

Specificity landscapes of DNA binding molecules elucidate biological function

Clayton D. Carlson^{a,1}, Christopher L. Warren^{a,1}, Karl E. Hauschild^a, Mary S. Ozers^a, Naveeda Qadir^a, Devesh Bhimsaria^a, Youngsook Lee^b, Franco Cerrina^c, and Aseem Z. Ansari^{a,d,2}

^aDepartment of Biochemistry, University of Wisconsin, 433 Babcock Drive, Madison, WI 53706; ^bDepartment of Anatomy, University of Wisconsin, 1300 University Avenue, Madison, WI 53706; ^cDepartment of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706; and ^dThe Genome Center of Wisconsin, University of Wisconsin, 425G Henry Mall, Madison, WI 53706

Communicated by Sankar Adhya, National Institutes of Health, NCI, Bethesda, MD, December 11, 2009 (received for review August 21, 2009)

Evaluating the specificity spectra of DNA binding molecules is a nontrivial challenge that hinders the ability to decipher gene regulatory networks or engineer molecules that act on genomes. Here we compare the DNA sequence specificities for different classes of proteins and engineered DNA binding molecules across the entire sequence space. These high-content data are visualized and interpreted using an interactive “specificity landscape” which simultaneously displays the affinity and specificity of a million-plus DNA sequences. Contrary to expectation, specificity landscapes reveal that synthetic DNA ligands match, and often surpass, the specificities of eukaryotic DNA binding proteins. The landscapes also identify differential specificity constraints imposed by diverse structural folds of natural and synthetic DNA binders. Importantly, the sequence context of a binding site significantly influences binding energetics, and utilizing the full contextual information permits greater accuracy in annotating regulatory elements within a given genome. Assigning such context-dependent binding values to every DNA sequence across the genome yields predictive genome-wide binding landscapes (genomespaces). A genomespace of a synthetic DNA binding molecule provided insight into its differential regulatory activity in cultured cells. The approach we describe will accelerate the creation of precision-tailored DNA therapeutics and uncover principles that govern sequence-specificity of DNA binding molecules.

chemical genomics | Cognate Site Identification | DNA binders | genomespaces | Energy Landscapes

A major goal at the interface of synthetic biology, chemistry, and personalized medicine is to create molecules that can specifically regulate genes and thereby dictate cell fate (1–4). An approach toward this goal has been to engineer natural DNA binding molecules to target desired DNA sequences. In a widely used strategy, the DNA binding residues of zinc-finger (ZF) proteins have been substituted to generate molecules with altered DNA specificities (5–7). Another approach has focused on engineering DNA binding small molecules, imidazole/pyrrole hairpin polyamides, to increase their sequence specificity (8–11). Both classes of engineered DNA binders have shown tremendous promise in regulating gene expression, but a critical hurdle in their application as therapeutic agents or genomics tools is controlling the precision with which engineered molecules target a desired site when faced with millions of binding sites of varying affinities within a genome (12–14).

As a representative of the synthetic DNA binders we examined the binding characteristics of a hairpin polyamide (PA-1, Fig. 1) (15, 16). This class of molecules has been shown to regulate target genes in human cells and in mouse models of disease (13, 17, 18). Hairpin polyamides are typically composed of pyrrole and imidazole rings and can be engineered, using simple recognition rules (16), to target specific 6–12 base pair (bp) sequences (Fig. 1*A* and Fig. S1*A* and *B*) (15, 16). These molecules bind in the minor groove of DNA with high affinity recognizing naked or chromatinized DNA with modest perturbation of the DNA groove geometry

(19). We also examined the specificity of an engineered ZF protein (ZFP-1, Fig. 1*B*). This protein domain has been utilized, not only to create molecules that regulate gene expression (5–7), but also to generate site-specific nucleases that greatly improve homologous recombination in mammalian cells (20, 21). To directly compare engineered molecules with natural DNA binding proteins, we examined members of different classes of DNA binding domains, representing ~75% of the human DNA binding proteome (Fig. 1*B*) (22, 23). Gata4 (C₄-type zinc finger), Nkx-2.5 (homeodomain), and Jumonji/Jarid2 (AT-Rich Interaction Domain) play a central role in cardiac development and disease (24, 25); p53 (Immunoglobulin-fold) is an antioncogene that regulates genes involved in apoptosis (26); c-Abl (HMG domain) is implicated in cellular signaling and cancer (27); and TBP with its unusual minor groove binding fold is an essential component of the transcriptional machinery (28).

To elucidate the specificity of engineered DNA binders and compare them with natural DNA binding proteins, we performed Cognate Site Identifier (CSI) analysis (Fig. 1) (29–31). With CSI, the sequence preferences of DNA binding molecules are comprehensively examined by determining their affinity for nearly half a million duplex DNA sequences that collectively display the entire sequence space of a binding site on a microarray (Fig. 1*A*) (31). The highest intensity sequences are interrogated by motif-finding algorithms to derive consensus motifs which are based on underlying position weight matrices (PWMs) (29, 32). To determine if PWM-derived motifs fully described the DNA binding preferences, we displayed the entire dataset of binding intensities as unique Sequence Specificity Landscapes (SSLs). An important advantage of this display, as opposed to two-dimensional bar graphs, is that the landscape can be adjusted to optimize the cognate site motif(s) and thereby uncover major binding motifs and effects of flanking sequence.

Using the CSI-SSL approach, we determined the optimal DNA binding motifs for several transcription factors (TFs) and compared them to engineered molecules. Our findings indicate that consensus motifs, derived from PWMs, often mask or compress important binding specificity information generated from large datasets. Solution-based measurements validate that the CSI intensity values are strongly correlated with binding affinities, thus SSLs also function as comprehensive energy landscapes. Moreover, the landscape displays reveal that 6-ring hairpin polyamides discriminate between optimal and suboptimal sequences as

Author contributions: C.D.C., C.L.W., K.E.H., M.S.O., D.B., and A.Z.A. designed research; C.D.C., C.L.W., K.E.H., M.S.O., and D.B. performed research and analyzed data; C.L.W., D.B., K.E.H., N.Q., Y.L., and F.C. contributed new reagents/analytic tools; C.D.C., C.L.W., K.E.H., M.S.O., and A.Z.A. wrote the paper.

The authors declare a conflict of interest. A.Z.A. is a founder/proprietor and C.L.W. and M.S.O. are part-time employees of VistaMotif, and M.S.O. owns Invitrogen stock.

¹C.D.C. and C.L.W. contributed equally to this work.

²To whom correspondence should be addressed. E-mail: ansari@biochem.wisc.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0914023107/DCSupplemental.

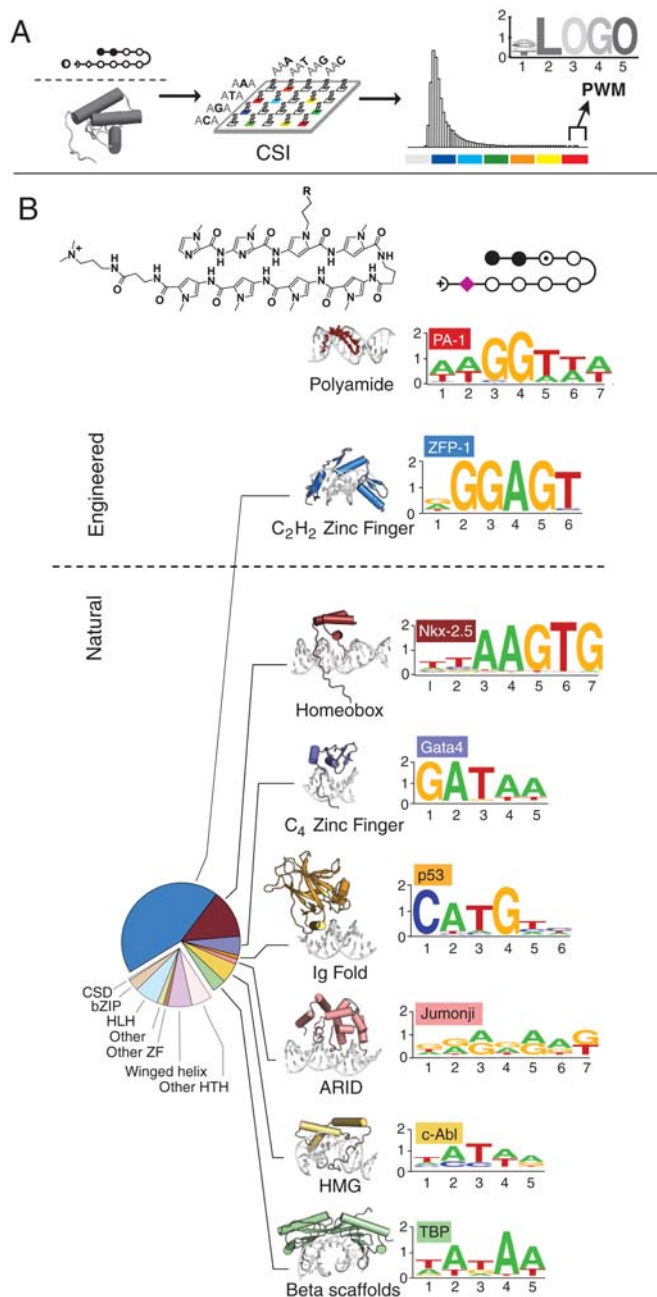


Fig. 1. CSI binding motifs for engineered and natural DNA-binding molecules. (A) Each feature on the array displays a unique sequence as a DNA hairpin, with all permutations of 10 bp DNA represented on the array (~1 million sequences). A protein or small molecule is applied to the microarray to obtain a comprehensive ligand-binding profile. The resulting binding profile yields a histogram of many weak binding features (Gray-Blue), some moderate binding features (Green-Yellow), and a few strong binding features (Red). PWMs are generated from the highest intensity data and displayed as a Logo. (B) The *Upper* displays CSI-determined PWMs for a hairpin polyamide (PA-1) and C₂H₂ zinc finger. The pie chart represents the distribution of DNA-binding folds across all TFs in the human genome. The *Lower* displays CSI-determined PWMs for six major classes of DNA-binding folds. PDB codes are listed in supplementary material. For polyamides, filled circle = N-methyl imidazole; open circle = N-methyl pyrrole; open circle with inner dot = pyrrole with attached Cy5 dye; turn = γ -aminobutyric acid; diamond = β -alanine; half-circle with a positive charge = dimethylaminopropylamide; and R = Cy5 dye.

efficiently as 8-ring structures, a property that has important implications for transcription-based therapeutics. Finally, our approach was applied to predict the genome-wide binding characteristics of engineered TFs.

Results

Cognate Site Identification and Sequence-Specificity Landscapes (CSI-SSL). CSI analysis of the natural and engineered DNA binding molecules yielded a comprehensive binding profile across the entire sequence space of a binding site (Fig. S1C). The highest intensity binding sites were compiled to identify consensus binding motifs (Fig. S1D) which are displayed as sequence Logos (33) in Fig. 1. This comparative analysis indicates that the consensus motif bound by polyamide PA-1 is similar to motifs recognized by natural proteins, both in length and information content.

While consensus motifs derived from PWMs summarize the best binding sequences, they overlook significant insights embedded within the complete sequence recognition profiles captured by CSI analysis (Fig. 2A left panel) (34–36). To surmount this limitation, we created the SSL as an adaptable tool to display and interpret the full recognition preferences of DNA binding molecules (Fig. 2). SSLs present the entire binding dataset through a series of concentric rings or a linear format (Fig. 2A). In a circular SSL, the innermost ring displays sequences that contain a perfect match to a given seed motif (0-mismatch). The subsequent circles, going outward, represent increasing mismatches from the seed motif. The height of each color-coded peak on the individual rings corresponds to the CSI binding intensity data. In this way, the entire sequence space can be displayed in a comprehensive yet easily interpreted format. To generate an SSL, a seed motif is used as the starting input; this motif may be generated from a known PWM, although a motif from any source may be used to initialize the sequence alignment. Through iterative optimizations the seed motif is refined until one obtains an SSL with high affinity sequences restricted to the 0-mismatch ring, and moderate-to-low affinity sequences assigned to the appropriate outer mismatch rings. The sequences in the center ring are sorted first alphabetically by motif (or submotif) and then alphabetically by the flanking sequences. This provides a consistent ordering of sequences in the center ring. The sequences represented in the outer rings are organized first by the position of the mismatch and then alphabetically (e.g. in the motif G₁A₂T₃A₄A₅, G₁ is substituted by A₁, C₁, T₁, followed by A₂ to C₂, G₂, T₂, and onwards through all positions of the motif). In the linear format, the display permits vertical alignment of the sequences by mismatch for each position in the motif (Fig. 2A and Fig. S2B). In essence, the SSL format reduces the high-dimensionality problem of displaying nonlocal sequence interdependencies of nearly a million different permutations into a readily interpretable graph that reports the entire specificity-spectrum of a DNA binding molecule.

A key advantage of the landscape display is that all data points are displayed without sequence compressions that are necessary to generate PWMs and consensus motifs. The absence of sequence compression in SSL facilitates the identification of optimal motif(s). If the seed motif, used to initiate an SSL, is too restrictive then several high-intensity sequences appear in the mismatch rings, and conversely if the motif is too inclusive then low-intensity peaks (i.e. valleys) would invade the innermost match ring (Fig. S2C). Importantly, if a chosen seed motif ignores the existence of alternative modes of binding to different sequences, then outer mismatch rings would display clustered peaks which would reveal the unique binding motifs (see below). Moreover, because the motifs identified by CSI (Fig. 1) are shorter than the permuted DNA duplexes on the array (5–7 bp motifs embedded within a 14–20 bp duplex), the landscapes identify the contributions of flanking sequences on the binding energetics of a given motif. In contrast to binding profiles predicted by PWMs (Fig. S2A), the flanking sequence influences binding properties to an unexpected degree, yielding highly textured binding profiles and rugged landscapes even in the 0-mismatch ring (Fig. 2). This additional contextual information greatly improves the accuracy of regulatory element annotation across genomes.

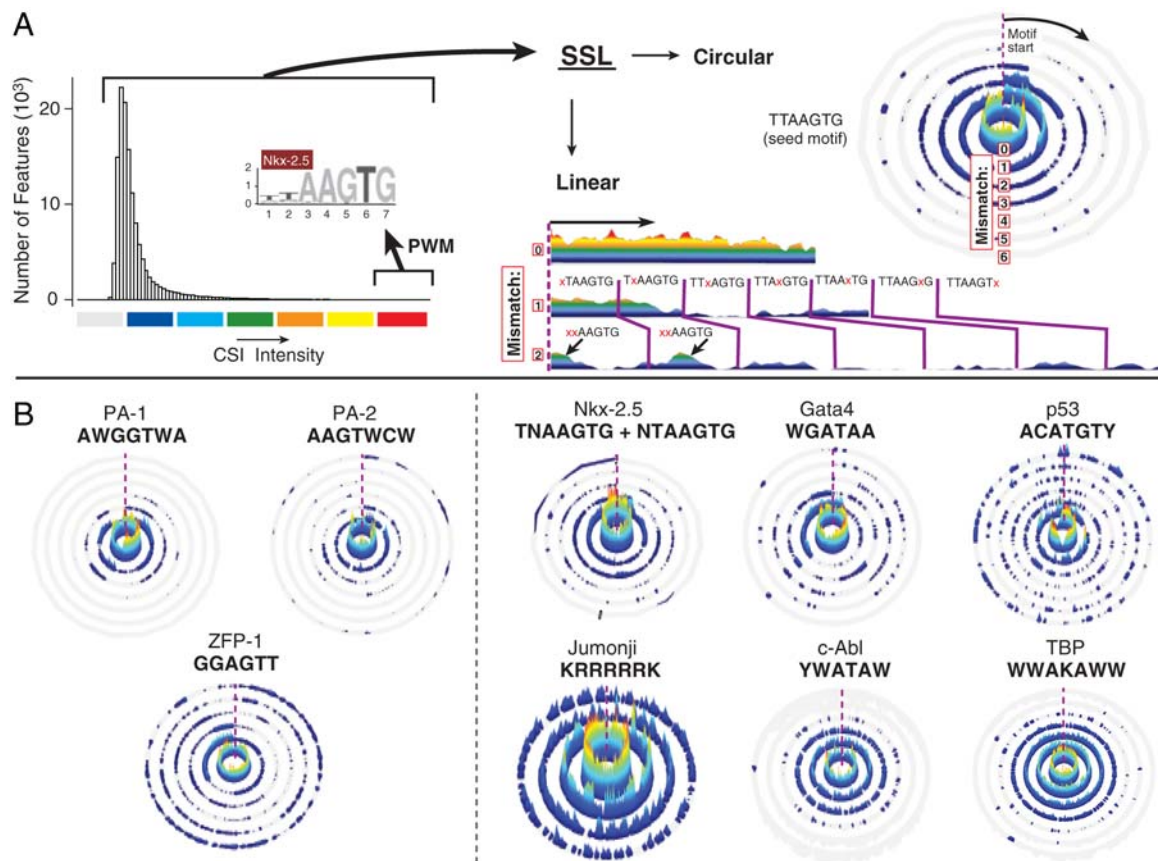


Fig. 2. Sequence-Specificity Landscapes present comprehensive DNA binding data. (A) The complete distribution (*Upper Left*) of intensities from the CSI analysis are shown for Nkx-2.5. Whereas PWMs only utilize the best binding sequences (*Red* portion of the histogram), SSLs represent every sequence assayed (*Gray-Red*) in either a circular (*right*) or linear (*bottom*) format. Circular SSLs display the binding intensities of a given DNA ligand across all DNA permutations on the CSI array, with every sequence displayed on the plot. We note the number of mismatches from the seed motif on each concentric circle. For these specificity landscapes, the seed motif TTAAGTG is used. The dashed purple line indicates the start of sequences bearing a mismatch in the first position. For linear SSLs (*Right* and Fig. S2), each row corresponds to a ring in the circular display. Unlike the circular SSL, sequences bearing multiple mismatches from the motif are plotted multiple times to maintain the vertical alignment between panels in the linear SSL. (B) The SSLs of the engineered DNA-binding molecules (*Left*) are compared to the SSLs for TFs representing a diverse set of DNA binding folds (*Right*). The optimized composite SSL for the Nkx-2.5 submotifs is shown rather than the single motif TTAAGTG in Fig. 2A. The chemical structure and Logo of PA-2 is shown in Fig. S1B.

Specificity Landscapes of Natural and Engineered DNA Binders. The SSLs of natural and engineered DNA binders immediately reveal several interesting results (Fig. 2B). Hairpin polyamides display a high-degree of sequence specificity as compared to natural DNA binding proteins. For PA-1, as well as previously studied PA-2 (chemical structure of PA-2 shown in Fig. S1B) (31), the highest intensity peaks are clustered within the perfect match ring (innermost circle). While few strong binding sites are present in the 1-mismatch ring, virtually none are present in the 2-mismatch ring. By contrast, widely differing specificity spectra are evident among the protein-based DNA binding domains. Notably, the C₄-type ZF from Gata4 shows high-fidelity binding to sequences bearing the known motif. Although p53 prefers its consensus site, binding to other biologically relevant deviations of this sequence is also detected by CSI-SSL analysis (37, 38). TBP prefers an AT stretch but permits some deviation from this sequence. The SSL displays also shed light on the specificity of weak protein binders and resolve inconsistencies among reports in the literature. While Jumonji was shown to bind AT-rich sequences (24) others find that Jumonji prefers additional sequences (39). Our findings indicate that Jumonji binds with low specificity to A/G-rich sequences, thus resolving the two conflicting views.

We also analyzed ZFP-1, an engineered C₂H₂-type ZF protein. This protein was designed to target the sequence 5'-GGA-GTT-AAG-3' using three linked ZF modules. Although the PWM-based Logo (from a CSI array with 10 permuted bp) suggests that

the best binding sequences include 5'-RGGAGT-3', the SSL analysis indicates that ZFP-1 recognizes 5'-GGAGTT-3' (Fig. S2D). The 6-bp site indicates that only two of the three linked fingers are binding DNA specifically (6). The third “nonspecific” finger, when placed between flanking ZF modules, targets the 5'-AAG-3' trinucleotide (40). However placing this finger at the N terminus leads to a significant diminution of its sequence specificity. This observation is consistent with previous observations of nonadditive behavior of these modules (34) and recent reports that the predictability of sequence recognition by engineered ZFPs is less than desired (41). Taken together, SSLs show that hairpin polyamides are among the most specific DNA binders.

Deconvoluting Submotifs That Are Compressed into a “Consensus.”

The SSL of a homeodomain protein, Nkx-2.5, shows that the consensus motif 5'-T₁T₂A₃A₄G₅T₆G₇-3' provides an incomplete view of the true binding specificity (Fig. 1B). The submotifs T₁N₂AAGTG and N₁T₂AAGTG are strong binding sites with TTAAGTG being the best (29). These compressed submotifs are readily deconvoluted as high-intensity peaks in the 1-mismatch ring of an SSL generated with the TTAAGTG seed motif (compare Fig. 2A with dual submotif landscape in Fig. 2B). To explore the generality of this phenomenon, we examined motifs of murine homeodomain proteins that were recently reported (42). Evaluating all 18 classes of homeodomains using the SSL displays revealed that nearly 40% of the motifs could be optimized to better define the

specificity of the proteins (Fig. S2E). For example, in the case of Barx1 the reported consensus motif compresses two submotifs and inaccurately includes poorly binding sequences within the motif. The SSLs of the same dataset yield a more accurate view of the cognate sites preferred by Barx1 as well as several other homeodomains (Fig. S2E).

Precision-Tailoring the Specificity of Engineered DNA Ligands. CSI analysis combined with SSL display permits the rapid and rigorous evaluation of engineered DNA ligands with altered chemical composition and architecture. We synthesized a 6-ring hairpin polyamide (PA-3) rather than the larger 8-ring hairpin polyamide (Fig. 3A and Fig. S3). The CSI-SSL analysis shows that this minimized molecule binds a shorter core sequence but its specificity profile is comparable to 8-ring hairpin polyamides. This has significant implications in engineering smaller molecules that retain specificity yet have improved cell permeability properties.

The SSL profiles also indicate that modular recognition between heterocycle rings of polyamides and individual base steps of DNA governs recognition in this class of molecules. The results strongly support the recognition rules described by Dervan and others (9, 11, 15). We also used the SSL format to display CSI data of a hairpin polyamide (PA-4) containing an atypical 3-chlorothiophene ring (Fig. 3A and Fig. S3) (30). The SSL of this bioactive molecule validates the increased preference for Thymine by the thiophene ring. This result highlights the importance of the CSI-SSL approach in evaluating specificities of new chemical entities that target DNA. Finally, the SSL of a linear polyamide, PA-5, (30) that targets sites abnormally repeated in Friedrich's ataxia patients shows binding to a larger 9 bp site (Fig. 3A and Fig. S3) (17). In the landscape, significant binding is detected in the outer mismatch rings. Examination of the outer-ring peaks identified three related submotifs (Fig. 3B) that were concealed in consensus motifs derived from PWMs (30). This example of relaxed specificity displayed by a linear polyamide (PA-5), along with the proteins described above, serve to highlight the ability of SSLs to identify multiple binding submotifs that are compressed into a single consensus by motif-finding algorithms. These results also reemphasize the utility of the CSI-SSL approach in creating and evaluating molecules that target DNA with the desired degree of specificity.

Landscapes Define Binding Energetics Across the Entire Sequence Space. SSLs directly translate to binding energy landscapes for hairpin polyamides because their CSI intensities correlate strongly with equilibrium binding energies measured in solution (29, 30). To evaluate if this relationship holds for a natural DNA binding protein, we measured affinities of Nkx-2.5 for six DNA sequences that span a range of CSI intensities. As a control we refined the previously reported (29) affinity of PA-1 for a similar range of sites identified by CSI. The binding isotherms for these sequences were determined by nuclease protection assays for PA-1 and by EMSA for Nkx-2.5 (Fig. 4 and Fig. S4). The CSI intensities of both molecules were highly correlated with equilibrium binding energy. Not only do we observe a clear correlation at strong-to-moderate binding sites, the CSI intensities indicate differential binding at suboptimal sites that are too weak to be evaluated by EMSA analysis. The linear correlation observed, indicates that the fractional occupancy on the array is sufficiently low so as to still correlate with solution binding affinity. The low fractional occupancy is likely due to a combination of multiple factors. Thus, SSLs directly translate to energy landscapes and provide unprecedented information on binding energetics across broad sequence space (see Fig. 5).

Genomescales: Genome-Wide Binding Landscapes Explain the Differential Bioactivity of an Engineered DNA Binder. A compelling need for developing CSI-SSL is to map the regulatory elements across the genome. Based on the CSI scores, we assigned binding probabilities across the entire human genome (Fig. 5). We designate these genome-wide binding landscapes, as "genomescales." Here, we focused on the CSI-SSL scores of a hairpin polyamide (PA-4) that targets the binding sites of the hypoxia induced transcription factor, HIF-1 α . PA-4 binding to HIF-responsive-elements (HREs) blocks the activation of the hypoxia induced genes (18), including VEGF, a cytokine that is implicated in cancer and angiogenesis (43). Blocking HIF-1 α dependent expression of VEGF suffices for averting tumorigenesis, thus marking this binding event as a target for therapeutic intervention. Previous genome-wide expression profiles indicated that genes bearing an HRE could be targeted in cells by the designed polyamide (PA-4) (Fig. 5B) (18). However, in those studies it was not clear why the endothelin gene, ET-2, bearing a moderate HRE was robustly inhibited by the compound.

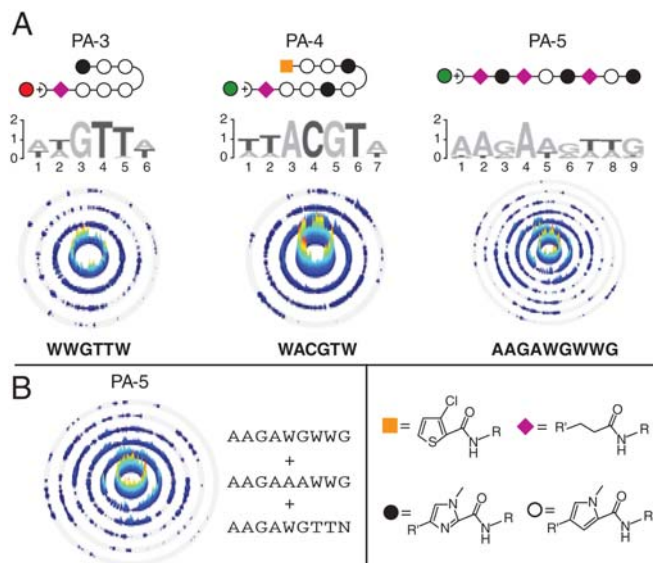


Fig. 3. Specificity landscapes of polyamide structures with diverse composition and architecture. (A) (Upper) Schematics of three additional polyamides examined on the CSI microarray. A key displaying the structure of each ring symbol is shown on the bottom and the conjugated Cy3 or Cy5 dyes are displayed as green or red circles respectively. (Center) Logos generated from PWMs of the highest intensity sequences on the CSI microarray. (Bottom) SSL for each of the polyamides, using a single motif. (B) Optimized SSL for PA-5. Multiple submotifs optimally partition the sequences to yield an even distribution of high-affinity sites in the center circle.

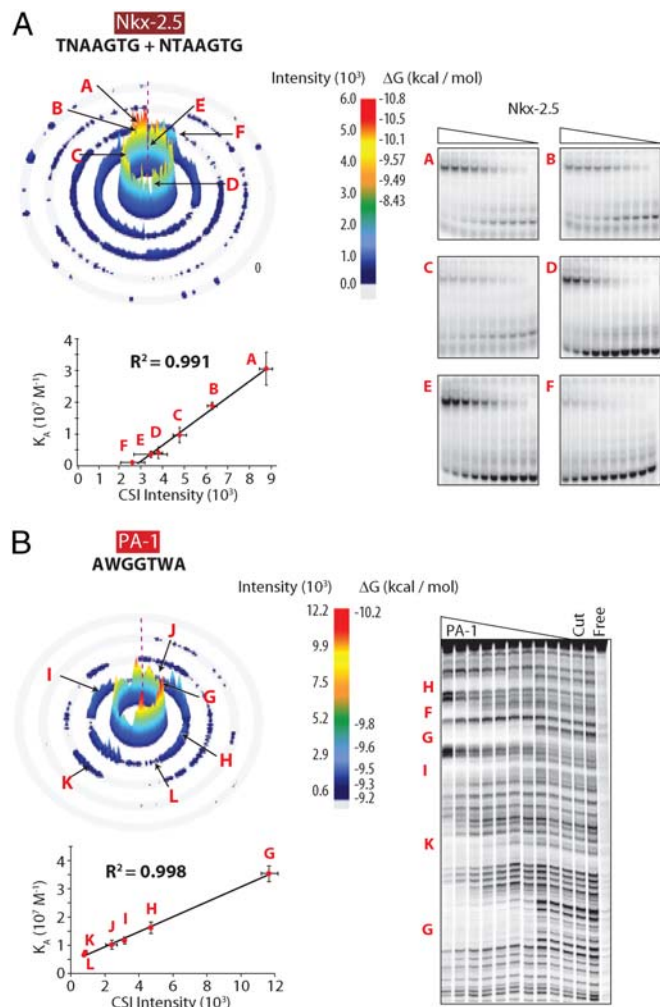


Fig. 4. Sequence-Specificity Landscapes as Energy Landscapes. (A) Six sequences of different intensities were identified (labeled A–F), and their binding was measured in solution by EMSA. The color scale (Center) shows the relationship of SSL-CSI intensities to their corresponding ΔG values ($\Delta G = -RT \ln K_A$). The linear correlation plot between K_A and CSI intensity is shown on the bottom. This correlation allows the binding affinity for all sequences to be determined using the corresponding CSI intensity. (B) Nuclease protection for six sequences (labeled G–L) from the CSI microarray refined from (29) for PA-1. The linear correlation plot between K_A and CSI intensity is shown on the bottom. Each point represents the average of three measurements and error bars indicate one standard deviation. The concentration of Nkx-2.5 ranged from 2 μ M to 3 nM (in twofold increments) and for PA-1 they ranged from 300 nM to 0.5 nM (in threefold increments).

To resolve this outstanding issue, we examined the genomic DNA for nonobvious binding sites that may exert additional inhibition. The CSI genomescape of ET-2 promoter regions identifies moderate PA-4 sites across the known HIF-1 α binding site. More importantly, in contrast to other HIF-1 α regulated genes, ET-2 also has several moderate PA-4 binding sites across the Transcription Start Site (TSS) and into the coding region (Fig. 5C and Fig. S5). At 1 μ M concentrations of the polyamide in the media, ET-2 gene is robustly down-regulated. CSI genomescapes suggest that this unusually high level of inhibition might be due to multiple synergistic inhibitory events where both the transcription factor and the transcriptional machinery are occluded from their binding sites. The binding energetics for this compound suggests that nanomolar binding affinities are required for effective action of the compound in living cells. We further anticipate that the CSI genomescapes will greatly aid in predicting how DNA binding molecules localize

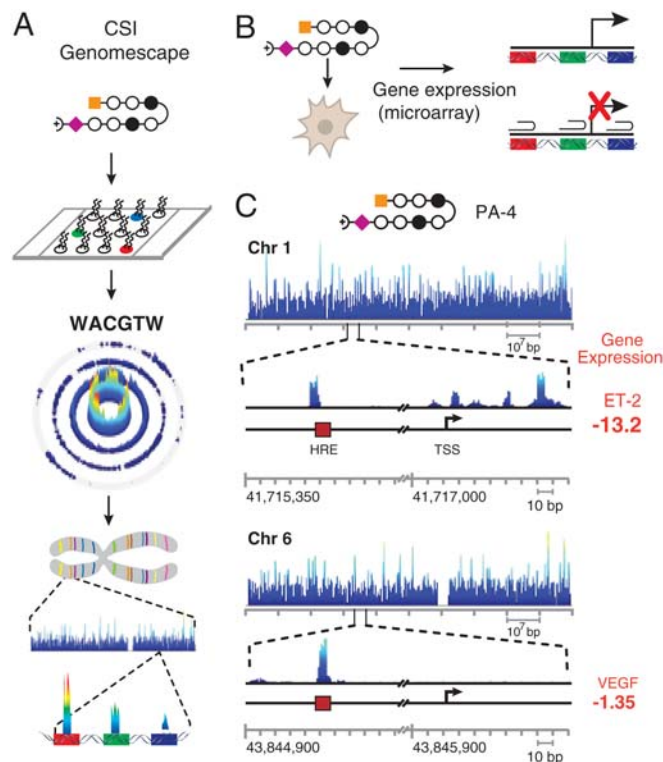


Fig. 5. Genomescapes of CSI microarray data. (A) Genomescapes are generated by assigning an intensity to every 10 bp sequence in the genome from the CSI data. (B) RNA microarray expression levels from untreated cells and cells incubated with PA-4 were compared to determine the degree of inhibition induced by the polyamide (18). (C) The genomescape for PA-4 data was obtained from CSI analysis and the top of each subpanel displays 100 Mbp of the chromosome surrounding either the VEGF or ET-2 gene. The expanded regions show two 100 bp regions containing the HRE binding site and TSS for each gene. The highest extent of inhibition is found at the ET-2 gene by PA-4 which binds at both the activating HRE and several points along the TSS.

across the genome and in elucidating seemingly off-target events in living cells.

Discussion

The combined CSI and SSL analyses of natural and engineered molecules lead to the unexpected conclusion that synthetic DNA binders can achieve specificities that match or exceed natural DNA binding proteins. The SSL displays overcome the limitations of consensus motifs derived from PWMs and reveal an unprecedented view of the entire specificity and energy landscape of DNA binding molecules. The varying specificity constraints imposed by different protein or small molecule structural folds offer insights into the mechanisms for modulating specificity in molecular recognition. Elucidating the effects of sequence context on binding energetics permits an accurate annotation of regulatory elements across the genome. This is particularly true for synthetic DNA binders, and the resulting genomescapes predict the biological outcomes with a high degree of accuracy. The CSI-SSL approach will be invaluable in designing, evaluating, and refining unique chemical entities that target genomic sites with the desired degree of precision. This is a critical step in the generation of molecules that will have therapeutic utility and also serve as powerful genomic tools.

In the near future, high-throughput data for most DNA binding proteins from various organisms will become publicly available. As demonstrated with the homeodomain family of proteins (Fig. S2E), high-dimensionality data when viewed through SSLs reveal unique modes of sequence recognition. SSL analysis of such datasets will address how natural proteins target cognate

sites in the genome, how sequence specificity evolved within different DNA binding folds, and how regulatory elements and networks are evolutionarily retained and organized.

The Sequence-Specificity Landscapes that we describe here are not limited to displaying DNA binding data. They can be readily applied to any high-dimensional data that define interaction interfaces of biopolymers or other molecules, including protein-peptide or protein-small molecule interactions. An integrative perspective that combines landscapes along with other physical and biological analyses will greatly enhance our understanding of the elusive principle of “specificity.”

Materials and Methods

CSI Microarray Analysis—Polyamide. The fluorescently-labeled polyamide was diluted to a final concentration (250 nM for PA-1, PA-2, PA-5; 20 nM for PA-3; and 10 nM for PA-4) in hybridization buffer (1 M NaCl, 100 mM MES pH 7.5, 20 mM EDTA, 0.01% Tween-20). The polyamide was added to the hybridization chamber (Grace BioLabs) on the array and rotated for 1 h at 23 °C. The arrays were washed with nonstringent wash buffer (6X SSPE pH 7.5, 0.01% Tween-20), dried, and scanned using an Axon 4000B (Molecular Devices). Data was viewed using GenePix™ Pro 6.0 (Molecular Devices).

CSI Microarray Analysis—Protein. Arrays were blocked with 2.5% nonfat dried milk for 1.5 h. The protein was diluted to a final concentration in buffer and mixed with a directly-labeled fluorescent antibody to either the protein or a tag on the protein. The protein-antibody mixture was added to the hybrid-

ization chamber on the array and rotated for 1 h. The arrays were washed, dried, and scanned as above.

Generation of Specificity Landscapes. To order the binding data for presentation in the SSL, the binding site on each sequence is determined using the best match to the seed motif. Landscapes are optimized by maximizing the number of high-intensity sequences and minimizing the number of low-intensity sequences in the 0-mismatch ring. The data is parsed according to the number of mismatches from the motif being plotted. If the seed motif has multiple submotifs, we use the minimum number of mismatches to the best submotif. Each ring is sorted first by the position of the mismatch and then alphabetically. The intensities are then plotted using MatLab R2006a (The MathWorks, Inc.). The intensities are smoothed using 0.5% of the data on each ring.

Additional details are included in *SI Text*.

ACKNOWLEDGMENTS. We thank P.B. Dervan, S. Keles, J. Mitchell, and S. Cunningham for invaluable discussions; S. Foister and P. Dervan for the gift of PA1 (SFIII-02); and J. Markley and the Center for Eukaryotic Structural Genomics for plasmids. This work was supported by grants from the National Institutes of Health (GM069420), March of Dimes, U.S. Department of Agriculture, Innovation and Economic Development Research Program, and Vilas Associate and Shaw scholar awards (to A.Z.A.). C.D.C. was supported by the American Heart Association predoctoral fellowship (0615615Z), C.L.W. was supported by a National Institutes of Health/National Library of Medicine predoctoral fellowship (T15LM007359), K.E.H. was supported by a Natural Sciences and Engineering Research Council fellowship. D.B. was supported by the Khorana program. Y.L. was supported by the National Institutes of Health (HL67050).

- Darnell JE, Jr (2002) Transcription factors as targets for cancer therapy. *Nat Rev Cancer*, 2:740–749.
- Gottesfeld JM, Turner JM, Dervan PB (2000) Chemical approaches to control gene expression. *Gene Expression*, 9(1-2):77–91.
- Hauschild KE, Carlson CD, Donato L, Moretti R, Ansari AZ (2008) Transcription Factors. *Wiley Encyclopedia of Chemical Biology*, ed Begley T (John Wiley & Sons, Inc, New York), 4, pp 566–584.
- Ptashne M, Gann A (2002) *Genes & Signals* (Cold Springs Harbor Laboratory Press, New York).
- Choo Y, Klug A (1994) Toward a code for the interactions of zinc fingers with DNA: Selection of randomized fingers displayed on phage. *Proc Natl Acad Sci USA*, 91:11163–11167.
- Mandell JG, Barbas CF 3rd (2006) Zinc Finger Tools: Custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Res*, 34:W516–523.
- Rebar EJ, Pabo CO (1994) Zinc finger phage: Affinity selection of fingers with new DNA-binding specificities. *Science*, 263:671–673.
- Dervan PB, Poulin-Kerstien AT, Fechter EJ, Edelson BS (2005) Regulation of Gene Expression by Synthetic DNA-Binding Ligands. DNA Binders and Related Subjects, Topics in Current Chemistry. (Springer Berlin/Heidelberg, Berlin/Heidelberg), 253, pp 1–31.
- Reddy BS, Sharma SK, Lown JW (2001) Recent developments in sequence selective minor groove DNA effectors. *Curr Med Chem*, 8:475–508.
- Tse WC, Ishii T, Boger DL (2003) Comprehensive high-resolution analysis of hairpin polyamides utilizing a fluorescent intercalator displacement (FID) assay. *Bioorg Med Chem*, 11:4479–4486.
- Wemmer DE (2000) Designed sequence-specific minor groove ligands. *Annu Rev Biophys Biomol Struct*, 29:439–461.
- Buchmueller KL, et al. (2005) Extending the language of DNA molecular recognition by polyamides: Unexpected influence of imidazole and pyrrole arrangement on binding affinity and specificity. *J Am Chem Soc*, 127(2):742–750.
- Nickols NG, Jacobs CS, Farkas ME, Dervan PB (2007) Suppression of androgen receptor-mediated gene expression by a sequence-specific DNA-binding polyamide. *Proc Natl Acad Sci USA*, 104:10418–10423.
- Supekova L, et al. (2002) Genomic effects of polyamide/DNA interactions on mRNA expression. *Chem Biol*, 9:821–827.
- Dervan PB, Edelson BS (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr Opin Struct Biol*, 13:284–299.
- White S, Szcwyczyk JW, Turner JM, Baird EE, Dervan PB (1998) Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands. *Nature*, 391:468–471.
- Burnett R, et al. (2006) DNA sequence-specific polyamides alleviate transcription inhibition associated with long GAA.TTC repeats in Friedreich's ataxia. *Proc Natl Acad Sci USA*, 103:11497–11502.
- Olenyuk BZ, et al. (2004) Inhibition of vascular endothelial growth factor with a sequence-specific hypoxia response element antagonist. *Proc Natl Acad Sci USA*, 101:16768–16773.
- Suto RK, et al. (2001) Crystal structures of nucleosome core particles in complex with minor groove DNA-binding ligands. *J Mol Biol*, 326:371–80.
- Porteus MH, Carroll D (2005) Gene targeting using zinc finger nucleases. *Nat Biotechnol*, 23:967–973.
- Wu J, Kandavelou K, Chandrasegaran S (2007) Custom-designed zinc finger nucleases: What is next?. *Cell Mol Life Sci*, 64:2933–2944.
- Garvie CW, Wolberger C (2001) Recognition of specific DNA sequences. *Mol Cell*, 8:937–946.
- Venter JC, et al. (2001) The sequence of the human genome. *Science*, 291:1304–1351.
- Kim TG, Kraus JC, Chen J, Lee Y (2003) JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor. *J Biol Chem*, 278:42247–42255.
- Nemer G, Nemer M (2001) Regulation of heart development and function through combinatorial interactions of transcription factors. *Ann Med*, 33:604–610.
- Murray-Zmijewski F, Slee EA, Lu X (2008) A complex barcode underlies the heterogeneous response of p53 to stress. *Nat Rev Mol Cell Biol*, 9:702–712.
- Zhu J, Wang JY (2004) Death by Abl: A matter of location. *Curr Top Dev Biol*, 59:165–192.
- Burley SK, Roeder RG (1998) TATA box mimicry by TFIID: Autoinhibition of pol II transcription. *Cell*, 94:551–553.
- Keles S, Warren CL, Clayton CD, Ansari AZ (2008) CSI-Tree: A regression tree approach for modeling binding properties of DNA-binding molecules based on Cognate Site Identification (CSI) data. *Nucleic Acids Res*, 36:3171–3184.
- Puckett JW, et al. (2007) Quantitative microarray profiling of DNA-binding molecules. *J Am Chem Soc*, 129:12310–12319.
- Warren CL, et al. (2006) Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci USA*, 103:867–872.
- Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 2 pp:28–36.
- Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol*, 188:415–431.
- Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein-DNA interactions: How good an approximation is it?. *Nucleic Acids Res*, 30:4442–4451.
- Frank DE, et al. (1997) Thermodynamics of the interactions of lac repressor with variants of the symmetric lac operator: Effects of converting a consensus site to a non-specific site. *J Mol Biol*, 267:1186–1206.
- Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315:233–237.
- Smeenk L, et al. (2008) Characterization of genome-wide p53-binding sites upon stress response. *Nucleic Acids Res*, 36:3639–3654.
- Wei CL, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124:207–219.
- Patsialou A, Wilsker D, Moran E (2005) DNA-binding properties of ARID family proteins. *Nucleic Acids Res*, 33:66–80.
- Dreier B, Beerli RR, Segal DJ, Flippin JD, Barbas CF 3rd (2001) Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem*, 276:29466–29478.
- Ramirez CL, et al. (2008) Unexpected failure rates for modular assembly of engineered zinc fingers. *Nat Methods*, 5:374–375.
- Berger MF, et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133:1266–1276.
- Underiner TL, Ruggeri B, Gingrich DE (2004) Development of vascular endothelial growth factor receptor (VEGFR) kinase inhibitors as anti-angiogenic agents in cancer therapy. *Curr Med Chem*, 11:731–745.