight of the extracted DNA was verified by running a 1% agarose gel and by assessing the 260/280 and 260/230 ratios of samples on a NanoDrop spectrophotometer. Concentration was determined using a Qubit 2.0 fluorometer (Invitrogen).

Generation of WGBS libraries. For every sample, 1% unmethylated lambda DNA (Promega, D1521) was spiked in to monitor bisulfite conversion efficiency. Genomic DNA was fragmented to an average size of 350 bp using a Covaris S2 sonicator. Bisulfite sequencing libraries were constructed using the Illumina TruSeq DNA Library Preparation kit protocol (primers included) or NEBNext Ultra (NEBNext Multiplex Oligos for Illumina module, New England BioLabs, E7535L) according to the manufacturer's instructions. Both protocols use a Kapa HiFi Uracil+ PCR system (Kapa Biosystems, KK2801).

For Illumina TruSeq DNA libraries, gel-based size selection was performed to enrich for fragments in the 300- to 400-bp range. For NEBNext libraries, size selection was performed using modified AMPure XP bead ratios of 0.4× and 0.2×, aiming also for an insert size of 300–400 bp. After size selection, the samples were bisulfite converted and purified using the EZ DNA Methylation Gold kit (Zymo Research, D5005). PCR-enriched products were cleaned up using 0.9× AMPure XP beads (Beckman Coulter, A63881).

Final libraries were run on the 2100 Bioanalyzer (Agilent) using the High-Sensitivity DNA assay for quality control purposes. Libraries were then quantified by qPCR with the Library Quantification kit for Illumina sequencing platforms (KK4824, Kapa Biosystems), using the 7900HT Real-Time PCR System (Applied Biosystems), and sequenced on the Illumina HiSeq 2000 (2 × 100-bp read length, v3 chemistry according to the manufacturer's protocol with 10× PhiX spike-in) and HiSeq 2500 (2 × 125-bp read length, v4 chemistry according to the manufacturer's protocol with 10× PhiX spike-in).

Quality control and alignment. FASTQ files were processed using Trim Galore v0.3.6 (Babraham Institute) to perform single-pass adaptor and quality trimming of reads, as well as running FastQC v0.11.2 for a general quality check of sequencing data. Reads were then aligned to the hg19/GRCh37 genome using Bismark v0.12.3 and Bowtie2 v2.1.0. Separate mbias plots for read 1 and read 2 were generated by running the Bismark methylation extractor using the 'mbias_only' flag. These plots were used to determine how many bases to remove from the 5′ end of reads. The number was generally higher for read 2, which is known to have poorer quality. The amount of 5′ trimming ranged from 4 bp to 25 bp, with the most common values being around 10 bp. BAM files were subsequently processed with SAMtools v0.1.19 for sorting, merging, duplicate removal, and indexing.

FASTQ files associated with the EBV sample were processed using the same pipeline described for the in-house samples. BAM files associated with the normal colon and liver samples, obtained from Ziller *et al.*[12], could not be assessed using the Bismark methylation extractor because of incompatibility of the original alignment tool (MAQ) used on these samples. We therefore followed the advice of the authors and trimmed 4 bp from all reads for those files.

Genomic features and annotations. Files and tracks bear genomic coordinates for hg19. CGIs were obtained from Wu *et al.*[40]. CGI shores were defined as the 2-kb flanking sequences on either side of islands, shelves were defined as the 2-kb flanking sequences beyond the shores, and open seas were defined as everything else. The R Bioconductor package 'TxDb.Hsapiens.UCSC.hg19.knownGene' was used to define 3′ UTRs, 5′ UTRs, exons, introns, and TSSs. Promoter regions were defined as the 2-kb regions flanking either side of TSSs. A curated list of enhancers was obtained from the VISTA enhancer browser[41] by downloading all human (hg19) positive enhancers that had reproducible expression in at least three independent transgenic mouse embryos. Hypomethylated blocks (colon and lung cancer) were obtained from Timp *et al.*[32], whereas H1 stem cell LOCKs and human pulmonary fibroblast (HPF) LOCKs were obtained from Wen *et al.*[42]. LAD tracks associated with Tig3 cells derived from embryonic lung fibroblasts were obtained from Guelen *et al.*[43]. Gene bodies were obtained from the UCSC genome browser, H1 and IMR90 TAD boundaries were obtained from the laboratory of B. Ren at the University of California, San Diego, and BED files for Hi-C data processed into compartments A and B were provided by J.-P. Fortin and K. Hansen (Johns Hopkins University).

Estimation and display of potential energy landscapes. We partitioned the genome into consecutive nonoverlapping regions of equal size and estimated the parameters θ of the potential energy landscape within a region by maximizing the average log-likelihood $\frac{1}{M}\sum_{m=1}^{M}\ln\left[P(\mathbf{x}_m|\theta)\right]$, with $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_M$ being M independent observations of the methylation state (i.e., WGBS sequencing reads) within the region. To take into account partially observed methylation states, we replaced $P(\mathbf{x}_m \mid \theta)$ by the joint probability distribution over only those sites at which methylation information was available, which we calculated by marginalizing $P(\mathbf{x}_m \mid \theta)$ over these sites. After extensive experimentation, we considered 3-kb estimation regions by striking a balance between estimation and computational performance. To avoid statistical overfitting, we did not model regions with fewer than ten CpG sites. We also ignored regions with not enough data for which less than two-thirds of the CpG sites were observed or the average depth of coverage was less than 2.5 observations per CpG site. We finally performed optimization using the multilevel coordinate search (MCS) algorithm[44], which was chosen because of its superior performance among the derivative-free global optimization algorithms we tested (such as simulated annealing).

To visualize a potential energy landscape as a 3D plot, we used the 2D version of the Gray code[45]. According to this method, we placed all possible 2^N binary-valued methylation states within a genomic region with N CpG sites on a 2D plane in a manner such that states located adjacent to each other in the east–west and north–south directions differed by only 1 bit. We then obtained a 3D plot by assigning to each state its potential value.

Computation of probability distribution of methylation level. We calculated the probability distribution $P(l)$ of the methylation level $L = \frac{1}{N}\sum_{n=1}^{N} X_n$ within a genomic unit with N CpG sites from the Ising probability distribution $P(\mathbf{x})$ of the methylation patterns within the genomic unit using $P(l) = \sum_{\mathbf{x}\in Q(Nl)} P(\boldsymbol{x})$ where $Q(Nl)$ is the set of all methylation patterns with exactly Nl methylated CpG sites. In the rarer case when N was too large to make direct summation tractable, we used the method of maximum entropy[46] to approximate $P(l)$ by estimating the first four noncentral moments of the methylation level L using Monte Carlo.

Normalized methylation entropy. We quantified methylation stochasticity within a genomic unit with N CpG sites using the normalized methylation entropy $h = H/\log_2(N+1)$, where $H = -\sum_l P(l)\log_2 P(l)$ is the informational (Shannon) entropy[20] of the methylation level. The normalized methylation entropy ranges between 0 and 1, taking its maximum value when all methylation

doi:10.1038/ng.3811

levels within a genomic unit are equally likely (fully disordered state) regardless of the number of CpG sites, and achieving its minimum value when only a single methylation level is observed (perfectly ordered state).

Quantifying differential behavior in methylation level. To quantify differences in the probability distributions of the methylation level within a genomic unit between two samples, we employed the Jensen–Shannon distance[47]

$$D_{\mathrm{JS}} = \sqrt{\frac{1}{2}\left[D_{\mathrm{KL}}\left(P_1, \bar{P}\right) + D_{\mathrm{KL}}\left(P_2, \bar{P}\right)\right]}$$

where P_1 and P_2 are the probability distributions of the methylation level within the genomic unit in the first and second samples, $\bar{P} = (P_1 + P_2)/2$ is the average of the two probability distributions, and

$$D_{\mathrm{KL}}\left(P, \bar{P}\right) = \sum_l P(l) \log_2 \left[\frac{P(l)}{\bar{P}(l)}\right]$$

is the relative entropy or Kullback–Leibler divergence[20]. The Jensen–Shannon distance simultaneously encapsulates any differences in the probability distributions of methylation level within a genomic unit across two samples by measuring dissimilarities between these distributions (including the mean methylation and entropy). It is a normalized distance metric, taking values between 0 and 1, which equals 0 only when the two probability distributions P_1 and P_2 are identical and reaches its maximum value of 1 when the supports of the two distributions do not intersect each other.

Epigenetic distances, multidimensional scaling, and gene ranking. We quantified the epigenetic discordance between two samples by calculating a dissimilarity value defined as the average of all Jensen–Shannon distance values computed genome-wide. To visualize epigenetic similarities or dissimilarities between samples, we computed the epigenetic distances between all pairs of samples, formed the corresponding dissimilarity matrix, and employed a 2D representation, using multidimensional scaling based on Kruskal's nonmetric method, to find a 2D configuration of points whose inter-point distances approximately corresponded to the epigenetic dissimilarities among the samples.

To rank genes based on the absolute difference in mean methylation levels within their promoters, we centered a 4-kb window on the TSS of each gene in the genome, computed the absolute difference in mean methylation levels within each genomic unit that overlapped this window, and scored the gene by averaging these values. We used the same method to rank genes based on the Jensen–Shannon distance. We also ranked genes using a relative scheme that assigned a higher score to genes with larger Jensen–Shannon distances but smaller absolute differences in mean methylation level. We did so by scoring a gene using the ratio of its ranking in the mean-based list to its ranking in the list obtained by the Jensen–Shannon distance.

Computation of entropy blocks. Computation of entropy blocks requires detection of ordered and disordered blocks, i.e., large genomic regions of consistently low or high normalized methylation entropy values. To effectively summarize methylation entropy in a single sample, we computed the normalized methylation entropy h within each genomic unit and classified it into one of three classes: ordered ($0 \le h \le 0.44$), weakly ordered/disordered ($0.44 < h < 0.92$), and disordered ($0.92 \le h \le 1$). We determined the threshold values by investigating the relationship between the normalized methylation entropy within a genomic unit that contained one CpG site and the ratio of the probability p of methylation to the probability $1 - p$ of unmethylation at that site. To this end, we focused on the odds ratio $r = p/(1 - p)$ and considered the methylation level to be ordered if $r \ge 10$ or $r \le 1/10$ (i.e., if the probability of methylation is at least ten times larger than the probability of unmethylation, and likewise for the probability of unmethylation), in which case, $p \ge 0.9091$ or $p \le 0.0909$, corresponding to a maximum normalized methylation entropy threshold of 0.44. Moreover, we considered the methylation level to be disordered if $1/2 \le r \le 2$ (i.e., if the probability of methylation was no more than two times the probability of unmethylation, and likewise for the probability of unmethylation), in which case, $0.3333 \le p \le 0.6667$, corresponding to a minimum normalized methylation threshold of 0.92.

To compute entropy blocks, we slid a window of 500 genomic units (75 kb) along the genome and labeled the window as being ordered or disordered if at least 75% of its genomic units were effectively classified as being ordered or disordered, respectively. We then determined ordered or disordered blocks by taking the union of all ordered or disordered windows while excluding regions of overlap between ordered and disordered windows.

Prediction of TAD boundaries. Using entropy blocks computed for a given sample, we identified predictive regions of the genome that might contain TAD boundaries by detecting the space between successive entropy blocks with distinct labels (ordered or disordered). For example, if an ordered block located at chr. 1: 1–1,000 were followed by a disordered block at chr. 1: 1,501–2,500, then chr. 1: 1,001–1,500 would be deemed to be a predictive region. To reduce false identification of predictive regions, we did not consider successive entropy blocks of the same type, as the genomic space between two such entropy blocks might be due to missing data or other unpredictable factors. To control the resolution of locating a TAD boundary, we considered only gaps smaller than 50 kb. This resulted in resolution an order of magnitude smaller than the mean TAD size (~900 kb). To combine predictive regions obtained from methylation analysis of several distinct epigenotypes, we computed the predictive coverage of each base pair by counting the number of predictive regions that contained the base pair. We then combined predictive regions by grouping consecutive base pairs whose predictive coverage was at least 4. We subsequently applied this method on WGBS data corresponding to 17 distinct cell and tissue types (stem, colon normal, colon cancer, liver normal-1, liver cancer-1, liver normal-2, liver cancer-2, liver normal-3, liver cancer-3, lung normal-1, lung cancer-1, lung normal-2, lung cancer-2, lung normal-3, lung cancer-3, brain-1, and brain-2), and analyzed our results using GenometriCorr[48], a statistical package for evaluating the correlation of genome-wide data with given genomic features. Finally, we considered a boundary prediction to be correct when the distance of a true TAD boundary from the center of a predictive region was less than the first quartile of the true TAD width distribution (**Fig. 5c**).

A and B compartment prediction and analysis. Genome-wide prediction of A and B compartments was performed by a random forest regression model. We trained this model using a small number of available Hi-C data sets associated with EBV and IMR90 samples[49], as well as A and B tracks produced by the method of Fortin and Hansen using long-range correlations computed from pooled 450k array data associated with colon cancer, liver cancer, and lung cancer samples[30]. Because of the paucity of currently available Hi-C data, we included the Fortin–Hansen data to increase the number of training samples and improve the accuracy of performance evaluation. We first paired the Hi-C and Fortin–Hansen data with WGBS EBV, fibro-P10, and colon cancer samples, as well as with samples obtained by pooling WGBS liver cancer (liver cancer-1, liver cancer-2, and liver cancer-3) and lung cancer (lung cancer-1, lung cancer-2, and lung cancer-3) data. We subsequently partitioned the entire genome into 100-kb bins (to match the available Hi-C and Fortin–Hansen data), and computed eight information-theoretic features of methylation maintenance within each bin (median values and interquartile ranges of information capacity, relative dissipated energy, normalized methylation entropy, and mean methylation level). By using all feature–output pairs, we trained a random forest model using the R package randomForest with its default settings, except that we increased the number of trees to 1,000. We then applied the trained random forest model on each WGBS sample and produced A and B tracks that approximately identified A and B compartments associated with the samples. Because regression took into account only information within a 100-kb bin, we averaged the predicted A and B values using a three-bin smoothing window and removed from the overall A and B signal its genome-wide median value, as suggested by Fortin and Hansen[30].

To test the accuracy of the resulting predictions, we employed fivefold leave-one-out cross-validation, which involved training using four sample pairs and testing on the remaining pair for all five combinations. We evaluated performance by computing the average correlation as well as the average percentage agreement between the predicted and each of the ground-truth A and B signals within 100-kb bins at which the absolute values of the predicted and ground-truth signals were both greater than a calling margin, where we

used a nonzero calling margin to remove unreliable predictions. We finally calculated agreement by testing whether the predicted and ground-truth A and B values within a 100-kb bin had the same sign.

For each pair of WGBS samples, we computed the percentage of A–B compartment switching by dividing the number of 100-kb bin pairs for which an A prediction was made in the first sample and a B prediction was made in the second sample by the total number of bins for which A and B predictions were available in both samples, and similarly for the case of B–A switching. We summed these percentages and formed a matrix of dissimilarity measures, which we then used as an input to a Ward error sum of squares hierarchical clustering scheme[50], which we implemented using the R package hclust by setting the method variable to ward.D2.

To test the significance of overlap of hypomethylated blocks, LADs, and LOCKs with compartment B, we used available hypomethylated blocks, LOCKs, and LADs and predicted compartment B data for the lung normal-1, lung normal-2, and lung normal-3 samples, which best matched the previous tracks. To evaluate enrichment of hypomethylated blocks (and similarly for LADs and LOCKs) within compartment B, we defined two binary (0 and 1) random variables R and B for each genomic unit of the genome such that $R = 1$ if the genomic unit overlaps a block and $B = 1$ if the genomic unit overlaps compartment B. We then tested against the null hypothesis that R and B are statistically independent by applying the χ^2 test on the 2×2 contingency table for R and B and calculated the odds ratio (OR) as a measure of enrichment.

Entropy sensitivity and gene ranking. We ranked genes on the basis of entropic sensitivity and its differences between a test and a reference sample within their promoters. We did so by centering a 4-kb window on the TSS of each gene in the genome and computing the values or the absolute difference in the values of the entropic sensitivity index within each genomic unit that 'touched' this window, and scored the gene by averaging these values.

Software availability. Source code is available at https://github.com/GarrettJenkinson/informME/.

Data availability. H1 and IMR90 TAD boundaries were from http://chromosome.sdsc.edu/mouse/hi-c/download.html; BED files for Hi-C data processed into compartments A and B are available at https://github.com/Jfortin1/HiC_AB_Compartments. WGBS data and bigWig files of relevant features have been deposited in the Gene Expression Omnibus under accession code GSE86340.

37. Schlaeger, T.M. *et al.* A comparison of non-integrating reprogramming methods. *Nat. Biotechnol.* **33**, 58–63 (2015).
38. Vandiver, A.R. *et al.* Age and sun exposure–related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biol.* **16**, 80 (2015).
39. Hansen, K.D. *et al.* Large-scale hypomethylated blocks associated with Epstein–Barr virus–induced B-cell immortalization. *Genome Res.* **24**, 177–184 (2014).
40. Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A. & Feinberg, A.P. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**, 499–514 (2010).
41. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
42. Wen, B. *et al.* Euchromatin islands in large heterochromatin domains are enriched for CTCF binding and differentially DNA-methylated regions. *BMC Genomics* **13**, 566 (2012).
43. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
44. Huyer, W. & Neumaier, A. Global optimization by multilevel coordinate search. *J. Glob. Optim.* **14**, 331–355 (1999).
45. Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. *Numerical Recipes. The Art of Scientific Computing* (Cambridge University Press, 2007).
46. Mohammad-Djafari, A. in *Maximum Entropy and Bayesian Methods* (eds. Smith, C.R., Erickson, G.J. & Neudorfer, P.O.) 221–234 (Kluwer Academic Publishers, 1991).
47. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
48. Favorov, A. *et al.* Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput. Biol.* **8**, e1002529 (2012).
49. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
50. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).

doi:10.1038/ng.3811