Research

Identification and utilization of arbitrary correlations in models of recombination signal sequences

Lindsay G Cowell*, Marco Davila*, Thomas B Kepler† and Garnett Kelsoe*

Addresses: *Department of Immunology, Duke University Medical Center, Durham, NC 27710, USA. †Center for Bioinformatics and Computational Biology, Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA.

Correspondence: Garnett Kelsoe. E-mail: ghkelsoe@duke.edu

Published: 21 November 2002

Genome ***Biology*** 2002, **3(12)**:research0072.1–0072.20

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0072

Received: 24 June 2002
Revised: 4 September 2002
Accepted: 10 October 2002

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: A significant challenge in bioinformatics is to develop methods for detecting and modeling patterns in variable DNA sequence sites, such as protein-binding sites in regulatory DNA. Current approaches sometimes perform poorly when positions in the site do not independently affect protein binding. We developed a statistical technique for modeling the correlation structure in variable DNA sequence sites. The method places no restrictions on the number of correlated positions or on their spatial relationship within the site. No prior empirical evidence for the correlation structure is necessary.

Results: We applied our method to the recombination signal sequences (RSS) that direct assembly of B-cell and T-cell antigen-receptor genes via V(D)J recombination. The technique is based on model selection by cross-validation and produces models that allow computation of an information score for any signal-length sequence. We also modeled RSS using order zero and order one Markov chains. The scores from all models are highly correlated with measured recombination efficiencies, but the models arising from our technique are better than the Markov models at discriminating RSS from non-RSS.

Conclusions: Our model-development procedure produces models that estimate well the recombinogenic potential of RSS and are better at RSS recognition than the order zero and order one Markov models. Our models are, therefore, valuable for studying the regulation of both physiologic and aberrant V(D)J recombination. The approach could be equally powerful for the study of promoter and enhancer elements, splice sites, and other DNA regulatory sites that are highly variable at the level of individual nucleotide positions.

Background

Modeling variable DNA sequence sites

The set of binding sites for a single DNA-binding protein can be highly variable [1,2], and the degree of nucleotide variation tolerated generally differs from position to position within a site [3-5]. This sequence diversity can have important functional consequences as the affinity between regulatory proteins and their binding sites is modulated by changes in binding-site sequence [1]. Large datasets of related DNA sites are now available [3,4,6,7]. Currently, more than 100 prokaryotic and eukaryotic genomes have been completely sequenced, permitting cross-species comparisons of even larger sequence

assemblies than collected here (see, for example [8,9]). Computational approaches can therefore be used to detect and model the sequence patterns present within the binding-site variability. The most useful models of variable DNA sites provide classification algorithms for identifying functional sites as well as insights that can help elucidate the relationship between the structure and function of these sites.

Variable DNA sequence sites are frequently described by a consensus sequence [10] or a weight matrix [11-13]. Consensus sequences have limited utility for even moderately variable sites because they preserve little or no information about the variability within the sequences they characterize. Differences from consensus are quantified by counting the number of mismatched positions, making no distinction between mismatches that abolish function of the site and those that modulate function.

Weight matrices were introduced to characterize transcription and translation initiation sites in *Escherichia coli* [11] and have become standard for characterizing binding sites for transcription factors. A weight matrix is a two-dimensional matrix in which each row corresponds to one of the four nucleotides and each column corresponds to one position in the site being described [11]. The elements of the matrix are the observed counts for each nucleotide at each position in an alignment of the DNA sites [11], or some function of these counts (reviewed in [13,14]). Any sequence can be scored by summing the matrix elements corresponding to the sequence [11], thereby quantitatively rating the sequence's functional potential.

Weight matrices can be used to scan genomic DNA and determine statistically significant matches to the sequence pattern described by the matrix [14]. Putative sites are scored by the weight matrix, so the number of false positives and false negatives can be balanced by choosing an appropriate threshold score defining functional sites. The scores of functional sites are sometimes correlated with their level of function [11,14-23].

Weight matrices are typically based on order zero Markov chains, meaning they assume that individual base pairs in the binding site affect binding affinity independently. Weight-matrix models based on order one Markov chains assume that adjacent positions are correlated [15,24-27]. Ponomarenko *et al.* [28] constructed weight matrices for many different transcription factor binding sites using mono-, di-, and trinucleotide motifs allowing correlation between non-adjacent positions by specifying X_1NX_3 and $X_1NX_3NX_5$ motifs, where X_i is the nucleotide at position i of the motif and N is any nucleotide. While this approach allows for correlation between more than two non-adjacent positions, the motifs may still be unnecessarily restrictive. The apposition of distant binding-site positions can be important to recruitment of the binding protein, for example, through formation of DNA secondary structure [29-32], or to the interaction of the binding site with its binding protein [33]. Therefore correlation between distant positions is expected.

Burge and Karlin [34] used hypothesis testing via χ^2 tests to detect significant correlations between any two positions and introduced a model-building procedure, maximal dependence decomposition, to account for the dependencies. For each significant correlation, this method partitions the dataset into subsets of sequences exhibiting no correlation; the final model is a composite of weight matrices (order zero), one for each subset. Our model-building technique is based on cross-validation criteria rather than on hypothesis testing *per se*, and we use a more general model class for DNA sequences.

Models that allow for pairwise correlations between positions were developed in the context of Bayes networks by Cai *et al.* [35]. Bayes networks represent a large class of models, and in fact, our model can be recast as a Bayes network, although it is not now formulated in those terms.

We have developed an approach to determine the correlation structure present in a set of variable sequence sites. For each position in the site, we identify all other correlated positions placing no restrictions on the number or spatial relationship of correlated positions in the site. Our approach determines, from all possible combinations of disjoint probability distributions, the set of distributions that most effectively distinguishes functional sites from non-functional sequences. Although the family of models we consider is very large, we proceed by model selection rather than hypothesis testing. The selected probability function is the product of these disjoint probability distributions. The natural logarithm of this probability function can be used as a score for any sequence, thereby recognizing the presence of evolutionarily conserved nucleotide associations. We have applied this approach to recombination signal sequences (RSS), a set of DNA binding sites necessary for specific immunity.

Recombination signal sequences

Specific immunity depends on the ability of B and T lymphocytes to recognize antigens, molecular identifiers of pathogens [36]. The immune system does not anticipate which antigen will be encountered, but remarkable genetic mechanisms have evolved that generate diverse antigen-receptor repertoires. The primary generative mechanism is V(D)J recombination, the rearrangement of B-cell receptor (BCR) or T-cell receptor (TCR) V, D, and J gene segments to form functional genes; the resulting combinatorial and junctional diversity can produce around 10^{14} BCR specificities and around 10^{18} TCR specificities in one animal [37]. This extraordinary plasticity has a price: V(D)J recombination can produce self-reactive receptors leading to autoimmune disease [38,39] and may err to create oncogenic chromosomal translocations [40].

V(D)J recombination proceeds by the introduction of double-strand DNA breaks (reviewed in [41]). To prevent DNA breaks that would cause cellular damage, the V(D)J recombinase must be specifically targeted to the *Bcr* and *Tcr* loci. This targeting is regulated primarily by binding of RAG1 [42], an enzyme in the recombinase complex, to the RSS adjacent to each V, D and J gene segment [43]. RSS were initially identified as a conserved heptamer (consensus CACAGTG) and a conserved nonamer (consensus ACAAAAACC) separated by a less conserved spacer of either 12±1 or 23±1 base-pairs (bp) [44]. The initial CAC of the heptamer is highly conserved, but all remaining positions show moderate to very low levels of conservation (reviewed in [45]).

RSS variability has important functional consequences: the efficiency with which each RSS mediates recombination depends on its sequence [46]. There is strong evidence that differing recombination efficiencies of RSS result in the biased utilization of their associated gene segments (reviewed in [47,48]). In addition, the biases in gene segment use observed in the post-selection TCR repertoire are consistent with the biases observed before selection, suggesting that the primary immune repertoire may be genetically patterned [48].

The high level of diversity among RSS makes it very difficult to evaluate their recombinogenic potential. RSS not associated with a V, D or J gene segment but participating in aberrant (illegitimate) V(D)J recombination are particularly difficult to recognize. These non-physiologic RSS are important because of their potential involvement in oncogenic chromosomal translocation and receptor editing, a process that alters the specificity of autoreactive receptors. Non-physiologic RSS that have simply arisen by chance can be referred to as fortuitous RSS whereas those thought to have descended from the same ancient transposon as physiologic RSS can be referred to as cryptic (cRSS) [49,50]. Generally, the origin of a non-physiologic RSS is not known; for simplicity, we will refer to both types as cryptic.

To understand the role of RSS variability in the formation of the primary immune repertoire and of cRSS in illegitimate V(D)J recombination, it is necessary to quantify the genetic variability among physiologic RSS and the relationship between this variability and the regulation of recombination. Allowing for each of the 4 nucleotides at just 10 of the 28 or 39 RSS positions results in over 10^6 different signals - it is not feasible to measure the efficiency of all potential RSS experimentally. Statistical models of RSS variability that allow prediction of recombination efficiency are essential to furthering our understanding of the biological function of RSS variability.

One estimate of cRSS frequency in mammalian genomes is one in every 600 bp [49]. Promiscuous recombination at this frequency, however, would result in significant damage to the cell. This suggests that complex patterns underlie the high level of nucleotide diversity observed at individual positions, thereby increasing the specificity of the signal governing recruitment of the V(D)J recombinase.

We find significant correlations between nucleotide positions in RSS, suggesting they act cooperatively and that nucleotide associations are conserved. We used our model selection procedure to determine the best models of RSS correlation structure, one model for 12-bp spacer RSS (12-RSS) and one for 23-RSS. We also modeled both types of RSS with order zero and order one Markov chains. The scores from all six models are highly correlated with measured recombination efficiencies, but the models arising from our technique are much better than the Markov models at identifying physiologic and cryptic RSS in genomic DNA.

Results

Mouse RSS show limited sequence conservation

We analyzed 356 physiologic mouse RSS (see Methods). While 62% (219/356) of these RSS contain a consensus heptamer, just 16% (57/356) contain a consensus nonamer and only 13% (48/356) contain both a consensus heptamer and a consensus nonamer. To measure the level of nucleotide diversity at individual RSS positions, we computed the entropy (H_i) [51] for each position i in an alignment of the 201 12-RSS and an alignment of the 155 23-RSS (Figure 1a). A strictly conserved position in which no variation is tolerated has $H_i = 0$, while a position in which the four nucleotides are observed at nearly uniform frequencies has $H_i = 1$.

The signal specified by the individual RSS positions is very low, especially in 23-RSS (Figure 1a). The average entropy is $\overline{H}_{12} = 0.50$ for 12-RSS and $\overline{H}_{23} = 0.67$ for 23-RSS; for a randomly generate string of A, G, C, and T characters, the expected value for $\overline{H}$ is one. The patterns of nucleotide diversity at individual positions are similar for both 12- and 23-RSS: the initial CAC of the heptamer is invariant, the remaining four positions of the heptamer are moderately diverse, the spacer is highly diverse, and the A tract of the nonamer (positions 5 through 7) is better conserved than the other nonamer positions (Figure 1a). The average entropy for 23-RSS is higher than for 12-RSS, primarily because the 23-RSS nonamers are much more diverse than 12-RSS nonamers ($\overline{H}_{N23} = 0.59$; $\overline{H}_{N12} = 0.35$; Figure 1a).

It is known from experimental studies that nucleotide substitutions at highly diverse positions influence recombination efficiency [46,52]. That a given RSS position can be both highly variable and exhibit strong effects on signal efficiency appears contradictory. The contradiction can be reconciled, however, by hypothesizing correlations between nucleotides in RSS: a nucleotide substitution at one position may be compensated for by a change at another position in the RSS.

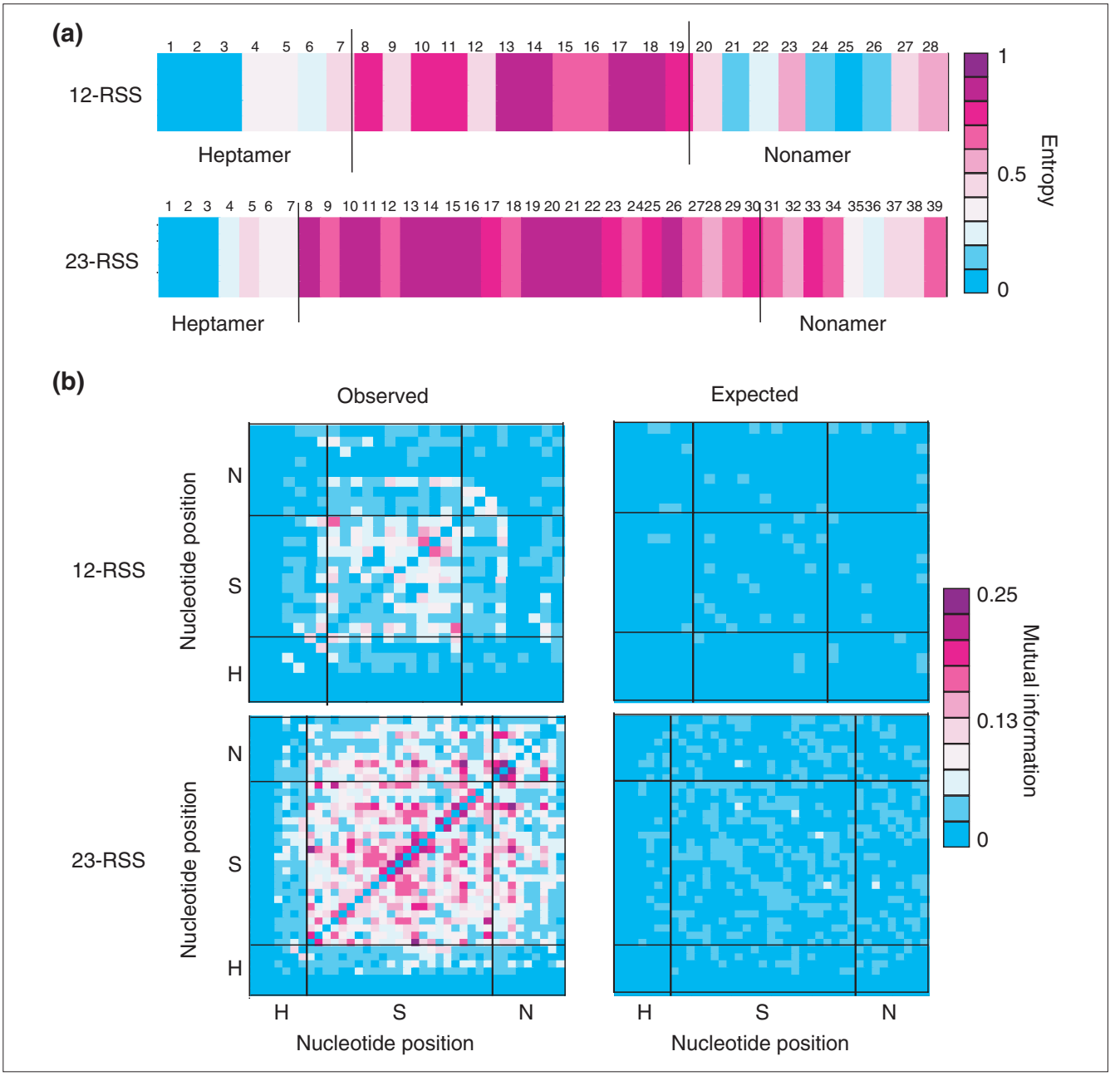

Figure 1
Entropy and mutual information in physiologic RSS. **(a)** Position-wise entropy H_i at each position in an alignment of physiologic 12-RSS (upper bar) and an alignment of physiologic 23-RSS (lower bar). Position in the alignment is shown above each bar, and the value of H_i at position i is indicated by the color of the bar at that position. **(b)** Mutual information ($MI_{i,i'}$) between pairs of positions in physiologic 12- and 23-RSS. $MI_{i,i'}$ values between positions in the 12-RSS are shown in the upper panel, and $MI_{i,i'}$ values between positions in the 23-RSS are shown in the lower panel. Location in the alignment is given on the *x*- and *y*-axes. The black lines mark the heptamer (H)-spacer (S) and spacer-nonamer (N) boundaries. The color at the intersection of an *x*-axis and a *y*-axis grid line gives the *MI* contained in the positions corresponding to the two grid lines. The actual *MI* computed from the RSS alignments are shown in the left panels, and one of 300 permutations is shown on the right as an example of the level of *MI* observable between each pair of positions in the absence of any correlation between them.

Significant pairwise correlations exist between positions in the RSS

To test for correlation between pairs of positions in the RSS, we computed the mutual information, $MI_{i,i}$, [51] between all pairs of positions i and i' in an alignment of the 201 physiologic mouse 12-RSS and in an alignment of the 155 physiologic mouse 23-RSS (see Methods) (Figure 1b). Statistically significant $MI_{i,i'}$ values result when there is a

correlation between nucleotide frequencies at the two positions, indicating selection pressure to maintain (or avoid) a particular dinucleotide.

In 12- and 23-RSS, we detect statistically significant correlations between positions in those regions of the RSS exhibiting high levels of position-wise entropy. These are the spacer, especially the nonamer-proximal half, and the positions of the heptamer and nonamer that are adjacent to the spacer (Figure 1b). In general, correlation between two positions decreases as a function of distance, but there are examples of strong correlation between widely separated positions (Figure 1b). For example, we observe 12-RSS with A at positions 8 and 19 at a much higher frequency (0.38) than expected (0.11) from the independent frequencies of A at each position. In addition, we find that the levels of correlation are higher in 23-RSS than in 12-RSS (Figure 1b); this is an especially interesting observation given that 23-RSS are much more variable than 12-RSS (Figure 1a).

Statistical models of 12- and 23-RSS determine the relevant correlation structure

The *MI* computations give evidence for strong pairwise correlations, including between distant positions, and correlations may exist between any number of positions. We therefore developed statistical models, one for 12-RSS and one for 23-RSS, that can account for correlations between arbitrary positions without regard for their distance from each other or the number of correlated positions. The models were developed using a procedure that selects from all possible combinations of disjoint probability distributions the set of marginal (for one position) and joint (for multiple positions) distributions that most distinguishes physiologic RSS from other nucleotide sequences of the same length. For example, positions 8 and 19 in 12-RSS are highly correlated (Figure 1b). They could be included in the model either through the two marginal probability distributions - the distribution over the four nucleotides at position 8 and the distribution over the four nucleotides at position 19, or through one joint probability distribution - the distribution over the 16 pairs of nucleotides at the pair of positions. Each model computes a score for RSS information content: RIC_{12} for 28-bp sequences and RIC_{23} for 39-bp sequences; the formation of joint probability functions is determined by maximizing the mean score for physiologic RSS.

The selected 12-RSS model is:

$RIC_{12} = \ln[P_1\ P_2\ P_{3,15,25}\ P_{4,5}\ P_{6,28}\ P_{7,8,19}\ P_{9,26}\ P_{10,12}\ P_{11,27}\ P_{13,14,23}\ P_{16,17,18}\ P_{20,21,22}\ P_{24}]$

where P_1 is the marginal probability function for position 1 and $P_{3,15,25}$ is the joint probability function for positions 3, 15 and 25. The presence of the joint probability function in the model indicates that these three positions are mutually correlated. The selected 23-RSS model is:

$RIC_{23} = \ln[P_1\ P_2\ P_3\ P_{4,14}\ P_{5,39}\ P_6\ P_{7,24,25}\ P_{8,9,21}\ P_{10,16}\ P_{11,12}\ P_{13,22}\ P_{15,23}\ P_{17,18}\ P_{19,27,30,31,32,33,37}\ P_{20,26}\ P_{28,29}\ P_{34,38}\ P_{35,36}]$.

Most of the marginal probabilities have been merged to form joint probability functions, but the order in which the joint functions were formed reflects the relative strength of the correlations between the corresponding positions: positions with large *MI* are grouped early in model selection (data not shown). The groups of positions with the strongest correlations are (7,8,19), (16,17,18), and (20,21,22) in 12-RSS (Table 1) and (19,27,30,31,32,33,37), (8,9,21), and (7,24,25) in 23-RSS (Table 2). For both 12- and 23-RSS, the positions exhibiting the strongest cooperative influence lie in the nonamer-proximal half of the spacer and in the heptamer and nonamer positions adjacent to the spacer (Figure 2). Interestingly, these positions overlap substantially with positions exhibiting ethylation/methylation interference in RSS complexed with RAG1/RAG2 (Figure 2 and [53]) suggesting that the correlations detected by the models are relevant to RAG-RSS interaction.

The selected models were compared with order zero and order one Markov models

To determine whether the correlation structure detected by our model selection procedure improves RSS recognition and evaluation, we constructed weight matrix models based on order zero or order one Markov chains. Order zero Markov models (WMo_{12} for 12-RSS and WMo_{23} for 23-RSS) assume that all RSS positions are independent and are therefore computed from marginal probability distributions. The score for an *N*-bp sequence is computed as

$$WMo_N = \ln\left[\prod_{i=1}^{N} P_i\right]$$

where P_i is the probability of observing nucleotide *X* at position *i*. Order one Markov models ($WM1_{12}$ and $WM1_{23}$) assume that adjacent positions are correlated (the identity of the nucleotide at position *i* depends on the nucleotide at position *i* - 1) and are computed from conditional probability distributions:

$$WM1_N = \ln\left[P_1 \prod_{i=2}^{N} P(i|i-1)\right]$$

where P_1 is the marginal probability distribution for position 1, and $P(i|i-1)$ is the probability of observing nucleotide *X* at position *i* given the nucleotide at position *i* - 1.

For physiologic 12- and 23-RSS, the score distribution is shifted toward higher scores for models that include correlation (*RIC* and *WM1*) (Table 3, Figure 3). The difference is most striking for 23-RSS (Table 3, Figure 3). The mean scores for the physiologic 12-RSS are $\overline{WMo}_{12}$ = -19.84, $\overline{WM1}_{12}$ = -18.49, and $\overline{RIC}_{12}$ = -18.47 (Table 3); the mean

Table 1

Strongest nucleotide associations in 12-RSS

Correlated positions	Associated nucleotides	Count	Frequency
7:8:19	G:A:A	73	0.365
	G:C:T	23	0.115
	G:C:A	17	0.085
	G:A:G	11	0.055
	A:G:G	10	0.05
	G:A:T	9	0.045
	A:G:C	6	0.03
	A:T:T	6	0.03
	G:T:T	6	0.03
	G:G:A	5	0.025
	A:A:T	4	0.02
	A:C:T	4	0.02
	A:G:T	3	0.015
	G:A:C	2	0.01
	G:C:G	2	0.01
	G:T:G	2	0.01
	C:A:T	2	0.01
	C:T:G	2	0.01
	C:T:C	2	0.01
	T:A:T	2	0.01
	A:A:A	1	0.005
	G:G:C	1	0.005
	G:G:T	1	0.005
	G:C:C	1	0.005
	C:A:A	1	0.005
	C:G:G	1	0.005
	C:C:C	1	0.005
	C:C:T	1	0.005
	C:T:T	1	0.005
16:17:18	C:T:T	34	0.169
	C:A:T	32	0.159
	T:G:G	20	0.1
	T:C:C	15	0.075
	C:T:G	14	0.07
	C:C:T	11	0.055
	A:G:C	9	0.045
	T:T:C	8	0.04
	C:T:C	7	0.035
	T:A:G	7	0.035
	C:C:A	5	0.025
	C:A:G	4	0.02
	T:C:T	4	0.02
	T:G:A	3	0.015
	A:T:C	2	0.01
	G:G:G	2	0.01
	C:A:C	2	0.01
	C:C:G	2	0.01
	C:C:C	2	0.01
	T:A:C	2	0.01
	T:A:T	2	0.01
	T:G:T	2	0.01
	T:C:A	2	0.01
	A:A:C	1	0.005
	A:C:T	1	0.005
	A:T:A	1	0.005
	G:C:T	1	0.005
	G:T:C	1	0.005
	C:A:A	1	0.005
	C:T:A	1	0.005
	T:A:A	1	0.005
	T:T:A	1	0.005
	T:T:G	1	0.005
20:21:22	A:C:A	157	0.785
	G:C:A	18	0.09
	T:C:A	6	0.03
	C:C:C	3	0.015
	A:A:A	2	0.01
	G:A:C	2	0.01
	G:G:T	2	0.01
	G:T:C	2	0.01
	A:C:T	1	0.005
	G:A:G	1	0.005
	G:A:T	1	0.005
	C:A:A	1	0.005
	C:A:G	1	0.005
	C:C:A	1	0.005
	T:C:G	1	0.005
	T:T:C	1	0.005

For the three groups of positions in the 12-RSS model exhibiting the strongest correlations, the nucleotide motifs present in our dataset at those positions, the number of RSS containing that motif, and the frequency in the dataset are shown. Nucleotide motifs not shown do not occur in our dataset. For example, three positions making up one group (for example, 7, 8 and 19) would be occupied in each RSS by one of the 64 triplets, but not every one of the 64 triplets will be present.

scores for physiologic 23-RSS are $\overline{WM0}_{23}$ = -37.07, $\overline{WM1}_{23}$ = -31.51, and $\overline{RIC}_{23}$ = -32.39 (Table 3). There is relatively little change in the value of low scores across models, so as the correlation structure of the models increases in complexity, the increase in high scores results in an increase in the score range (Table 3 and Figure 3).

Functional RSS are best discriminated by *RIC* scores

To test whether functional RSS can be recognized by the sequence properties captured in the models, we characterized the score distributions for non-RSS DNA (Figure 4). These background distributions are characterized by their means and ranges. *WM0*, *WM1* and *RIC* scores were computed for all 28- and 39-bp segments in a 212,128-bp fragment of mouse chromosome 8 (AC084823) containing no known RSS. We also characterized the background distributions by computing scores for a pseudorandom string (PRS) of A, T, C and G the same length as sequence AC084823 and having the same nucleotide-usage frequencies.

Table 2

Strongest nucleotide associations in 23-RSS

Correlated positions	Associated nucleotides	Count	Frequency
19:27:30:31:32:33:37	T:G:T:C:A:G:C	16	0.108
	C:C:T:A:C:C:A	12	0.081
	A:C:A:A:C:A:A	9	0.061
	G:C:G:A:C:A:A	8	0.054
	C:C:T:A:C:A:A	6	0.041
	T:T:A:A:C:A:A	6	0.041
	C:C:C:A:C:A:A	5	0.034
	T:G:T:C:A:G:T	5	0.034
	A:C:A:G:C:A:A	4	0.027
	A:C:C:A:C:A:A	4	0.027
	A:C:C:A:T:A:A	4	0.027
	T:G:T:C:A:G:A	4	0.027
	A:C:T:A:C:A:A	3	0.020
	G:G:T:C:A:G:A	3	0.020
	C:C:T:G:C:A:A	3	0.020
	C:T:A:A:C:T:A	3	0.020
	T:C:T:A:C:A:A	3	0.020
	A:T:A:A:C:A:A	2	0.014
	C:A:A:A:C:A:A	2	0.014
	C:A:C:C:T:C:G	2	0.014
	C:C:C:A:C:C:A	2	0.014
	C:C:T:A:A:C:A	2	0.014
	C:C:T:A:C:T:A	2	0.014
	C:C:T:C:C:T:A	2	0.014
	C:C:T:T:C:A:A	2	0.014
	T:G:T:A:C:A:A	2	0.014
	T:C:G:A:C:A:A	2	0.014
	T:C:T:T:A:C:A	2	0.014
	T:T:A:C:T:A:C	2	0.014
	A:A:G:G:A:C:G	1	0.007
	A:G:C:A:G:A:A	1	0.007
	A:C:G:C:A:C:T	1	0.007
	A:C:T:G:C:A:A	1	0.007
	A:T:G:A:C:A:A	1	0.007
	A:T:T:A:C:A:A	1	0.007
	G:C:A:G:C:G:A	1	0.007
	G:C:G:A:A:A:A	1	0.007
	G:C:G:C:C:C:A	1	0.007
	G:C:C:A:C:A:A	1	0.007
	G:C:T:A:A:G:A	1	0.007
	G:C:T:G:C:A:A	1	0.007
	G:C:T:C:T:C:A	1	0.007
	G:T:A:C:A:A:C	1	0.007
	C:C:A:C:A:A:C	1	0.007
	C:C:G:A:C:A:A	1	0.007
	C:C:T:A:T:G:A	1	0.007
	C:T:A:G:C:A:A	1	0.007
	C:T:C:A:C:A:A	1	0.007
	T:G:T:C:A:C:C	1	0.007
	T:G:T:T:A:G:C	1	0.007
	T:C:C:A:C:A:A	1	0.007
	T:C:T:G:C:A:A	1	0.007
	T:T:G:G:C:A:A	1	0.007
	T:T:G:C:A:G:C	1	0.007
	T:T:G:T:C:A:A	1	0.007
8:09:21	T:T:C	34	0.222
	A:G:G	24	0.157
	C:T:T	21	0.137
	G:T:C	7	0.046
	G:T:G	6	0.039
	A:G:T	5	0.033
	T:G:G	5	0.033
	T:T:A	5	0.033
	A:C:G	4	0.026
	A:T:T	4	0.026
	G:C:G	4	0.026
	T:G:T	4	0.026
	C:T:A	3	0.020
	T:C:C	3	0.020
	T:C:T	3	0.020
	T:T:G	3	0.020
	A:C:A	2	0.013
	G:C:C	2	0.013
	C:T:G	2	0.013
	T:C:A	2	0.013
	T:T:T	2	0.013
	A:G:A	1	0.007
	A:C:T	1	0.007
	A:T:A	1	0.007
	A:T:G	1	0.007
	A:T:C	1	0.007
	G:A:C	1	0.007
	C:T:C	1	0.007
	T:G:A	1	0.007
7:24:25	G:A:G	65	0.419355
	G:A:A	24	0.154839
	G:G:A	11	0.070968
	G:C:C	11	0.070968
	G:C:A	6	0.03871
	A:A:G	5	0.032258
	G:A:C	5	0.032258
	G:T:G	4	0.025806
	C:T:C	4	0.025806
	C:T:T	3	0.019355
	A:A:A	2	0.012903
	A:C:T	2	0.012903
	A:T:T	2	0.012903
	G:G:G	2	0.012903
	T:T:C	2	0.012903
	A:C:C	1	0.006452
	A:T:A	1	0.006452
	A:T:G	1	0.006452
	G:T:A	1	0.006452
	G:T:C	1	0.006452
	T:G:A	1	0.006452
	T:T:T	1	0.006452

Details as in Table 1.

As expected, the *WMo* score distributions for AC084823 and the PRS are almost identical for both 28- and 39-bp segments (Table 4, Figure 4). For the models that account for correlation, however, scores from genomic DNA are higher

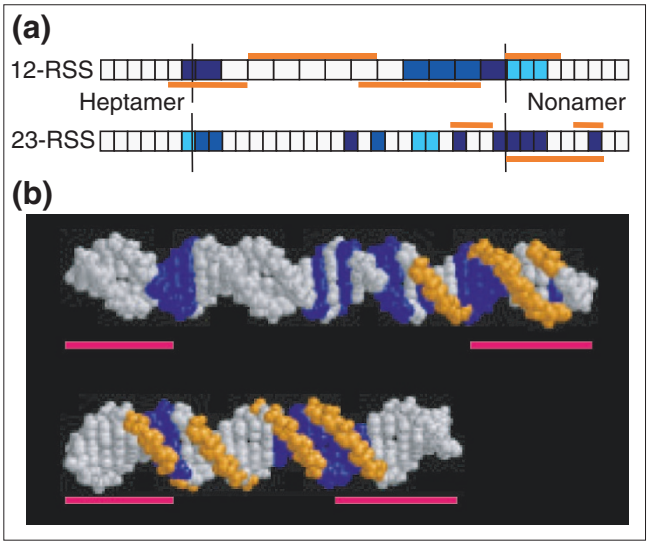

Figure 2
Location of the most highly correlated positions in 12- and 23-RSS and sites of ethylation/methylation interference in RSS complexed with RAG. **(a)** Positions in 12-RSS are shown in the upper row of boxes and those in 23-RSS are shown in the lower row. Boxes filled with the same color indicate positions that are correlated. For each model, the three associations with the highest level of correlation are shown. The intensity of the blue indicates the relative strength of the correlation with the darkest being the most correlated. Sites of ethylation/methylation interference in RSS complexed with RAG are shown in orange [53]. **(b)** Overlap between the strongest correlations (shown in blue) and sites of ethylation/methylation interference (shown in orange) [53]. The relative positions of the heptamer and nonamer are indicated by the red lines. Figure generated in RasMol [69].

Table 3

Distribution of scores for physiologic RSS

Score level	12-RSS			23-RSS		
	WM0	*WM1*	*RIC*	*WM0*	*WM1*	*RIC*
0	0	0	0	0	0	0
-5	0	16	19	0	0	0
-10	56	70	83	0	0	0
-15	68	44	25	0	21	26
-20	30	24	26	0	38	18
-25	26	32	26	25	21	29
-30	13	9	12	50	17	24
-35	5	4	8	31	27	21
-40	2	1	0	27	16	16
-45	1	1	2	11	7	8
-50	0	0	0	4	5	6
-55	0	0	0	5	1	4
-60	0	0	0	2	1	1
-65	0	0	0	0	1	2
-70	0	0	0	0	0	0
Total	201	201	201	155	155	155
Maximum	-10.52	-8.16	-8.02	-26.89	-15.85	-15.83
Mean	-19.84	-18.49	-18.47	-37.07	-31.51	-32.39
Minimum	-46.86	-46.32	-48.16	-63.22	-65.24	-69.69

For each model and each RSS type, the maximum score, the mean score, and the minimum score for physiologic RSS is given, along with the number of physiologic RSS scoring at the level indicated by column one. The number shown at score level -5 is the number of scores between -5 and -9.999; the number shown at score level -10 is the number of scores between -10 and -14.999.

on average than those computed for the PRS (Table 4, Figure 4). This is especially true for the *WM1* models, indicating that, whereas both the *WM1* and *RIC* models are influenced by correlations present in physiologic DNA, the correlations detected by the *RIC* models are more specific to RSS. Importantly, the distributions for *RIC* scores are always shifted toward lower scores than the corresponding *WM0* and *WM1* distributions (Table 4, Figure 4).

We next compared across models the frequency of signal-length segments in AC084823 that score above a given threshold. To determine test thresholds for a given model, we ranked the physiologic RSS by score in ascending order and took the lowest 5% of scores under each model as the test thresholds for that model (Table 5, Figure 5). For example, the lowest RIC_{12} is -48.16. It has rank zero, and zero physiologic RSS have a lower RIC_{12}. Of 212,101 (0.00414) 28-bp segments in AC084823, 859 have RIC_{12} > -48.16. Similarly, threshold zero under the $WM1_{12}$ model is -46.32, and 2,372 of 212,101 (0.01118) 28-bp segments in AC084823 have $WM1_{12}$ > -46.32. Under each model, the frequency of physiologic RSS scoring below threshold estimates the probability of failing to recognize a functional RSS and is a measure of the model's sensitivity. The frequency of non-RSS scoring above threshold estimates the probability of classifying non-RSS as functional and is a measure of the model's specificity. Both numbers should be small. The high-scoring non-RSS segments may support recombination, however, and, if so, would be classified as cRSS. Those not supporting recombination would be counted as false positives.

In general we find that the *WM0* models predict the highest number of RSS in AC084823, and the *RIC* models predict the fewest (Table 5, Figure 5). Of 20 test thresholds, there are three for which *RIC* does not predict the smallest number of RSS: for the 12-RSS models, the score for the ninth-lowest ranking physiologic 12-RSS and for the 23-RSS models, the scores for the second- and eighth-lowest ranking physiologic 23-RSS (Table 5, Figure 5).

From Figure 5, we determine thresholds at which excluding just one more physiologic RSS results in a large drop in the models' predicted number of cryptic RSS. We select -38.81 for RIC_{12} and -58.45 for RIC_{23} (Figure 5). Only two of the 201 (0.01) 12-RSS have RIC_{12} < -38.81, and just 54 of 212,101 (2.5×10^{-4}) 28-bp segments in sequence AC084823 achieve RIC_{12} > -38.81. The frequency of RIC_{12} > -38.81 from the PRS is 9.9×10^{-5} (21/212,102). Similarly, of the 155 physiologic

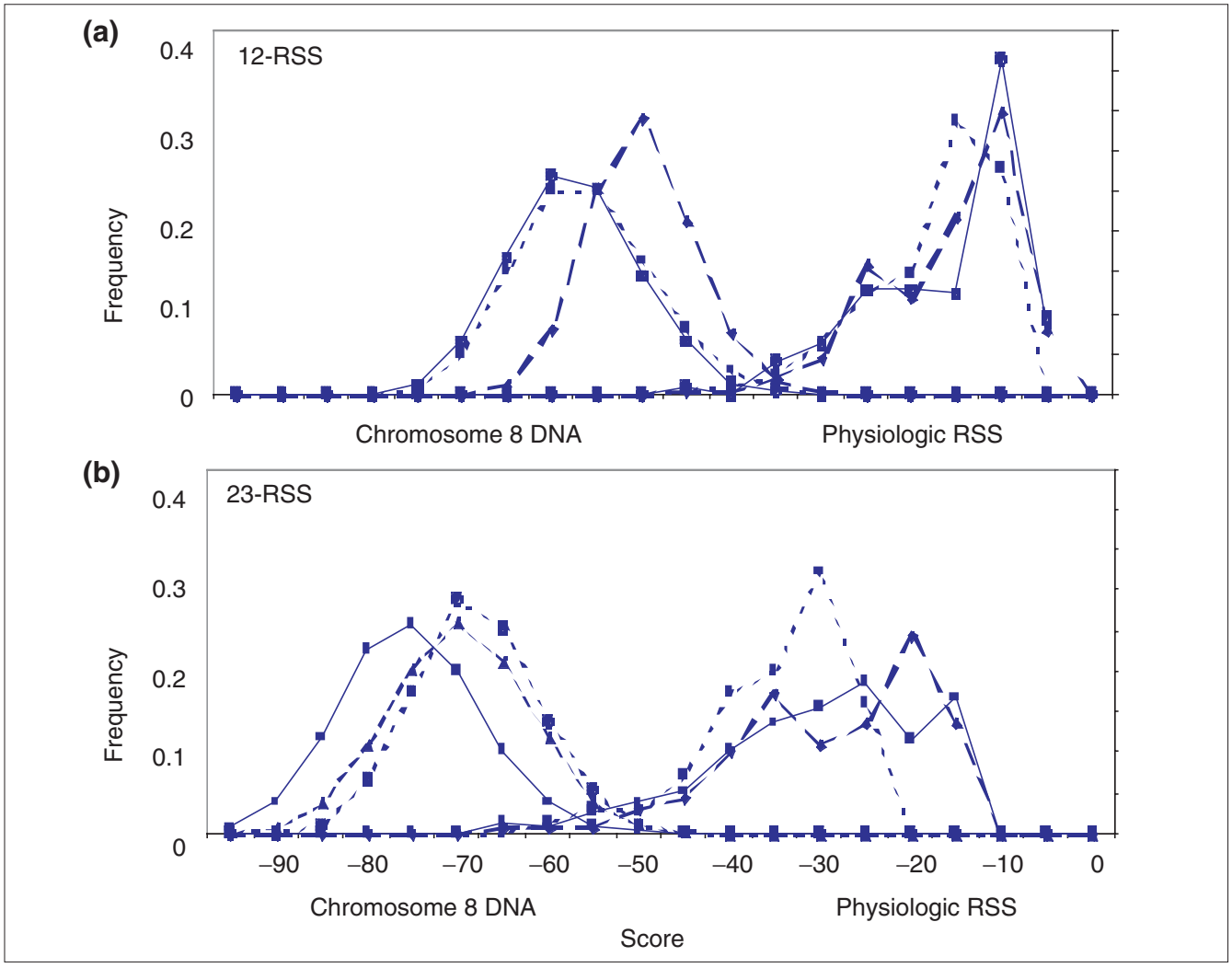

Figure 3
Plots of scores computed for physiologic RSS, and for 28- or 39-bp segments taken from chromosome 8 DNA. **(a)** 12-RSS; **(b)** 23-RSS. The *y*-axis indicates the frequency of physiologic RSS or of segments from chromosome 8 with a finite score that score at the level given on the *x*-axis. Solid line, *RIC*; dashed line, *WM1*; dotted line, *WM0*.

23-RSS, only three (0.02) have RIC_{23} < -58.45. Just 100 of the 212,090 (4.7 x 10^{-4}) 39-bp segments from sequence AC084823 and only 58 of the 212,091 segments from the PRS (2.7 x 10^{-4}) have RIC_{23} > -58.45. We therefore set -38.81 and -58.45 as RIC_{12} and RIC_{23} thresholds, respectively.

The *RIC* models reliably recognize physiologic RSS

We searched genomic DNA containing physiologic RSS to determine if *RIC* scores can resolve 12- and 23-RSS. Sequences X58411 (7,360 bp) and X58414 (5,867) encompass the mouse Jλ locus and contain four 12-RSS, three associated with functional Jλ gene segments and one associated with the pseudogene Jλ4 [54]. We computed RIC_{12} scores for 13,173 28-bp segments in X58411 and X58414; the resulting RIC_{12} distributions are almost identical to that for the chromosome 8 sequence with means well below threshold (Table 6). The RIC_{12} for RSS associated with functional Jλ gene segments lie well outside this distribution (Table 6, Figure 6a); all three score > -38.81 ($\overline{RIC_{12}}$ = -21.16). Analogous searches of mouse D_H (AF018146) and DJβ (AE000665) regions also demonstrated RIC_{12} values for physiologic RSS above threshold (Table 6). Of the 21 physiologic 12-RSS included in the search, only the RSS associated with the Jλ4 pseudogene [54] scored below threshold (Table 6).

Scans of the sequence containing Dβ and Jβ gene segments (AE000665, described above) and a 250,611-bp sequence containing 16 Vβ genes (AE000663) showed that physiologic 23-RSS are also easily resolved by their *RIC* scores (Table 6). The two Dβ 23-RSS have RIC_{23} well above threshold (-35.85 and -49.56; Table 6), and the Vβ 23-RSS associated

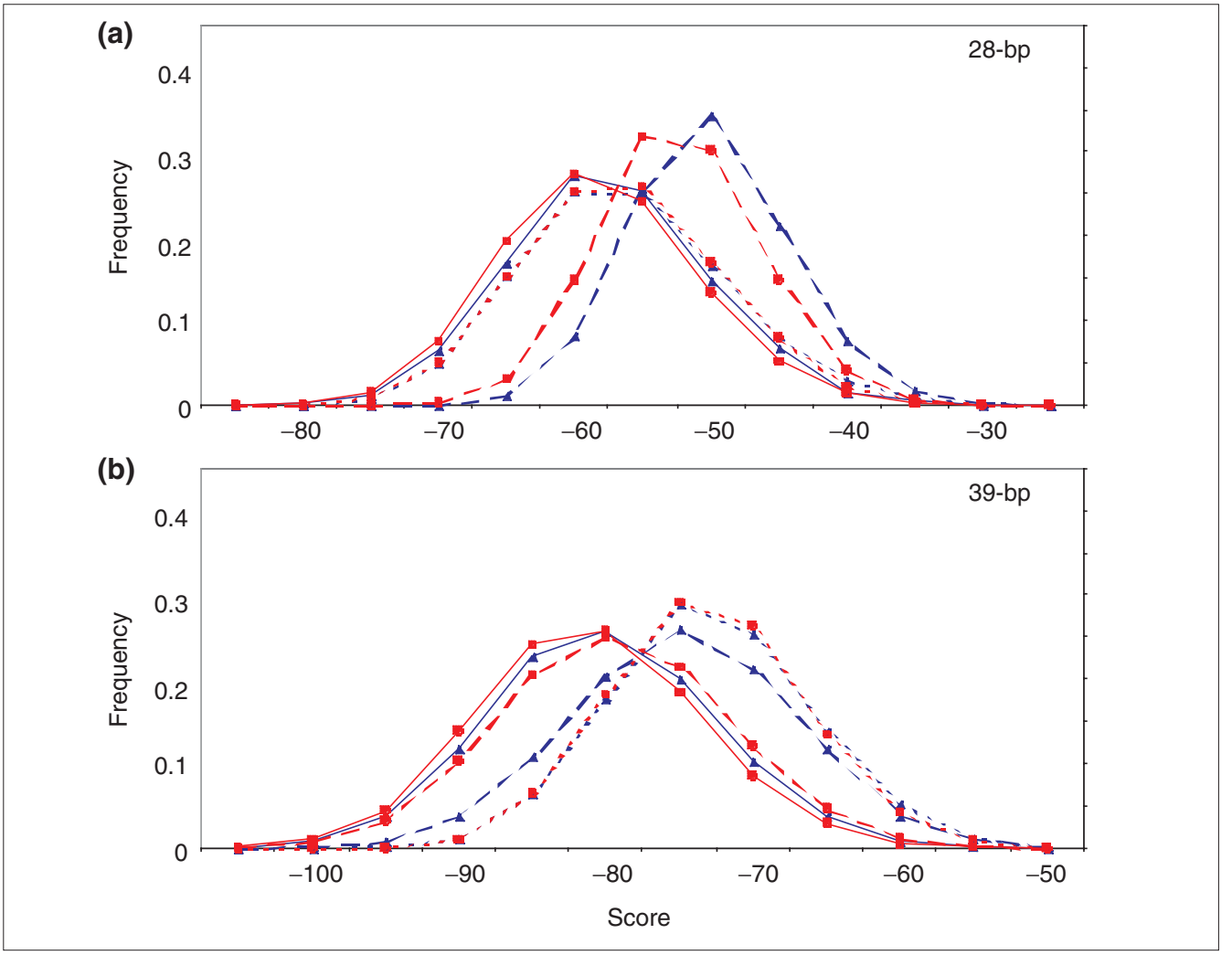

Figure 4
Score distributions for non-RSS segments. **(a)** Results for 28-bp segments (12-RSS model); **(b)** results for 39-bp segments (23-RSS model). Score level is given by the *x*-axis, and the frequency of finite scores is given by the *y*-axis. Distributions for segments of chromosome 8 DNA are shown in blue, and distributions for segments from a pseudorandom string of A, G, C and T are shown in red. Solid line, *RIC*; dashed line, *WM1*; dotted line, *WM0*.

with functional gene segments are also easily identified (mean -34.17; Table 6, Figure 6b). Again, the only RSS not scoring above threshold are associated with pseudogenes [55-57] (Vβ11 -61.78; Vβ12-3 -61.76; Vβ8 -58.87); nevertheless, all three RIC_{23} scores fall above the background mean (-77.75). RSS flanking pseudogenes cannot be stringently selected, so we expect their *RIC* to be below threshold but above background.

For comparison, we scanned the five sequences (AE000663, AE000665, AF018146, X58411 and X58414) using the *WM0* and *WM1* models. In all cases the mean score for non-RSS is lower than the mean score for physiologic RSS associated with functional gene segments, but the disparity between the two sets of scores is always greatest for the *RIC* models (Table 6), showing that discrimination between RSS and non-RSS is clearest using the *RIC* models. This wider disparity may be important when searching for functional but degenerate signals such as cryptic and pseudogene-associated RSS.

RIC_{23} scores for AE000663 are directly compared with $WM0_{23}$ or $WM1_{23}$ scores in Figure 7. The majority of scores for non-RSS fall below the $y = x$ line, indicating that they receive lower scores under the RIC_{23} model than under either of the Markov models (Figure 7). The scores for RSS associated with functional gene segments tend to fall above $y = x$, indicating their better discrimination by the RIC_{23} model (Figure 7). Of the seven pseudogene-associated RSS in AE000663, scores for three fall in the cloud of background scores under all three models whereas scores for the remaining four do not (Figure 7). Although the scores for

Table 4

Distribution of scores for non-RSS DNA and a pseudorandom string of A, G, C and T

	28-bp segments					
	Chromosome 8 DNA			Pseudorandom string		
Score level	*WM0*	*WM1*	*RIC*	*WM0*	*WM1*	*RIC*
0	0	0	0	0	0	0
-5	0	0	0	0	0	0
-10	0	0	0	0	0	0
-15	0	0	0	0	0	0
-20	0	0	0	0	0	0
-25	1	3	1	0	1	1
-30	26	36	6	8	10	3
-35	154	298	68	75	92	23
-40	521	1335	245	287	586	180
-45	1426	3730	1152	1121	2082	719
-50	2878	5978	2572	2352	4230	1856
-55	4380	4485	4429	3627	4458	3384
-60	4415	1414	4730	3552	2076	3837
-65	2669	193	2935	2144	429	2716
-70	861	7	1135	705	33	1052
-75	136	1	189	119	1	208
-80	12	0	18	8	0	19
-85	1	0	0	0	0	0
-90	0	0	0	0	0	0
-95	0	0	0	0	0	0
-100	0	0	0	0	0	0
-∞	194621	194621	194621	198104	198104	198104
Total	212101	212101	212101	212102	212102	212102
Maximum	-28.62	-25.09	-29.03	-30.04	-27.51	-31.44
Mean	-58.9	-52.64	-60.07	-59.18	-54.81	-61.13
Minimum	-85.62	-75.54	-83.66	-82.19	-75.42	-84.68

	39-bp segments					
	Chromosome 8 DNA			Pseudorandom string		
Score level	*WM0*	*WM1*	*RIC*	*WM0*	*WM1*	*RIC*
0	0	0	0	0	0	0
-5	0	0	0	0	0	0
-10	0	0	0	0	0	0
-15	0	0	0	0	0	0
-20	0	0	0	0	0	0
-25	0	0	0	0	0	0
-30	0	0	0	0	0	0
-35	0	0	0	0	0	0
-40	1	3	0	2	1	0
-45	22	26	3	10	5	1
-50	206	182	28	124	20	14
-55	919	681	148	600	162	80
-60	2433	2079	651	1876	673	409
-65	4421	3689	1785	3690	1701	1197
-70	5069	4544	3500	4079	3006	2588
-75	3073	3536	4509	2549	3520	3611
-80	1151	1907	3983	907	2880	3370
-85	170	680	2052	150	1452	1955
-90	13	124	676	8	464	617
-95	0	23	131	1	94	141
-100	0	4	12	0	18	13
-∞	194612	194612	194612	198095	198095	198095
Total	212090	212090	212090	212091	212091	212091
Maximum	-41.85	-43.13	-46.77	-42.79	-42.16	-47.65
Mean	-70.49	-72.28	-77.76	-70.73	-76.94	-78.64
Minimum	-92.37	-101.2	-103.03	-95.21	-104.23	-104.32

A pseudorandom string of A, G, C and T was generated the same length as sequence AC084823 from chromosome 8 and having the same nucleotide-usage frequencies. The 12-RSS models scored all 28-bp segments and the 23-RSS models scored all 39-bp segments. The distribution of scores is shown. The number of segments not beginning with CA is shown at score level -∞; other details as in Table 3.

these four RSS are higher under both Markov models than under the RIC_{23} model, they are better discriminated by the RIC_{23} model because the background scores are higher under the Markov models (Figure 7).

Parameters for all six models were estimated from the full set of RSS (201 12-RSS and 155 23-RSS). Of the 39 RSS contained in the search contigs, 27 of the RSS associated with functional gene segments were part of the estimation set. In contrast, the pseudogene-associated RSS (Jλ4, Jβ2.6, and seven Vβ) and three of the functional Jβ RSS were not part of the estimation set.

RIC scores identify functional cryptic RSS

12-cRSS in 3′ → 5′ orientation and embedded near the 3′ end of V gene segments can mediate receptor editing, the replacement of the V gene segment portion of a rearranged variable-region gene with an upstream V gene segment (reviewed in [58]). To test whether our model can recognize these cRSS, we computed WMO_{12}, $WM1_{12}$, and RIC_{12} scores for all 28-bp segments in 3′ → 5′ orientation in V_H gene segments known to participate in receptor editing: the V_H2S1*01 gene segment [59], the V_H14S1 gene segment [60], and the 3H9 transgene [61]. The cRSS were not part of the RSS set used for parameter estimation. While the cRSS in V_H2S1*01 and V_H14S1 have higher scores than any other 28-bp segment in their respective gene segments under all three models, the RIC_{12} and WMO_{12} models are better at identifying them because these models have lower background scores than the $WM1_{12}$ model (Table 7). The cRSS in 3H9, however, is best recognized by its RIC_{12} score (Figure 8). While it has a higher score under the *WM1* model, its RIC_{12} score is more separated from the background *RIC* scores than its $WM1_{12}$ or WMO_{12} score

Table 5

Frequency of above-threshold scores for non-RSS

			Number of physiologic RSS scoring below the indicated score										
			0	1	2	3	4	5	6	7	8	9	10
12-RSS	*RIC*	Score	-48.16	-45.81	-38.81	-38.81	-38.21	-37.77	-37.48	-36.03	-36	-35.01	-34.44
		Count	859	449	54	54	48	40	34	16	16	10	10
		Frequency	0.00414	0.00212	0.00025	0.00025	0.00023	0.00019	0.00016	0.00008	0.00008	0.00005	0.00005
	WM1	Score	-46.32	-40.87	-37.75	-37.6	-36.33	-35.29	-34.63	-33.73	-32.73	-32.73	-32.69
		Count	2372	458	169	161	76	43	35	22	12	12	12
		Frequency	0.01118	0.00216	0.00080	0.00076	0.00036	0.00020	0.00017	0.00010	0.00006	0.00006	0.00006
	WM0	Score	-46.86	-44.61	-42.83	-39.33	-39.06	-38.58	-38.58	-35.23	-34.59	-34.45	-34.21
		Count	1067	632	405	115	99	83	83	30	19	17	15
		Frequency	0.00503	0.00298	0.00191	0.00054	0.00047	0.00039	0.00039	0.00014	0.00009	0.00008	0.00007
23-RSS	*RIC*	Score	-69.69	-66.37	-64.84	-58.45	-57.45	-57.41	-55.61	-54.35	-54.3		
		Count	2460	1147	791	100	71	69	37	26	26		
		Frequency	0.01160	0.00541	0.00373	0.00047	0.00033	0.00033	0.00017	0.00012	0.00012		
	WM1	Score	-65.24	-63.84	-57.58	-54.55	-54.52	-54.16	-52.96	-52.55	-48.5		
		Count	3126	2309	471	190	189	167	109	97	14		
		Frequency	0.01474	0.01089	0.00222	0.00090	0.00089	0.00079	0.00051	0.00046	0.00007		
	WM0	Score	-63.22	-61.52	-58.9	-58.74	-56.46	-56.46	-55.91	-52.41	-52.17		
		Count	2481	1691	816	777	365	365	317	82	72		
		Frequency	0.01170	0.00797	0.00385	0.00366	0.00172	0.00172	0.00149	0.00039	0.00034		

Scores for physiologic RSS were ranked in ascending order so that the lowest score was ranked 0, the second lowest score was ranked 1, and so on. For each score, the number of physiologic RSS with a lower score is equal to this rank. We took the lowest 5% of scores under each model as the test thresholds for that model: ranks 0 through 10 were used for the three 12-RSS models and ranks 0 through 8 were used for the three 23-RSS models. For each of these threshold scores, we counted the number of non-RSS in the chromosome 8 sequence AC084823 that scored above the threshold. The threshold score, the count, and its relative frequency are shown.

(Figure 8). All receptor-editing events in 3H9 involved the cRSS identified by *RIC* [61].

RIC scores are correlated with RSS recombination efficiencies

To quantify any correlation between *RIC* and RSS function, we computed Spearman's rank correlation coefficient (r_S) between *RIC* and published recombination frequencies [46]. The correlation coefficients are r_S = 0.81 for 12-RSS (Figure 9) and r_S = 0.86 for 23-RSS (Figure 10); for these RSS, *RIC* explains 67-74% of the variation in recombination efficiency. *WMO* and *WM1* are equally well correlated with recombination efficiency: WMO_{12}, r_S = 0.80; $WM1_{12}$, r_S = 0.82, WMO_{23}, r_S = 0.88, $WM1_{23}$, r_S = 0.91. Only 1 of 27 12-RSS and 1 of 13 23-RSS included in this study [46] are part of our parameter estimation set.

Discussion

We developed models of the correlation structure in 12- and 23-RSS using a model selection procedure that finds the models with the most power to predict the population of functional RSS. The procedure determines the groups of positions such that the predictive power of the model is increased by including the correlation between the positions within each group. Positions not correlated with any other position are included in the models by the probability distribution over the four nucleotides at that position; groups of correlated positions are included by the probability distribution for the set of motifs prescribed by the number of positions within the group; for example, a group of four positions would be modeled by the probability distribution for the 256 possible quadruplet motifs. The models compute an RSS information content score, *RIC*, for any RSS-length sequence, that is, 28-bp segments are scored by the 12-RSS model and 39-bp segments are scored by the 23-RSS model.

Model selection evaluates all possible combinations of probability distributions in a stepwise fashion, so the correlation structure of RSS is not specified nor assumed, but rather detected. Interestingly, the positions exhibiting the strongest correlation overlap substantially with the RSS nucleotides that contact the recombinase (Figure 2 and [53]), offering support for the hypothesis that positions acting cooperatively to influence binding by the recombinase coevolve. This overlap also suggests that the correlation structure detected by our model selection procedure is relevant to RSS function.

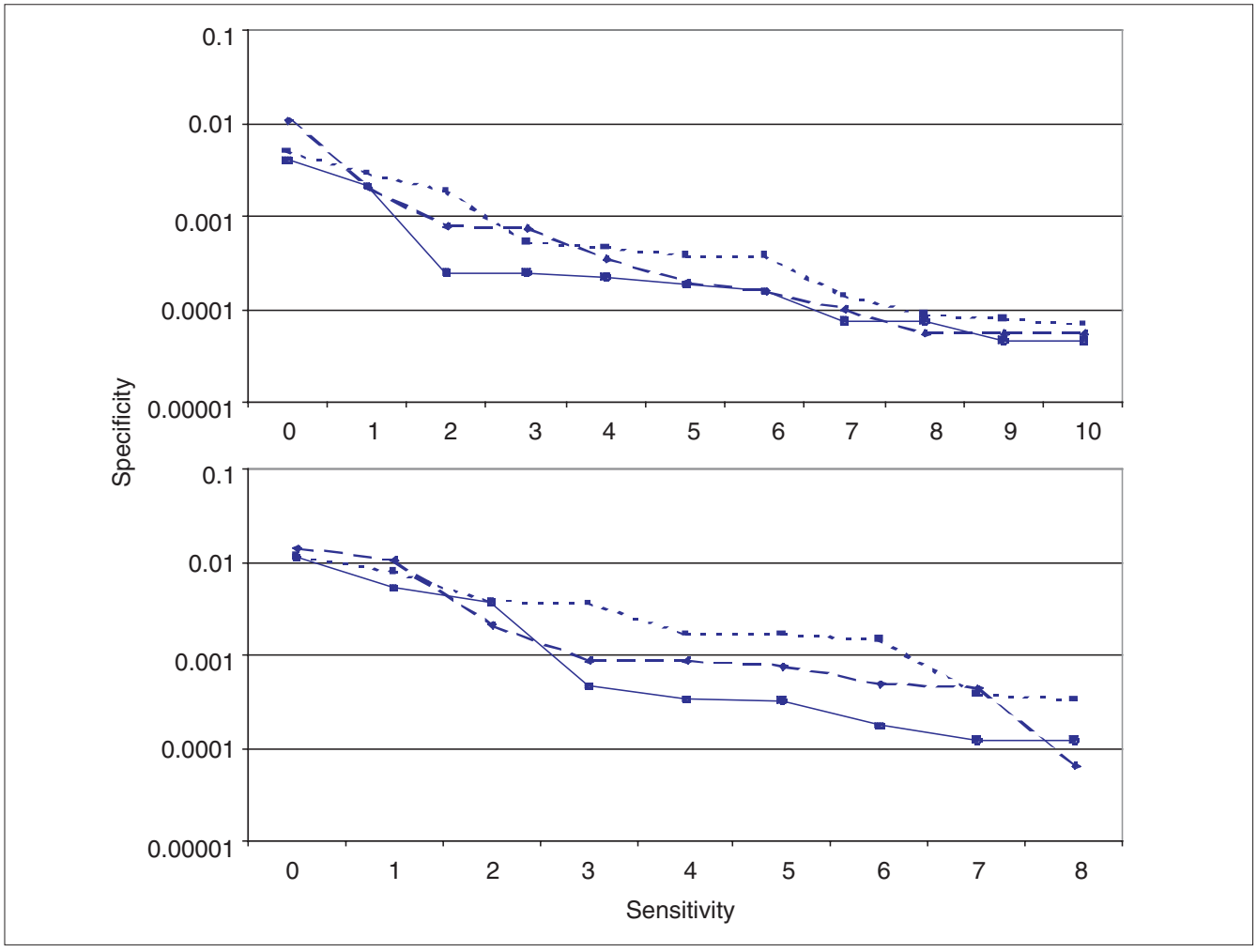

Figure 5
Specificity of the models as a function of their sensitivity. Scores for physiologic RSS were ranked in ascending order so that the lowest score was ranked 0, the second lowest score was ranked 1, and so on. For each score, the number of physiologic RSS with a lower score is equal to this rank. We used the lowest 5% of scores as test thresholds: ranks 0 through 10 were used for the three 12-RSS models and ranks 0 through 8 were used for the three 23-RSS models. **(a)** Results for the 12-RSS models; **(b)** results for the 23-RSS models. The number of physiologic RSS scoring below a given threshold score (Table 5) is given on the *x*-axis, and the frequency of non-RSS segments in chromosome 8 sequence AC084823 scoring above the threshold is shown on the *y*-axis. Solid line, *RIC*; dashed line, *WM1*; dotted line, *WM0*.

To determine if modeling the correlation structure specific to RSS improves RSS recognition and evaluation, we compared the *RIC* models with models that assume RSS positions are independent (order zero Markov models, *WM0*) and models that assume adjacent positions are correlated (order one Markov models, *WM1*). In general we find that, while all models predict RSS function equally well, with Spearman's rank correlation coefficients around 0.81 for 12-RSS and 0.88 for 23-RSS, the *RIC* models are better at discriminating between RSS and non-RSS segments. This is especially true for the degenerate relatives of RSS - cRSS and pseudogene-associated RSS.

Under each model, we compared the score distribution for physiologic RSS with that for non-RSS segments. We find that the two distributions are most disparate under the *RIC* models (Figure 3). While including correlation in the models increases the scores for most physiologic RSS (Figure 3), assuming adjacent positions are correlated increases scores nonspecifically, raising the background scores (Figures 3, 4). The distribution of scores for non-RSS was approximated by computing the score for all RSS-length segments in a 200-kb region of chromosome 8 containing no known RSS and also in a pseudorandom string of A, G, C and T characters the same length as the chromosome 8 region and having the same nucleotide-usage frequencies.

We determined threshold scores to discriminate between functional RSS and non-RSS. The lowest 5% of scores for physiologic RSS under each model were used as test

Table 6

Mean scores for physiologic RSS and non-RSS segments in genomic sequence

				Mean scores					
				WM0		*WM1*		*RIC*	
Accession	Length (bp)	RSS type	Number of RSS	Non-RSS	RSS	Non-RSS	RSS	Non-RSS	RSS
AE000665	199101	Dβ, Jβ 12-RSS	15	-58.36	-32.13	-53.16	-27.73	-59.87	-28.48
AF018146	3926	D_H 12-RSS	2	-59.05	-18.4	-52.84	-17.41	-60.08	-14.58
X58411	7360	Jλ1, Jλ3, 12-RSS	2	-57.63	-23.04	-52.51	-22.33	-58.91	-20.46
X58414	5867	Jλ2, Jλ4, 12-RSS	2*	-58.67	-26.01	-53.04	-26.91	-60.04	-22.58
AE000663	250611	Vβ 23-RSS	16†	-70.77	-39.88	-73.42	-34.6	-77.75	-34.17
AE000665	199101	Dβ, Jβ 23-RSS	2	-70.8	-35.21	-73.2	-37.23	-77.81	-38.64

Location of physiologic RSS by their *RIC* score. Scores were computed for all 28- and 39-bp segments in mouse sequences containing physiologic RSS. The sequence accession number, its length and the physiologic RSS it contains are shown. The mean score for all non-RSS in the sequence is compared to the mean score for physiologic RSS associated with functional gene segments. *The Jλ4 gene segment is a pseudogene. †Seven of the 16 Vβ gene segments are pseudogenes.

thresholds; for 17 of 20 thresholds, the frequency of non-RSS scoring above threshold was lowest under the *RIC* models (Figure 5). We selected threshold *RIC* scores (-38.81 and -58.45 for 12- and 23-RSS respectively), at which more than 98% of physiologic RSS score above threshold, and the predicted frequency of functional signals in the mouse genome is on the order of 10^{-4} (Table 6). Many of the signals identified by the models may mediate recombination, and so this frequency overestimates the false-positive rate. The false-positive rate can only be estimated by testing a random sample of these putative cRSS for function.

We show that the threshold scores can be effectively used to screen genomic DNA for functional RSS. We screened over 650 kb of genomic DNA containing 39 physiologic RSS (Table 6), 12 of which are not part of the RSS set used for parameter estimation. Under all models, the mean score for non-RSS is lower than the mean score for physiologic RSS associated with functional gene segments, but the disparity between the two sets of scores is always greatest for the *RIC* models (Table 6). Only four RSS, one 12-RSS and three 23-RSS have *RIC* scores below their respective threshold scores, and all four are associated with pseudogenes. The remaining four pseudogene-associated RSS included in our search are better recognized by *RIC* than by *WM0* or *WM1* (Figure 7). Furthermore, we computed the RIC_{12} scores for all potential $3' \rightarrow 5'$ oriented 12-cRSS in three V_H gene segments observed to mediate receptor editing. Importantly, cRSS were not included in the estimation set. Two cRSS are recognized by all three models, but the third is recognized only by *RIC* (Figure 8). We expect cRSS and pseudogene-associated RSS to be under less stringent selection pressure than physiologic RSS and therefore to have lost some of their sequence similarity to RSS. An important future test of the *RIC* models will be to scan genomic DNA prospectively for cRSS and show that the cRSS identified mediate recombination by the V(D)J recombinase.

We have successfully used the model selection procedure introduced here to develop models for two sets of highly variable and complex binding sites. An advantage of this procedure is that the correlation structure of the site is determined from the statistical properties of the sequence set alone; therefore, the sequence properties governing recognition of the site by the binding protein(s) can be identified in the absence of experimental evaluation of function. Often there is an abundance of sequence data that has not yet been, or perhaps cannot be, experimentally evaluated. Information about the correlation structure of a binding site can give important insight into how the binding site and binding protein(s) interact, and may suggest important experiments. In addition, by choosing the positions to be correlated independently of their relative positions, we are able to model highly complex patterns with a relatively reduced increase in model complexity. Inclusion of highly complex correlation patterns in models of DNA-sequence sites can improve the precision with which sites are identified and functional levels are predicted.

Materials and methods

RSS sequence set

We analyzed 356 physiologic mouse RSS from all *Tcr* and *Ig* loci available [62]. About 96% (340/356) of the RSS are associated with functional V, D or J gene segments (Immunogenetics database [57,63,64]). Two are associated with pseudogenes known to rearrange, and two are associated with open reading frames (ORFs) also known to rearrange (see Immunogenetics database [57,63,64]). The ability of the remaining RSS to rearrange is unconfirmed.

Calculation of position-wise entropy and mutual information

For a DNA sequence alignment, the position-wise entropy H_i [65] measures the level of nucleotide diversity: minimum H is

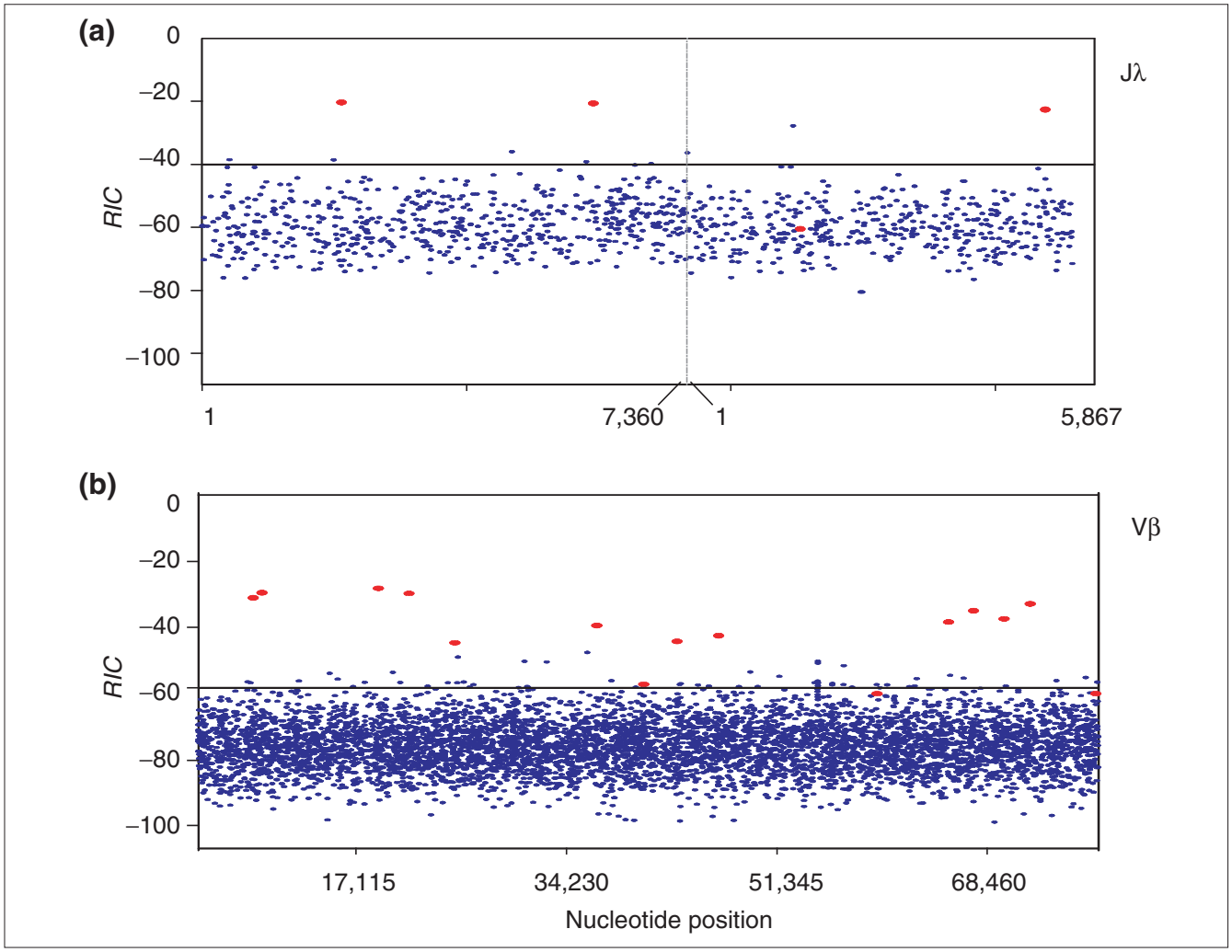

Figure 6
RIC values. **(a)** RIC_{12} values for 28-bp segments from genomic regions containing Jλ gene segments. RIC_{12} values are given on the *y*-axis, and position in the sequence is shown on the *x*-axis. Each blue or red point represents the RIC_{12} for the 28-bp segment beginning at that position. Points corresponding to physiologic RSS are red. The horizontal black line indicates *RIC* = -40. **(b)** RIC_{23} values for 39-bp segments from genomic regions containing Vβ gene segments. The horizontal black line indicates *RIC* = -60. Other details as in (a).

0 and indicates strict sequence conservation; maximum H is 1 and indicates that the four nucleotides occur at nearly uniform frequencies. The estimated entropy at the ith position in an alignment is given by $H_i = -\sum_j p_{i,j} \log_4 p_{i,j}$ where $p_{i,j}$ is the estimated probability of nucleotide j at position i.

The estimated mutual information (MI) [51] between two positions, i and i', is computed as: $MI_{i,i'} = H_i + H_{i'} - H_{i,i'}$, where H_i is the estimated entropy computed from the frequency of the four nucleotides at position i and $H_{i,i'}$ is the entropy computed using the frequency of the 16 pairs of nucleotides at the two positions.

Nucleotide and nucleotide pair probabilities are themselves estimated using the Bayesian posterior mean [66]

$$\hat{P}_i(s) = \frac{m_{i,s} + {}^{2}/_{r}}{N_i + 2},$$

where $m_{i,s}$ is the frequency of nucleotide or nucleotide pair s in position(s) i of the alignment, N_i is the total number of sequences in the alignment at position(s) i, and r is the number of distinct classes in the probability distribution (for single nucleotides r = 4, for nucleotide pairs, r = 16, and so on).

Under the null hypothesis of no correlation between nucleotide positions in RSS, positive $MI_{i,i'}$ values may still be observed. To test if the $MI_{i,i'}$ values differ significantly from 0, we randomized the order of the sequences in position i' and recomputed $MI_{i,i'}$. By randomizing the order of the

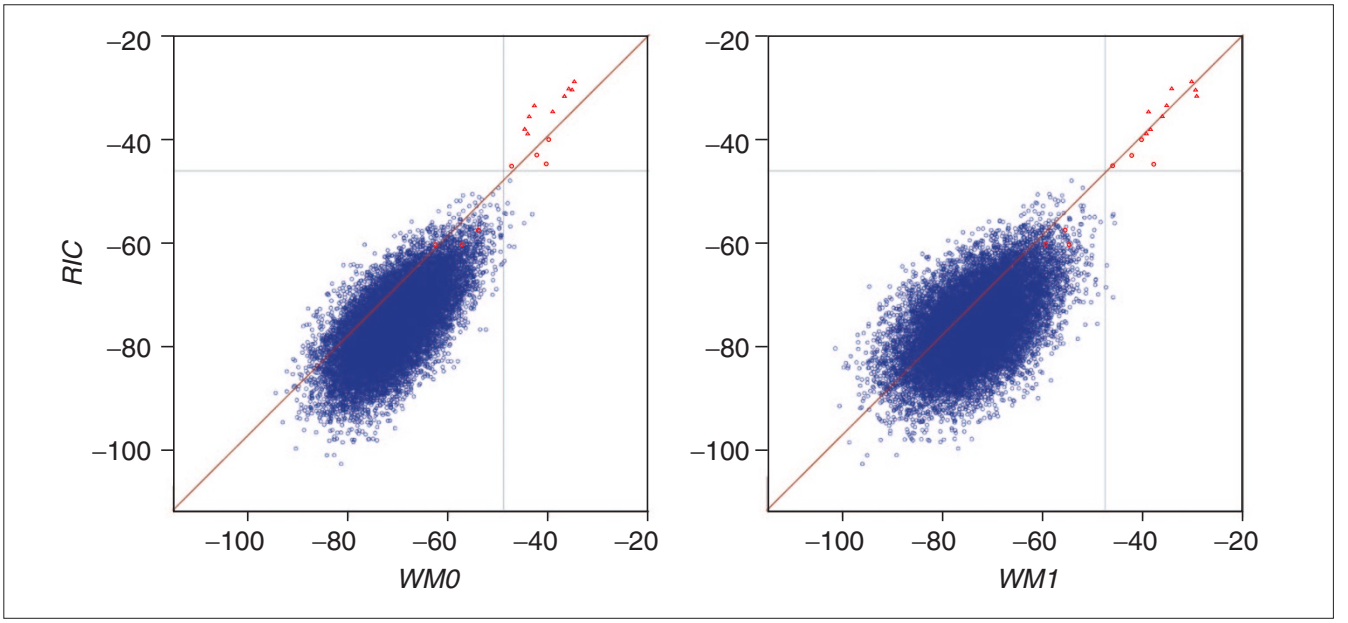

Figure 7
RIC models are better at recognizing pseudogene-associated RSS. The scores for sequence AE000663 under the RIC_{23} model are plotted against the scores under the $WM0_{23}$ model (left panel) or the $WM1_{23}$ model (right panel). RIC_{23} scores are shown on the *y*-axis, and $WM1_{23}$ or $WM0_{23}$ scores are shown on the *x*-axis. The $y = x$ line is shown in red. AE000663 contains seven Vβ pseudogenes and nine functional Vβ gene segments. Blue points indicate scores for non-RSS, red points indicate scores for pseudogene-associated RSS, and the red triangles indicate scores for RSS associated with functional gene segments. The horizontal gray line is halfway between the lowest RIC_{23} score for a physiologic RSS distinguishable from non-RSS scores (-46.08.) and the highest RIC_{23} score for a non-RSS (-49.09): -46.08 - 1.505 = -47.585 = -49.09 + 1.505. The vertical gray line is the same distance below the corresponding $WM1_{23}$ or $WM0_{23}$ score: -46.02 - 1.505 or -47.52 - 1.505, respectively.

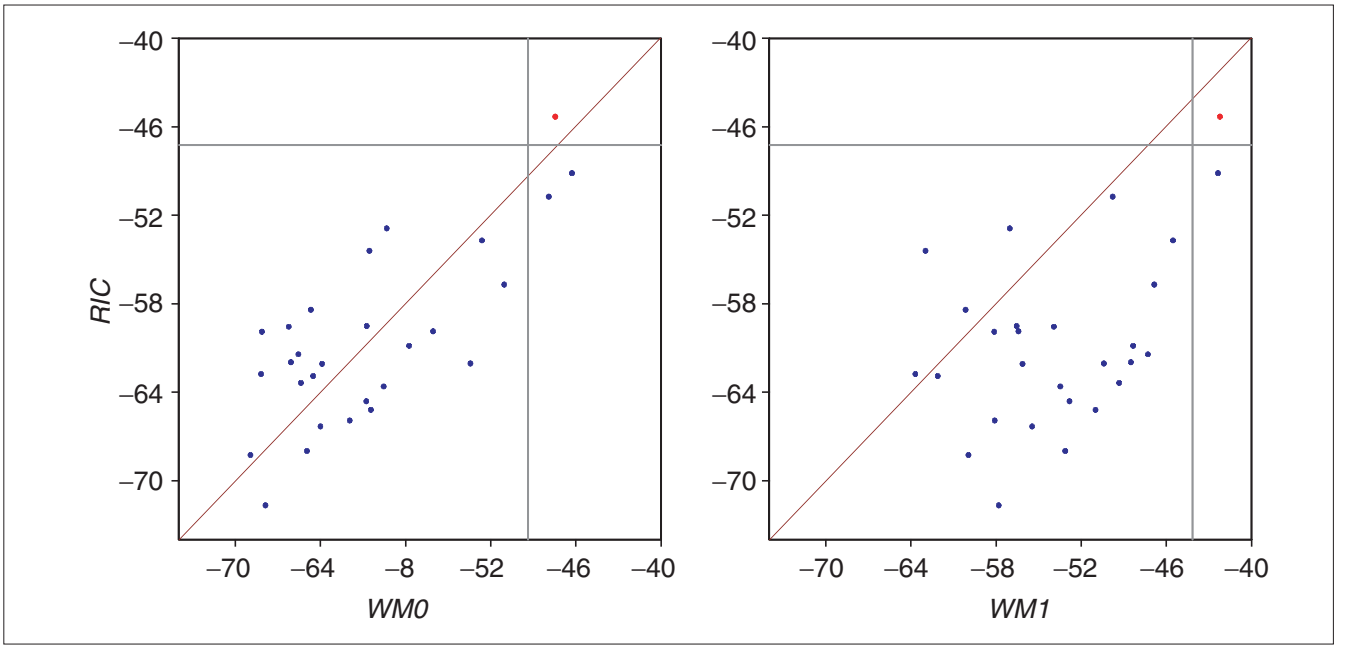

Figure 8
Low-scoring cRSS are better recognized by RIC_{12} than by $WM0_{12}$ or $WM1_{12}$. RIC_{12} scores for 28-bp segments in 3′ → 5′ orientation in the 3H9 transgene [62] are plotted against $WM0_{12}$ or $WM1_{12}$ scores for the same segments. The cRSS score is shown in red, and the non-RSS scores are shown in blue. The horizontal gray line is at -47.23, halfway between the cRSS RIC_{12} score (-45.32) and the highest scoring non-RSS (-49.15). The vertical gray line is the same distance below the cRSS $WM0_{12}$ (-47.46) or $WM1_{12}$ (-42.23) score. Other details as in Figure 7.

sequences in one position but not the other, we preserve both marginal distributions but disrupt any correlations. The proportion of 300 permutations giving $MI_{i,i'}$ values higher than that observed in the data is our estimate for the *p* value under the null hypothesis.

Statistical models of RSS structure

Single models for 12-RSS and 23-RSS define the set of probability distributions that contain the most information about RSS structure. We developed the models by stepwise model enlargement and selection. We begin with the smallest models, order zero Markov models estimated under the assumption of pairwise site independence. These models take as the probability of observing a sequence *S*, the product of the individual marginal probabilities over the RSS positions, $P(S) = \prod_i P_i(s)$ where *s* is the nucleotide observed in sequence *S* at position *i*. The marginal probability distribution for a position defines the probability of observing each of the four nucleotides at that position and is estimated from the RSS dataset using the Bayesian estimator for $P_i(s)$ described above. Each potential model enlargement is accomplished by replacing the product of an independent marginal probability and *k*-variate joint probability by a corresponding *k*+1-variate joint probability.

For example, the first step of model enlargement for RSS of length *N* compares all possible combinations of one joint probability distribution for two positions and marginal probability distributions for the remaining *N*-2 positions: for every pair of positions *i* and *i'*, there is the model

$$P(S) = P(s_i, s_{i'}) \prod_{j \neq i,i'} P_j(s).$$

Under each of these models we compute scores by taking the natural logarithm of *P*(*S*) (ln *P*(*S*)) for each RSS in the dataset using leave-one-out cross-validation [67]. Briefly, for each model, we exclude one RSS from the dataset, estimate the probability distributions used in the model from the remaining RSS, and compute ln *P*(*S*) for the excluded RSS. We perform this computation for each RSS and take the average ln *P*(*S*) over all RSS. We then select the combination of probability distributions giving the largest average ln *P*(*S*).

The second step of model enlargement expands the model selected in step one by comparing all models within two classes: those based on two joint probability distributions, the one formed in step one and one for an additional pair of positions, and marginal probability distributions for the remaining *N* - 4 positions, and those based on one joint probability distribution for a group of three positions, the pair joint selected in step one expanded to include one additional position, and marginal probability distributions for the remaining *N* - 3 positions. We again compute ln *P*(*S*) under each model for each RSS using leave-one-out cross-validation and select the model giving the largest average ln *P*(*S*).

Table 7

Location of cRSS by their *RIC* score

	WM0	*WM1*	*RIC*
V_H2S1*01	-42.02	-36.12	-39.86
V_H14S1	-42.75	-40.78	-41.05
Background mean	-61.15	-52.28	-60.22

cRSS in the V_H2S1*01 and V_H14S1 gene segments can be identified by their scores. The score under all three 12-RSS models is shown for the cRSS in V_H2S1*01 and V_H14S1 which are known to undergo receptor editing. For comparison, the mean score for all other 28-bp segments in the two gene segments is also given.

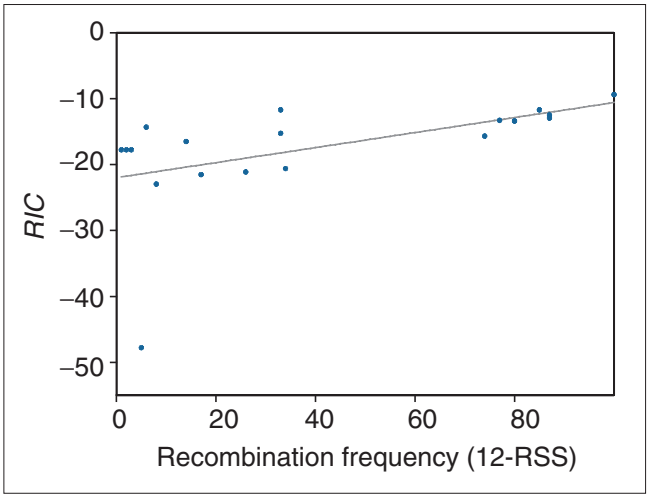

Figure 9
Correlation between recombination efficiency (*x*-axis) as measured by Hesse *et al.* [46] and *RIC* (*y*-axis) for 12-RSS.

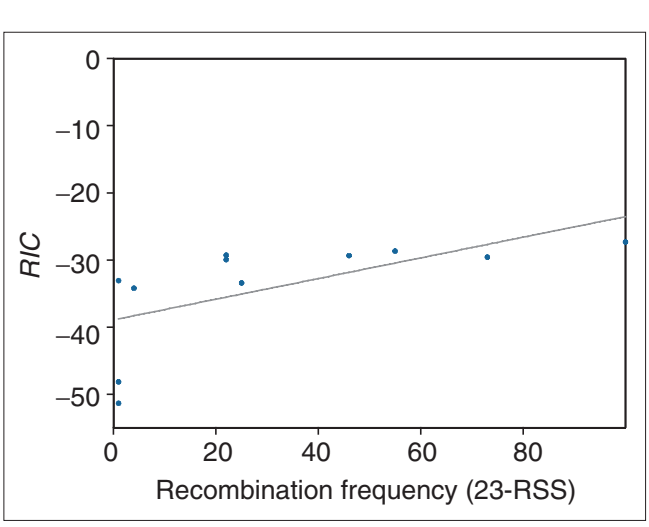

Figure 10
Correlation between recombination efficiency (*x*-axis) as measured by Hesse *et al.* [46] and *RIC* (*y*-axis) for 23-RSS.

We continued the step-wise model enlargement and selection, iteratively expanding the models until the mean ln $P(S)$ ceased improving. The final model is not necessarily optimal, however; we perform the maximization in a stepwise manner and so, as in any stepwise statistical procedure, are not guaranteed to achieve the global maximum.

Once the final set of probability distributions has been selected (the mean ln $P(S)$ no longer improves), the parameters of each probability distribution are estimated on the complete set of RSS. ln $P(S)$ for an RSS is a value between $-\infty$ and 0. If RSS were strictly conserved, the consensus RSS would have ln $P(S) = 0$. We define the ln $P(S)$ of a sequence computed from the final model as its RSS information content (*RIC*). The final models take the form:

$$RIC_{12} = \ln [P_1\ P_2\ P_{3,15,25}\ P_{4,5}\ P_{6,28}\ P_{7,8,19}\ P_{9,26}\ P_{10,12}\ P_{11,27}\ P_{13,14,23}\ P_{16,17,18}\ P_{20,21,22}\ P_{24}]$$

and

$$RIC_{23} = \ln [P_1\ P_2\ P_3\ P_{4,14}\ P_{5,39}\ P_6\ P_{7,24,25}\ P_{8,9,21}\ P_{10,16}\ P_{11,12}\ P_{13,22}\ P_{15,23}\ P_{17,18}\ P_{19,27,30,31,32,33,37}\ P_{20,26}\ P_{28,29}\ P_{34,38}\ P_{35,36}].$$

The CA dinucleotide at positions 1 and 2 of the heptamer is required for rearrangement [46,52,68]. Therefore, the models assign a probability of 0 to any RSS not beginning with CA.

RSS in the dataset receive higher *RIC* values (and *WM0* and *WM1* values, see below) during genome searches than during model development, because the sequence for which *RIC* is calculated is excluded from the dataset during model development, a property of leave-one-out cross-validation [67]. In contrast, nucleotide frequencies used in the searches are estimated on the complete set of RSS. The difference between the two *RIC* values is greatest for rare sequences and, in general, the differences are larger for 23-RSS. The difference between the two scores is < 1 for 78% of 12-RSS and for 39% of 23-RSS.

Markov models

We also modeled RSS with weight matrices based on an order zero or order one Markov model (reviewed in [27]). The order zero Markov model assumes that all RSS positions are independent, so the probability of observing a sequence S that is N-bp long is:

$$P(S) = \prod_{i=1}^{N} P_i(s)$$

where $P_i(s)$ is the probability of observing nucleotide s at RSS position i. $P_i(s)$ is estimated as described above from the set of physiologic 12- or 23-RSS. The score for sequence S is then the natural logarithm of the probability $P(S)$,

$$WM0_N = \sum_{i=1}^{N} \ln P_i(s).$$

The order one Markov model assumes that all adjacent positions are correlated, so the probability of observing sequence S is based on conditional probabilities:

$$P(S) = P_1(s) \prod_{i=2}^{N} P(s_i|s_{i-1})$$

where $P(s_i|s_{i-1})$ is the probability of observing nucleotide s_i in position i given that nucleotide s_{i-1} occupies position i-1. From Bayes rule,

$$P(s_i|s_{i-1}) = \frac{P(s_{i-1},s_i)}{P_{i-1}(s)}$$

where $P(s_{i-1},s_i)$ is the joint probability distribution for the dinucleotide $s_{i-1}s_i$ occurring at the pair of positions i-1 and i. Again, the probability distributions are estimated from the physiologic 12- or 23-RSS, and we take the natural logarithm of the probability $P(S)$ as the score for sequence S:

$$WM1_N = \ln P_1(s) + \sum_{i=2}^{N} [\ln P(s_{i-1},s_i) - \ln P_{i-1}(s)].$$

Acknowledgements

This work was supported in part by grants AI24335 and AI49326 (G.K.). L.G.C. received a Bioinformatics and Genome Technology Postdoctoral Fellowship from Duke University.

References

1. Mitchison A: **Partitioning of genetic variation between regulatory and coding gene segments: the predominance of software variation in genes encoding introvert proteins.** *Immunogenetics* 1997, **46:**46-52.
2. Purugganan MD: **The molecular population genetics of regulatory genes.** *Mol Ecol* 2000, **9:**1451-1461.
3. Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG: **Transcription Regulatory Regions Database (TRRD): its status in 2002.** *Nucleic Acids Res* 2002, **30:**312-317.
4. Zhu J, Zhang MQ: **SCPD: a promoter database of the *yeast Saccharomyces cerevisiae*.** *Bioinformatics* 1999, **15:**607-611.
5. Vanet A, Marsan L, Sagot MF: **Promoter sequences and algorithmical methods for identifying them.** *Res Microbiol* 1999, **150:**779-799.
6. Perier RC, Praz V, Junier T, Bonnard C, Bucher P: **The eukaryotic promoter database (EPD).** *Nucleic Acids Res* 2000, **28:**302-303.
7. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Res* 2000, **28:**316-319.
8. Bucher P: **Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.** *J Mol Biol* 1990, **212:**563-578.
9. Walker M, Pavlovic V, Kasif S: **A comparative genomic method for computational identification of prokaryotic translation initiation sites.** *Nucleic Acids Res* 2002, **30:**3181-3191.
10. Day WH, McMorris FR: **Critical comparison of consensus methods for molecular sequences.** *Nucleic Acids Res* 1992, **20:**1093-1099.
11. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: **Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*.** *Nucleic Acids Res* 1982, **10:**2997-3011.
12. Harr R, Haggstrom M, Gustafsson P: **Search algorithm for pattern match analysis of nucleic acid sequences.** *Nucleic Acids Res* 1983, **11:**2943-2957.
13. Stormo GD: **Consensus patterns in DNA.** *Methods Enzymol* 1990, **183:**211-221.

14. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
15. Stormo GD, Schneider TD, Gold L: **Quantitative analysis of the relationship between nucleotide sequence and functional activity.** *Nucleic Acids Res* 1986, **14:**6661-6679.
16. Berg OG, von Hippel PH: **Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.** *J Mol Biol* 1987, **193:**723-750.
17. Fickett JW: **Quantitative discrimination of MEF2 sites.** *Mol Cell Biol* 1996, **16:**437-441.
18. Lustig B, Jernigan RL: **Consistencies of individual DNA base-amino acid interactions in structures and sequences.** *Nucleic Acids Res* 1995, **23:**4707-4711.
19. Roulet E, Fisch I, Junier T, Bucher P, Mermod N: **Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA.** *In Silico Biol* 1998, **1:**21-28.
20. Roulet E, Bucher P, Schneider R, Wingender E, Dusserre Y, Werner T, Mermod N: **Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites.** *J Mol Biol* 2000, **297:**833-848.
21. Stormo GD, Strobl S, Yoshioka M, Lee JS: **Specificity of the Mnt protein. Independent effects of mutations at different positions in the operator.** *J Mol Biol* 1993, **229:**821-826.
22. Takeda Y, Sarai A, Rivera VM: **Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments.** *Proc Natl Acad Sci USA* 1989, **86:**439-443.
23. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266:**231-245.
24. Bulyk ML, Johnson PL, Church GM: **Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.** *Nucleic Acids Res* 2002, **30:**1255-1261.
25. Salzberg SL: **A method for identifying splice sites and translational start sites in eukaryotic mRNA.** *Comput Appl Biosci* 1997, **13:**365-376.
26. Zhang MQ, Marr TG: **A weight array method for splicing signal analysis.** *Comput Appl Biosci* 1993, **9:**499-509.
27. Burge CB: **Modeling dependencies in pre-mRNA splicing signals.** In *Computational Methods in Molecular Biology*. Edited by Salzberg SL, Searls DB Kasif S. New York: Elsevier; 1998: 371.
28. Ponomarenko MP, Ponomarenko JV, Frolov AS, Podkolodnaya OA, Vorobyev DG, Kolchanov NA, Overton GC: **Oligonucleotide frequency matrices addressed to recognizing functional DNA sites.** *Bioinformatics* 1999, **15:**631-643.
29. Zazopoulos E, Lalli E, Stocco DM, Sassone-Corsi P: **DNA binding and transcriptional repression by DAX-1 blocks steroidogenesis.** *Nature* 1997, **390:**311-315.
30. Bianchi ME, Beltrame M, Paonessa G: **Specific recognition of cruciform DNA by nuclear protein HMG1.** *Science* 1989, **243:**1056-1059.
31. Robbe K, Bonnefoy E: **Non-B-DNA structures on the interferon-beta promoter?** *Biochimie* 1998, **80:**665-671.
32. Wadkins RM: **Targeting DNA secondary structures.** *Curr Med Chem* 2000, **7:**1-15.
33. Sadofsky MJ: **The RAG proteins in V(D)J recombination: more than just a nuclease.** *Nucleic Acids Res* 2001, **29:**1399-1409.
34. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
35. Cai D, Delcher A, Kao B, Kasif S: **Modeling splice sites with Bayes networks.** *Bioinformatics* 2000, **16:**152-158.
36. Zinkernagel RM, Hengartner H: **Regulation of the immune response by antigen.** *Science* 2001, **293:**251-253.
37. Janeway C, Travers P, Walport M, Shlomchik M: *Immunobiology: The Immune System in Health and Disease*. New York: Garland; 2001.
38. Radic MZ, Weigert M: **Origins of anti-DNA antibodies and their implications for B-cell tolerance.** *Ann NY Acad Sci* 1995, **764:**384-396.
39. Sprent J, Kishimoto H: **T cell tolerance and the thymus.** *Ann NY Acad Sci* 1998, **841:**236-245.
40. Davila M, Foster S, Kelsoe G, Yang K: **A role for secondary V(D)J recombination in oncogenic chromosomal translocations?** *Adv Cancer Res* 2001, **81:**61-92.
41. Fugmann SD, Lee AI, Shockett PE, Villey IJ, Schatz DG: **The RAG proteins and V(D)J recombination: complexes, ends, and transposition.** *Annu Rev Immunol* 2000, **18:**495-527.
42. Difilippantonio MJ, McMahan CJ, Eastman QM, Spanopoulou E, Schatz DG: **RAG1 mediates signal sequence recognition and recruitment of RAG2 in V(D)J recombination.** *Cell* 1996, **87:**253-262.
43. Sakano H, Huppi K, Heinrich G, Tonegawa S: **Sequences at the somatic recombination sites of immunoglobulin light-chain genes.** *Nature* 1979, **280:**288-294.
44. Tonegawa S: **Somatic generation of antibody diversity.** *Nature* 1983, **302:**575-581.
45. Lewis SM: **The mechanism of V(D)J joining: lessons from molecular, immunological, and comparative analyses.** *Adv Immunol* 1994, **56:**27-150.
46. Hesse JE, Lieber MR, Mizuuchi K, Gellert M: **V(D)J recombination: a functional definition of the joining signals.** *Genes Dev* 1989, **3:**1053-1061.
47. Feeney AJ, Tang A, Ogwaro KM: **B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization.** *Immunol Rev* 2000, **175:**59-69.
48. Livak F, Petrie HT: **Somatic generation of antigen-receptor diversity: a reprise.** *Trends Immunol* 2001, **22:**608-612.
49. Lewis SM, Agard E, Suh S, Czyzyk L: **Cryptic signals and the fidelity of V(D)J joining.** *Mol Cell Biol* 1997, **17:**3125-3136.
50. Marculescu R, Le T, Simon P, Jaeger U, Nadel B: **V(D)J-mediated translocations in lymphoid neoplasms: a functional assessment of genomic instability by cryptic sites.** *J Exp Med* 2002, **195:**85-98.
51. Kullback S: *Information Theory and Statistics*. New York: Wiley; 1959.
52. Akamatsu Y, Tsurushita N, Nagawa F, Matsuoka M, Okazaki K, Imai M, Sakano H: **Essential residues in V(D)J recombination signals.** *J Immunol* 1994, **153:**4520-4529.
53. Swanson PC, Desiderio S: **V(D)J recombination signal recognition: distinct, overlapping DNA- protein contacts in complexes containing RAG1 with and without RAG2.** *Immunity* 1998, **9:**115-125.
54. Miller J, Selsing E, Storb U: **Structural alterations in J regions of mouse immunoglobulin lambda genes are associated with differential gene expression.** *Nature* 1982, **295:**428-430.
55. Chen F, Rowen L, Hood L, Rothenberg EV: **Differential transcriptional regulation of individual TCR V segments before gene rearrangement.** *J Immunol* 2001, **166:**1771-1780.
56. Chou HS, Anderson SJ, Louie MC, Godambe SA, Pozzi MR, Behlke MA, Huppi K, Loh DY: **Tandem linkage and unusual RNA splicing of the T-cell receptor beta-chain variable-region genes.** *Proc Natl Acad Sci USA* 1987, **84:**1992-1996.
57. Lefranc MP: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 2001, **29:**207-209.
58. Kouskoff V, Nemazee D: **Role of receptor editing and revision in shaping the B and T lymphocyte repertoire.** *Life Sci* 2001, **69:**1105-1113.
59. Kleinfield R, Hardy RR, Tarlinton D, Dangl J, Herzenberg LA, Weigert M: **Recombination between an expressed immunoglobulin heavy-chain gene and a germline variable gene segment in a Ly 1$^+$ B-cell lymphoma.** *Nature* 1986, **322:**843-846.
60. Usuda S, Takemori T, Matsuoka M, Shirasawa T, Yoshida K, Mori A, Ishizaka K, Sakano H: **Immunoglobulin V gene replacement is caused by the intramolecular DNA deletion mechanism.** *EMBO J* 1992, **11:**611-618.
61. Chen C, Nagy Z, Prak EL, Weigert M: **Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing.** *Immunity* 1995, **3:**747-755.
62. **Downloads for *statistical models of recombination signal sequences that identify cryptic signals and predict recombination efficiencies*** [http://www.duke.edu/~lgcowell]
63. Lefranc MP, Giudicelli V, Ginestoux C, Bodmer J, Muller W, Bontrop R, Lemaitre M, Malik A, Barbie V, Chaume D: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 1999, **27:**209-212.
64. Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh SG, Bontrop R, Lemaitre M, Lefranc G, *et al.*: **IMGT, the international ImMunoGeneTics database.** *Nucleic Acids Res* 2000, **28:**219-221.
65. Shannon CE, Weaver W: *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.

66. Jaynes ET: *Probability Theory: The Logic Of Science.* Edited by Bretthorst GL. Cambridge: Cambridge University Press; in press.
67. Stone M: **Cross-validatory choice and assessment of statistical predictions (with discussion).** *J Roy Stat Soc B* 1974, **36:**111-147.
68. Ramsden DA, Baetz K, Wu GE: **Conservation of sequence in recombination signal sequence spacers.** *Nucleic Acids Res* 1994, **22:**1785-1796.
69. **RasMol** [http://www.umass.edu/microbio/rasmol/]