Field-theoretic density estimation for biological sequence space with applications to 5′ splice site diversity and aneuploidy in cancer

Wei-Chia Chen[a], Juannan Zhou[a], Jason M. Sheltzer[b], Justin B. Kinney[a], and David M. McCandlish[a,1]

[a]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; and [b]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited by Curtis G. Callan Jr., Princeton University, Princeton, NJ, and approved August 29, 2021 (received for review December 22, 2020)

Density estimation in sequence space is a fundamental problem in machine learning that is also of great importance in computational biology. Due to the discrete nature and large dimensionality of sequence space, how best to estimate such probability distributions from a sample of observed sequences remains unclear. One common strategy for addressing this problem is to estimate the probability distribution using maximum entropy (i.e., calculating point estimates for some set of correlations based on the observed sequences and predicting the probability distribution that is as uniform as possible while still matching these point estimates). Building on recent advances in Bayesian field-theoretic density estimation, we present a generalization of this maximum entropy approach that provides greater expressivity in regions of sequence space where data are plentiful while still maintaining a conservative maximum entropy character in regions of sequence space where data are sparse or absent. In particular, we define a family of priors for probability distributions over sequence space with a single hyperparameter that controls the expected magnitude of higher-order correlations. This family of priors then results in a corresponding one-dimensional family of maximum a posteriori estimates that interpolate smoothly between the maximum entropy estimate and the observed sample frequencies. To demonstrate the power of this method, we use it to explore the high-dimensional geometry of the distribution of 5′ splice sites found in the human genome and to understand patterns of chromosomal abnormalities across human cancers.

field theory | spectral graph theory | maximum entropy | bioinformatics | molecular evolution

Biological data are often discrete and combinatorial. We observe, for instance, some collection of macromolecular sequences that take the form of a string of nucleotides or amino acids. Or we make a multichannel neural recording, resulting in a collection of strings composed of zeroes and ones corresponding to the set of neurons that are firing at each instant in time. A natural question given a collection of such strings is whether we can estimate the probability distribution that these sequences are drawn from (1–5).

Estimating such a probability distribution can be challenging because the number of possible sequences grows exponentially in sequence length, and even if the subset of biologically active or relevant sequences is small compared with the entirety of the space, this biologically relevant subset can still easily contain thousands of sequences. As a result, estimating the frequency of each possible sequence becomes impractical, and we require some prior or set of simplifying assumptions in order to make progress.

Among the most common simplifying assumptions is that the true distribution takes the form of a maximum entropy distribution, defined as the most uniform (i.e., highest-entropy) distribution compatible with certain summary statistics of the sample (6). These summary statistics are often taken to be the frequency of each possible letter at each position. In that case, the resulting maximum entropy model is the well-known position weight matrix, which represents the distribution of sequences as the product of independent position-specific probability distributions (7, 8). Matching the correlations between positions in addition to the site-specific frequencies results in pairwise maximum entropy models, also known as Potts models (1, 9–18). Such pairwise interaction models have seen great success in a variety of applications, including identifying functional elements (9), predicting residues or positions that contact each other or interact (14, 19, 20), and predicting the effects of mutations (21).

Here, we provide a generalization of these maximum entropy models that can achieve greater expressivity in well-sampled, high-probability regions of sequence space while still providing parsimonious density estimates in low-probability regions, where data are by necessity sparse or absent. We do this by deriving a one-parameter family of Bayesian priors for probability distributions over sequence space, with the single hyperparameter controlling the expected deviation from the local geometry implied by the maximum entropy assumption. The resulting family of maximum a posteriori (MAP) estimates matches the same moments as the corresponding maximum entropy model and includes the maximum entropy model and the histogram of observed frequencies for each sequence as limiting cases. In nonlimiting cases, these models can capture correlations of all orders, and they produce estimates that resemble the histogram

Significance

Often in computational biology, our data take the form of a set of sequences: for example, a collection of nucleic acid sequences with a shared function. Given such a collection of sequences, how can we estimate the probability distribution that these sequences were drawn from? This problem is important both for identifying other sequences with similar function and because the structure of these probability distributions is often informative about the biophysical and evolutionary processes that generated the observed sequences. Here, we present a method for estimating such probability distributions that generalizes the popular maximum entropy approach. This method is capable of capturing fine details of the distribution in high-density regions where data are plentiful but still behaves conservatively in low-density regions.

Author contributions: W.-C.C., J.Z., J.M.S., J.B.K., and D.M.M. designed research; W.-C.C., J.Z., and J.B.K. performed research; and W.-C.C., J.Z., and D.M.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mccandlish@cshl.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2025782118/-/DCSupplemental.

Published October 1, 2021.

of observed frequencies in densely sampled regions where the data overwhelm the prior but also exhibit the smooth behavior typical of a maximum entropy model in sparsely sampled regions. At a more technical level, the method we propose is the discrete multivariate analog of recently developed field-theoretic density estimation techniques (22–28) for real-valued random variables. One such approach (28) is known as density estimation using field theory (DEFT), and we therefore refer to our method as SeqDEFT.

In what follows, we describe the formal characteristics of our method and then apply the method to two biological datasets. The first of these datasets is the collection of all annotated 5′ pre-mRNA splice sites in the human genome (RNA sequences of length nine). Because of the relatively large number of annotated splice sites (305,106) compared with the number of possible sequences ($4^9 = 262{,}144$), this dataset allows us to use subsampling to assess the characteristics and performance of our model on complex distributions under varying amounts of training data. For the second dataset, we consider the distribution of chromosomal copy number abnormalities observed across human cancers (29). Here, there are far fewer observations (10,522 samples) than there are possible karyotypic patterns ($2^{22} = 4{,}194{,}304$ under our scoring scheme). However, our method is still able to recover the signatures of several complex aneuploid states corresponding to sets of multiple chromosomes that are frequently altered together. Using an evolutionarily motivated visualization strategy (30), we further explore and discuss the major features of the inferred probability distributions for each of these two datasets, with an emphasis on understanding the biological basis for their complex, multimodal structure.

Results

We consider probability distributions defined on the Hamming graph of sequences with length ℓ and α alleles per position, where two sequences are adjacent if they differ in exactly one position. To define a probability distribution Q on this graph, we first define a field ϕ on the graph and then let the probability of drawing sequence i be given by

$$Q_i = \frac{e^{-\phi_i}}{\sum_{j=1}^{\alpha^\ell} e^{-\phi_j}}. \quad [1]$$

Thus, we can define a prior over probability distributions by imposing a prior on this field ϕ. In what follows, we are particularly interested in the "shape" or "geometry" of the field ϕ with respect to the adjacency structure of the Hamming graph, both in terms of its global structure such as the number of distinct regions with low values of ϕ (i.e., "modes" of the probability distribution Q) as well as its local features such as its behavior on specific sub-Hamming graphs (which correspond to conditional distributions defined by restricting the possible alleles at a subset of positions).

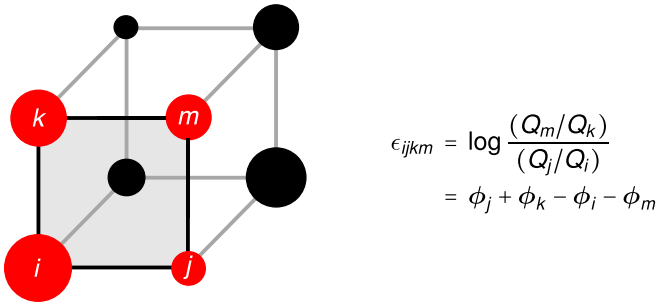

Fig. 1. Illustration of a simple Hamming graph and a face thereof ($\alpha = 2$, $\ell = 3$). Each node represents a sequence, and each edge corresponds to a point mutation. The size of a node is proportional to the probability of that sequence. A face consists of four sequences: one wild type (*i*), two single mutants (*j* and *k*), and one double mutant (*m*). The conditional log odds ratio for this face ϵ_{ijkm} quantifies the association of the mutation from *i* to *j* and the mutation from *i* to *k* on this face. This conditional odds ratio can also be thought of as measuring how much the presence vs. absence of a mutation at one of these two sites increases the frequency of the other mutation [i.e., $\epsilon_{ijkm} = (\phi_k - \phi_m) - (\phi_i - \phi_j) = (\phi_j - \phi_m) - (\phi_i - \phi_k)$, where $\phi_i - \phi_j = -(\phi_j - \phi_i) = \log(Q_j/Q_i)$].

Extending the Independent Model. The maximum entropy model based on the marginal frequencies of the alleles at each position is equivalent to making an independent draw of the allele for each position from the observed marginal frequencies. For concreteness, we will derive our method for this basic form of the maximum entropy model before turning to the general case.

As noted earlier, our overall strategy is to think geometrically about ϕ as a function on the Hamming graph. In particular, let us consider the behavior of ϕ on one particular "face" of the Hamming graph, where a face is defined by selecting two specific positions out of the ℓ positions, a mutant and a wild-type allele at each of these positions, and a specific sequence for the other $\ell - 2$ positions. That is, we define a face as a choice of wild-type sequence i, two single mutants j and k, and a double mutant m (Fig. 1). Moreover, if we consider only sequences drawn from this face, then we can capture the conditional association between these two mutations in terms of the log odds ratio:

$$\epsilon_{ijkm} = \phi_j + \phi_k - \phi_i - \phi_m. \quad [2]$$

For the special case of the independent model, the values of ϕ are additive such that the double mutant ϕ_m is given by ϕ evaluated at the wild type plus the effects of each of the two single mutants on ϕ: that is,

$$\phi_m = \phi_i + (\phi_j - \phi_i) + (\phi_k - \phi_i).$$

Rearranging this expression, it is easy to see that this implies that for the independent sites maximum entropy model, the conditional log odds ratio ϵ_{ijkm} is zero for every face in the Hamming graph.

To build a model that allows deviations from the maximum entropy assumption while tending to make these deviations small, we can thus construct a prior on functions ϕ where the probability of ϕ is determined by the extent of the deviation from the perfectly additive local geometry implied by the maximum entropy model. In particular, since under the maximum entropy model ϵ is zero for every face in the Hamming graph, we can quantify the extent of the deviation from local independence by considering the average squared conditional log odds ratio $\overline{\epsilon^2}$, where this average is taken over all faces of the Hamming graph. In fact, it is possible to derive a simple formula for $\overline{\epsilon^2}$ in terms of the graph Laplacian L of the Hamming graph, where L is defined as

$$L(i,j) = \begin{cases} -1 & i \text{ and } j \text{ are at Hamming distance 1} \\ \ell(\alpha - 1) & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if we let $\Delta = (L^2 - \alpha L)/2$ and let $s = \binom{\ell}{2}\binom{\alpha}{2}^2 \alpha^{\ell-2}$ be the number of distinct faces, then this average is given by the positive semidefinite quadratic form $\overline{\epsilon^2} = \phi^T \Delta \phi / s$ (31).

Using this expression for the mean squared log odds ratio, we can then define a family of improper Gaussian priors on ϕ where the prior probability is maximized for ϕ compatible with the maximum entropy model but which also allows a controllable degree of nonadditivity. In particular, we use the prior

$$p(\phi|a) \propto e^{-S_a^0[\phi]} \quad [3]$$

in terms of the action

$$S_a^0[\phi] = \frac{a}{2s}\phi^T \Delta \phi, \quad [4]$$

where $S_a^0[\phi] = 0$ for any ϕ that results in a maximum entropy (i.e., independent sites) model and a is a hyperparameter that controls the expected magnitude of deviations from independence, with larger values of a producing smaller deviations. In *SI Appendix*, we show that the expectation of $\overline{\epsilon^2}$ under the prior is given by $\mathrm{rank}(\Delta)/a$, where in this case, $\mathrm{rank}(\Delta) = \alpha^\ell - (\alpha-1)\,\ell - 1$. We also show that this prior is equivalent to independently drawing the value of ϵ for each face in the Hamming graph from a zero-mean Gaussian with variance s/a and then conditioning on these local correlations being globally consistent with each other (cf., ref. 32) in the sense of being simultaneously realizable by some choice of ϕ.

Given this prior distribution and a sample of size N with realized frequencies given by the vector R, we can then derive the corresponding posterior distribution, which in general, will be non-Gaussian. Specifically, we find that under a multinomial likelihood, the posterior distribution has $p(\phi|\text{data}, a) \propto e^{-S_a[\phi]}$ where the posterior action $S_a[\phi]$ is given by

$$S_a^0[\phi] + N\sum_{i=1}^{\alpha^\ell} R_i\phi_i + N\sum_{i=1}^{\alpha^\ell} e^{-\phi_i}. \quad [5]$$

The MAP estimate is then found by minimizing this action. We show in *SI Appendix* that the MAP estimate approaches the maximum entropy solution in the limit as $a \to \infty$, approaches the empirically observed frequencies in the limit as $a \to 0$, and for all values of a matches the same moments as the maximum entropy solution.

Extending Pairwise and Higher-Order Maximum Entropy Models. So far for concreteness we have concentrated on providing a nonparametric extension to the independent model, which is the maximum entropy model that matches the observed frequencies of the alleles at each position. However, we can also generalize the above approach to provide analogous results for pairwise and higher-order (33) maximum entropy models by making a corresponding change to the prior.

In particular, let us consider the maximum entropy model that matches the moments of our observations up to order $P-1$. As we have seen for $P=2$ (the independent model), the key geometrical feature of the maximum entropy ϕ is that each mutation (i.e., set of parallel edges in the Hamming graph) has a constant effect on ϕ, so that the conditional log odds ratio defined on each face of the Hamming graph is zero. In the preceding section, we constructed a relaxation of this maximum entropy model by considering the average squared conditional log odds ratio, where this average is taken over all faces of the Hamming graph.

To extend this same idea to $P=3$, which is the pairwise interaction model that matches the site-specific allele frequencies and pairwise correlations between sites, we note that in this case the conditional log odds ratio takes the same value for any two faces that are defined by the same pair of mutations, regardless of the identity of the states at the other $\ell-2$ sites. Thus, for a pair of adjacent parallel faces (such as the red and black faces in Fig. 1), the difference between the conditional log odds ratios must be zero. These adjacent parallel faces form 3-faces or cubes, and we can measure the local departure from the implied maximum entropy geometry by the difference in conditional log odds ratios on these pairs of adjacent faces. To get a global summary of this departure, we can average the squared value of this difference over all such subcubes of the Hamming graph and define a family of priors parametrized by the expected value of this mean squared departure from the local geometry of a pairwise maximum entropy model. Similarly, for a maximum entropy model that matches the first $P-1$ moments, our extension is based on defining a prior in terms of the expected mean squared deviation from the local geometry implied by the corresponding maximum entropy assumption, where the average is taken over all P faces of the Hamming graph.

More formally, to define this prior, consider the operator

$$\Delta^{(P)} = \frac{1}{P!}\prod_{k=0}^{P-1}(L - \alpha k I), \quad [6]$$

where I is the identity matrix, and let $s = \binom{\ell}{P}\binom{\alpha}{2}^P \alpha^{\ell-P}$ be the number of P-dimensional faces of the Hamming graph. Then, $\phi^T\Delta^{(P)}\phi/s$ gives the mean squared value of the log conditional P-way association (*SI Appendix*), and so, we define the prior action to be

$$S_a^0[\phi] = \frac{a}{2s}\phi^T\Delta^{(P)}\phi. \quad [7]$$

Under this prior, 1) the MAP estimate always matches the first $P-1$ moments of the observations, 2) the limit as $a \to \infty$ results in the maximum entropy model, 3) the limit as $a \to 0$ results in the empirically observed distribution, 4) the expected mean squared conditional log P-association under the prior is given by $\mathrm{rank}(\Delta^{(P)})/a$, and 5) we can likewise construct this prior by drawing the conditional P-association for each P-face from a zero-mean normal distribution with variance s/a (*SI Appendix* has details).

Maximum Entropy and the Eigendecomposition of *L*. These general P results are largely explicable in terms of the eigendecomposition of the graph Laplacian L of our Hamming graph, whose eigenspaces have a close relationship with maximum entropy models over sequence space. In particular, L has only $\ell+1$ distinct eigenvalues, which are of the form $\lambda_k = \alpha\, k$ for $k=0$ to ℓ, and the eigenspace associated with λ_k has dimension $\binom{\ell}{k}(\alpha-1)^k$. In fact, finding a maximum entropy ϕ that matches the first $P-1$ moments of a probability distribution on the Hamming graph corresponds exactly to finding the ϕ satisfying these moment conditions within the linear subspace generated by the eigenvectors associated with λ_0 through λ_{P-1}. Moreover, this same linear subspace is also the null space of the operator $\Delta^{(P)}$, consisting of the ϕ for which $\phi^T\Delta^{(P)}\phi = 0$. Thus, in sequence space, finding the ϕ that satisfies the maximum entropy assumption not only implies a specific local geometry of ϕ, but this local geometry $\phi^T\Delta^{(P)}\phi = 0$ is actually equivalent to the maximum entropy assumption itself.

More specifically, a key result in the general theory of maximum entropy distributions is that for a set of constraints of the form $\mathbb{E}_Q f^{(1)} = c^{(1)}, \mathbb{E}_Q f^{(2)} = c^{(2)}, \ldots$ (where $\mathbb{E}_Q f$ is the expected value of the function f under draws from the distribution Q), any distribution that satisfies these constraints specified by a ϕ of the form $\phi = \sum_i \theta^{(i)} f^{(i)}$ for these same functions $f^{(i)}$ and some set of coefficients $\theta^{(i)}$ is the unique maximum entropy distribution (e.g., theorem 12.1.1 in ref. 6) (*SI Appendix*). In the case of maximum entropy distributions over sequence space, the constraints are typically that the inferred distribution matches the first $P-1$ moments of the empirical distribution or equivalently, the $(P-1)$th-order marginals. These marginals can be set as constraints by letting the $f^{(i)}$ be indicator functions for matching a specific set of states at a specific set of $P-1$ positions and letting $c^{(i)}$ be the corresponding marginal. Importantly, these indicator functions span a subspace identical to the subspace spanned by the eigenvectors associated with λ_0 through λ_{P-1} of the graph Laplacian (*SI Appendix*). Now, from Eq. **6**,

we observe that the operator $\Delta^{(P)}$ is a polynomial in L, and one can show (*SI Appendix*) that each eigenvalue–eigenvector pair of L with $\lambda_k = \alpha\, k$ is transformed to an eigenvalue–eigenvector pair of $\Delta^{(P)}$ with eigenvalue $\alpha^P \binom{k}{P}$ and an identical eigenvector. Since $\binom{k}{P}$ is zero for $k < P$ and positive for $k \geq P$, we see that the null space of $\Delta^{(P)}$ is simply the subspace of maximum entropy ϕ. Thus, finding a ϕ that satisfies the moment constraints and maximizes the entropy is the same as finding a ϕ that satisfies these constraints and respects the equivalent local geometry $\phi^T \Delta^{(P)} \phi = 0$.

The above analysis also clarifies several other aspects of the prior. For example, $\mathrm{rank}(\Delta^{(P)})$ is just α^ℓ minus the dimensionality of the subspace of maximum entropy models, $\sum_{k=0}^{P-1} (\alpha-1)^k \binom{\ell}{k}$, and $a\,\Delta^{(P)}/s$ is the precision matrix of an improper Gaussian distribution with infinite variance in the directions corresponding to the ϕ that form valid maximum entropy models based on the first $P-1$ moments. Intuitively, this flat prior for the subspace of maximum entropy models explains why the SeqDEFT MAP estimate exactly matches the first $P-1$ moments of the empirical distribution. However, our prior also specifies a (proper) Gaussian prior on the orthogonal complement of this maximum entropy subspace, and the magnitude of this second component is controlled by setting the expected value of $\phi^T \Delta^{(P)} \phi / s$, which is the mean squared deviation from the local geometry that defines the corresponding maximum entropy model. Moreover, the value of this non-maximum entropy component is conditionally independent for sequences i and j with Hamming distance greater than P since the precision matrix $a\Delta^{(P)}/s$ has nonzero entries only for pairs of sequences with Hamming distance less than or equal to P (*SI Appendix*). Thus, intuitively, the family of priors we propose is the family where the direct dependencies in the complement of the maximum entropy subspace are restricted to being as local as possible.

Practical Implementation. The posterior action $S_a[\phi]$ given by Eq. **5** is a nonlinear function of α^ℓ variables, and there is no explicit formula for the vector ϕ that achieves the minimum. Nonetheless, following ref. 27, we show that the minimization problem is convex (*SI Appendix*), and at a practical level, we can calculate the minimum numerically for sequence spaces containing up to low millions of sequences by exploiting sparsity. In particular, due to the product form for $\Delta^{(P)}$ given in Eq. **6**, most of the calculation can be implemented via repeated matrix multiplication by the sparse matrix L.

Another issue is how to choose the value of the hyperparameter a. Here, we choose this value by maximizing the k-fold cross-validated log likelihood with $k = 5$ and refer to the resulting optimal value as a^*. In *SI Appendix*, we describe an alternative approach for finding a^* (for small sequence spaces) using the evidence ratio.

Finally, while our method is defined in terms of a specific Gaussian prior, the posterior is non-Gaussian, and we implement posterior sampling using Hamiltonian Monte Carlo (34); *SI Appendix* has details.

Distribution of Human 5′ Splice Sites

In most eukaryotes, the sequence of the final processed form of a messenger RNA (mRNA) transcript is not encoded contiguously in the genome but rather, appears as several discrete segments known as exons that are separated by other DNA segments known as introns. During transcription, both the intronic and exonic DNA are transcribed into RNA, after which the intronic RNA is removed in a process known as pre-mRNA splicing (35). To demonstrate the characteristics of our SeqDEFT method, we first considered 5′ splice sites, the RNA sequences at the boundary between each intron and its upstream exon (36). Because there are hundreds of thousands of such sites in the human genome, 5′ splice sites provide a relatively well-sampled model system for understanding the complexity and geometry of high-dimensional distributions in sequence space, as well as an opportunity to subsample real data in order to investigate the performance of our method when less data are available. Moreover, the modeling and identification of splice sites was one of the early successes of maximum entropy models with pairwise interactions, and maximum entropy remains a common method for scoring splice sites (9). On the other hand, the vast majority of sequences that are annotated to have splicing activity are annotated only a small number of times across the genome (e.g., 89.3% of unique 5′ splicing sequences are annotated 20 or fewer times), so that formal density estimation techniques are still needed to accurately determine the frequencies of most sequences with observed splicing activity.

For each annotated intron in the human genome, we extracted the last three positions of the upstream exon (which are also generally under selection for their amino acid coding activity and typically labeled as -3, -2, -1) and the first six positions of intron itself (labeled $+1$ through $+6$). In total, this resulted in a collection of 305,106 nine-nucleotide sequences, which we modeled using SeqDEFT with $P = 2$, so that our model is a nonparametric extension of the independent sites model. To understand the qualitative behavior of SeqDEFT on datasets of different sizes, we further performed a rarefaction analysis, where we trained models on 25%, 5%, or 1% of the data. In Fig. 2, *Top*, we see that at very low sampling, the SeqDEFT MAP estimate Q^* behaves very similarly to the independent sites maximum entropy model but becomes substantially different as the amount of data increases. Fig. 2, *Middle* compares the SeqDEFT estimate from the subsampled data with the estimate using the full dataset, and we see that the SeqDEFT distribution has taken a relatively similar form to its fit on the full dataset by the time we have given it 25% of the data. Fig. 2, *Bottom* shows the predicted frequency vs. the observed frequency (*SI Appendix*, Fig. S1*A* shows credible intervals based on posterior sampling, and *SI Appendix*, Fig. S2*A* shows hyperparameter tuning). We see that the MAP estimate under SeqDEFT closely approximates the empirical frequency for states with greater than on the order of 10 observations, and thus, the smoothness included by the prior essentially only influences our predictions in the more sparsely sampled regions of sequence space (as shown by deviations from the line $y = x$). With increasing data, the breakpoint between these two regimes moves to proportionally lower-frequency states.

To gain some intuition for why SeqDEFT behaves differently in well-sampled vs. poorly sampled regions of sequence space, it is helpful to consider the form of the posterior action. The first term $\frac{a}{2s}\phi^T \Delta^{(2)} \phi$ is the prior action and favors conditional log odds ratios that are as close to zero as possible. The second and third terms, $N \sum_{i=1}^{\alpha^\ell} R_i \phi_i + N \sum_{i=1}^{\alpha^\ell} e^{-\phi_i}$, measure the match between ϕ_i and the observed data and are minimized by setting ϕ_i equal to the negative logarithm of the observed frequency R_i. However, the value of these last two terms is relatively insensitive to the value of ϕ_i when the number of observations NR_i is zero (e.g., if $NR_i = 0$, then the corresponding ϕ_i does not contribute to the second term and contributes minimally to the third term as long as $Q_i \propto e^{-\phi_i}$ is small). Thus, the prior dominates in regions of sequence space where the number of observations is small, producing an MAP estimate with small conditional associations in these regions, whereas well-sampled sequences are predicted to have frequencies similar to those that are empirically observed.

Another useful comparison is with the pairwise maximum entropy fit (Fig. 3, *Left*). We see that SeqDEFT and the

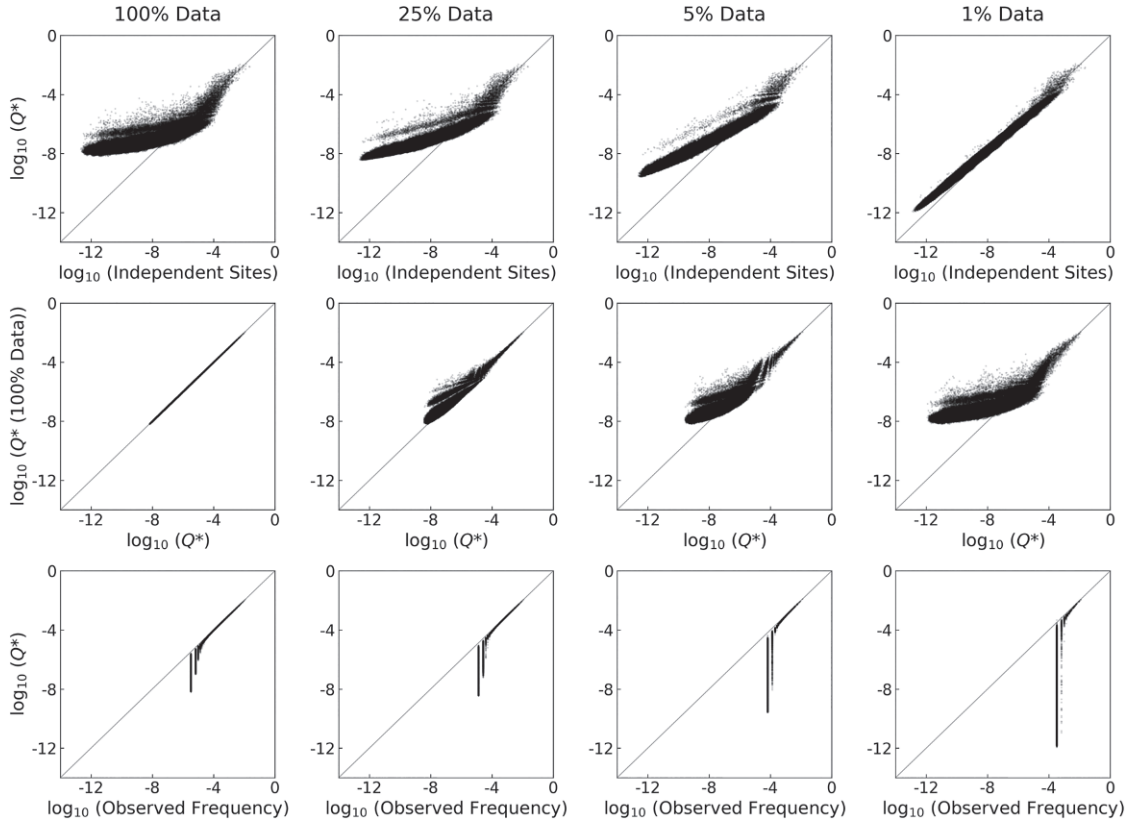

Fig. 2. Behavior of SeqDEFT with changing sample size. *Top* plots the SeqDEFT MAP estimate vs. the independent sites maximum entropy model. *Middle* shows the SeqDEFT estimate using the full dataset vs. the estimates for the smaller samples. *Bottom* shows the SeqDEFT estimate vs. the sampled frequencies. In order to show the class of sequences that were not observed in the sample, we added a pseudocount of one to the sequence counts. Thus, the vertical striations indicate sequences that were observed zero, one, two times, etc. Column headings indicate the proportion of annotated 5′ splice sites from the human genome used while fitting the models.

pairwise maximum entropy model are in closer agreement for high-frequency sequences than SeqDEFT and the independent model. However, for low- and intermediate-frequency sequences, SeqDEFT still produces a much narrower range of estimates.

Why do the independent sites and pairwise maximum entropy models produce such a wide range of estimates in low-frequency regions of sequence space? The key observation is that for these maximum entropy models there is an assumption that some feature of the data remains constant over all of sequence space (e.g., for the independent model, each possible mutation has a constant multiplicative effect on frequency, and for the pairwise model, the conditional log odds ratio between any given pair of mutations is constant over all of sequence space). Moreover, these constants are determined by the moments of the sample, which are primarily influenced by the high-frequency sequences. Thus, if for example a pair of mutations is associated among the high-frequency sequences, the pairwise maximum entropy model will assume that they remain associated even in regions of sequence space where we have no data to support this association. In contrast, SeqDEFT allows these associations to decay in regions of sequence space where there are no data to support them. Such behavior is also in better accordance with biological intuition in that, for example, correlations between positions in functional sequences are due to natural selection on that function, and thus, these correlations should not be observed in regions of sequence space that consist of nonfunctional sequences.

Another important difference between the maximum entropy models and SeqDEFT concerns SeqDEFT's ability to learn components of the probability distribution that are at lower frequencies (e.g., in treating a multimodal distribution, the maximum entropy solution will tend to fit the largest of the modes while providing a poor fit for other modes that might have strong statistical support but contain a small absolute fraction of the total probability). The 5′ splice sites provide a good illustration of this principle. The vast majority of splice sites have a G in the +1 position and a U or C in the +2 position, but a small fraction (1.68% in our dataset) has other nucleotides (37, 38), and A in the +1 position in particular can be recognized by a different splicing machinery known as the minor spliceosome (39). Fig. 3, *Center* shows SeqDEFT's fit to these atypical non +1 G or non +2 U/C sequences, while Fig. 3, *Right* shows the fit of the pairwise maximum entropy model to these same sequences. We see that SeqDEFT is able to learn the density for these atypical sequences, whereas the pairwise maximum entropy model produces a qualitatively incorrect fit. In fact, both the independent model and the pairwise maximum entropy model show substantial deviations between the observed and learned sequence frequencies for many relatively high-frequency sequences, far beyond what can be accounted for by the binomial variability inherent in counts-based frequency estimation, whereas the MAP SeqDEFT estimates and posterior samples essentially match both the observed counts and the expected binomial variability (*SI Appendix*, Fig. S3).

For pairwise maximum entropy models, in order to identify positions that may be interacting, it is common to construct heat maps showing the magnitude of the inferred coupling parameters between any given pair of positions, which are meant to

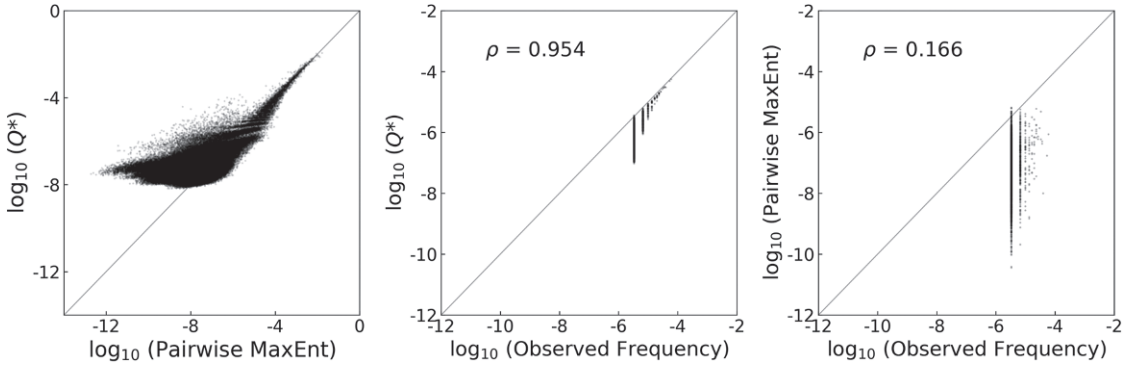

Fig. 3. Comparison of SeqDEFT and pairwise maximum entropy models fit to the distribution of human 5′ splice sites. *Left* shows the SeqDEFT MAP estimate vs. the pairwise maximum entropy fit. *Center* shows the SeqDEFT MAP estimate vs. the empirically observed frequency for the subset of 5′ splice sites that do not have the canonical +1 G and +2 U/C nucleotides. *Right* shows the pairwise maximum entropy fit vs. the observed frequency for the subset of 5′ splice sites that do not have the canonical +1 G and +2 U/C nucleotides. ρ denotes the Pearson correlation.

quantify the direct influence of one position in a sequence on another. The local interaction coefficients ϵ_{ijkm} in our SeqDEFT framework play a similar role to these couplings and quantify the direct influence of one mutation on another by determining their interaction on a fixed genetic background (in fact, in the maximum entropy limit, the ϵ_{ijkm} can be easily derived from the couplings via Eq. **2**, and the couplings can be recovered from the ϵ_{ijkm} by simple averaging, see *SI Appendix*). We can thus construct similar plots by displaying an appropriate summary statistic for the distribution of local interaction strengths, such as the root mean square (RMS) conditional log odds ratio for specific pairs of mutations (i.e., pick a pair of mutations at a pair of positions, average the squared log odds ratio for that pair of mutations over all possible choices for the other $\ell - 2$ positions, and then, take the square root of this value). Fig. 4, *Left* shows a heat map of this type, and it is easy to see that the strongest interactions are between mutations altering the consensus +1 G and mutations altering the consensus G at the –1 position or the consensus U at +2 position, but it is also clear that there are interactions between many other pairs of mutations.

However, unlike the pairwise maximum entropy model, which must produce a single conditional log odds ratio for any given pair of mutations, SeqDEFT allows the strength and direction of the association between each pair of mutations to vary with the genetic background. Thus, rather than look at a single summary statistic of association strength between pairs of mutations, we can also take a more detailed view by plotting histograms of the log odds ratios inferred for different genetic backgrounds (i.e., how the log odds ratio varies over the different faces of the Hamming graph). Importantly, while 12.6% of the faces of the Hamming graph contain at least one observed splice site, only 0.5% of the faces are composed entirely of observed splice sites, so that some inference of the underlying density is essential to examine this variability in the strength of local associations.

Fig. 4, *Upper Right* shows these distributions for mutations to the consensus nucleotides at the +1 and +2 positions. We see that these histograms are strongly right skewed, a pattern that likely arises because mutations to either of these nucleotides typically render a functional splice site nonfunctional (producing a strong positive association in the conditional distribution) but have little effect on an already nonfunctional splice site. The solid vertical lines in the plots indicate the mean conditional log odds ratio (averaged over all faces in the Hamming graph), while the dashed vertical lines indicate the constant log odds ratio for all such faces assigned by the pairwise maximum entropy model. Fig. 4, *Lower Right* shows a similar set of distributions for interactions between the –1 and –2 positions. A careful examination of these histograms shows that they are typically trimodal, with a large central mode and two smaller side modes, indicating that a subset of sites shows a substantial interaction between these two mutations but that the sign of this association differs depending on the genetic background. Thus, our SeqDEFT estimate suggests that the sign and strength of the association between a pair of mutations can vary in a complex manner depending on the genetic background, an observation that is qualitatively incompatible with the assumptions of a pairwise maximum entropy model.

Visualizing the Inferred Geometry Using an Evolutionary Model. A key strength of nonparametric approaches such as SeqDEFT is that, provided sufficient data, they can capture whatever complex geometry is present in the data. However, this comes at the expense of interpretability because we can no longer express the inferred distribution in terms of a small number of parameters. We have already explored one way of overcoming this difficulty in the form of the summary statistics and histograms shown in Fig. 4. A different solution is to attempt to visualize or represent the inferred distribution in such a way that the visualizations allow us to identify the major qualitative features of the distribution and explore the underlying causes of these features.

The visualization approach we take here (30) is based on considering our inferred probability distributions over sequence space as being the result of an evolutionary process (*SI Appendix* has details). The main idea is that biological evolution can be viewed as the process of a population taking a random walk over sequence space, where each step in the random walk consists of the replacement of one sequence in the population by another, and the role of natural selection is to bias the probability that any given mutational neighbor of the current sequence becomes fixed (40, 41). With this idea in mind, given an inferred distribution over sequence space we can write down a model of molecular evolution as a reversible Markov chain that takes the inferred distribution as its stationary distribution, and then use the subdominant eigenvectors of the rate matrix of the Markov chain to construct an embedding of the graph where clusters of vertices correspond to sets of initial states from which the Markov chain approaches stationarity in a similar manner; we refer to these axes as diffusion axes (42) because they capture the slow modes of the diffusion of the probability distribution describing the location of the population in sequence space. Importantly, inference of the underlying density over sequence space is

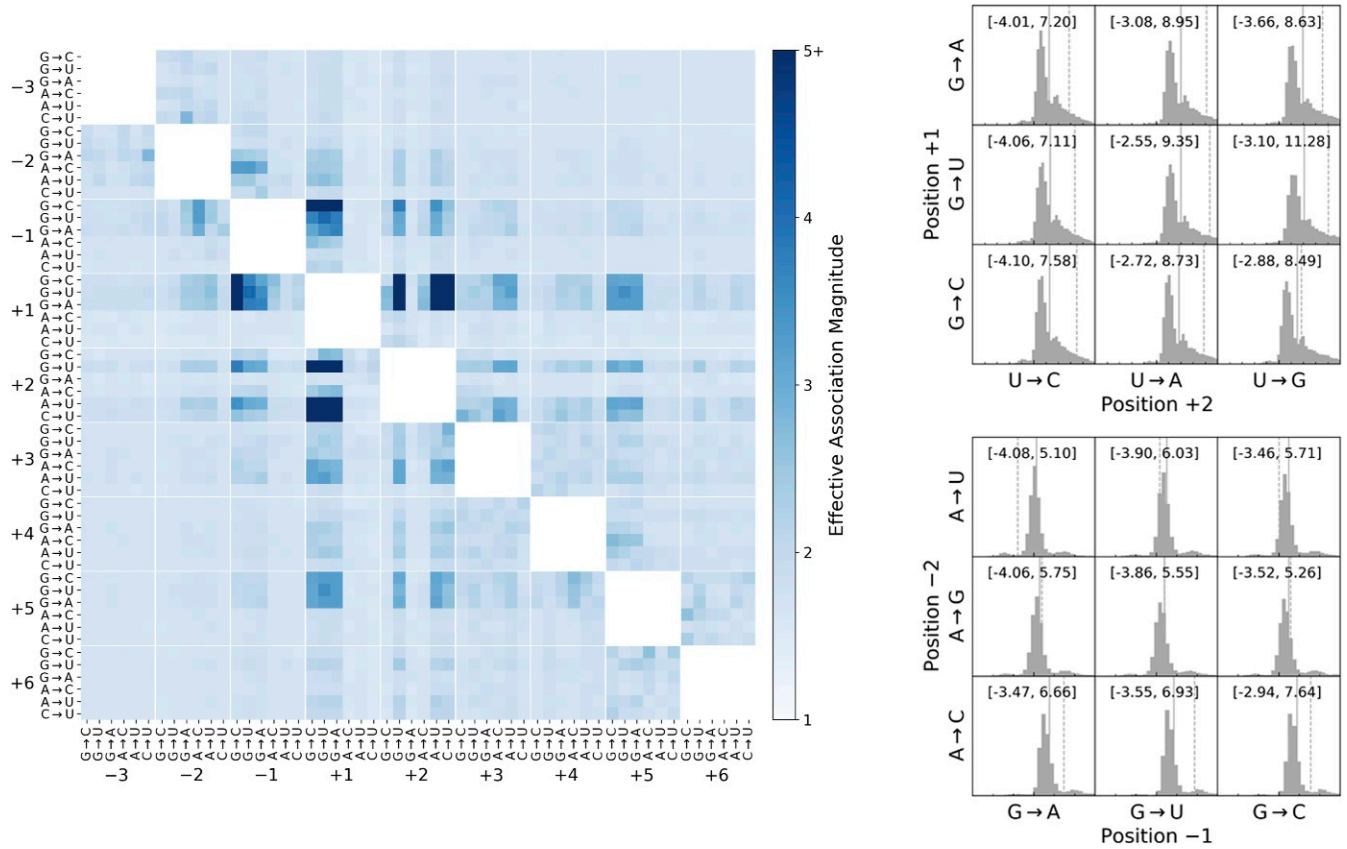

Fig. 4. Interactions between pairs of mutations in the SeqDEFT fit. *Left* shows the exponentiated RMS conditional log odds ratio between each possible pair of mutations under the MAP SeqDEFT fit. The RMS conditional log odds ratio is calculated by averaging the squared conditional log odds ratio over the faces of the Hamming graph corresponding to a given pair of mutations and is exponentiated to express the effective strength of the association between any given pair of mutations on the same scale as the odds ratio. To provide a more fine-grained view of the variation in the strength of these associations between different faces of the Hamming graph, *Right* shows histograms of the distribution of log odds ratios for mutations away from the consensus bases at the +1 and +2 positions (*Upper Right*) and at the –2 and –1 positions (*Lower Right*). The major tick on the *x* axis of each histogram indicates a log odds ratio of zero, and the other ticks are placed at intervals of 1 unit on the log odds scale. To aid comparison, the histograms are shown with a limited range of conditional log odds ratios on the *x* axis, but the full range of inferred conditional log odds ratios is shown by the bracketed values at the top of each panel. Mutations are polarized from the more preferred to the less preferred state, so that positive associations typically indicate that either of the single mutants results in loss of function, while a negative association indicates that the two mutations are tolerable individually but not jointly. The solid vertical lines indicate the mean of the log odds ratios, while the dashed vertical lines indicate the uniform conditional log odds ratio assigned to that pair of mutations under the pairwise maximum entropy model.

essential for implementing an approach of this type due to the sparsity of our observations; for instance, the set of observed splicing sequences contains over 1,000 disconnected components, and so, we need a method of predicting the frequencies of unobserved sequences in order to be able to define a model of molecular evolution over the full range of biologically active sequences.

To briefly describe the visualization method at a more technical level, we first note that in certain standard models of molecular evolution, the quantity $\log Q_i^*$ is equal to the product of the fitness of state i and the effective population size, so that population genetic theory predicts that if each possible point mutation occurs at rate 1, a population currently fixed for sequence i becomes fixed for sequence j at rate

$$T_{ij} = \frac{\log Q_j^* - \log Q_i^*}{1 - e^{-(\log Q_j^* - \log Q_i^*)}}$$

for mutationally adjacent sequences i and j (40, 43); we set the leaving rate from sequence i, T_{ii}, so that the row sums of the matrix T are all zero. Since the Markov chain generated by the rate matrix T satisfies detailed balance, we can construct the eigendecomposition $T = -\sum_{k=1}^{\alpha^\ell} \lambda_k \, r^{(k)} (l^{(k)})^{\mathrm{T}}$ where $l^{(k)}$ and $r^{(k)}$ are the left and right eigenvectors, respectively, of T associated with the eigenvalue $-\lambda_k$; the $l^{(k)}$ and $r^{(k)}$ are normalized such that $(l^{(k)})^{\mathrm{T}} r^{(k)} = \sum_i (r_i^{(k)})^2 / Q_i^* = 1$; and we order the eigenvalues so that $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \ldots \leq \lambda_{\alpha^\ell}$. For a d-dimensional visualization of our Markov chain, we can then plot each sequence i with coordinates $\sqrt{1/\lambda_2}\, l_i^{(2)}, \ldots, \sqrt{1/\lambda_{d+1}}\, l_i^{(d+1)}$. Thus, the coordinates for sequences along any axis are just the entries of an eigenvector describing one of the slow modes of the system, where these eigenvectors have been rescaled by the square root of the corresponding relaxation time $1/\lambda_k$ (this rescaling results in a connection between distances in the visualization and the expected waiting time to evolve from sequence i to j, see *SI Appendix* for details). In the case at hand, we are considering a population evolving under selection to have a functional splice site at a particular location in the genome under the assumption that the stationary distribution for this process is given by the inferred distribution of 5′ splice site sequences given by SeqDEFT. Fig. 5 shows the resulting visualization, where we have fixed the +1 and +2 positions to be the canonical GU

nucleotides in order to best display the geometrical features of the subset of functional sequences (*SI Appendix*, Fig. S4 shows the corresponding visualizations over all of sequence space).

Fig. 5 shows the first three diffusion axes. Sequences are colored according to their inferred frequency, and edges connect sequences that differ by a single point mutation. We see that there is a connected cluster of sequences similar to the canonical binding motif CAG/GUAAGU that is stretched along diffusion axis 1 and then, two clusters of moderate frequency with extreme values on diffusion axis 2 or diffusion axis 3. To understand the structure of the main streak, it is helpful to know that the canonical 5′ splice site is recognized and bound by a small nuclear RNA (snRNA) known as the U1 snRNA with which it forms a series of adjacent base pairs resulting in a helical structure (36). Moreover, it has been previously observed that 5′ splice sites show a pattern of 5′/3′ compensation or "see-saw" linkage where mismatches in the exonic portion of the binding site are associated with consensus nucleotides in the intronic portion and vice versa (44–46). The long axis of the streak turns out to correspond to this pattern in the locations of mismatches between the 5′ splice site and U1 snRNA, with sequences that primarily form base pairs in the exonic portion of the splice site having negative values on diffusion axis 1 and sequences that bind via base pairs in the intronic portion having positive values (*SI Appendix*, Fig. S5 shows the average position of consensus bases for sequences in the streak as a function of position along diffusion axis 1). These two sides of the streak are broadly separated because it would take many mutations to transform a sequence forming primarily exonic base pairs into a sequence forming primarily intronic base pairs with mismatches in the exonic portion of the binding site.

Other than the main streak of sequences that appear to bind via minor variations on the canonical 5′ splice site motif, there are two smaller clusters of high-frequency sequences plotted far from the main streak. The cluster with a negative value on diffusion axis 2 corresponds to sequences that are recognized by the minor spliceosome (39). This machinery recognizes the 5′ splice site via binding with the U11 rather than U1 snRNA, and this recognition occurs in conjunction with another protein known as 48K, where the 48K protein contacts the +1 and +2 positions and binding with U11 is completely intronic beginning at the +3 position (47). The other small cluster of high-frequency sequences turns out to correspond to a previously characterized noncanonical binding mode known as the shifted +1 register (48) where U1 snRNA binding forms a gapped helical structure that is shifted one base pair over from the canonical motif but where the position of the splicing reaction (transesterification) itself remains unchanged.

Now that we have identified the major geometric features of the inferred probability distribution and identified each of them as corresponding to a distinct biophysical mechanism of splice site recognition, we can also use these visualizations to ask questions about the evolution of binding by considering different paths that populations can take through sequence space. For example, it has been previously observed in genomic comparisons between different species that individual splice sites can be "converted" from being recognized by the minor spliceosome to being recognized by the major spliceosome and vice versa (49, 50). In our visualizations, we see that the sequences recognized by the minor spliceosome have large negative values on diffusion axis 1 and that for sequences similar to the canonical motif, diffusion axis 1 separates sequences where U1 binds primarily to the exonic portion of the splice site (negative values on diffusion axis 1) from those where it binds primarily in the intronic portion (positive values on diffusion axis 1). This suggests that conversion is most likely to occur via a transition from minor spliceosome recognition to a sequence capable of being recognized by U1 via the gain or loss of a CAG motif in positions –3 to –1 (i.e., paths leading up or down the left side of Fig. 5).

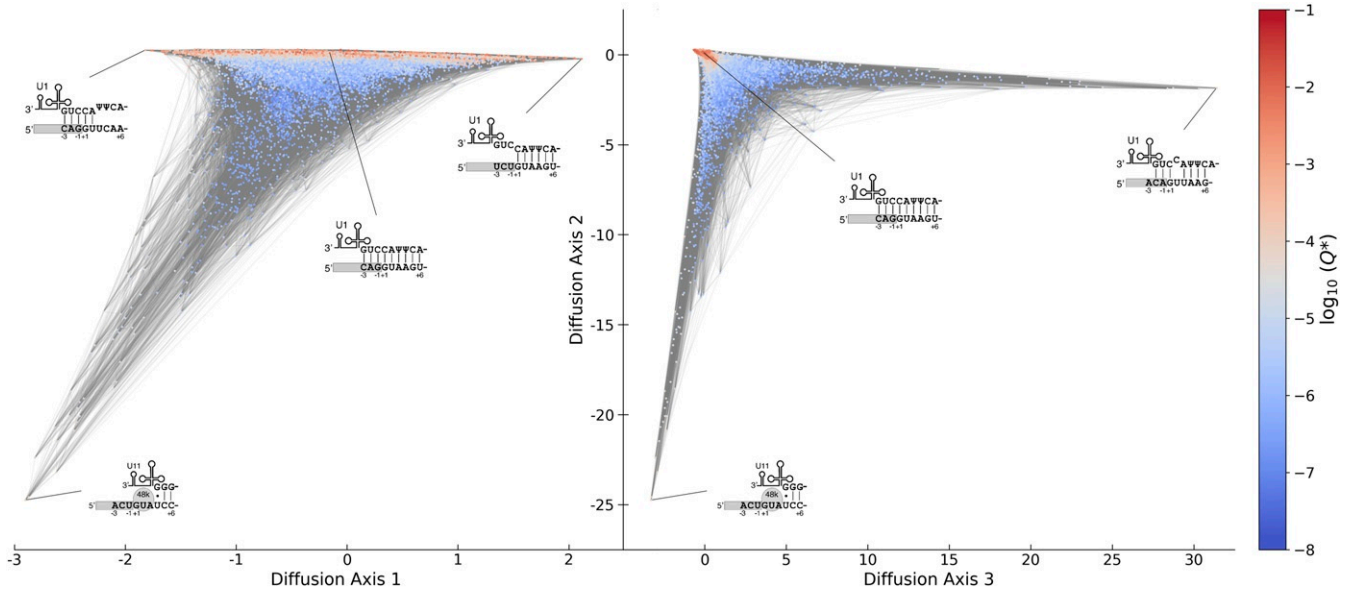

Fig. 5. Visualization of the distribution of 5′ splice sites inferred by SeqDEFT. Visualization uses the method of ref. 30, where $\log Q_i^*$ is equated with the scaled fitness of sequence *i*, and this quantity is used to define an evolutionary Markov chain whose stationary distribution equals the MAP SeqDEFT estimate (conditioned on the +1 and +2 positions being the canonical GU nucleotides). Sequences are colored according to their estimated frequency, and edges connect sequences that are adjacent under point mutations. Under the visualization method, squared Euclidean distances optimally approximate the expected time to evolve from one sequence to another, and we scale time so that each possible point mutation occurs at rate 1. The figure shows the first three diffusion axes, which are rescaled subdominant eigenvectors of the transition matrix for the Markov chain. The cartoons show hypothesized binding mechanisms. The analogous visualization based on the pairwise maximum entropy density estimate does not show any obvious structure (*SI Appendix*, Fig. S6).

Distribution of Karyotypic Abnormalities in Cancer

Our exploration of the distribution of human 5′ splice sites both demonstrated SeqDEFT's behavior in the well-sampled regime and highlighted the rich geometry of biological distributions that can be captured via a nonparametric approach. We now turn to an example in the poorly sampled regime, where the number of sequences far exceeds the number of observations. In particular, we consider the problem of karyotypic abnormalities in human cancers, where cancerous cells frequently exhibit a bewildering array of changes to the structure and number of chromosomes ranging from losses and duplications of small portions of individual chromosomes to duplications and losses of chromosome arms, translocations that attach a portion of a chromosome to another, duplications and losses of whole chromosomes, and (multiple) whole-genome duplications (51). Moreover, the root causes of these changes in genomic structure remain poorly understood because chromosomal copy number changes can either promote or inhibit cellular proliferation depending on the specific alterations involved (52–54).

To better understand the distribution of karyotypic states exhibited by human cancers, we considered karyotypes inferred for 10,522 tumors (29) collected as part of The Cancer Genome Atlas (55). Starting with the simplest possible approach, we considered each of the 22 autosomes and scored each autosome as being either euploid or aneuploid, where we scored a chromosome as aneuploid if ref. 29 reported that chromosome as exhibiting large-scale alterations from the background cellular ploidy. Under this scoring scheme, we observe 7,443 distinct karyotypic states in our dataset, which is only a tiny fraction (0.18%) of the $2^{22} = 4{,}194{,}304$ possible states whose frequencies we seek to estimate. Moreover, even among the sampled sequences, our data are highly sparse in that the vast majority (91.9%) of observed karyotypic states were observed only once.

In fitting these data with SeqDEFT, we observed that the pairwise maximum entropy model had a greater cross-validated log likelihood than our model using $\Delta^{(2)}$ (i.e., the pairwise maximum entropy model provided a higher likelihood fit than a model based on a perturbation of the independent model) (*SI Appendix*, Fig. S2 *C* and *D*). This was likely caused by a particularly strong global pattern of nonindependence between chromosomal states wherein observed karyotypes were roughly uniformly (rather than binomially) distributed in terms of their number of aneuploid chromosomes (*SI Appendix*, Fig. S7*A*), indicating a strong enrichment for chromosomal configurations that are either perturbations of the standard euploid genome with a handful of altered chromosomes or else nearly completely aneuploid (*SI Appendix*, Fig. S7*B*). This observation is consistent with the well-known phenomenon of chromosomal instability, where deviations from a euploid karyotype result in further mitotic errors and hence an increasingly high degree of aneuploidy (56). We therefore proceeded with an analysis based on $\Delta^{(3)}$, which relaxes the constraints of the pairwise model. Importantly, the pairwise maximum entropy model treated all chromosomes in an approximately uniform manner, where the marginal frequencies of aneuploidy for each chromosome have a mean of 0.41 and SD of 0.06 for each chromosome and the conditional log odds ratios have a mean of 0.23 and an SD of 0.17 so that the presence of any one chromosomal alteration increases the probability of each of the others by an approximately constant factor. Thus, while the maximum entropy model does a good job capturing the overall bimodality of the data, it does not appear to be capturing more detailed interactions between specific pairs or subsets of chromosomes.

Using $\Delta^{(3)}$, we find that $a^* = 7.9 \times 10^5$, resulting in an increase of 422.4 in the cross-validated log likelihood relative to the pairwise maximum entropy model (*SI Appendix*, Fig. S1*B* shows posterior variance estimates, and *SI Appendix*, Fig. S2*D* shows hyperparameter tuning). Although the SeqDEFT predictions are mostly similar to the pairwise maximum entropy model, there are also relatively dramatic differences for a subset of sequences that SeqDEFT predicts to be at much higher frequencies than the pairwise maximum entropy model (*SI Appendix*, Fig. S7*C*).

To better understand what these high-frequency sequences are, we turn to our visualization technique (30). Here, our application of the visualization technique is somewhat less principled since a cancer's exploration of sequence space is not stationary but rather ends in either the patient's death or eradication of the tumor, and there are a number of other differences such as, for example, gains or losses of multiple chromosomes at once (so that evolution is not restricted to the edges of the Hamming graph) and polyclonality within the tumor (so that the tumor contains cells with a number of different karyotypic states rather than a single state for each time). Nonetheless, the visualizations do provide insight into the basic geometry of the inferred probability distribution and hence indirectly into the process that generated it.

Fig. 6 shows the resulting visualization. Here, diffusion axis 1 picks out the number of chromosomes that are altered, with the wild-type euploid karyotype having a large negative value on this axis and the karyotype where all chromosomes are aneuploid having a large positive value (faint striations are also visible, which correspond to the Hamming distance from the wild-type sequence). However, diffusion axis 2 reveals two sets of unusually high-frequency sequences found in a region of sequence space where most other sequences have much lower frequency. In particular, the tip sequences for these protrusions are karyotypes with simultaneous copy number changes at chromosomes 1, 2, 6, 10, 13, 17, and 21 or at chromosomes 6, 7, 9, 10, 19, and 20. Diffusion axis 3 then reveals the geometric relationship between these high-frequency sets, showing that they are two distinct regions in sequence space that both branch off the main arc that connects the wild type to the all-aneuploid state. Fig. 6, *Right* also shows a protrusion of high-frequency sequences around the state with chromosomes 2, 3, 7, 12, 16, 17, 20, and 21 simultaneously altered, which turns out to appear as a third branch-like protrusion when we include diffusion axes 4 and 5 in our visualizations (*SI Appendix*, Fig. S8).

What do these visualizations tell us about the geometry of our inferred probability distribution? Our evolutionary Markov chain treats log frequency as a measure of evolutionary fitness. Thus, it treats the wild-type and all-aneuploid sequences as two major fitness peaks, separated by the broad valley of partially aneuploid sequences, so that populations typically stay at one of these two peaks or the other but occasionally stochastically transition from one to the other. However, within this valley, we have observed three clusters of high-frequency sequences, with local frequency (and hence, fitness) maxima at states with chromosomes {1, 2, 6, 10, 13, 17, 21}, {6, 7, 9, 10, 19, 20}, or {2, 3, 7, 12, 16, 17, 20, 21} altered. Populations that wander into the basin of attraction of these local fitness maxima can become stuck there, leading to long waiting times for these populations to visit the other maxima and hence, large distances between these maxima in our visualizations. Importantly, using the pairwise maximum density estimate for the visualization does not reveal any of this fine-scale structure (*SI Appendix*, Fig. S9), which demonstrates the power of our nonparametric approaches to capture the qualitative features of this complex dataset even when the size of sequence space is orders of magnitude larger than the number of observed sequences.

To better understand the biological basis of these local maxima in our inferred probability distribution, we also considered which specific tissue types and specific modes of chromosomal

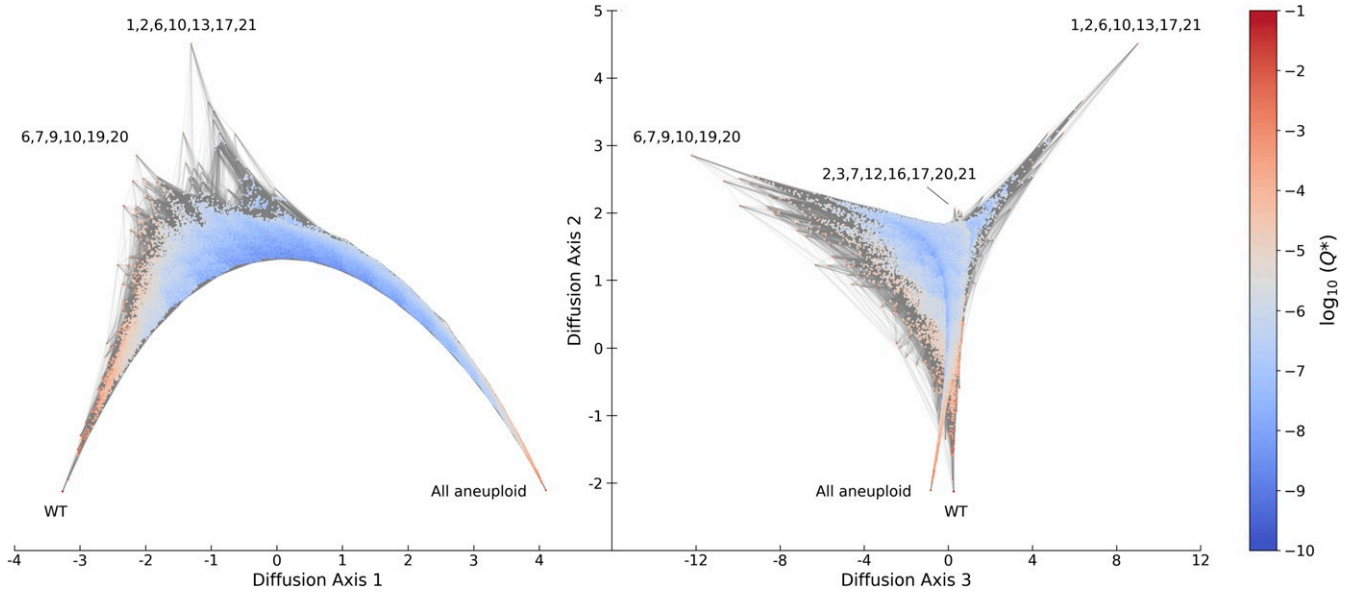

Fig. 6. Visualization of the distribution of human karyotypes based on data from ref. 29. Visualization uses the method of ref. 30, where $\log Q_i^*$ is equated with the scaled fitness of karyotype *i*, and this quantity is used to define an evolutionary Markov chain whose stationary distribution corresponds to the MAP SeqDEFT estimate. Sequences are colored according to their estimated frequency, and edges connect sequences that can be transformed one into the other by changing the state of a single chromosome. The figure shows the first three diffusion axes for a binary encoding of the karyotypic state, where each of the 22 autosomes is scored as being euploid or aneuploid; the pattern of aneuploid chromosomes is indicated for key local maxima in the estimated probability distribution. *SI Appendix*, Fig. S8 shows diffusion axes 4 and 5. WT, wild type.

alteration (e.g., chromosomal gain or loss) contributed to these maxima. We found that the regions around {1, 2, 6, 10, 13, 17, 21} and {2, 3, 7, 12, 16, 17, 20, 21} were composed of cancers that originated in kidney tissue and that the region around {6, 7, 9, 10, 19, 20} was based in observations of brain cancers, with other tissue types making little or no contribution to these local maxima (*SI Appendix*, Fig. S10). Furthermore, we found that the simultaneous chromosomal alterations at chromosomes {1, 2, 6, 10, 13, 17, 21} or {2, 3, 7, 12, 16, 17, 20, 21} were in fact largely attributable to different kidney cancers, with the heightened frequency around {1, 2, 6, 10, 13, 17, 21} primarily due to chromophobe renal cell carcinoma and the heightened frequency around {2, 3, 7, 12, 16, 17, 20, 21} primarily due to kidney renal papillary cell carcinoma; kidney clear cell carcinoma made some contributions to both regions (*SI Appendix*, Fig. S11). Indeed, for {1, 2, 6, 10, 13, 17, 21}, the specific attracting pattern appears to be losses of all seven of these chromosomes, a pattern that has previously been recognized as a signature of chromophobe renal cell carcinoma (57), whereas for {2, 3, 7, 12, 16, 17, 20, 21}, the observed pattern is simultaneous copy number gains for all these chromosomes, a pattern that has also been noted in the literature as being a signature of renal papillary cell carcinoma (58). For the region around {6, 7, 9, 10, 19, 20}, the pattern is more complicated with the main attracting state being gains of chromosomes 7, 19, and 20 together with loss of chromosome 10, where simultaneous gains of chromosomes 7 and 19 together with loss of chromosome 10 are widely known to be common in glioblastomas (59) and coamplification of chromosomes 19 and 20 has been identified as a marker of positive prognosis (60). This pattern at chromosomes 7, 10, 19, and 20 is frequently complemented, particularly in glioblastomas, by either loss or complex aneuploidy of chromosome 9, and then, in the presence of this additional chromosomal change, we also frequently see loss or complex aneuploidy of chromosome 6 in glioblastoma. *SI Appendix*, Fig. S12 shows a more detailed set of visualizations concentrating on these key interacting subsets of chromosomes. *SI Appendix*, Fig. S13 shows that we obtain qualitatively similar results under two alternative binary scoring schemes.

Discussion

Probability distributions in molecular biology are often complex and idiosyncratic because they inherit the complexity and idiosyncrasy of the chemical, historical, and evolutionary processes from whose confluence they arise. Likewise, the character of these probability distributions is often discrete and combinatorial because the organization of biological information typically takes this form, either in the guise of informational heteropolymers (RNA, DNA, proteins) or because biological complexity often arises from a collection of subunits where each subunit can be in a certain number of states (a collection of neurons, each of which is firing or not; a collection of chromosomes that each appear with a certain number of copies). Here, we have proposed a flexible method for estimating probability distributions over these types of discrete combinatorial spaces that is capable of capturing the detailed idiosyncrasy typical of these naturally occurring distributions.

Many familiar probability distributions, such as the exponential and normal distributions, can be characterized as the highest-entropy distributions compatible with some set of constraints on the moments (e.g., the normal distribution is the maximum entropy distribution given a fixed mean and variance). Recent advances in Bayesian field theoretic density estimation (22–28) have shown that it is possible to elaborate on such maximum entropy estimates by defining a suitable prior over the space of possible probability distributions. The theory we have developed here has a completely analogous structure but transferred to a discrete multivariate setting (*SI Appendix*, Table S1). For example, in the continuous case, the key quantity for determining the prior probability of a particular probability distribution is the integral of the squared Pth-order derivative of the log density, which is a measure of the average local roughness of the probability distribution. Here, the corresponding quantity is, e.g., the average squared value of the conditional log odds ratio. This

makes sense since the conditional log odds ratio is just the discrete analog of a second-order mixed partial derivative of the log density, in that it measures how changing the letter at one position alters the effect on the log probability of changing a letter at another position.

The work proposed here is also closely related to minimum epistasis interpolation, a technique that we recently proposed for regression, rather than density estimation, in biological sequence space (31). The relationship between these techniques is that if we view $-\phi_i$ as a phenotype, then the conditional log odds ratio exactly corresponds to the classical notion of a double-mutant epistatic coefficient. Minimum epistasis interpolation works to estimate the relationship between genotype and phenotype by taking a set of known phenotypic values and estimating the remaining values by minimizing the value of the average squared double-mutant epistatic coefficient (i.e., minimizing $\phi^T\Delta^{(2)}\phi$ subject to an equality constraint at known genotypes). This can produce a very complex reconstruction in regions of sequence space where data are plentiful but relaxes toward a nonepistatic (i.e., additive) reconstruction in regions of sequence space where data are sparse or absent. In the regression case, the solution to this problem is given by solving a system of linear equations, unlike the nonlinear problem explored here. However, at an intuitive level, the SeqDEFT problem is very similar to minimum epistasis interpolation in that sequences with a large number of observations have essentially known log frequencies, whereas sequences with zero or low counts have essentially unknown log frequencies, and the prior works to ensure that the model displays relatively simple behavior in these poorly sampled regions while matching the empirical log frequencies in well-sampled regions.

Our results here also extend our previous results on minimum epistasis interpolation. First, we gain an intuitive perspective on the Bayesian generalization of minimum epistasis interpolation by observing that its prior is equivalent to drawing epistatic coefficients independently for each face of the Hamming graph and then conditioning on mutual consistency [i.e., being a valid solution to the bookkeeping problem (32)]. Second, we see that the operators $\Delta^{(P)}$ for $P>2$ enable higher-order analogs of the original minimum epistasis interpolation technique via constrained minimization of $\phi^T\Delta^{(P)}\phi$.

Our results are also related to the theory of spin glasses, in that our prior can be viewed as a model of quenched disorder on a collection of ℓ spins that can each take α states as in a Potts model (1) but with stochastic interactions of all orders (61) rather than interactions being fixed at order 2 (or more precisely, since our prior is improper, a limit of such models). Despite this complex interpretation at the level of spin–spin interactions, the interpretation is simple in terms of the distribution of energy landscapes over configuration space. Specifically, for the proper component of our prior (consisting of those correlations that cannot be captured by the corresponding maximum entropy model), our method here is to use the unique family of isotropic Gaussian distributions such that the energies of configurations that differ at greater than P spins are conditionally independent. In particular, we show in *SI Appendix* that up to a multiplicative factor $\Delta^{(P)}$ is the unique precision matrix with the properties that 1) its null space is identical to the space of maximum entropy models based on the first $P-1$ moments, 2) its entries depend only on the Hamming distance between sequences, and 3) it takes the value zero (indicating conditional independence) for all pairs of sequences with distance greater than P.

Although SeqDEFT can capture complex probability distributions in the bulk of the data while exhibiting simpler behavior in poorly sampled regions of sequence space, this flexibility comes at a cost in terms of computational complexity. The main issue is the nonquadratic nature of the posterior action, which results in a numerical minimization problem that in practice we can only currently solve for sequence spaces with a few million genotypes or less. This limitation corresponds to a maximum length of 11 for DNA sequences or 5 for amino acid sequences, which prevents the application of SeqDEFT at the whole-protein scale where pairwise maximum entropy models have shown such impressive performance (14, 19–21). While more work is needed to develop nonparametric inference and exploratory data analysis techniques applicable to these larger-sequence spaces, the techniques developed here are still applicable to smaller genomic elements such as transcription factor binding sites, $3'$ splice sites, transcriptional and translational initiation motifs, protein phosphorylation motifs, covariation at key catalytic positions, etc. Much remains to be understood about sequence distributions in spaces, such as these, that are small enough to allow comprehensive estimates of sequence frequencies but large enough to allow distributions exhibiting rich and hitherto unexplored geometries.

Materials and Methods

Our implementation of SeqDEFT is available at https://github.com/davidmccandlish/SeqDEFT. Sparse matrices and their manipulations were computed using the SciPy "sparse" package. MAP solutions were estimated by minimizing the posterior action using the SciPy "optimize" package. The optimal hyperparameter was determined by maximizing the k-fold cross-validated log likelihood with $k=5$. Probability distributions were visualized using the dimensionality reduction technique in ref. 30. Details are given in *SI Appendix*.

Data Availability. Previously published data were used for this work. The dataset of annotated human 5′ splice sites was extracted from GENCODE Release 34 (GRCh38.p13) available at https://www.gencodegenes.org/human/. The dataset of karyotypic abnormalities in human cancer is from ref. 29 and is available as part of the supplementary material at https://doi.org/10.1016/j.ccell.2018.03.007.

ACKNOWLEDGMENTS. This work was supported by NIH Cancer Center Support Grant 5P30CA045508, NIH Grants 5R35GM133777 (to J.B.K.) and 5R35GM133613 (to D.M.M.), a Cold Spring Harbor Laboratory/Northwell Health Alliance grant (to J.B.K.), an Alfred P. Sloan research fellowship (to D.M.M.), computational support from NIH Grant S10OD028632-01, and additional funding from The Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

1. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, M. Weigt, Inverse statistical physics of protein sequences: A key issues review. *Rep. Prog. Phys.* **81**, 032601 (2018).
2. E. Schneidman, M. J. Berry II, R. Segev, W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
3. J. Humplik, G. Tkačik, Probabilistic models for neural populations that naturally capture global coupling and criticality. *PLOS Comput. Biol.* **13**, e1005763 (2017).
4. R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998).
5. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
6. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 1999).
7. R. Staden, Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* **12**, 505–519 (1984).
8. G. D. Stormo, Modeling the specificity of protein-DNA interactions. *Quant. Biol.* **1**, 115–130 (2013).
9. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
10. A. S. Lapedes, B. Giraud, L. Liu, G. D. Stormo, "Correlated mutations in models of protein sequences: Phylogenetic and structural effects" in *Statistics in Molecular Biology and Genetics* (Institute of Mathematical Statistics, 1999), pp. 236–256.
11. W. Bialek, R. Ranganathan, Rediscovering the power of pairwise interactions. arXiv [Preprint] (2007). https://arxiv.org/abs/0712.4397 (Accessed 26 November 2020).
12. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).

13. T. Mora, A. M. Walczak, W. Bialek, C. G. Callan, Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5405–5410 (2010).
14. F. Morcos *et al*., Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
15. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, C. J. Langmead, Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
16. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **87**, 012707 (2013).
17. E. van Nimwegen, Inferring contacting residues within and between proteins: What do the probabilities mean?*PLOS Comput. Biol.* **12**, e1004726 (2016).
18. R. M. Levy, A. Haldane, W. F. Flynn, Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* **43**, 55–62 (2017).
19. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
20. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
21. T. A. Hopf *et al*., Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
22. W. Bialek, C. G. Callan, S. P. Strong, Field theories for learning probability distributions. *Phys. Rev. Lett.* **77**, 4693–4697 (1996).
23. I. Nemenman, W. Bialek, Occam factors and model independent Bayesian learning of continuous distributions. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **65**, 026137 (2002).
24. T. A. Enßlin, M. Frommert, F. S. Kitaura, Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis. *Phys. Rev. D* **80**, 105005 (2009).
25. T. Enßlin, Information field theory. *AIP Conference Proceedings* **1553**, 184–191 (2013).
26. J. B. Kinney, Estimation of probability densities using scale-free field theories. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **90**, 011301 (2014).
27. J. B. Kinney, Unification of field theory and maximum entropy methods for learning probability densities. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **92**, 032107 (2015).
28. W. C. Chen, A. Tareen, J. B. Kinney, Density estimation on small data sets. *Phys. Rev. Lett.* **121**, 160605 (2018).
29. A. M. Taylor *et al*, Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
30. D. M. McCandlish, Visualizing fitness landscapes. *Evolution* **65**, 1544–1558 (2011).
31. J. Zhou, D. M. McCandlish, Minimum epistasis interpolation for sequence-function relationships. *Nat. Commun.* **11**, 1782 (2020).
32. A. C. Nagel, P. Joyce, H. A. Wichman, C. R. Miller, Stickbreaking: A novel fitness landscape model that harbors epistasis and is consistent with commonly observed patterns of adaptive evolution. *Genetics* **190**, 655–667 (2012).
33. I. J. Good, Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Stat.* **34**, 911–934 (1963).
34. S. Brooks, A. Gelman, G. L. Jones, X. L. Meng, *Handbook of Markov Chain Monte Carlo* (Chapman & Hall/CRC, Boca Raton, FL, 2011).
35. P. A. Sharp, The discovery of split genes and RNA splicing. *Trends Biochem. Sci.* **30**, 279–281 (2005).
36. X. Roca, A. R. Krainer, I. C. Eperon, Pick one, but be quick: 5′ Splice sites and the problems of too many choices. *Genes Dev.* **27**, 129–144 (2013).
37. G. E. Parada, R. Munita, C. A. Cerda, K. Gysling, A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.* **42**, 10564–10578 (2014).
38. S. Erkelenz *et al*., Ranking noncanonical 5′ splice site usage by genome-wide RNA-seq analysis and splicing reporter assays. *Genome Res.* **28**, 1826–1840 (2018).
39. J. J. Turunen, E. H. Niemelä, B. Verma, M. J. Frilander, The significant other: Splicing by the minor spliceosome. *Wiley Interdiscip. Rev. RNA* **4**, 61–76 (2013).
40. G. Sella, A. E. Hirsh, The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9541–9546 (2005).
41. D. M. McCandlish, A. Stoltzfus, Modeling evolution using the probability of fixation: History and implications. *Q. Rev. Biol.* **89**, 225–252 (2014).
42. R. R. Coifman, S. Lafon, Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
43. A. L. Halpern, W. J. Bruno, Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol. Biol. Evol.* **15**, 910–917 (1998).
44. C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
45. I. Carmel, S. Tal, I. Vig, G. Ast, Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA* **10**, 828–840 (2004).
46. X. Roca *et al*., Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics. *Genome Res.* **18**, 77–87 (2008).
47. J. J. Turunen, C. L. Will, M. Grote, R. Lührmann, M. J. Frilander, The U11-48K protein contacts the 5′ splice site of U12-type introns and the U11-59K protein. *Mol. Cell. Biol.* **28**, 3548–3560 (2008).
48. X. Roca, A. R. Krainer, Recognition of atypical 5′ splice sites by shifted base-pairing to U1 snRNA. *Nat. Struct. Mol. Biol.* **16**, 176–182 (2009).
49. C. B. Burge, R. A. Padgett, P. A. Sharp, Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2**, 773–785 (1998).
50. D. C. Moyer, G. E. Larue, C. E. Hershberger, S. W. Roy, R. A. Padgett, Comprehensive database and evolutionary dynamics of U12-type introns. *Nucleic Acids Res.* **48**, 7066–7078 (2020).
51. L. Sansregret, C. Swanton, The role of aneuploidy in cancer evolution. *Cold Spring Harb. Perspect. Med.* **7**, a028373 (2017).
52. J. M. Sheltzer, A. Amon, The aneuploidy paradox: Costs and benefits of an incorrect karyotype. *Trends Genet.* **27**, 446–453 (2011).
53. J. M. Sheltzer *et al*., Single-chromosome gains commonly function as tumor suppressors. *Cancer Cell* **31**, 240–255 (2017).
54. U. Ben-David, A. Amon, Context is everything: Aneuploidy in cancer. *Nat. Rev. Genet.* **21**, 44–62 (2020).
55. J. N. Weinstein *et al*., The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
56. L. Sansregret, B. Vanhaesebroeck, C. Swanton, Determinants and clinical implications of chromosomal instability in cancer. *Nat. Rev. Clin. Oncol.* **15**, 139–150 (2018).
57. M. V. Yusenko *et al*., High-resolution DNA copy number and gene expression analyses distinguish chromophobe renal cell carcinomas and renal oncocytomas. *BMC Cancer* **9**, 152 (2009).
58. W. M. Linehan *et al*., Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).
59. A. Cohen *et al*., DNA copy number analysis of Grade II-III and Grade IV gliomas reveals differences in molecular ontogeny including chromothripsis associated with IDH mutation status. *Acta Neuropathol. Commun.* **3**, 34 (2015).
60. C. Geisenberger *et al*., Molecular profiling of long-term survivors identifies a subgroup of glioblastoma characterized by chromosome 19/20 co-gain. *Acta Neuropathol.* **130**, 419–434 (2015).
61. B. Derrida, Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B* **24**, 2613 (1981).