BMC Immunology

BioMed Central

Research article

The κB transcriptional enhancer motif and signal sequences of V(D)J recombination are targets for the zinc finger protein HIVEP3/KRC: a site selection amplification binding study

Carl E Allen[1], Chi-ho Mak[2] and Lai-Chu Wu*[2,3,4]

Address: [1]Department of Pediatrics, College of Medicine and Public Health, The Ohio State University, Columbus, OH, 43210, USA, [2]Ohio State Biochemistry Program, College of Medicine and Public Health, The Ohio State University, OH, 43210, USA, [3]Department of Molecular and Cellular Biochemistry, College of Medicine and Public Health, The Ohio State University, Columbus, OH, 43210, USA and [4]Department of Internal Medicine, Division of Immunology, College of Medicine and Public Health, The Ohio State University, Columbus, OH 43210, USA

E-mail: Carl E Allen - allen.111@osu.edu; Chi-ho Mak - chmak@hkuspace.hku.hk; Lai-Chu Wu* - wu.39@osu.edu

*Corresponding author

Published: 22 August 2002

BMC Immunology 2002, **3**:10

Received: 20 June 2002

Accepted: 22 August 2002

This article is available from: http://www.biomedcentral.com/1471-2172/3/10

Abstract

Background: The ZAS family is composed of proteins that regulate transcription via specific gene regulatory elements. The amino-DNA binding domain (ZAS-N) and the carboxyl-DNA binding domain (ZAS-C) of a representative family member, named κB DNA binding and recognition component (KRC), were expressed as fusion proteins and their target DNA sequences were elucidated by site selection amplification binding assays, followed by cloning and DNA sequencing. The fusion proteins-selected DNA sequences were analyzed by the MEME and MAST computer programs to obtain consensus motifs and DNA elements bound by the ZAS domains.

Results: Both fusion proteins selected sequences that were similar to the κB motif or the canonical elements of the V(D)J recombination signal sequences (RSS) from a pool of degenerate oligonucleotides. Specifically, the ZAS-N domain selected sequences similar to the canonical RSS nonamer, while ZAS-C domain selected sequences similar to the canonical RSS heptamer. In addition, both KRC fusion proteins selected oligonucleoties with sequences identical to heptamer and nonamer sequences within endogenous RSS.

Conclusions: The RSS are *cis*-acting DNA motifs which are essential for V(D)J recombination of antigen receptor genes. Due to its specific binding affinity for RSS and κB-like transcription enhancer motifs, we hypothesize that KRC may be involved in the regulation of V(D)J recombination.

Background

The ZAS gene family is an emerging family of important transcriptional proteins that have been implicated in the regulation of gene expression of the HIV-1 long terminal repeat [1], and genes encoding αA-crystallin [2], somatostatin receptor type II [3], the small calcium binding protein S100A4/mts1 [4], and type II collagen [5] via specific promoter or enhancer elements. Three human genes, *HIVEP1/Mbp1/PRDII-BF1* [6–9], *HIVEP2/Mbp2* [10–12], and *HIVEP3* [13], and their respective mouse counterparts *αACRYBP1* [2], *MIBP1* [3], and *KRC* [14,15], as well as rat *AGIE-BP1/MIBP1* [16,17] have been cloned and character-

ized. In addition, a distant relative *Schnurri* (*Shn*) has been identified in *Drosophila* [18–20]. Although little is known about the physiological functions of the mammalian ZAS proteins, Shn has been shown to be an important transcription regulator during embryonic development. Shn modulates transcription by relieving the repression of the nuclear protein Brinker and, in association with SMAD, mediates transcription response of the decapentaplegic pathway [21,22].

Each ZAS gene encodes large sequence-specific DNA-binding proteins with Mr >250,000 that contain two widely separated of C_2H_2-type zinc finger pairs. Smaller protein isoforms with a single zinc finger pair or with no zinc finger pairs can be generated by alternative RNA splicing [23,24]. The amino acid sequence and relative location of the two zinc finger pairs are highly conserved among ZAS proteins from invertebrate to vertebrate [Reviewed in [25]]. Although the zinc finger is a major structural motif involved in protein-nucleic acid interactions and is present in the largest superfamily of transcription factors, few proteins contain separate zinc finger pairs. The ZAS proteins (with two zinc finger pairs), tramtrack (with one finger pair), and basonuclin (with three finger pairs), constitute a unique class of C_2H_2 zinc finger transcription factors. [Reviewed in [26]]. In addition, each ZAS protein contains a sequence similar to the serine stripe present in basonuclin, in which eight serines are located on one side of a putative α-helix [13,27].

The ZAS domain is a protein structure unique to the ZAS protein family. A ZAS domain denotes a composite protein structure consisting of a pair of C_2H_2 **z**inc fingers, an **a**cidic region, and a **s**erine/threonine-rich sequence [15,25]. Here, we name the amino-DNA binding domain ZAS-N and the carboxyl-DNA binding domain ZAS-C. The DNA binding specificity of the ZAS-N or ZAS-C domains from several ZAS members have been characterized by electrophoretic mobility shift assays, methylation interference experiments, and DNAse I footprinting experiments. The cumulative data show that individual ZAS domains bind a κB-like consensus sequence, $GGGN_{(4-5)}CC$ [25]. However, mouse KRC, αACRYBP1 and mouse MIBP1 have also been shown to bind distinct DNA sequences. KRC binds to the signal sequences of V(D)J recombination (RSS) [14,28]. αACRYBP1 binds to a sequence in the type II collagen gene enhancer [5]. MIBP1 binds to a TC-rich element present in the somatostatin type II receptor gene enhancer [3].

This is the first study to evaluate DNA targets of both ZAS domains from a single protein independently. We used site selection amplification binding assays to select specific DNA targets recognized by KRC fusion proteins containing ZAS-N or ZAS-C from an initial oligonucleotide pool containing degenerate 25-mers. After cloning and DNA sequencing, the KRC-selected sequence datasets were analyzed by the computer program Multiple Expectation Maximum for Motif Elicitation (MEME version 3.0) to generate sets of position specific scoring matrices (PSSMs) or motifs [30]. When the program was set to identify wider sequences (= 9 nucleotides) the PSSMs were homologous to Sb [29], the κB motif [31], and the canonical heptamer and nonamer elements of the RSS [32]. However, shortening the width of the motifs to 5 nucleotides, the target length for the C_2H_2 zinc finger pairs of the tramtrack [33], the ZAS-N dataset yielded two motifs, "GGTAT" and "T(T/C)TT(T/G)G" and the ZAS-C dataset yielded a single motif, "TGTGG". Juxtaposition of the two pentamers of ZAS-N forms a sequence homologous to the canonical RSS nonamer, "GGTTTTTGT". Similarly, the ZAS-C pentamer together with its complement form the canonical RSS heptamer palindrome "CACTGTG".

The computer program Multiple Alignment Sequence Tool (MAST version 3.0) [34] was used to search human and mouse genome databases for DNA elements matching the KRC-selected PSSMs. The hits included sequence matching DNA regions located within or close to mobile genetic elements, including the diversity (D) gene segments of the variable region of the immunoglobulin (Ig) heavy chain [35], and break points of chromosomal translocation between Ig DH2-2 and the B cell lymphoma 1 *BCL-1* gene [36], and between the *ENL/MLLT1/LTG19* gene and myeloid lymphoid leukemia *MLL* gene [37].

Results

Amplification of KRC's DNA targets with a site selection amplification binding assay

In this study, sequences bound by the DNA binding domains of KRC were identified in a site selection PCR amplification DNA binding assay. KRC/ZAS-N or KRC/ZAS-C (100 μg each; Fig. 1A) were initially incubated with an pool of ^{32}P-labeled degenerate oligonucleotides and non-specific competitor DNA poly(dI-dC) (10 μg). DNA-protein complexes and unbound DNA were then resolved on a 5% polyacrylamide gel, and the protein-bound DNA was purified and amplified. The oligonucleotides in the degenerate pool were composed of twenty-five random nucleotides (25-mer) in the middle flanked by a specific sequence BSS1 at one end and the complementary sequence of BSS2 at the other end. Subsequently, the primer set BSS1 and BSS2 was used to amplify the recovered oligonucleotides by PCR. The sequence of binding, selection and amplification was repeated several times before protein-selected oligonucleotides were cloned, sequenced and analyzed. To select optimal binding sequences, the stringency of succeeding rounds of the selection procedures was increased by using successively less (0.5×) fu-

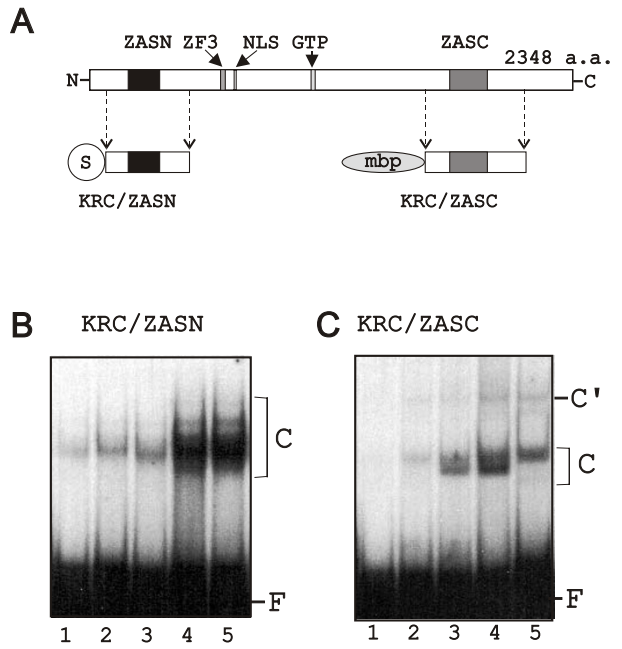


Figure 1

KRC fusion proteins and site-selection EMSA Figure 1A. KRC *fusion proteins*. (Top) The full-length KRC protein is described schematically. In the ZAS-N and ZAS-C DNA-binding domains the **z**inc-fingers, **a**cidic regions, and serine-threonine-rich regions highlighted. ZASN, ZAS-N domain; ZF3, zinc finger 3; NLS, nuclear localization signal; GTP, GTPase motif; ZASC, ZAS-C motif. (Bottom) KRC fusion proteins, KRC/ZAS-N and KRC/ZAS-C are described schematically. KRC/ZAS-N is a S-tag fusion protein containing the ZAS-N DNA-binding domain (nt 949–2167) KRC/ZAS-C is an Mbp fusion protein containing the ZAS-C DNA-binding domain (nt 5544–7015). These are the fusion proteins used in the site-selection assay described in this paper. Figure 1B and 1C. Electrophoretic mobility shift assays of the site selection procedures. (Bottom) A portion of the oligonucleotides (~0.2 ng and 5000 cpm) recovered from each round of site selection was 32P-labeled and incubated with KRC fusion proteins (~0.5 μg), (**B**) KRC/ZAS-N and (**C**) KRC/ZAS-C, in the presence of 10 μg poly(dI-dC). DNA-protein complexes and free probes were resolved in 6% polyacrylamide gels and visualized by exposing dried gels to X-ray films. The probes used in lanes 1 through 5 were derived from aliquots of DNA recovered from round one through five of site selection, respectively. C, DNA-protein complexes; and F, free probes.

sion proteins and more (4×) non-specific competitor DNA in each round.

The formation of protein-DNA complexes was monitored throughout the site selection experiments (Fig. 1B and Fig). Analytical EMSAs were performed under more stringent conditions than in EMSAs used to purify protein-bound oligonucleotides in the site selection experiments, using much less fusion protein (~0.1 to 0.5 μg) and an excess non-specific DNA poly(dI-dC) (10 μg). Initially, the DNA-protein complexes formed between the degenerate oligonucleotide pool and KRC/ZAS-N or KRC/ZAS-C were barely detectable, indicating that both fusion proteins bound DNA selectively (Figs. 1B and 1C, lane 1). In the subsequent rounds, the yield of the DNA-protein complexes increased, suggesting successful enrichment of KRC binding sites in the recovered oligonucleotides during the selection procedures. After the fourth rounds of selection and amplification, no further increase in the amount of DNA-protein binding complexes was observed. The experiment, therefore, was stopped at the fifth round for both fusion proteins. Furthermore, in rounds four and five, a cluster of close migrating DNA-protein complexes were observed for KRC/ZAS-N (Fig. 1B, lanes 4 and 5). In EMSA, the gel mobility of DNA-protein complexes depends on the overall mass of the binding proteins [38] and on the possible protein induced bending angle of DNA [39]. Since a single fusion protein was used in each binding reaction, the slight variation in the gel mobility of the DNA-protein complexes may reflect that KRC has more than one target, or that the targets were located at different positions within the 25-mer DNA. Similarly, two closely migrating DNA-protein complexes were clearly seen for KRC/ZAS-C at rounds three through five (labeled C, Fig. 1C, lanes 3, 4, and 5). In addition, another complex, labeled C', which was minor and had significantly slower gel mobility was observed in round four and round five (Fig. 1C, lanes 4 and 5). Previously, we showed that KRC/ZAS-C bound DNA as dimers, tetramers, and multiple of tetramers [28]. The significant difference in the gel mobility between complex C and complex C' suggested that they were likely composed of KRC/ZAS-C dimer and tetramer, respectively. These data show that the site selection amplification binding assays using both KRC DNA-binding domains were efficient in selecting KRC targets and that KRC/ZAS-C readily formed highly ordered DNA-protein structures.

DNA oligonucleotides recovered from the fifth rounds of site selections were cloned into plasmid vectors. We obtained fifty-three KRC/ZAS-N selected sequences and forty-nine KRC/ZAS-C selected sequences. The 25-mer sequences of individual site selected sequences from each fusion protein are shown in the BSS1-N_{25}-BSS2 orientation (Figure 2). These sequences were named tentatively after ZAS-N or ZAS-C correspondingly, and a suffix, a number given in the order of plasmid DNA preparation. Gaps in the numberings represented clones with "empty vectors" and therefore were excluded from Figure 2 as they most likely resulted from cloning artifacts. Among the protein-selected sequences, 6 out of 53 ZAS-N-sequences (ZAS-N-9 and ZAS-N-10; ZAS-N-11 and ZAS-N-16; and

ZAS 1 Site Select (n=53)		ZAS 2 Site select (N=49)	
ZAS 1-1	GGTATTTTTGTGTTTATTTACGCGT	ZAS 2-1	GAGATGTTGTGTTTTATTTGTTTGT
ZAS 1-3	GCAGCGATGATTTGATGTTGTCGTG	ZAS 2-2	TGGTTGATATTCTTTATGTAATTGG
ZAS 1-4	AGGGAATGCTCATTGTGGCTCTTGG	ZAS 2-7	GGTGATTACATCTTTGGTTTATGTG
ZAS 1-5	GTAAGGAGTTCTGTTGCTATTTATG	ZAS 2-12	TGAAGAGTTACTCACTCTTTGTGGG
ZAS 1-7	GGGAATTTTGTCTCTATTTTTGTCG	ZAS 2-13	TTGGTTTCCTGGTTTACTATATCGG
ZAS 1-8	AGTGGAATTTTGCTTTTTGTTTGTG	ZAS 2-16	AGAATTGTGTTTTGTTGTTCTGTGG
ZAS 1-9	GTAGGTTGTTCCTTGAATTTTCTGG	ZAS 2-17	AATATGTGTGTGGTTATTGATTTTA
ZAS 1-10	GTAGGTTGTTCCTTGAATTTTCTGG	ZAS 2-18	CAATTAACTGGCCTATGGTGTCAGG
ZAS 1-11	GGATTTGTTTATTGGATTGTAGTGT	ZAS 2-20	GGAGTGATTACCCAATAGTGTGGGG
ZAS 1-12	GTATTCCTCTTAGCTCTGTTGAGGG	ZAS 2-21	GGATTTGTTATACTAGTTGATAGTA
ZAS 1-13	GGGACGTTTCCATTTGTTAATTTGG	ZAS 2-22	AGCATTACACTGTTATTGTTCAGTG
ZAS 1-16	GGATTTGTTTATTGGATTGTAGTGT	ZAS 2-23	AGTATTTTCTTCGATTAGACTAGTG
ZAS 1-17	GGTTGGTATTTTAGCCTTTGGATGG	ZAS 2-27	AGGTTTTAATTGACAAATAGTCAGG
ZAS 1-18	AGGTATTTTCCTTCTAGTGAGTGGT	ZAS 2-28	GGGGTAATTGCTTTAATGCTTTTGG
ZAS 1-19	GGGATTCTTATCATTTGTATGTTGT	ZAS 2-29	GATATTTTAATTCTCTATGTGTTGG
ZAS 1-21	GGATTTGTTGTTTTGCCTTGCTCTG	ZAS 2-31	ATTCTCTTATTTCTCAATTCTGTGG
ZAS 1-22	GGAGGTTTGTATTTAATGCATTTGG	ZAS 2-32	ATTCATGTTCTGCCTTAATGCTGAG
ZAS 1-23	TGGGAATTTCTTTTTGCTAGTGTGG	ZAS 2-33	ATTCATTCTAGCTTTTGTTATGTGG
ZAS 1-25	ATGTCCTCTTTCTTCTACTTGTGTG	ZAS 2-37	GAAGATTTTCCTATCATGGTCTAGG
ZAS 1-26	GGTGGTGTGGTCTTTTTCTATTGAG	ZAS 2-38	TTGCGGTGTTTTCTGTATATTGACG
ZAS 1-27	CGAATTTTAGTGTATTGTTATTTGG	ZAS 2-39	GGTGATTCGTCAATTGTTCGTTGGG
ZAS 1-30	CGTACGTTTGTTTTGATGTTTCTCG	ZAS 2-41	GTAGGTTTTACCTCATTATTGTGGG
ZAS 1-31	GCATAAATATTACAACTGCTTTTTG	ZAS 2-42	ATGTAGGAATTTATATTTGACTATG
ZAS 1-33	GTGTGTTTTGTTTTATTTCTTGCGG	ZAS 2-44	GATAGCTTCTAATTAATTGTGCGAG
ZAS 1-34	GGTGATACGTGGTCTCTGTCTGTGG	ZAS 2-45	GATAGCTTCTAATTAATTGTGCGAG
ZAS 1-35	ATATTCTAGTTGCCTTTTTTAGTGG	ZAS 2-46	CAGGAGTTTTGTTACAAGATTGGTG
ZAS 1-37	GACGGTTGAGTGTTTGTTTTTTTGG	ZAS 2-47	GTAACTTAGGGGTTTTTGTTGTCTG
ZAS 1-39	GTAATATTTGTTTGAGCTGGTTTGG	ZAS 2-49	GCAGGTAGTTGCCGAGTAACTTTGG
ZAS 1-40	GTAATATTTGTTTGAGCTGGTTTGG	ZAS 2-50	ATATTAATTTTTTATAATACTGTGG
ZAS 1-43	GCAGAGGTAGTTTCGTCTTCGATGT	ZAS 2-51	GTGTATTATTCCTCAATTTCGTGGG
ZAS 1-45	GAATGTGCTTGGTATTTGTTTTCGG	ZAS 2-53	GGTGAAGTCATCATGTTAGGTTGGG
ZAS 1-46	AGGATATGTTTTGGTTGATTCGTGG	ZAS 2-54	AATTTAAATATATTTCATTATTGGG
ZAS 1-47	GGAGGAAATTTGATCAGTTTGATTG	ZAS 2-56	ATACTATCTATATATGTTAATGTAG
ZAS 1-48	GGCCTACATATCTAGTTTTGTTTGG	ZAS 2-58	AGTTATGGTAATTCTAATCTCATGG
ZAS 1-49	ATGTGCTTGTGTTGTTCTTATGTGG	ZAS 2-60	GGCAGTCGATTTCTCAGTTCTATGT
ZAS 1-50	GAAGGTATTCGTAACTTGCTATTGT	ZAS 2-62	GGAGTTTGTCTTATGGATATTGTGG
ZAS 1-51	GGGAATTTCATATTTATGTGTTTGG	ZAS 2-63	GTGGTATTTTACCTAAATATTGTGG
ZAS 1-52	GGTATTCTCTAGTTTCTTTACTGGT	ZAS 2-68	GTTGTGTACTTTCTTCGTTCTATTG
ZAS 1-54	GGTATTACTTAGTCTCTTTACTTGG	ZAS 2-71	GGGTGTGGAGTTGTCATATTTTTGG
ZAS 1-56	CGTAGGTATTGCTGTGATGTGTTGG	ZAS 2-73	ATGGTGTGGTGTTTTCCTTTTCGAG
ZAS 1-57	CGTAATGTCTAATCTTATTTTGTTG	ZAS 2-74	ACAGGTTGAAGTGTTTAGAATTAAG
ZAS 1-58	AGGTATTTTTAGGCGTTGGCTTTGG	ZAS 2-75	ACAGGTTTAATTGTTTAGAACTAAG
ZAS 1-60	GGATTGCTCATGGTATTTGTATGGG	ZAS 2-79	AGAGGAGTTTTGCTTTTATAGAGTG
ZAS 1-61	GTAGTAGTGAATTTTGCTTGTTGCG	ZAS 2-81	TAGATTTTGGATATTGTGTTATGAG
ZAS 1-62	ATGTGAATTTTGTGACTTATGTTCG	ZAS 2-82	GACTTGTAAAGGGATTGCCTGTGCT
ZAS 1-64	GGTATTGCTTTTAATGATGATGTTG	ZAS 2-83	ATTTCGACAATGGTTCGCAGTTGTG
ZAS 1-65	GGATAGTGGATTTTACTGTGTGTGG	ZAS 2-85	TATGGATACTGTTGATGTGGATTGT
ZAS 1-67	GGATGTGTGAATTCTGATTTGCTTG	ZAS 2-87	GGATTATTCGATGATACGATCTATG
ZAS 1-68	ATGCAGTATTCTTGTACATTGTTGG	ZAS 2-88	GGATTGTTACACAGGTTTTTCTGTG
ZAS 1-70	GTATCACTAGTTGTTCATTTTGTGC		
ZAS 1-71	GGAATTGTTGGTTCGTCATACGTGG		
ZAS 1-72	GGGGATTTTGGACTGTTTTTGTCGG		
ZAS 1-73	GGATTCTGGTATTCTCATTTTTTGG		

Figure 2
KRC-bound sequences recovered from site-selection experiments. The sequences were isolated from the pool of oligonucleotides remaining after five rounds of site selection with either the N-terminal KRC/ZAS-N or the C-terminal KRC/ZAS-C fusion proteins. 53 sequences were obtained from the KRC/ZAS-N-bound pool, and 49 sequences were obtained from the KRC/ZAS-C-bound pool. Sequences are shown as sequenced with a 5' BSS1 primer in the forward orientation: BSS1(N_{25})BSS2.

ZAS-N-39 and ZAS-N-40), and 2 out of 49 ZAS-C-selected sequences (ZAS-C-44 and ZAS-C-45) were identical. The redundancy observed during DNA amplification appeared to be minimal, and therefore, the complexity of the protein-selected DNA sequences in the datasets should be high.

Motif discovery by MEME: ≥ 6 W ≤ 25 nucleotides

The fifty-three KRC/ZAS-N-selected sequences (ZAS-N dataset) and 49 KRC/ZAS-C-selected sequences (ZAS-C dataset) were first analyzed by the Motif Expectation Maximum for Motif Elicitation (MEME) computer program. MEME analyzes input sequences for similarities and produces a PSSM or motif for each pattern it discovers [30]. We set the parameters of MEME as follows: (i) zero or one occurrences of a single motif per sequence; (ii) five as the maximum number of motifs to identify; (iii) 5 to 25 nucleotides as the range of motif size; and (iv) both DNA strands as input.

The first pass of MEME for ≥ 6 W ≤ 25 nucleotides generated a single 25-mer TG-rich motif found in all 53 sequences in the ZAS-N-dataset (Fig 3). The log likelihood ratio (llr), the logarithm of the ratio of the probability of the occurrences of the motif given the motif model (likelihood given the motif) versus their probability given the background model (likelihood given the null model), was calculated to be 370. The E-value, which is an estimate of the expected number of motifs with the given log likelihood ratio, and with the same width and number of occurrences, that one would find in a similarly sized set of random sequences, was calculated to be $1.1e^{-90}$. The llr and E-value scores suggested that the PSSMs discovered were statistically significant. Using the ZAS-C-dataset, a similar pass of MEME also generated a TG-rich motif with llr of 222 and an E-value of $2.3e^{-37}$ (Fig. 4). A sequence comparison showed that the ZAS-N-PSSM was generally homologous to that of the ZAS-C-PSSM, with 2–3 guanines at both ends and a T-rich sequence in the middle. In fact, when the two datasets were combined for a pass of MEME, a 25-mer motif with llr of 526 and an E-value of $2.9e^{-134}$ was obtained (data not shown), suggesting ZAS-N and ZAS-C have similar DNA targets. Furthermore, we were able to align the κB motif, the Sb sequence, the RSS nonamer, and part of the RSS heptamer "TGTG" with both PSSMs (Figs. 3 and 4). As a control, several sets of 50 random 25-mers were generated by a random number generator (G = 1, A = 2, T = 3, C = 4) and none yielded any statistically significant PSSMs (where E-values <1.0) when analyzed by MEME under the same settings (data not shown).

Motif discovery by MEME: ≥ 6 W ≤ 15 nucleotides

A motif of 25-nucleotides was obtained in the above MEME analysis when widths of ≥ 6 W ≤ 25 nucleotides were set. This is longer than known transcription factor binding sites. In addition, the information content (measured in bits), which reflects the degree of conservation of each column (or position) in those PSSMs was relatively low, ranging from 0 to 1.7, with an overall average <0.5 per position (Figs. 3 and 4). To elucidate more biologically relevant motifs with higher information content, a second pass of MEME was performed with motif widths set to shorter lengths, ranging from 6 to 15 nucleotides. The PSSMs discovered for each width all had significantly higher overall bits per position and obtained a TG-rich core sequence (data not shown).

Representative results of passes of MEME with W = 9 nucleotides are presented in Figures 5,6,7,8,9. In those passes, two PSSMs were discovered in the ZAS-N dataset. One motif with a consensus sequence (G/T)-G-(T/A)-(A/T)-T-T-T-(T/G)-(T/G) was found in 50 of the 53 sequences of the ZAS-N-dataset, with a llr of 276 and an E-value of $1.5e^{-18}$. This PSSM was more conserved than the 25-mer PSSM described above, with bits for all positions ranging from 0.2 to 1.9 and an average bits/position of 0.9 (Figure 5). We were able to align the canonical RSS nonamer and Sb with this PSSM. The second motif "GGTTGTTC" was found only in two input sequences, had a llr of 26 and E-value of $9.6e^{3}$, and was similar to the κB motif (Figure 6).

A similar pass of MEME discovered three motifs in the ZAS-C dataset. The major motif had a consensus sequence: (A/T)-T-(T/A)-T-T-G-T-G-G, with a llr of 125 and an E-value of $1.5e^{-2}$ (Figure 7). This consensus sequence aligns with canonical RSS nonamer. Notably, the terminal 5 nucleotides, "T-G-T-G-G", each had an information content of >1.5 bits and were nearly invariant in that sequence alignment. With respect to the heptamer (canonical sequence: CACAGTG), the CAC sequence bordering the recombination site is the most conserved segment of the sequence [32], and mutation of these nucleotides has been found to decrease V(D)J joining in transfection assays using recombination substrates [40]. Because the sequence of the canonical RSS heptamer is palindrome, we speculate that the TGTG sequence may be sufficient for KRC/ZAS-C binding. A second motif also contained a "TGTG" core sequence (Figure 8). A third sequence was homologous to the RSS nonamer, κB or Sb sequences (Figure 9). In general, PSSMs generated from MEME passes looking for shorter motifs yielded more conserved sequences.

Motif discovery by MEME: W = 5 nucleotides

KRC and tramtrack (TTK) share the same class of C_2H_2 zinc finger pairs. The crystal structure of the zinc finger pairs of tramtrack-DNA duplex revealed that the two fingers together contacted 5 base-pairs "$A^1G^2G^3A^4T^5$" in the major groove of DNA: The first finger interacts with

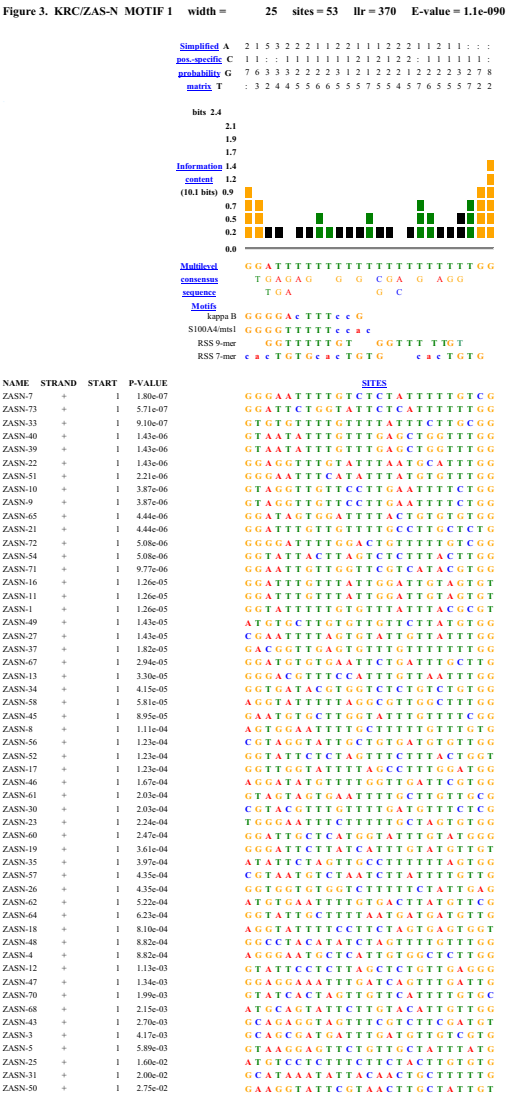
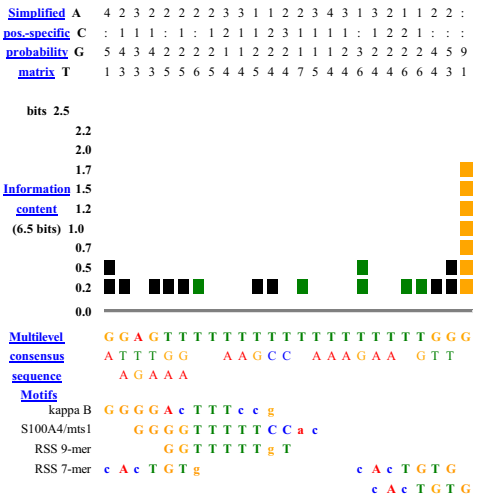


Figure 3

ZAS-N-selected sequence: MEME motif search and alignment, n = 25. ***Simplified position-specific probability matrix*** represents the probability of each possible letter appear a each possible position in an occurrence of the motif. Numbers are described as n/10 (eg 50 is represented as 5), with "0" described as ":". ***Information Content Diagram*** provides information of which positions in the motif are most (and least) highly conserved. Columns in the information content diagram are shaded according to the majority category of the letters occurring in that column of alignment. If no letter category has a frequency >0.5, the column is black. A = red, C = blue, G = orange, T = green. Summing of information content for each position gives the total information content of the motif, which is approximately equal to the log likelihood ratio divided by the number of occurrences times ln(2). ***Multilevel Consensus Sequence*** is calculated from the motif position-specific matrix where the most likely nucleotide is printed at the top of a column. Only letters with probabilities of 0.2 or higher are included. ***Summary Information*** is printed below the multilevel consensus sequence. "Width" describes the length of the motif. "Sites" describes the number of sequences in the dataset which contributed to the consensus sequence. The "log likelihood ratio" is the logarithm of the ratio of the probability of the occurrences of the motif given the motif model versus their probability given the null model. The "E-value" is an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and number of occurrences, that one would find in a similarly sized set of random sequences. ***Motif Alignment*** displays the occurrences of the motif in the dataset. Each site is identified by the name of the sequence where it occurs, the strand (+ or -), and the position in the sequence where the site begins. The p-value of a site is computed from the match score of the site with the position specific scoring matrix for the motif, and gives the probability of a random string having the same match score or higher. (Above is described in detail at MEME [http://www.sdsc.edu/meme][30]).

Figure 4. KRC/ZAS-C MOTIF 1 width = 25 sites = 49 llr = 222 E-value = 2.3e-037

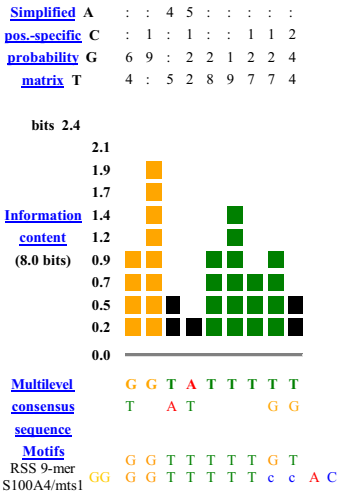


NAME	STRAND	START	P-VALUE	SITES
ZASC-63	+	1	4.70e-06	GTGGTATTTTACCTAAATATTGTGG
ZASC-29	+	1	6.82e-06	GATATTTTAATTCTCTATGTGTTGG
ZASC-20	+	1	1.17e-05	GGAGTGATTACCCAATAGTGTGGGG
ZASC-28	+	1	1.39e-05	GGGGTAATTGCTTTAATGCTTTTGG
ZASC-62	+	1	2.31e-05	GGAGTTTGTCTTATGGATATTGTGG
ZASC-41	+	1	2.72e-05	GTAGGTTTTACCTCATTATTGTGGG
ZASC-71	+	1	3.76e-05	GGGTGTGGAGTTGTCATATTTTTGG
ZASC-7	+	1	4.40e-05	GGTGATTACATCTTTGGTTTATGTG
ZASC-16	+	1	5.14e-05	AGAATTGTGTTTTGTTGTTCTGTGG
ZASC-1	+	1	8.08e-05	GAGATGTTGTGTTTTATTTGTTTGT
ZASC-51	+	1	1.08e-04	GTGTATTATTCCTCAATTTCGTGGG
ZASC-31	+	1	1.44e-04	ATTCTCTTATTTCTCAATTCTGTGG
ZASC-68	+	1	1.65e-04	GTTGTGTACTTTCTTCGTTCTATTG
ZASC-40	+	1	1.65e-04	GGTGATTCGTCAATTGTTCGTTGGG
ZASC-33	+	1	1.89e-04	ATTCATTCTAGCTTTTGTTATGTGG
ZASC-53	+	1	2.48e-04	GGTGAAGTCATCATGTTAGGTTGGG
ZASC-2	+	1	2.83e-04	TGGTTGATATTCTTTATGTAATTGG
ZASC-73	+	1	3.66e-04	ATGGTGTGGTGTTTTCCTTTTCGAG
ZASC-23	+	1	3.66e-04	AGTATTTTCTTCGATTAGACTAGTG
ZASC-37	+	1	4.15e-04	GAAGATTTTCCTATCATGGTCTAGG
ZASC-50	+	1	4.71e-04	ATATTAATTTTTTATAATACTGTGG
ZASC-79	+	1	5.32e-04	AGAGGAGTTTTGCTTTTATAGAGTG
ZASC-46	+	1	1.50e-03	CAGGAGTTTTGTTACAAGATTGGTG
ZASC-12	+	1	1.50e-03	TGAAGAGTTACTCACTCTTTGTGGG
ZASC-88	+	1	1.67e-03	GGATTGTTACACAGGTTTTTCTGTG
ZASC-54	+	1	2.06e-03	AATTTAAATATATTTCATTATTGGG
ZASC-75	+	1	2.79e-03	ACAGGTTTAATTGTTTAGAACTAAG
ZASC-27	+	1	2.79e-03	AGGTTTTAATTGACAAATAGTCAGG
ZASC-74	+	1	3.09e-03	ACAGGTTGAAGTGTTTAGAATTAAG
ZASC-49	+	1	3.09e-03	GCAGGTAGTTGCCGAGTAACTTTGG
ZASC-13	+	1	3.09e-03	TTGGTTTCCTGGTTTACTATATCGG
ZASC-22	+	1	5.96e-03	AGCATTACACTGTTATTGTTCAGTG
ZASC-87	+	1	7.11e-03	GGATTATTCGATGATACGATCTATG
ZASC-38	+	1	7.75e-03	TTGCGGTGTTTTCTGTATATTGACG
ZASC-47	+	1	8.45e-03	GTAACTTAGGGGTTTTTGTTGTCTG
ZASC-45	+	1	8.45e-03	GATAGCTTCTAATTAATTGTGCGAG
ZASC-44	+	1	8.45e-03	GATAGCTTCTAATTAATTGTGCGAG
ZASC-58	+	1	9.19e-03	AGTTATGGTAATTCTAATCTCATGG
ZASC-81	+	1	9.99e-03	TAGATTTTGGATATTGTGTTATGAG
ZASC-42	+	1	1.18e-02	ATGTAGGAATTTATATTTGACTATG
ZASC-32	+	1	1.27e-02	ATTCATGTTCTGCCTTAATGCTGAG
ZASC-17	+	1	1.38e-02	AATATGTGTGTGGTTATTGATTTTA
ZASC-21	+	1	1.49e-02	GGATTTGTTATACTAGTTGATAGTA
ZASC-56	+	1	1.74e-02	ATACTATCTATATATGTTAATGTAG
ZASC-18	+	1	1.74e-02	CAATTAACTGGCCTATGGTGTCAGG
ZASC-60	+	1	2.68e-02	GGCAGTCGATTTCTCAGTTCTATGT
ZASC-83	+	1	5.13e-02	ATTTCGACAATGGTTCGCAGTTGTG
ZASC-85	+	1	6.87e-02	TATGGATACTGTTGATGTGGATTGT
ZASC-82	+	1	2.48e-01	GACTTGTAAAGGGATTGCCTGTGCT

Figure 4
ZAS-C-selected sequence: MEME motif search and alignment, n = 25 Figure legend as above (Figure 3).

Figure 5. KRC/ZAS-N

MOTIF 1 width = 9 sites = 50 llr = 276 E-value = 1.5e-018

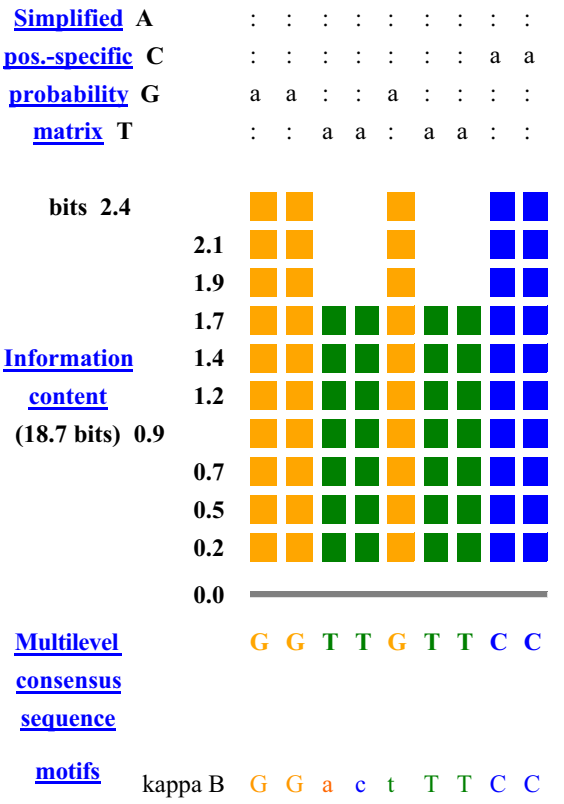


NAME	STRAND	START	P-VALUE	SITES	
ZASN-8	+	4	1.20e-05	AGT	G G A A T T T T G CTTTTTGTTT
ZASN-7	+	2	1.20e-05	G	G G A A T T T T G TCTCTATTTT
ZASN-58	+	2	2.14e-05	A	G G T A T T T T T AGGCGTTGGC
ZASN-1	+	1	2.14e-05		G G T A T T T T T GTGTTTATTT
ZASN-18	+	2	4.05e-05	A	G G T A T T T T C CTTCTAGTGA
ZASN-62	+	4	1.02e-04	ATG	T G A A T T T T G TGACTTATGT
ZASN-61	+	8	1.02e-04	GTAGTAG	T G A A T T T T G CTTGTTGCG
ZASN-60	+	12	1.56e-04	GATTGCTCAT	G G T A T T T G T ATGGG
ZASN-45	+	11	1.56e-04	GAATGTGCTT	G G T A T T T G T TTTCGG
ZASN-71	+	1	2.07e-04		G G A A T T G T T GGTTCGTCAT
ZASN-26	+	9	2.25e-04	GGTGGTGT	G G T C T T T T T CTATTGAG
ZASN-33	+	2	2.63e-04	G	T G T G T T T T G TTTTATTTCT
ZASN-10	+	14	3.50e-04	GGTTGTTCCT	T G A A T T T T C TGG
ZASN-9	+	14	3.50e-04	GGTTGTTCCT	T G A A T T T T C TGG
ZASN-73	+	8	3.83e-04	GGATTCT	G G T A T T C T C ATTTTTTGG
ZASN-52	+	1	3.83e-04		G G T A T T C T C TAGTTTCTTT
ZASN-43	+	6	3.83e-04	GCAGA	G G T A G T T T C GTCTTCGATG
ZASN-72	+	3	4.18e-04	GG	G G A T T T T G G ACTGTTTTTG
ZASN-37	+	15	5.52e-04	GTTGAGTGTT	T G T T T T T T T GG
ZASN-27	+	11	5.52e-04	CGAATTTTAG	T G T A T T G T T ATTTGG
ZASN-22	+	1	5.52e-04		G G A G G T T T G TATTTAATGC
ZASN-5	+	5	5.52e-04	GTAA	G G A G T T C T G TTGCTATTTA
ZASN-67	+	8	6.73e-04	GGATGTG	T G A A T T C T G ATTTGCTTG
ZASN-21	+	1	6.73e-04		G G A T T T G T T GTTTTGCCTT
ZASN-16	+	1	6.73e-04		G G A T T T G T T TATTGGATTG
ZASN-11	+	1	6.73e-04		G G A T T T G T T TATTGGATTG
ZASN-50	+	4	7.65e-04	GAA	G G T A T T C G T AACTTGCTAT
ZASN-51	+	17	8.32e-04	TTCATATTTA	T G T G T T T G G
ZASN-23	+	3	9.10e-04	TG	G G A A T T T C T TTTTGCTAGT
ZASN-19	+	16	1.01e-03	TCTTATCATT	T G T A T G T T G T
ZASN-49	+	8	1.10e-03	ATGTGCT	T G T G T T G T T CTTATGTGG
ZASN-64	+	1	1.77e-03		G G T A T T G C T TTTAATGATG
ZASN-56	+	5	1.77e-03	CGTA	G G T A T T G C T GTGATGTGTT
ZASN-17	+	5	1.77e-03	GGTT	G G T A T T T T A GCCTTTGGAT
ZASN-13	+	2	1.77e-03	G	G G A C G T T T C CATTTGTTAA
ZASN-47	+	4	2.71e-03	GGA	G G A A A T T T G ATCAGTTTGA
ZASN-46	+	7	2.71e-03	AGGATA	T G T T T T G G T TGATTCGTGG
ZASN-31	+	17	2.71e-03	ATATTACAAC	T G C T T T T T G
ZASN-30	+	14	2.71e-03	ACGTTTGTTT	T G A T G T T T C TCG
ZASN-4	+	3	2.71e-03	AG	G G A A T G C T C ATTGTGGCTC
ZASN-40	+	16	3.04e-03	ATTTGTTTGA	G C T G G T T T G G
ZASN-39	+	16	3.04e-03	ATTTGTTTGA	G C T G G T T T G G
ZASN-65	+	17	3.58e-03	TGGATTTTAC	T G T G T G T G G

Figure 5
ZAS-N-selected sequence: MEME motif search and alignment, n = 9. Motif #1. Figure legend as above (Figure 3).

Figure 6. KRC/ZAS-N

MOTIF 2 width = 9 sites = 2 llr = 26 E-value = 9.6e+003



NAME	STRAND	START	P-VALUE	SITES	
ZASN-10	+	4	2.42e-06	GTA	G G T T G T T C C TTGAATTTTC
ZASN-9	+	4	2.42e-06	GTA	G G T T G T T C C TTGAATTTTC

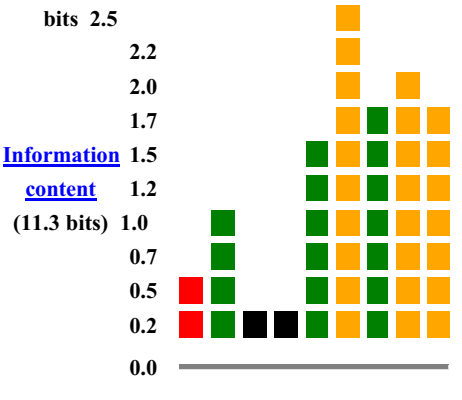
Figure 6
ZAS-N-selected sequence: MEME motif search and alignment, n = 9. Motif #2. Figure legend as above (Figure 3).

$A^1G^2G^3$ while the second finger interacts with $G^3A^4T^5$[33]. By inference, each zinc finger pair of KRC might also bind to a pentamer. To test this hypothesis, a pass of MEME was performed with a fixed width of 5 nucleotides. Two PSSMs were obtained for the ZAS-N dataset. One motif was an invariant "GGTAT" (Figure 10), and the other was "T(T/G)T(T/G)G (Figure 11). Both motifs when superimposed form a sequence that is homologous to the RSS nonamer "GGTTTTTGT". It is possible that two KRC molecules may be needed to interact with an RSS nonamer: one protein whose ZAS-N domains may bind to the 5'-half of an RSS nonamer while a second protein's ZAS-N domain may bind to the 3'-half of an RSS. For the ZAS-C-dataset, a single motif "TGTG(G/T)" was obtained (Figure 12). Since the RSS heptamer is a palindrome, it is possible for two KRC molecules to bind to an RSS heptamer, with one ZAS-C binding to the top strand and the other ZAS-C binding to the bottom strand. This notion is consistent with previous observation that two molecules of KRC-ZAS-C are required for DNA binding [28]. Further passes of MEME with W = 3 were too short to generate statistically significant motifs (data not shown). These data suggest that with respect to the RSS canonical elements, KRC/ZAS-N binds the nonamer more efficiently while

Figure 7. KRC/ZAS-C

MOTIF 1 width = 9 sites = 16 llr = 125 E-value = 1.7e-002

Simplified pos.-specific probability matrix	A	6	:	4	1	:	:	:	:
	C	:	1	:	2	1	:	:	:
	G	1	1	2	2	:	a	:	9	9
	T	3	9	4	5	9	:	a	1	1



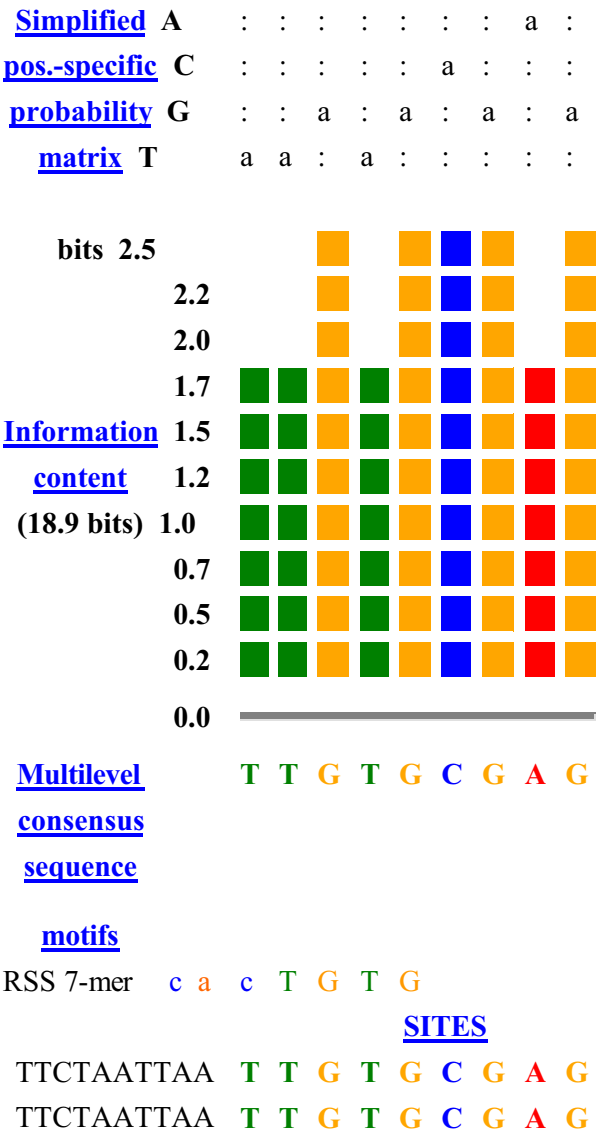
Multilevel consensus sequence		A	T	T	T	T	G	T	G	G
		T		A						
motifs										
RSS 9-mer	G	g	T	T	T	T	G	T		
RSS 7-mer			c	A	c	T	G	T	G	

NAME	STRAND	START	P-VALUE		SITES								
ZASC-63	+	17	1.25e-05	TTTTACCTAA	A	T	A	T	T	G	T	G	G	
ZASC-62	+	17	1.25e-05	TGTCTTATGG	A	T	A	T	T	G	T	G	G	
ZASC-31	+	17	2.31e-05	TTATTTCTCA	A	T	T	C	T	G	T	G	G	
ZASC-50	+	17	3.65e-05	ATTTTTTATA	A	T	A	C	T	G	T	G	G	
ZASC-20	+	15	3.65e-05	TGATTACCCA	A	T	A	G	T	G	T	G	G	GG
ZASC-73	+	1	4.05e-05		A	T	G	G	T	G	T	G	G	TGTTTTCCTT
ZASC-41	+	16	4.67e-05	TTTTACCTCA	T	T	A	T	T	G	T	G	G	G
ZASC-16	+	17	8.16e-05	GTGTTTTGTT	G	T	T	C	T	G	T	G	G	
ZASC-51	+	16	1.27e-04	TTATTCCTCA	A	T	T	T	C	G	T	G	G	G
ZASC-85	+	12	1.38e-04	ATGGATACTG	T	T	G	A	T	G	T	G	G	ATTGT
ZASC-33	+	17	1.49e-04	TCTAGCTTTT	G	T	T	A	T	G	T	G	G	
ZASC-12	+	16	1.95e-04	AGTTACTCAC	T	C	T	T	T	G	T	G	G	G
ZASC-81	+	11	2.27e-04	TAGATTTTGG	A	T	A	T	T	G	T	G	T	TATGAG
ZASC-1	+	4	2.41e-04	GAG	A	T	G	T	T	G	T	G	T	TTTATTTGTT
ZASC-17	+	5	2.68e-04	AATA	T	G	T	G	T	G	T	G	G	TTATTGATTT
ZASC-47	+	13	5.14e-04	AACTTAGGGG	T	T	T	T	T	G	T	T	G	TCTG

Figure 7
ZAS-C-selected sequence: MEME motif search and alignment, n = 9. Motif #1. Figure legend as above (Figure 3).

Figure 8. KRC/ZAS-C

MOTIF 2 width = 9 sites = 2 llr = 26 E-value = 5.6e+003



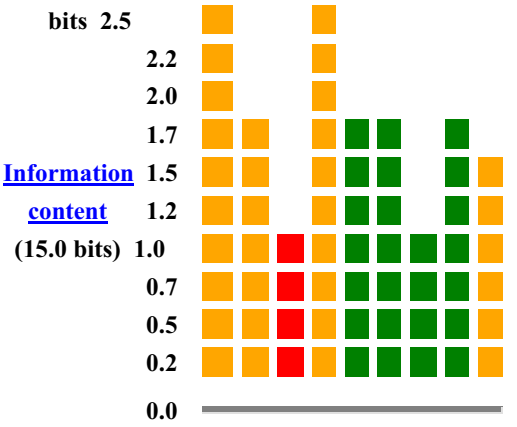
NAME	STRAND	START	P-VALUE	SITES
ZASC-45	+	17	2.01e-06	TTCTAATTAA
ZASC-44	+	17	2.01e-06	TTCTAATTAA

Figure 8
ZAS-C-selected sequence: MEME motif search and alignment, n = 9. Motif #3. Figure legend as above (Figure 3).

Figure 9. KRC/ZAS-C

MOTIF 3 width = 9 sites = 5 llr = 52 E-value = 3.9e+003

Simplified pos.-specific probability matrix	A	:	:	8	:	:	:	:	:
	C	:	2	:	:	:	:	:	4
	G	a	8	:	a	:	:	4	:	6
	T	:	:	2	:	a	a	6	a	:



Multilevel consensus sequence	G	G	A	G	T	T	T	T	G
		C	T				G		C

motifs

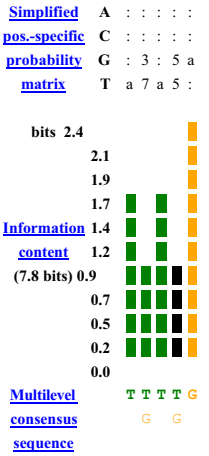
S100A4/mts1	G	G	T	t	T	T	T	c	C
RSS 9-mer	G	G	T	t	T	T	T	g	t
kappa B	G	G	A	c	T	T	T	c	C

NAME	STRAND	START	P-VALUE		SITES									
ZASC-79	+	4	5.56e-06	AGA	G	G	A	G	T	T	T	T	G	CTTTTATAGA
ZASC-46	+	3	5.56e-06	CA	G	G	A	G	T	T	T	T	G	TTACAAGAT
ZASC-71	+	7	7.57e-06	GGGTGT	G	G	A	G	T	T	G	T	C	ATATTTTTGG
ZASC-83	+	17	1.51e-05	ACAATGGTTC	G	C	A	G	T	T	G	T	G	
ZASC-38	+	5	3.34e-05	TTGC	G	G	T	G	T	T	T	T	C	TGTATATTGA

Figure 9
ZAS-N-selected sequence: MEME motif search and alignment, n = 9. Motif #1. Figure legend as above (Figure 3).

Figure 10. KRC/ZASN

MOTIF 1 width = 5 sites = 38 llr = 206 E-value = 6.6e-005

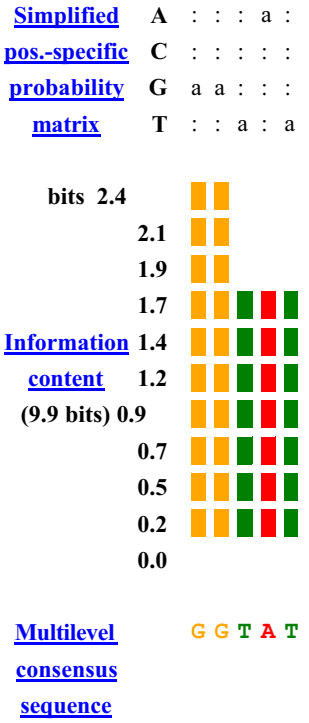


NAME	STRAND	START	P-VALUE		SITES	
ZASN-73	+	21	1.08e-03	ATTCTCATTT	T T T G G	
ZASN-72	+	7	1.08e-03	GGGGAT	T T T G G	ACTGTTTTTG
ZASN-58	+	21	1.08e-03	AGGCGTTGGC	T T T G G	
ZASN-51	+	21	1.08e-03	TATTTATGTG	T T T G G	
ZASN-48	+	21	1.08e-03	TCTAGTTTTG	T T T G G	
ZASN-46	+	10	1.08e-03	AGGATATGT	T T T G G	TTGATTCGTG
ZASN-40	+	21	1.08e-03	TTTGAGCTGG	T T T G G	
ZASN-39	+	21	1.08e-03	TTTGAGCTGG	T T T G G	
ZASN-37	+	21	1.08e-03	TGTTTGTTTT	T T T G G	
ZASN-27	+	21	1.08e-03	TGTATTGTTA	T T T G G	
ZASN-22	+	21	1.08e-03	ATTTAATGCA	T T T G G	
ZASN-17	+	17	1.08e-03	TATTTTAGCC	T T T G G	ATGG
ZASN-13	+	21	1.08e-03	CATTTGTTAA	T T T G G	
ZASN-65	+	21	1.77e-03	TTTTACTGTG	T G T G G	
ZASN-49	+	21	1.77e-03	GTTGTTCTTA	T G T G G	
ZASN-34	+	21	1.77e-03	GGTCTCTGTC	T G T G G	
ZASN-26	+	6	1.77e-03	GGTGG	T G T G G	TCTTTTTCTA
ZASN-4	+	14	1.77e-03	GAATGCTCAT	T G T G G	CTCTTGG
ZASN-70	+	18	3.46e-03	TAGTTGTTCA	T T T T G	TGC
ZASN-62	+	8	3.46e-03	ATGTGAA	T T T T G	TGACTTATGT
ZASN-61	+	12	3.46e-03	TAGTAGTGAA	T T T T G	CTTGTTGCG
ZASN-57	+	18	3.46e-03	TCTAATCTTA	T T T T G	TTG
ZASN-33	+	6	3.46e-03	GTGTG	T T T T G	TTTTATTTCT
ZASN-31	+	21	3.46e-03	TACAACTGCT	T T T T G	
ZASN-30	+	11	3.46e-03	CGTACGTTTG	T T T T G	ATGTTTCTCG
ZASN-23	+	12	3.46e-03	GGGAATTTCT	T T T T G	CTAGTGTGG
ZASN-21	+	11	3.46e-03	GGATTTGTTG	T T T T G	CCTTGCTCTG
ZASN-8	+	8	3.46e-03	AGTGGAA	T T T T G	CTTTTTGTTT
ZASN-7	+	6	3.46e-03	GGGAA	T T T T G	TCTCTATTTT
ZASN-1	+	6	3.46e-03	GGTAT	T T T T G	TGTTTATTTA
ZASN-71	+	6	4.54e-03	GGAAT	T G T T G	GTTCGTCATA
ZASN-68	+	20	4.54e-03	TCTTGTACAT	T G T T G	G
ZASN-64	+	21	4.54e-03	TTAATGATGA	T G T T G	
ZASN-56	+	20	4.54e-03	TGCTGTGATG	T G T T G	G
ZASN-19	+	20	4.54e-03	ATCATTTGTA	T G T T G	T
ZASN-12	+	17	4.54e-03	CTCTTAGCTC	T G T T G	AGGG
ZASN-5	+	12	4.54e-03	TAAGGAGTTC	T G T T G	CTATTTATG
ZASN-3	+	16	4.54e-03	GATGATTTGA	T G T T G	TCGTG

Figure 10
ZAS-N-selected sequence: MEME motif search and alignment, n = 5. Motif #1. Figure legend as above (Figure 3).

Figure 11. KRC/ZASN

MOTIF 2 width = 5 sites = 12 llr = 82 E-value = 3. 2e+003

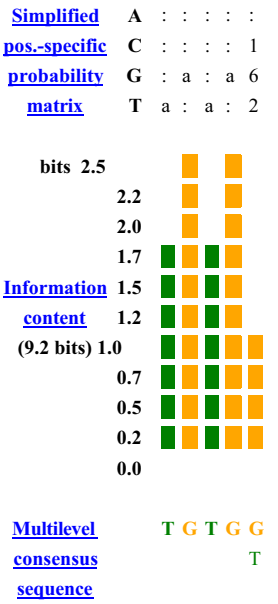


NAME	STRAND	START	P-VALUE		SITES	
ZASN-73	+	8	1.08e-03	GGATTCT	G G T A T	TCTCATTTTT
ZASN-64	+	1	1.08e-03		G G T A T	TGCTTTTAAT
ZASN-60	+	12	1.08e-03	GATTGCTCAT	G G T A T	TTGTATGGG
ZASN-58	+	2	1.08e-03	A	G G T A T	TTTTAGGCGT
ZASN-56	+	5	1.08e-03	CGTA	G G T A T	TGCTGTGATG
ZASN-54	+	1	1.08e-03		G G T A T	TACTTAGTCT
ZASN-52	+	1	1.08e-03		G G T A T	TCTCTAGTTT
ZASN-50	+	4	1.08e-03	GAA	G G T A T	TCGTAACTTG
ZASN-45	+	11	1.08e-03	GAATGTGCTT	G G T A T	TTGTTTTCGG
ZASN-18	+	2	1.08e-03	A	G G T A T	TTTCCTTCTA
ZASN-17	+	5	1.08e-03	GGTT	G G T A T	TTTAGCCTTT
ZASN-1	+	1	1.08e-03		G G T A T	TTTTGTGTTT

Figure 11
ZAS-N-selected sequence: MEME motif search and alignment, n = 5. Motif #2. Figure legend as above (Figure 3).

Figure 12. KRC/ZASC

MOTIF 1 width = 5 sites = 21 llr = 134 E-value = 9.4e-002



NAME	STRAND	START	P-VALUE		SITES	
ZASC-85	+	16	6.04e-04	ATACTGTTGA	T G T G G	ATTGT
ZASC-73	+	5	6.04e-04	ATGG	T G T G G	TGTTTTCCTT
ZASC-71	+	4	6.04e-04	GGG	T G T G G	AGTTGTCATA
ZASC-63	+	21	6.04e-04	ACCTAAATAT	T G T G G	
ZASC-62	+	21	6.04e-04	TTATGGATAT	T G T G G	
ZASC-50	+	21	6.04e-04	TTTATAATAC	T G T G G	
ZASC-41	+	20	6.04e-04	ACCTCATTAT	T G T G G	G
ZASC-33	+	21	6.04e-04	GCTTTTGTTA	T G T G G	
ZASC-31	+	21	6.04e-04	TTCTCAATTC	T G T G G	
ZASC-20	+	19	6.04e-04	TACCCAATAG	T G T G G	GG
ZASC-17	+	9	6.04e-04	AATATGTG	T G T G G	TTATTGATTT
ZASC-16	+	21	6.04e-04	TTTGTTGTTC	T G T G G	
ZASC-12	+	20	6.04e-04	ACTCACTCTT	T G T G G	G
ZASC-88	-	9	2.27e-03	CAGAAAAACC	T G T G T	AACAATCC
ZASC-82	+	20	2.27e-03	AGGGATTGCC	T G T G C	T
ZASC-81	+	15	2.27e-03	TTTTGGATAT	T G T G T	TATGAG
ZASC-68	+	3	2.27e-03	GT	T G T G T	ACTTTCTTCG
ZASC-45	+	18	2.27e-03	TCTAATTAAT	T G T G C	GAG
ZASC-44	+	18	2.27e-03	TCTAATTAAT	T G T G C	GAG
ZASC-29	+	18	2.27e-03	TAATTCTCTA	T G T G T	TGG
ZASC-1	+	8	2.27e-03	GAGATGT	T G T G T	TTTATTTGTT

Figure 12
ZAS-C-selected sequence: MEME motif search and alignment, n = 5. Motif #1. Figure legend as above (Figure 3).

KRC/ZAS-C binds the heptamer more efficiently, and that two KRC molecules may be needed to bind a single RSS element.

MAST analysis

The human and mouse genomes in the GenBank databases were searched with the KRC-bound sequences identified as PSSMs by the MEME program with the Multiple Alignment Search Tool (MAST). MAST is a program designed to search biological sequence databases for sequences that contain one or more of a group of known motifs [34]. Of a total of 923,310 sequences analyzed, only a total of 15 hits were obtained: 5 hits for KRC/ZAS-N (Figure 13) and 10 hits for KRC/ZAS-C (Figure 14). Significantly, 20% and 40% of the hits derived from KRC/ZAS-N and KRC/ZAS-C, respectively, came from the D gene segments of the variable region of human or mouse Ig heavy chains. For example, the KRC/ZAS-C consensus sequence "ATTTTGTGG" matches completely with 6 nucleotides of the RSS nonamer and 3 flanking nucleotides of the human IgH D1, D2, D3 and D4 gene segments [35]. Furthermore, the MAST search identified KRC-selected motifs near two chromosomal breakpoints: between a t(11:14) translocation of the Ig DH2-2 gene segment and the *B cell lymphoma 1* (*BCL-1*) gene in a mantle cell lymphoma [36], and at a t(11:19) translocation of the *myeloid lymphoid leukemia* (*MLL*) gene and the *ENL/MLLT1/LTG19* gene in a T-cell acute lymphoblastic leukemia [37]. The other hits were derived from loci of cellular genes, pseudogenes, or DNA fragments [41–44]. The expression of those genes has not been shown to be regulated by KRC or other family members, therefore, the biological significance of those MAST results is unknown. The result of the MAST analysis identifies probable endogenous KRC targets and suggests that KRC might interact with genetic elements involved in legitimate or illegitimate V(D)J recombination.

KRC-bound sequences match RSS elements within endogenous antigen receptor gene segments

Although RSS are evolutionarily conserved, the sequences of individual nonamer and heptamers vary [32,40]. To further determine the physiological significance of KRC's DNA binding, we compared the datasets with known endogenous RSSs of Ig and TCR loci. The mouse TCR αchain J gene segments (TCRAJ) and the human Igκ light chain genes (IgVκ) were analyzed, taking advantage of the fact that both loci have been sequenced and the location and sequence of their RSSs have been characterized [45,46]. Of the 97 IgVκ gene segments listed, 42 (43%) have a nonamer, a heptamer or both matching a sequence within the datasets (Figure 15). Similarly, 20 out of 59 (34%) of the TCRAJ RSS elements matched one or more sequences in the datasets (Figure 15). As controls, only 0 – 1.5% of random sequences from several data sets matched the endogenous RSS sequences (data not shown). Potentially, the DNA binding domains of KRC might interact with the range of heptamer and nonamer sequences found in endogenous antigen receptor loci.

Discussion

KRC was independently cloned due to its ability to bind the RSS [14] and the κB [4] motifs. Subsequently, sequence analysis identified KRC as a member of the ZAS family of proteins which share the ability to bind κB-like motifs [15]. DNA competition analysis showed that KRC fusion proteins containing the ZAS-C domain bind specifically to both the RSS and to the κB motif [14,28]. DNA footprinting analysis further showed that KRC/ZAS-C binds to specific nucleotides within the κB and the heptamer of the RSS [14]. In this study, using a PCR-based DNA-binding site-selection and amplification procedure, we demonstrated that both the N-terminal ZAS-N and the C-terminal ZAS-C domains are able to bind GT-rich DNA sequences, and confirmed that the RSS and κB motifs are the high-affinity targets of KRC.

In the site-selection experiment, the increasing yield of DNA-protein complexes in successive rounds of DNA amplification and purification suggest that KRC/ZAS-N and KRC/ZAS-C bound DNA specifically. Conceivably, repetitive binding, selection and amplification should have selected increasingly specific KRC targets as increasingly stringent binding conditions were established. As far as we know, this is the first DNA site-selection study to employ the MEME program to identify target consensus sequence. The program has been conventionally used to identify conserved motifs in proteins. It was chosen as a DNA motif search tool in this study due to its flexibility in recognizing several patterns within a set of sequences. It was able to identify multiple motifs that could not be recognized by other alignment programs, such as Pileup (GCG Software Package, [47]) or Clustal W [48] which have been used in other site-selection experiments to identify a single consensus sequence. The oligonucleotide pool presented in this study was composed of a random 25-mer flanked by specific primers. A relatively large target was used to accommodate ligands with a range of potential sizes and also to minimize the influence of flanking primer sequences. Inspection of the sequence alignments shows that all of the KRC-bound oligonucleotides align in the (+) orientation, suggesting that orientation of binding may have been influenced by the flanking sequence. However, the flanking sequences were constant throughout the oligonucleotide pool, which should have controlled for their relative contribution to consensus sequence.

Using the ZAS-N- and ZAS-C-selected DNA sequences as input, the MEME program discovered motifs containing sequences similar to the κB or Sb transcriptional enhancer

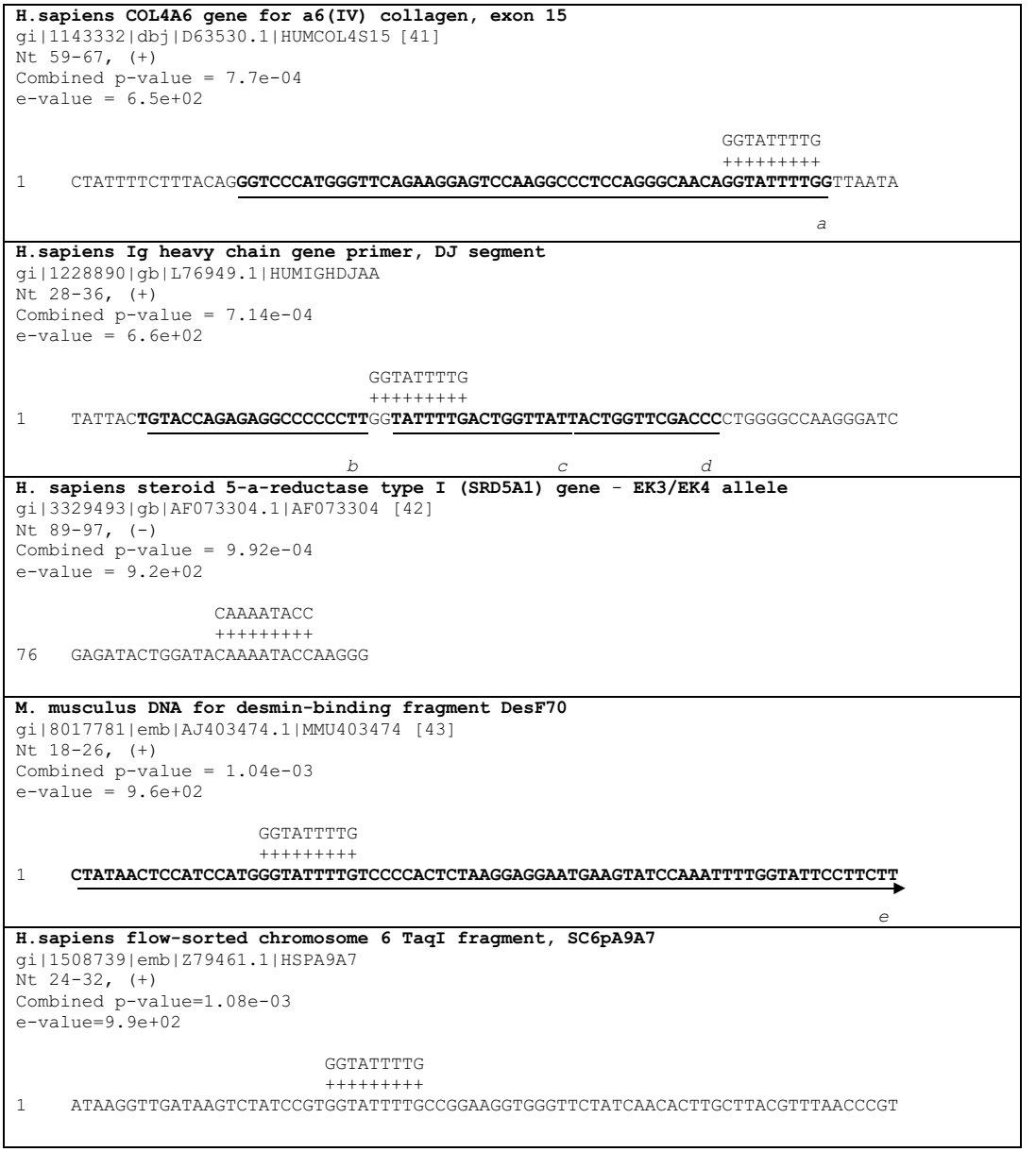


Figure 13

KRC/ZAS-N site selection consensus motif search: MAST results Human and murine genomic DNA identified by MAST by searching with the KRC/ZAS-N are listed. Features of DNA sequence from GenBank/NCBI are delineated by bar and footnote as follows: a. H. Sapiens COL4A6 gene, nt 16–60: COL4A6 exon 15 b. H. Sapiens IgH gene primer, DJ segment, 7–27: leukemia-specific primer c. H. Sapiens IgH gene primer, DJ segment, 30–46: D segment (Dxp1) d. H. Sapiens IgH gene primer, DJ segment, 47–59: J segment (J5) e. M. Musculus DNA for desmin-binding fragment DesF7, 1–107: LINE element The first line describes the GenBank-designated name of the DNA sequence. The next line lists the accession number and any publications associated with the sequence. The third line indicates the target sequence nucleotide positions of overlap as well as orientation. The ***combined p-value*** is defined as the probability of a randomly generated sequence of the same length having ***sequence p-values*** whose product is at least as small as the product of the sequence p-values of the matches of the motifs to the given sequence. The ***E-value*** of the match of a sequence in a database to a group of motifs is defined as the expected number of sequences in a random database of the same size that would match the motifs as well as the sequence does and is equal to the combined p-value of the sequence times the number of sequences in the database. (Above is described in detail at MAST [http://www.sdsc.edu/mast][34]).

Figure 14

KRC/ZAS-C site selection consensus motif search: MAST results Human and murine genomic DNA identified by MAST by searching with the KRC/ZAS-C are listed. Nomenclature is described above (Figure 13). Features of DNA sequence from GenBank/NCBI are delineated by bar and footnote as follows: f. H. Sapiens ATP7B pseudogene, 1–54: ATP7B exon 6 g. H. Sapiens ATP7B pseudogene, > = 55: ATP7B exon 7 h. H. Sapiens clone 19r DH2-2/bcl-1 gene fusion reciprocal breakpoint, 1–54: IgHD (14q32) i. H. Sapiens clone 19r DH2-2/bcl-1 gene fusion reciprocal breakpoint, 55: N-nucleotide j. H. Sapiens clone 19r DH2-2/bcl-1 gene fusion reciprocal breakpoint, > = 56: bcl-1 (11q13) k. M. Musculus DNA for GFAP-binding fragment GFAPA10, 1–82: LINE element k. M. Musculus DNA for GFAP-binding fragment GFAPC13, 1–83: LINE element l. H. Sapiens ENL translocation, 1–51: ENL gene (partial) m. H. Sapiens ENL translocation, 55–84: 3' ENL 11q23 n. H. Sapiens IgH D1-4, 1–9: nanomer o. H. Sapiens IgH D1-4, 22–28: heptamer p. H. Sapiens IgH D1-4, 29–59: D region q. H. Sapiens IgH D1-4, 60–66: heptamer r. M. Musculus aldose reductase gene, 5–75: exon 4

motifs as well as the conserved heptamer or nonamer elements of the RSS. Generally, the significance of the llr and E-values scores of a motif generated by MEME increased with its length whereas the information content per nucleotide position decreased (Figures 3,4,5,6,7,8,9,10,11,12). These are statistical values showing different parameters of a motif: llr and E-values reflect the likelihood of a motif being generated at random whereas the information content represents the degree of conservation. In our analysis, passes of MEME where the width was not set to a fixed length but to a given range of nucleotides always yielded the longest motif. The MEME program aims at generating motifs with the most probable occurrence, and the length of a motif may override other parameters in the algorithm. MEME is a useful tool with which to discover the best motif among DNA sequences provided the length is specified and determined experimentally.

Given that the crystal structure of the DNA-protein complex revealed that the C_2H_2 zinc finger pair of TTK, like the first two zinc fingers of DNA-Zif268, binds five base-pairs [27], we hypothesize that each zinc finger pair of KRC may also interact with five base-pairs. Passes of MEME for pentameric motifs yielded homologous TG-rich sequences for the KRC/ZAS-N and KRC/ZAS-C datasets: T(T/G)T(T/G)G and GGTAT for KRC/ZAS-N, and TGTGG/T for KRC/ZAS-C. We had previously shown by methylation interference analysis that KRC/ZAS-C bound specifically to the sequence TGTGG within the context of the canonical RSS heptamer plus the immediately flanking guanine [14]. Because the pentamer motif for KRC/ZAS-C predicted by MEME completely matched with the empirical results, we conclude that the two pentameric motifs discovered by MEME are likely authentic binding sites for KRC/ZAS-N as well.

A. Nonamers

GGTTTTTGT(ZASC-47)	IgVk 1-15d	IgVk 1-15p	IgVk 1-17d	IgVk 2-43d	IgVk 2-43p
	IgVk 2-54d	IgVk 2-54p	IgVk 3-14d	IgVk 3-14p	IgVk 3-21d
	IgVk 3-21p	IgVk 3-27d	IgVk 3-27p	IgVk 3-38d	IgVk 3-38p
	IgVk 3-45d	IgVk 3-45p	IgVk 5-3p	IgVk 7-6p	TCRAJ34
	TCRAJ49	TCRAJ61			
GGTTTATGT(ZASC-7)	IgVk 1-19p	IgVk 1-22d	IgVk 1-22p	IgVk 1-24d	IgVk 1-24p
	IgVk 1-28p	IgVk 1-28d	IgVk 1-30d	IgVk 1-30p	IgVk 1-66d
	IgVk 1-66p	IgVk 1-71p	IgVk 1-71d	IgVk 1-08p	TCRAJ45
TGTTTTTGT(ZASN-72)	TCRAJ03	TCRAJ33	TCRAJ50	TCRAJ52	TCRAJ53
GTTTTCTGT(ZASC-38)	TCRAJ46				
CCATTTTGT(ZASN-70)	TCRAJ08	TCRAJ09			
GGTTTGTAT(ZASN-22)	IgVk 2-35d	IgVk 2-35p			
GGTATTTGT(ZASN-45)	TCRAJ05				
AATTCTTGT(ZASN-65)	TCRAJ04				
--GTTTTGA(ZASN-30)	TCRAJ25				

B. Heptamers

CATTGTG(ZASN-4)	IgVk 2-20d	IgVk 2-20p	IgVk 2-42p	IgVk 2-42d	TCRAJ21
	TCRAJ56				
CGTGGTC(ZASN-34)	TCRAJ14				
GTCTGTG(ZASN-34)	TCRAJ33				
TGCTGTG(ZASN-56)	TCRAJ11	TCRAJ50			
TGATGTG(ZASN-56,85)	TCRAJ31				
TACTGTG(ZASC-50,65)	IgVk 3-60d	IgVk 3-60p			
GGGTGTG(ZASC-71)	TCRAJ57				
CACTCTA(ZASC-79)	IgVk 1-41d	IgVk 1-41p			
TCCTGTG(ZASC-88,31,16)	TCRAJ61				

Figure 15
Comparison of KRC-selected sequence with IgVκ and TCRAJ RSS elements. The KRC-bound sequences were compared to known RSS sequence in the murine TCR-alpha genes and the human immunoglobulin heavy chain genes, the loci of which have been fully sequenced [45,46]. Individual sequences in the training sets were compared manually with (A) RSS nonamers or (B) RSS hepamers defined in of the human Ig variable region gene segments of κ light chain (Table I of [45]) and the mouse TCRAJ gene segments (Fig. 2 of [46]). The nucleotide sequence of a RSS nonamer or heptamer was listed in the first column, followed by the name of the site selected sequence(s), and then by the gene segments with matching RSS elements. Sequence in the KRC fusion protein-bound oligonucleotides overlapped with heptamer, nonamer, or both in the RSS of 43.3% of human IgVκ genes and in 33.9% of murine TCRAJ genes. These calculations include only the IgVκ [45] and TCRAJ genes [46] which contain a nonamer, a heptamer, or both.

The putative DNA binding sequences of the ZAS-N and ZAS-C domains as determined by MEME (width = 5) were homologous. Specific DNA binding of separate-paired C_2H_2 zinc fingers depends on the amino acid sequences of the finger domains, the linker sequence between fingers, and the higher-ordered structure of fingers. The structure of individual C_2H_2 fingers as determined by 2D NMR methods has shown that each zinc finger consists of two N-terminal short anti-parallel β sheets followed by an α helix. The amino acid residues at position -1, 2, 3 and 6 of the α helix form base contacts with DNA [Reviewed in [49]]. The amino acid sequences of the zinc finger pairs among the ZAS proteins are highly conserved [Reviewed in [25]]. The conservation of the zinc fingers from invertebrate to vertebrate species and their common DNA binding target sequences suggest that these proteins may play similar physiological roles in diverse organisms. While the overall sequence identity between human and mouse KRC is 80%, their corresponding zinc finger pairs are completely identical [25] (Figure 16). Notably, for those critical amino acid residues within the α helical regions of the finger described above, they are identical at all corresponding positions between the first and second zinc fingers of the ZAS-N and ZAS-C domains except at position 3 of the first zinc finger (Val in ZAS-N; Met in ZAS-C) and at position 2 of the second zinc finger (Ser in ZAS-N; Gly in ZAS-C). Because Val and Met are both non-polar amino acids, and Ser and Gly are both polar and uncharged ami-

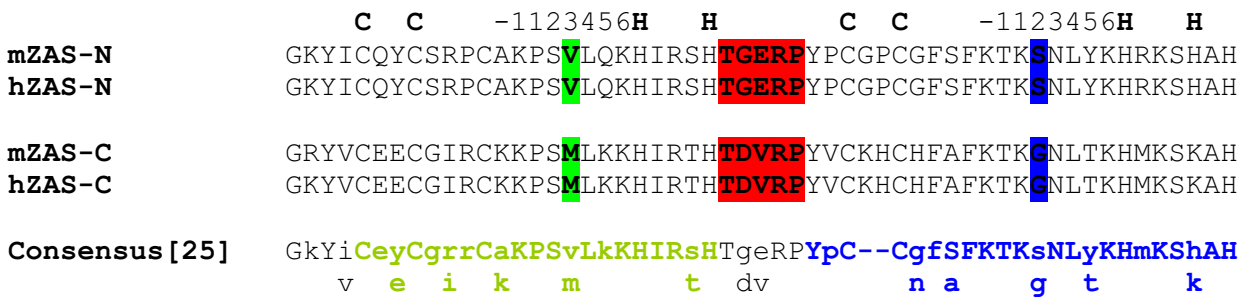


Figure 16
Comparison of KRC protein zinc finger domains. The zinc finger amino acid sequences of the KRC ZAS domains are compared. Mouse and human KRC ZAS-N amino acid sequences are aligned above the corresponding mouse and human KRC ZAS-C sequence. Above the sequences, the C and H represent the canonical Cys and His residues of the C_2H_2 domains, and the numbers represent the position within the alpha helix. The ***consensus sequence*** is derived from sequence analysis of all the known ZAS protein sequences [25]. The zinc finger 1 and zinc finger 2 domains within the ZAS domains are designated in green and blue font, respectively, in the consensus sequence. Amino acid residues that are common to all ZAS family members are capitalized. The critical amino acid residues within the α helical regions of the zinc finger regions are identical at all corresponding positions between the first and second zinc fingers of the ZAS-N and ZAS-C domains except at position 3 of the first zinc finger (outlined in gray) and at position 2 of the second zinc finger (outlined in gray). The linker region between the two zinc fingers is outlined in red.

no acids, those amino acid substitutions between ZAS-N and ZAS-C may result in minor changes in the tertiary structure of the zinc fingers which account for the subtle differences in the DNA binding properties of the ZAS-N and ZAS-C domains. The differences in the linker regions, TGERP for ZAS-N and TDVRP for ZAS-C, which presumably interact with the sugar-phosphate backbone of DNA, and the observation that KRC/ZAS-C more readily forms higher-ordered structures with DNA than KRC/ZAS-N may also contribute to the differences in the DNA binding of ZAS-N and ZAS-C.

Based on the results here, we hypothesize that each DNA binding domain of KRC binds to pentameric TG-rich sequences. Two KRC binding sites when put together can form some longer known KRC targets. For example a copy of GGTTT and its complement can form a sequence GG(N_{5-6})CC, fulfilling the minimal DNA binding requirement for the ZAS proteins other than a lack of the 5' guanine [25]. Similarly, the GT-rich RSS nonamer and the palindromic RSS heptamer can serve as binding sites of KRC. Furthermore, our hypothesis can explain why half-sites but not complete KRC targets were frequently found in the protein-selected datasets. Although previous results of protein titration experiments suggested that KRC/ZAS-C binds DNA in a cooperative manner for a given oligonucleotide, the presence of multiple binding sites might not be favored over a single site in our site selection assay which used a degenerate pool of oligonucleotides and limited rounds of amplification. The data suggest that both DNA binding domains of KRC are potentially capable of binding to either RSS heptamer or nonamer. Because the pentameric motifs derived from the KRC/ZAS-N dataset more closely resemble the canonical RSS nonamer and the motif derived from the KRC/ZAS-C dataset more closely resemble the canonical RSS heptamer and a canonical sequence was derived from the majority of sequences, we propose that the ZAS-N domain of KRC binds RSS nonamers more frequently than the ZAS-C domain, and *vice versa* for the RSS heptamer *in vivo*.

Conclusions

Our results suggest that KRC binds with individual endogenous RSS elements and transcriptional enhancer motifs. As the most abundant RSS-binding species detected in thymus [50], it is intriguing to propose a role for KRC in regulation of the V(D)J recombination process. Several studies have shown that the RSS themselves may act as *cis*-acting elements which influence recombination frequency [51–56]. Furthermore, affinity of KRC for the RSS has been shown to vary inversely with activation of the catalytic components of the V(D)J recombinase, RAG1 and RAG2 [57]. It is possible that differential affinity of KRC for individual RSS influences RSS utilization by the recombinase, allowing differential recombination of gene segments.

Our finding that KRC binds to the RSS as well as the κB motif may also provide a link between transcription and recombination in the context of the accessibility model [58,59]. Enhancer or promoter elements are important for the recombination process in cell lines and animal models [60]. Similarly, expression of transcription factors, in conjunction with the recombination activating genes, has

been shown to induce V(D)J recombination in non-lymphoid tissues by rendering RSS accessible to the recombinase [61]. The κB motif, first found in the Igκ light chain [62], and later in the TCR β2 locus [31], has been shown to promote V(D)J recombination by modulating locus accessibility [63]. In addition to influencing recombination by binding of RSS, KRC binding of the κB motif may modulate accessibility and transcription of target loci. The ability of KRC to promote transcription of target genes has been demonstrated for the S100/mts1 gene by binding at the Sb enhancer motif [4]. Similarly, binding of κB-like motifs by other ZAS proteins has also been implicated in transcriptional regulation [1–3,17]. Considering KRC's target sequences, the κB motif and the RSS, the two binding domains on a single KRC protein could theoretically bring together *cis*-acting DNA elements for gene regulation, V(D)J recombination, or both. Such a molecule could coordinate transcription of individual promoter or enhancer elements, and/or could physically connect different cellular machineries via distinct DNA elements. KRC could provide a link between the fundamental processes of DNA transcription and V(D)J recombination.

Methods

Oligonucleotides

Oligonucleotides were synthesized chemically (Life Technologies, Rockville, MD. BSS1: 5'-GACGGTATCGATAAGCTT-3'; BBS2: 5'-CCGGGCTGCAGGAATTC-3'; and BSS4: 5'-GACGGTATCGATAAGCTT$(N)_{25}$GAATTCCTGCAGCCCGG-3' where N is A, T, C, or G.

Fusion proteins

The fusion proteins KRC/ZAS-N and KRC/ZAS-C were produced in *E. coli* and purified by affinity chromatography as described previously [24,28]. The regions of KRC used to generate KRC/ZAS-N and KRC/ZAS-C are schematically shown in Fig. 1A.

Site selection amplification binding assay, DNA cloning and sequencing

Site selection amplification binding assay was performed as described [64] with modifications. In the first DNA-protein binding reaction, [^{32}P]-labeled double stranded oligonucleotides were generated by first annealing BSS2 (50 ng) to the BSS4 oligonucleotide (500 ng) then end-filling with 250 μM each of dATP, dGTP, and dTTP, 50 μCi of ^{32}P-dCTP and Klenow. The oligonucleotide pool (~500 ng) was incubated with KRC/ZAS-N or KRC/ZAS-C (100 μg each) and 10 μg of non-specific competitor DNA poly(dI-dC). DNA-protein complexes and free DNA were resolved on a 5% polyacryamide gel. After autoradiography, DNA-protein complexes were isolated from the gels and were eluted from the gel slices by incubation in 1 ml of 10 mM Tris (pH 8.0) and 1 mM EDTA at 42°C for 4 hours. DNAs were purified by phenol/chloroform extraction, followed by alcohol precipitation, then were amplified by PCR using the BSS1 and BSS2 primer set. Subsequently, a portion of the DNA was labeled with [^{32}P]dCTP and used for the next round of site-selection. After the first round, the stringency of each succeeding round of site selection was increased by using successively less (0.5×) fusion proteins and more (4×) non-specific competitor DNA. Protein-bound oligonucleotides from the fifth round of selection were purified and subcloned into plasmid vectors pCR 2.1 (Invitrogen, Carlsbad, CA). Plasmid DNA was prepared from cohorts of bacteria colonies using a kit (Qiagen, Carlsbad, CA). The nucleotide sequences of the inserts were determined using automated DNA sequencing procedures performed by the DNA Sequencing Core Facility at the Ohio State University.

Sequence analysis

Sequence analysis was performed using the computer programs MEME (version 3.0) [30] and MAST (version 3.0) [34]. The data of both programs were processed on the Cray T3E supercomputer at the San Diego Supercomputer Center accessed through the Internet: MEME [http://www.sdsc.edu/meme]. For MEME, the free parameters of the analysis were set as the following: (i) the occurrences of a single motif distributed among the sequences were zero or one per sequence; (ii) the maximum number of motifs to find was five; (iii) the optimum width of each motif ranged from 3 to 25 nucleotides; and (iv) both strands of DNA were searched. For MAST, only MEME PSSMs with an E-value < 1 were presented, and the reverse complement DNA strand was considered with the forward orientation in the search.

Authors' contributions

CEA carried out the site-selection experiments, sequence analysis, and drafted the manuscript. CHM prepared fusion proteins and assisted with experimental design of the site-selection assay. LCW conceived of the study, participated in data analysis, and finalized the manuscript. All authors read and approved the final manuscript.

List of abbreviations

Shn: schnurri

MEME: Motif Expectation Maximum for Motif Elicitation

MAST: Multiple Alignment Search Tool

PSSM: position specific scoring matrix

TTK: tramtrack

EMSA: electrophoretic mobility shift assay

Ig: immunoglobulin

TCR: T cell receptor

A/C/G/T: adenine/cytosine/guanine/thymine

llr: log likelihood ratio

nt: nucleotide

Acknowledgements

This research was supported in part by grant GM48798 (LCW) from the National Institutes of Health and by grant P30 CA16058 from National Cancer Institute. CEA was funded by a T-32 pre-doctoral fellowship (National Cancer Institute, Bethesda, MD). We thank Dr. Michael Gribskov for assistance with the MEME and MAST analysis.

References

1. Seeler JS, Muchardt C, Suessle A, Gaynor RB: **Transcription factor PRDII-BF1 activates human immunodeficiency virus type 1 gene expression.** *J Virol* 1994, **68:**1002-1009
2. Brady JP, Kantorow M, Sax CM, Donovan DM, Piatigorsky J: **Murine transcription factor α-A crystalline binding protein I.** *J Biol Chem* 1995, **270:**1221-1229
3. Dörflinger U, Pscherer A, Moser M, Rümmele P, Schüle R, Buettner R: **Activation of somatostatin receptor II expression by transcription factors MIBP1 and SEF-2 in the murine brain.** *Mol Cell Biol* 1999, **19:**3736-3747
4. Hjelmsoe I, Allen CE, Cohn MA, Tulchinsky EM, Wu LC: **The κB and V(D)J recombination signal sequence binding protein KRC regulates transcription of the mouse metastasis associated gene *S100A4/mts1*.** *J Biol Chem* 2000, **275:**913-920
5. Tanaka K, Matsumoto Y, Nakatani F, Iwamoto Y, Yamada Y: **A zinc finger transcription factor, alpha A-crystallin binding protein 1, is a negative regulator of the chondrocyte-specific enhancer of the alpha 1 (II) collagen gene.** *Mol Cell Biol* 2000, **20:**4428-4435
6. Singh H, LeBowitz JH, Baldwin AS, Sharp P: **Molecular cloning of an enhancer binding protein, isolation by screening of an expression library with a recognition site DNA.** *Cell* 1988, **52:**415-423
7. Maekawa T, Sakura H, Sudo T, Ishii S: **Putative metal finger structure of the human immunodeficiency virus type 1 enhancer binding protein HIV-EP1.** *J Biol Chem* 1989, **264:**14591-14593
8. Baldwin AS, LeClair KP, Singh H, Sharp PA: **A large protein containing zinc finger domains binds to related sequence elements in the enhancers of the class I major hisocompatibility complex and kappa immunoglobulin genes.** *Mol Cell Biol* 1990, **10:**1406-1414
9. Fan CM, Maniatis T: **A DNA-binding protein containing two widely separated zinc finger motifs that recognize the same DNA sequence.** *Genes Dev* 1990, **4:**29-42
10. Rustgi A, van't Veer LJ, Bernards R: **Two genes encode factors with NF-κB- and H2TF1-like DNA-binding properties.** *Proc Natl Acad Sci USA* 1990, **87:**8707-8710
11. Nomura N, Zhao MJ, Nagase T, Maidawa T, Ishizaki S Tabata R, Ishii S: **HIV-EP2, a new member of the gene family encoding the human immunodeficiency virus type 1 enhancer-binding protein. Comparison with HIV-EP1/PRDII-BF1/MBP-1.** *J Biol Chem* 1991, **266:**8590-8594
12. van't Veer LJ, Lutz PM, Isselbacher KJ, Bernards R: **Structure and expression of major histocompatibility complex-binding protein 2, a 275-kDa zinc finger protein that binds to an enhancer of major histocompatibility complex class I genes.** *Proc Natl Acad Sci USA* 1992, **89:**8971-8975
13. Hicar MD, Liu Y, Allen CE, Wu LC: **Structure of the human zinc finger protein KRC: Molecular cloning, expression, exon-intron structure, and comparison with paralogous genes HIV-EP1 and HIV-EP2.** *Genomics* 2001, **71:**89-100
14. Wu LC, Mak CH, Dear N, Boehm T, Foroni L, Rabbitts TH: **Molecular cloning of a zinc finger protein which binds to the heptamer of the signal sequence for V(D)J recombination.** *Nucleic Acids Res* 1993, **21:**5067-5073
15. Wu LC, Liu Y, Strandtmann J, Mak CH, Lee B, Li Z, Yu CY: **The mouse DNA binding protein Rc for the kappa B motif of transcription and for the V(D)J recombination signal sequences contains composite DNA-protein interaction domains and GTPase motifs.** *Genomics* 1996, **35:**415-424
16. Ron D, Brasier AR, Habener JF: **Angiotensinogen gene-inducible enhancer-binding protein 1, a member of a new family of large nuclear proteins that recognize nuclear factor kappa B-binding sites through a zinc finger motif.** *Mol Cell Biol* 1991, **11:**2887-2895
17. Makino R, Akiyama K, Yasuda J, Mashiyama S, Honda S, Sekiya T, Hayashi K: **Cloning and characterization of a c-myc intron binding protein (MIBP1).** *Nucleic Acids Res* 1994, **22:**5679-5685
18. Arora K, Dai H, Kazuko SG, Jamal J, O'Connor MB, Letsou A, Warrior R: **The *Drosophila schnurri* gene acts in the Dpp/TGF beta signalling pathway and encodes a transcription factor homologous to the human MBP family.** *Cell* 1995, **81:**781-790
19. Grieder NC, Nellen D, Burke R, Basler K, Affolter M: ***Schnurri* is required for *Drosophila* Dpp signalling and encodes a zinc finger protein similar to the mammalian transcription factor PRDII-BF1.** *Cell* 1995, **81:**791-800
20. Stahling-Hampton K, Laughon AS, Hoffman FM: **A *Drosophila* protein related to the human zinc-finger transcription factor PRFII/MIBP1/HIV-EP1 is required for dpp signaling.** *Development* 1995, **121:**3393-3403
21. Dai H, Hogan C, Gopaladrishnan B, Torres-Vasquez J, Nguyen J, Park S, Raferty LA, Warrior R, Arora K: **The zinc finger protein schnurri acts as a Smad partner in mediating the transcriptional response to decapentaplegic.** *Develop Biol* 2000, **227:**373-387
22. Torres-Vasquez J, Park S, Warrior R, Arora K: **The transcription factor Schnurri plays a dual role in mediating dpp signaling during embryogenesis.** *Development* 2001, **128:**1657-1670
23. Muchardt C, Seeler JS, Nirula A, Shurland DL, Gaynor RB: **Regulation of human immunodeficiency virus enhancer function by *PRDII-BF1* and *c-rel* gene products.** *J Virol* 1992, **66:**244-250
24. Mak CH, Li Z, Allen CE, Liu Y, Wu LC: ***KRC* transcripts: Identification of an unusual splicing event.** *Immunogenetics* 1998, **48:**32-39
25. Wu LC: **ZAS: C_2H_2 zinc finger proteins involved in growth and development.** *Gene Expression* 2002, **10:**
26. Iuchi S: **Three classes of C_2H_2 zinc finger proteins.** *Cell Mol Life Sci* 2001, **58:**625-635
27. Tseng H, Green H: **Basonuclin: a keratonicyte protein with multiple paired zinc fingers.** *Proc Natl Acad Sci USA* 1992, **89:**10311-10315
28. Mak CH, Strandtmann J, Wu LC: **The V(D)J recombination signal sequence and κB binding protein Rc binds DNA as dimers and forms multimeric structures with its DNA ligands.** *Nucleic Acids Res* 1994, **22:**383-390
29. Tulchinsky E, Prokhortchouk E, Georgiev G, Lukanidin E: **A kappa B-related binding site is an integral part of the mts1 gene composite enhancer element located in the first intron of the gene.** *J Biol Chem* 1997, **272:**4828-4835
30. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36
31. Sen R, Baltimore D: **Multiple nuclear factors interact with the immunoglobulin enhancer sequences.** *Cell* 1986, **46:**705-716
32. Akira S, Okazaki K, Sakano H: **Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining.** *Science* 1987, **238:**1134-1138
33. Fairall L, Schwabe JW, Chapman L, Finch JT, Rhodes D: **The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition.** *Nature* 1993, **366:**483-487
34. Bailey TL, Gribskov M: **Combining evidence using p-values, application to sequence homology searches.** *Bioinformatics* 1998, **14:**48-54
35. Siebenlist U, Ravetch JV, Korsmeyer S, Waldmann T, Leder P: **Human immunoglobulin D segments encoded in tandem multigenic families.** *Nature* 1981, **294:**631-635
36. Welzel N, Le T, Marculescu R, Mitterbauer G, Chott A, Pott C, Kneba M, Du MQ, Kusec R, Drach et al J: **Templated nucleotide addi-**

tion and immunoglobulin JH-gene utilization in t(11:14) junctions: Implications for the mechanism of translocation and the origin of mantle cell lymphoma. *Cancer Res* 2001, **6:**1629-1636
37. Chervinsky DS, Sait SN, Nowak NJ, Shows TB, Aplan PD: **Complex MLL rearrangement in a patient with T-cell acute lymphoblastic leukaemia.** *Genes Chromosomes Cancer* 1995, **14:**76-84
38. Badding H: **Determination of the molecular weight of DNA-bound protein(s) responsible for gel electrophoretic mobility shift of linear DNA fragments exemplified with purified viral myb protein.** *Nucleic Acids Res* 1988, **16:**5241-5248
39. Kim J, Zwieb C, Wu C, Adhya S: **Bending of DNA by gene-regulatory proteins: construction and use of a DNA bending vector.** *Gene* 1989, **85:**15-23
40. 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
41. Oohashi T, Ueki Y, Sugimoto M, Ninomiya Y: **Isolation and structure of the *COL4A6* gene encoding the human alpha 6 (IV) collagen chain and comparison with other type IV collagen genes.** *J Biol Chem* 1995, **270:**26863-26867
42. Eminovic I, Liovic M, Prezelj J, Kocijancic A, Rozman D: **New steroid 5-α-reductase type 1 (SRD5A1) homologous sequences on chromosomes 6 and 8.** *Pflugers Arch* 2001, **442:**R187-R189
43. Tostonog GV, Wang X, Shoeman R, Traub P: **Intermediate filaments reconstituted from vimentin, desmin, and the glial fibrillary acidic protein selectively bind repetitive and mobile DNA sequences from a mixture of mouse genomic DNA fragments.** *DNA Cell Biol* 2000, **19:**647-677
44. McGowan MH, Iwata T, Carper DA: **Characterization of the mouse *aldose reductase* gene and promoter in a lens epithelial cell line.** *Mol Vis* 2000, **4:**2
45. Kawasaki K, Minoshima S, Nakato E, Shibuya K, Shintani A, Asakawa S, Sasaki T, Klobeck HG, Combriato G, Zachau HG, Shimizu N: **Evolutionary dynamics of the human immunoglobulin κ locus and the germline repertoire of the Vκ genes.** *Eur J Immuno* 2001, **31:**1017-1028
46. Koop BF, Rowen L, Wang K, Kuo CL, Seto D, Lenstra JA, Howard S, Shan W, Deshpande P, Hood L: **The human T-cell receptor TCRAC/TCRDC (Cα/Cδ) region: Organization, sequence, and evolution of 97.6 kb of DNA.** *Genomics* 1994, **19:**478-493
47. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680
48. Devereux J, Haeberli P, Smithies O: **A comprehensive set of sequence analysis programs for the VAX.** *Nucleic Acids Res* 1984, **12:**387-395
49. Coleman JE: **Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins.** *Ann Rev Biochem* 1992, **61:**897-946
50. Hicar MD, Robinson ML, Wu LC: **Embryonic expression and regulation of the large zinc finger protein KRC.** *Genesis* 2002, **33:**8-20
51. Feeney AJ, Tang A, Ogwaro KM: **B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilitzation.** *Immunol Rev* 2000, **175:**59-69
52. Bassing CH, Alt FW, Hughes MM, D'Auteuil M, Wehrly TD, Woodman BB, Gartner F, While FM, Davidson L, Sleckman BP: **Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule.** *Nature* 2000, **495:**583-586
53. Larijani M, Yu CC, Golub R, Lam QL, Wu GE: **The role of components of recombination signal sequences in immunoglobulin gene segment usage: a V81x model.** *Nucleic Acids Res* 1999, **27:**2304-2309
54. Nadel B, Tang A, Lugo G, Love V, Escuro G, Feeney AJ: **Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus influenza type b disease.** *J Immunol* 1998, **161:**6068-6073
55. Pan PY, Lieber MR, Teale JM: **The role of recombination signal sequences in the preferential joining by deletion in DH-JH recombination and in the ordered rearrangement of the IgH locus.** *Int Immunol* 1997, **9:**515-522
56. VanDyk LF, Wise TW, Moore BB, Meek K: **Immunoglobulin D(H) recombination signal sequence targeting: effect of D(H) coding and flanking regions and recombination partner.** *J Immunol* 1996, **157:**4005-4015
57. Wu LC, Hicar MD, Hong JW, Allen CE: **The DNA binding ability of HIVEP3/KRC decreases upon activation of V(D)J recombination.** *Immunogenetics* 2001, **53:**564-571
58. Yancopoulos G, Alt F: **Developmentally controlled and tissue-specific expression of unrearranged V_H gene segments.** *Cell* 1985, **40:**271-281
59. Schlissel MS, Stanhope-Baker P: **Accessibility and the developmental regulation of V(D)J recombination.** *Seminars in Immunology* 1997, **9:**161-170
60. Sikes ML, Suarez CC, Oltz EM: **Regulation of V(D)J recombination by transcriptional promoters.** *Mol Cell Biol* 1999, **19:**2773-2781
61. Langerak AW, Wolvers-Tettero IL, van Gastel-Mol EJ, Oud ME, van Dongen JJ: **Basic helix-loop-helix proteins E2A and HEB induce immature T-cell receptor rearrangements in nonlymphoid cells.** *Blood* 2001, **98:**2456-2465
62. Kelley EE, Wiedemann LM, Pittet AC, Strauss S, Nelson KJ, Davis J, VanNess B, Perry RP: **Nonproductive kappa immunoglobulin genes: recombinational abnormalities and other lesions affecting transcription, RNA processing, turnover, and translation.** *Molecular and Cellular Biology* 1985, **5:**1660-1675
63. Jamieson C, Mauxion F, Sen R: **Identification of a functional NF-kappa B binding site in the murine T cell receptor beta 2 locus.** *J Exp Med* 1989, **170:**1737-1743
64. Chen CY, Schwartz RJ: **Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5.** *J Biol Chem* 1995, **270:**15628-15633