SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site

Liyang Zhang\textsuperscript{1,+}, Gabriella D. Martini\textsuperscript{3,4,+}, H. Tomas Rube\textsuperscript{3,4}, Judith F. Kribelbauer\textsuperscript{3,4}, Chaitanya Rastogi\textsuperscript{3,4}, Vincent D. FitzPatrick\textsuperscript{3,4}, Jon C. Houtman\textsuperscript{2}, Harmen J. Bussemaker\textsuperscript{3,4,*}, Miles A. Pufall\textsuperscript{1,*}

AFFILIATIONS

\textsuperscript{1}Department of Biochemistry, Carver College of Medicine, University of Iowa; \textsuperscript{2}Department of Immunology, Carver College of Medicine, University of Iowa; \textsuperscript{3}Department of Biological Sciences, Columbia University, New York, NY 10027; \textsuperscript{4}Department of Systems Biology, Columbia University, New York, NY 10032

\textsuperscript{+}These authors contributed equally
\textsuperscript{*}Corresponding authors:

Miles A. Pufall, PhD
University of Iowa, Dept. of Biochemistry
51 Newton Road, BSB 4-430
Iowa City, IA, 52242
(319) 384-1820
miles-pufall@uiowa.edu (M.A.P.)

Harmen J. Bussemaker, PhD
Columbia University, Department of Biological Sciences
1212 Amsterdam Ave, MC 2441
New York, NY10027
(212) 854-9932
hjb2004@columbia.edu (H.J.B.)

Running title: Sequence, shape distinguish AR and GR DNA binding
Keywords: Transcription, DNA-binding, Steroid hormone receptors, Prostate Cancer
ABSTRACT:

The DNA-binding interfaces of the androgen (AR) and glucocorticoid (GR) receptors are virtually identical, yet these transcription factors share only about a third of their genomic binding sites and regulate similarly distinct sets of target genes. To address this paradox, we determined the intrinsic specificities of the AR and GR DNA binding domains using a refined version of SELEX-seq. We developed an algorithm, SelexGLM, that quantifies binding specificity over a large (31 bp) binding-site by iteratively fitting a feature-based generalized linear model to SELEX probe counts. This analysis revealed that the DNA binding preferences of AR and GR homodimers differ significantly, both within and outside the 15bp core binding site. The relative preference between the two factors can be tuned over a wide range by changing the DNA sequence, with AR more sensitive to sequence changes than GR. The specificity of AR extends to the regions flanking the core 15bp site, where isothermal calorimetry measurements reveal that affinity is augmented by enthalpy-driven readout of poly-A sequences associated with narrowed minor groove width. We conclude that the increased specificity of AR is correlated with more enthalpy-driven binding than GR. The binding models help explain differences in AR and GR genomic binding, and provide a biophysical rationale for how promiscuous binding by GR allows functional substitution for AR in some castration-resistant prostate cancers.
INTRODUCTION:

Gene expression programs are precisely regulated by transcription factors (TFs), a class of DNA-binding proteins that orchestrate the activity of the RNA polymerase II and chromatin-modifying complexes. The DNA binding domains (DBDs) of TFs fall into families consisting of dozens or even hundreds of members (Weirauch et al. 2014), leading to similar DNA sequence preferences among family members. Nonetheless, subtle quantitative differences in DNA binding specificity between related TFs are associated with large qualitative differences in the sets of target genes they control (Maerkl and Quake 2007). To understand gene regulation and regulatory networks, it is therefore essential not only to accurately quantify these differences in DNA recognition but also determine the structural and physical basis of that specificity. The former can be done using comprehensive, unbiased experimental and computational methods; the latter requires more focused mechanistic analyses.

The intrinsic DNA binding specificities for hundreds of TFs have been profiled using a number of different high-throughput assays. These include (universal) protein binding microarrays (PBM) (Berger and Bulyk 2009), bacterial 1-hybrid (B1H) (Meng et al. 2005), and (solution-based) high-throughput SELEX (HT-SELEX) (Zhao et al. 2009; Jolma et al. 2010; 2013). These methods have been effective at defining core binding motifs (a sequence pattern that contains the main determinants of TF:DNA interaction), which can be represented as position weight matrices (Stormo 2000) or position-specific affinity matrices (Foat et al. 2006), and visualized as information content (Schneider et al. 1985) or energy/affinity logos (Foat et al. 2006). The models are available via databases such as Jaspar (Mathelier et al. 2015), UniProbe (Orenstein and Shamir 2014; Hume et al. 2015) and CisBP (Weirauch et al. 2014). Although for most TFs these approaches have been effective for defining a core sequence pattern required
for high-affinity DNA binding, the resulting motifs often fail to discriminate between TFs from the same structural family.

Interactions with base pairs outside the core binding site can contribute to differences in DNA sequence preference between paralogous TFs (Zhou and O'Shea 2011; Maerkl and Quake 2007; Fisher and Goding 1992). For example, Cbf1p and Tye7p, members of the basic helix-loop-helix (bHLH) family in S. cerevisiae, recognize similar E-box core sequences (CANNTG), but distinct flanking preferences became apparent when binding was assayed over a larger footprint using custom-designed genomic context protein binding microarrays (gcPBMs) (Gordân et al. 2013). Differences in specificity associated with flanking regions have also been observed in HT-SELEX data (Jolma et al. 2013), and for ETS proteins using PBMs (Wei et al. 2010). Moreover, SELEX-seq technology has been used to show that Hox proteins read out the spacer sequences between half-sites in distinct ways when binding as heterodimers with the cofactor Exd (Slattery et al. 2011; Abe et al. 2015). However, although these techniques have demonstrated that flanking sequences are important, they cannot accurately quantify sequence specificity over larger footprints (>15bp) without a specialized design (e.g. gcPBM). Thus, a more refined, general approach is needed to assess whether functional differences between TFs from the same family can be attributed to differences in intrinsic DNA binding specificity considered over the full TF:DNA interface.

The “missing specificity” issue – or the functional differences between transcription factor paralogs that cannot be explained by initial measurements of binding specificity – is brought into sharp relief by castration resistant prostate cancer (CRPC). Prostate cancer is driven by androgen signaling through regulation of gene expression by the androgen receptor (AR, AR) (Watson et al. 2015). Blocking of androgen synthesis and inhibition of ligand binding to AR have
both been effective treatments. Unfortunately, CRPC eventually arises due to alternative production of androgens or activation of AR (Feldman and Feldman 2001). In some cases, CRPC is accompanied by increased expression of the glucocorticoid receptor (GR, NR3C1), which then functionally substitutes for AR by activating a subset of the AR transcriptional program that drives cancer progression (Arora et al. 2013). Despite this overlap, chromatin immunoprecipitation followed by sequencing (ChIP-seq) in LNCaP-1F5, a cell line model of CRPC, has shown that AR and GR share only about a third of their genomic binding sites (Sahu et al. 2013; 2014). Although co-factors such as the transcription factor FoxA1 help distinguish between AR and GR binding (Belikov et al. 2016; Sahu et al. 2013), these two factors still bind distinct loci in their absence suggesting an intrinsic ability to distinguish sequences that is not reflected in previous measurements of their in vitro specificities (He et al. 2012; Jin et al. 2014; Pihlajamaa et al. 2014; Jolma et al. 2013).

It is not clear how AR and GR, both members of the steroid hormone receptor (SHR) family, are directed to different genomic loci. Existing ChIP-seq and HT-SELEX (Nelson et al. 1999; Jolma et al. 2013; Sahu et al. 2013) studies suggest that AR and GR bind indistinguishable 15bp core motifs composed of inverted hexameric half-sites separated by a three-base-pair spacer: RGAACA NNN TGTTCY. Crystallographic evidence indicates that AR and GR each bind DNA as head-to-head dimers, with the two monomer subunits each occupying a half-site and dimerizing over the spacer (Shaffer et al. 2004; Meijsing et al. 2009; Watson et al. 2013). Our detailed analysis of AR- and GR-DNA crystal structures indicates that they share an identical DNA binding interface (Fig. S1). Conserved residues make specific contacts in the major groove at positions 2, 4, and 5 (Luysi et al. 1991; Arbuckle and Luisi 1995) (Fig. S2) accounting for most of the binding energy, although other non-contacted base pairs
within the half-site have also been shown to affect GR affinity (La Baer and Yamamoto 1994). Additional energy is derived from backbone contacts along the 3bp spacer, which are sensitive to minor groove width in GR (Meijsing et al. 2009; Watson et al. 2013). Contacts made with DNA sequences flanking the core motif further contribute to GR affinity (Meijsing et al. 2009). In this work, we test whether, despite their conservation, AR and GR use this shared DNA-binding interface differently to distinguish sequences within and flanking the core 15bp motif.

Models that estimate the contribution of individual DNA features to the overall binding free energy by directly fitting PBM intensities have been shown to capture sequence specificity with greater accuracy than classification-based algorithms (Weirauch et al. 2013). The MatrixREDUCE algorithm (Foat et al. 2006) is an early example of this general approach that has been extended in other studies (Zhao and Stormo 2011; Gordân et al. 2013; Riley et al. 2015). Analysis of SELEX read counts, however, differs fundamentally from that of PBM intensities. and typically involves the tabulation of relative enrichment for all oligomers (“k-mers”) of a given length (Riley et al. 2014). Because the expected number of reads matching any particular k-mer decreases exponentially with \( k \), generation of comprehensive tables is typically limited by sequencing cost to a 12-15bp footprint (Abe et al. 2015; Slattery et al. 2011; Jolma et al. 2013). Feature-based models have been used to interpret the oligomer tables, either in the form of a simple position weight matrix (Jolma et al. 2013) or based on a combination of base and shape features (Abe et al. 2015). Although this approach has provided valuable insights, it is limited by the restriction on binding-site size inherited from the oligomer tables on which they are based. Thus, in order to distinguish between the specificities of AR and GR over their full footprints, more refined experimental and computation techniques are required.
RESULTS:

AR interacts with DNA over a larger footprint than GR. To determine the intrinsic specificity of AR and GR at high resolution, we performed SELEX-seq (Slattery et al. 2011; Riley et al. 2014) for homodimers of the DNA binding domain (DBD) of each factor (Fig. 1A, Fig. S1C). Although SELEX-seq (and lower throughput predecessors) have been developed over the years (Ogawa and Biggin 2011; Djordjevic 2010), increased sequencing and computational power have allowed some refinements. Purified proteins were incubated with a pool of DNA molecules, each containing a larger (23bp) random region than typical, flanked by Illumina adapters and tagged at one end with Cy5. Electrophoretic mobility shift assays (EMSA) were used to separate dimer-bound DNA sequences over eight rounds of affinity-based selection (Fig. 1A, Fig. S3). After each round, the concentration of the isolated DNA was quantified by qPCR, and then in parallel amplified for reselection and packaged into a sequencing library by adding Illumina flow cell adapters. Each library was then sequenced to a depth of ~10^7 reads. Preliminary analysis of the SELEX-seq data revealed unexpected differences between AR and GR. Following a previous study (Slattery et al. 2011), we estimated affinities as normalized oligomer enrichments, using the R/Bioconductor package SELEX (Riley et al. 2014). Biases in the initial round zero (R0) pool were estimated using a 5th-order Markov model, after which we computed the information gain between the initial (R0) and final round (R8) to estimate binding site size. For GR the information gain peaks at 15bp (Fig. 1C), consistent with the previously defined core motif. However, for AR information content continues to increase beyond 15bp (Fig. 1D), indicating a sensitivity to base identity over a larger binding site. One concern was that sequences would be over-selected after 8 rounds. Due to the diversity of the pool and lack of a strict consensus sequence for both AR and GR, very few 23-mers sequences are observed more than once in the sequenced libraries (Fig 1E). This indicates not only that the libraries were not over-
selected, but also that further rounds of selection might allow better discrimination of lower affinity sequences without over-selecting high-affinity sites (Djordjevic 2010). A cursory k-mer based analysis indicated that AR and GR differ in their specificities, sharing only about a third of their significantly enriched 15-mers (Fig. 1F, S4A,B), with substantially different preferences among those common 15-mers (Fig. 1G), and confirmed specificity for AR over a larger footprint (Fig. S4C). Quantitative EMSAs supported the rank order of these sequences in terms of their enrichment in the SELEX-seq experiment (Fig. S4E,F).

Further analysis indicated that both spacer sequence and flanking poly-A tracts contribute to AR affinity over at least 21bp (Fig. S4C). At that footprint size, it is not possible to accurately quantify differences in affinity using oligomer tables as in previous studies (Slattery et al. 2011; Jolma et al. 2013) for the following reasons. The most enriched 21-mer for AR (AAAAGAACACGATGACTTTT) is contained in ~4,000 reads out of the ~10^7 that make up the R8 library. Because the expected count for suboptimal 21-mers in R8 scales as the 8th power of affinity, a mutated 21-mer sequence with an affinity 3-fold lower than the optimal 21-mer would have a 3^8-fold (6,561) lower expected frequency in the pool, and therefore be unobservable. This is problematic, as much lower-affinity TF binding sites have recently been shown to be functionally important (Tanay 2006; Crocker et al. 2015; Farley et al. 2015). Furthermore, 15-mers covering only one of the two half-sites can be more strongly enriched than those centered at the dimer site for AR (see Table S1 and caption of Fig. S3F), a misleading effect that arises because individual 15-mers cannot cover the entire binding site but their presence nevertheless can imply favorable flanking sequence. Although the logos suggest a larger footprint for AR, the k-mer tables (Figure S4C,D) do not allow for biophysical interpretation. Together, these
limitations motivated us to develop a new computational technique to accurately describe DNA recognition over a large binding-site.

**Quantifying DNA binding specificity using feature-based modeling.** To overcome the limitations of oligomer tables outlined above, we developed *SelexGLM*, a flexible modeling strategy based on Poisson regression that allows us to estimate binding free-energy contributions throughout the protein-DNA interface directly from the SELEX-seq read counts (Fig. 2). We use a standard equilibrium thermodynamics description of protein-DNA interaction that was previously used to analyze protein binding microarray (PBM) data (Foat et al. 2006; Zhao and Stormo 2011; Gordân et al. 2013; Riley et al. 2015). The relative binding free energy for sequence $S$ is modeled as a sum of parameters $\Delta \Delta G_\phi$ associated with the DNA sequence features $\phi \in \Phi(S)$ that characterize the difference between $S$ and a reference sequence $S_{\text{ref}}$ (typically the highest-affinity sequence):

$$\Delta G(S) - \Delta G(S_{\text{ref}}) = \sum_{\phi \in \Phi(S)} \Delta \Delta G_\phi$$

In this study, we restricted ourselves to single-nucleotide features (e.g., $\phi = C_3$, denoting the presence of a C at position 3 within the binding site window). Our modeling assumptions imply that the combined effect of multiple mutations within the binding site is additive in terms of binding free energy or multiplicative in terms of relative affinity.

The model parameters are found through an iterative fitting process that starts from an initial estimate, or seed. To generate this seed, we construct a relative enrichment table using a footprint large enough to capture the binding site core but not so large that counts get too low. We chose a seed length of 15bp for both AR and GR, but the final model is only weakly dependent on this choice and can have a much larger footprint (31bp in our case; see below).
The negative logarithm of the relative enrichment of each mutated 15-mers is used as an initial estimate of $\Delta \Delta G_{\phi}/RT$ for each feature.

Once seeded, the model is refined by alternating between two steps. In the first step, we determine the highest-affinity binding site within each unique observed SELEX probe in the data ("affinity-based alignment"). This allows us to construct a design matrix $X$ defining each DNA feature (in this case each base pair) relative to the optimal binding window in each probe; only probes whose rate of selection is dominated by a single binding site offset are included (see METHODS). In the second step, the design matrix is used to fit a generalized linear model (GLM) to the read counts, leading to a re-estimated set of free-energy coefficients:

$$
\log(\lambda_i) = \log(p_i^0) + \beta_0 + \sum_{\phi} \beta_{\phi} X_{i\phi}
$$

Here $\lambda_i$ is the expected value of the read count $y$ for probe $i$:

$$
y_i \sim \text{Poisson}(\lambda_i)
$$

Each model coefficient $\beta_{\phi}$ is used as a re-estimate of $\Delta \Delta G_{\phi}/RT$, which are then used to update the design matrix, and the process repeats. Convergence is reached once the position of the optimal binding window no longer changes for any of the probes in the data set after re-estimation of the free-energy coefficients.

Earlier rounds of selection are sufficient for SelexGLM analysis. We performed SELEX-seq over eight rounds of selection in order to obtain linear enrichment of sequences appropriate for k-mer enrichment based analysis. When we analyzed each round of selection through SelexGLM, we were surprised to find that, although R8 provided the highest-resolution model without over selection (Fig. 1E, Fig. S5A-J), high-quality models could also be generated from
earlier rounds of selection (Fig. S5A-J). However, the accuracy of the models increases in the later rounds, particularly when it comes to distinguishing the base-pair substitutions with the most deleterious effect on binding (Fig. S5, S6). Thus we used models generated from R8 for both proteins in all subsequent analyses.

AR is distinguished from GR by a preference for poly-A sequences outside the 15bp core. We used SelexGLM to build recognition models as position specific affinity matrices (PSAMs) for AR and GR with a 31bp footprint, significantly larger than the 15bp binding site core and even the 23bp variable region of the SELEX probes. SelexGLM is capable of fitting such wide models because it considers offsets within the probe that partially cover the fixed sequences upstream and downstream of the variable region. The corresponding energy logos for these PSAMs confirm the previously modeled 15bp size of the GR binding site (Fig. 2B,D), but reveal AR’s preference for poly-A sequences flanking the 15bp core (Fig. 2A,C) and surprising differences within the 3bp central spacer (Fig. 2E, and more evident in Fig S4C,G,H). For AR, replacing the best flanking sequence (AAAA) with the worst (TCGC) on one side leads to a 1.6-fold reduction in affinity (or 2.6-fold when both flanks are replaced) (Fig. 2E). Validation by qEMSA confirmed that spacer sequence affects the affinity of both AR and GR, but that flanking sequences affect only AR (Fig. 2F). The observed change for AR (∆∆G = 0.60 ± 0.05 kcal/mol) was larger than predicted (∆∆G = 0.26 kcal/mol), perhaps because our model ignores dependencies between nucleotide positions within the binding site.

Further differences in AR and GR specificity are encoded throughout the 15bp core. Though the AR and GR binding models are similar at first glance, detailed analysis of the PSAMs highlights AR’s sensitivity to changes from the consensus sequence in the central 15bp core region (Fig. 3a). Based on these differences, we tested sequences that favor AR (Fig. 3B, Table S2,
Shape-1 and Shape-2) and GR binding (Fig. 3B, Table S2, Shape-4) and verified that the two proteins can differentiate between DNA sequences over an order of magnitude in affinity (Fig. 3B, Figure S5D). As the two proteins have identical base-reading chemistry at the DNA-binding interface, we examined whether their sequence preferences could be explained in terms of DNA shape readout. Indeed, the shape preferences of AR and GR contrast sharply: the AR benefits from a narrowed minor groove in the flanking regions (Fig. 3C) whereas the GR prefers binding to sites with widened minor grooves within the half-sites (Fig. 3D). To validate this finding, we tested a sequence with a half-site predicted to have a narrow minor groove (TTTTAT) (Zhou et al. 2013) and found that GR bound significantly worse than AR (Fig. 3B, S5D, Shape-3). Similarly, we tested a site predicted to have a wide minor groove (GGGACA, Fig. 3B, S5D, Shape-4) and found a preference for GR. More dramatic examples of sequences predicted to have disparate GR and AR affinities are plentiful in the low affinity range (K_D ≥ 5µM), however measurements in this range are near the non-specific binding limit and have not been reliable (data not shown). In addition to differences in core preference, the nearly inverted minor groove preferences in the flanks for AR and the half-site for GR suggest the two proteins have different recognition modes.

The thermodynamics of binding reflect the specificity of AR and GR. To understand the thermodynamic basis of the differing AR and GR binding modes, we measured the effect of varying flanking and spacer sequences using isothermal titration calorimetry (ITC). Our ITC data allowed us to accurately quantify the affinity (and thus ΔG) of bound sequences and the contribution of individual sequence features to the enthalpy (ΔH) and entropy (ΔS) of binding (Buurma and Haq 2007). Although the K_Ds for AR and GR sequences were consistent with the PSAMs derived from the SELEX data (Fig. S7C), including AR preference for poly-A flanks, it
was immediately apparent from the heat of binding (Fig. 4A) that AR and GR recognize the same DNA sequences differently. Overall, for the sequences tested, AR binding is more enthalpically driven (low ΔH) and GR binding more entropically driven (high TΔS) (Fig. 4C,D). This is consistent with previous findings; TF families engaged in direct readout via the major groove are more enthalpically driven whereas those engaged in readout via the minor groove or backbone contacts are more entropically driven by solvent exclusion (Privalov et al. 2011; 2007). Our data also identify an exception to this general pattern: AR’s recognition of the poly-A stretches in the minor groove is driven not entropically, but rather enthalpically. This observation contributes to our understanding of specificity: enthalpic contributions to binding are not solely the result of hydrogen bonds with individual base-pairs or backbone phosphates, but can also result from interaction with a DNA feature; the narrowed minor groove (Fig. 3).

Promiscuous GR specificity predicts binding at AR genomic loci. We next asked whether the SelexGLM models can explain why GR is able to functionally substitute for AR despite having non-overlapping binding sites in models of CPRC. To this end, we analyzed ChIP-seq data from LNCaP-1F5, a cell line model of prostate cancer engineered to overexpress GR and enable genomic mapping of both AR and GR binding under similar cellular conditions (Sahu et al. 2013; 2014). As a positive control, we confirmed that the SelexGLM models can differentiate ChIP-seq peaks from adjacent regions (area under ROC-curve 0.78 and 0.83 for AR and GR, respectively; Fig. 5A). Further, we observed a significant quantitative relationship between predicted affinity and degree of genomic occupancy (Fig. 5B), with ChIP-seq peaks in the top decile for in vitro affinity being significantly higher than those in the bottom decile (1.4-fold, p=7.4*10^{-49}, Wilcoxon rank-sum test, for AR; 1.6-fold, p=2.8*10^{-68}, for GR). To test whether the difference in intrinsic binding specificity between AR and GR discovered using SELEX-seq was
reflected in the genomic occupancy patterns probed using ChIP-seq, we let our AR and GR PSAMs compete in a multiple linear regression model. As expected, when analyzing the variation in GR peak height in this manner, we found that the regression coefficient for GR affinity was significantly larger than that for AR \( (p<10^{-6}, t\text{-test}; \text{Fig. 5C}) \) and that the latter did not deviate significantly from zero \( (p=0.97) \). However, when analyzing the variation in AR peak height, both the AR and GR coefficients were non-zero \( (p<10^{-5} \text{ and } p<10^{-7}, \text{respectively}) \) and did not significantly differ from each other \( (p=0.16) \), indicating that both AR and GR are likely to bind well at AR loci.

**DISCUSSION:**

In this study, we developed and validated a strategy for quantifying the DNA binding specificity of TFs with high precision over a much larger footprint than was previously feasible for SELEX data. This required experimental refinements to *SELEX-seq* and a new computational method, *SelexGLM*, to fit the free-energy parameters of a biophysically motivated model to the *SELEX-seq* data. Experimentally, we designed a library to accommodate the footprint of AR and GR defined by crystal structures (Meijsing et al. 2009; Shaffer et al. 2004), isolated dimer-bound DNA from a mixed population by EMSA, and performed qPCR between rounds of enrichment to avoid PCR artifacts that add noise to measurements. These refinements revealed that, despite their similarity, AR and GR have different sequence specificity both within and outside the core 15bp sequence that can be genetically tuned over a significant range, even in the absence of any co-factors. Compared to HT-SELEX assays performed on similar protein fragments, both of these refinements improved the accuracy of the model (Fig. S8). SelexGLM was able to distinguish specificity over a 23bp footprint from this data (Fig. S8A-D), whereas k-mer-based
analysis only revealed a partial binding site at lower resolution (Fig. S8E-G). Importantly, the models derived from HT-SELEX correlated poorly with the EMSA validation (Fig S8H-J), and were less consistent and accurate than those derived from our SELEX-seq data.

The intrinsic specificity, thermodynamics of binding, and genomic localization indicate that GR is more promiscuous than AR, allowing it to accommodate a wider range of sequences. These binding properties are consistent with the behavior of GR in some CRPCs. Anti-androgen therapy can result in de-repressed GR expression, increasing the amount of GR in the cell (Arora et al. 2013; Watson et al. 2015), which, in the presence of endogenous or administered glucocorticoids results in an increase in the active concentration of GR in the nucleus. Our data is consistent with the idea that the more relaxed specificity provided by entropy-driven binding allows the excess GR to bind related sequences with reasonable affinity, including those previously bound by AR (Fig. 5C), enabling regulation of AR-driven genes aberrantly by GR (Arora et al. 2013). Thus, the biophysical properties of GR provide a rationale for how it can effectively substitute for AR in the context of CPRC.

The differences between AR and GR that our analysis uncovered were surprising, since the DNA binding interfaces of AR and GR are chemically identical. However, the integration of SELEX analysis and thermodynamic quantification of binding energies using ITC make it clear that AR and GR use these interfaces differently. Having comprehensively measured the specificity of AR and GR using all possible DNA ligands, it is clear that GR derives more entropic energy for the same DNA than AR. Since the majority of entropic energy is derived from solvent and ion exclusion from an interface, and the protein and DNA surfaces at the interface are chemically identical, we expected this contribution to be similar for AR and GR. However, the differences between GR and AR suggest an increase in conformational entropy.
within GR (1994; Frederick et al. 2007) (Fig. 4D), which would allow it to accommodate more diverse DNA sequences and shapes, but decrease its specificity. AR, conversely, does not derive as much binding energy from entropy, suggesting that precise positioning of hydrogen bonding in the half sites is more important. AR also derives enthalpic energy from the recognition of narrowed minor grooves created by poly-A flanking sequences. Our results provide direct evidence that the increased negative electrostatic focusing (Rohs et al. 2009) associated with a narrowed minor groove (Yoon et al. 1988) in the presence of interactions with basic residues (Fig. S1, S2) drives an increase in affinity. These differences in the thermodynamics of recognition must be mediated by non-conserved amino acids within the fold of the protein, indicating a difference in how the interface is scaffolded. This phenomenon has also been observed in recent studies of the C2H2 Zinc finger family of TFs (Hughes 2011; Persikov et al. 2015), and indicates that both the residues displayed and how they are configured are critical to distinguishing how transcription factor family members recognize different DNA sequences. It also suggests that, much as in the case of RUNX1 (Yan et al. 2004) and Ets-1 (Pufall et al. 2005), the conformational entropy of GR may provide an opportunity for regulation that could direct it away from AR sites.

To enable analysis of the SELEX-seq data over a large enough footprint, it was necessary develop an improved algorithm for analyzing SELEX-seq data, which we named SelexGLM. In a self-consistent procedure that echoes Gibbs sampling for discriminative motif discovery (Lawrence et al. 1993), SelexGLM alternates between an “affinity based alignment” step that identifies the dominant binding site within each sequenced DNA ligand, and a GLM fitting step that estimates the binding free-energy change associated with each possible base substitution in the DNA binding site, until convergence is reached. The Poisson statistics that underlie the GLM fits allow us to obtain precise estimates of the free-energy parameters even
when individual unique DNA have very few counts, since each possible base substitution or “feature” can occur in many different contexts. Using a modest sequencing depth of \( \sim 10^7 \) reads per round, we were able to achieve precise, validated models over a footprint of 31bp that would accommodate many other transcription factors complexes.

*SelexGLM* also opens new avenues for studying the modulation of transcription factor function. Transcription factor binding and activity are regulated by post-translational modifications (Pufall et al. 2005), cofactor binding (Chodankar et al. 2014), and combinatorial control through interaction other transcription factors on DNA. The increased footprint size made possible by our method enables measurement of specificity for larger, heterologous complexes on DNA, and the effect of sequence on their occupancy. Further, *SelexGLM* allows us to quantify relatively modest effects on specificity distributed nature over the entire footprint, which evolution may exploit to fine-tune expression levels in a flexible manner, and which may be altered by phosphorylation of the TF or cofactor binding (Kumar and Calhoun 2008; Chodankar et al. 2014). Thus, our approach may enable us to begin to understand to what extent signal-dependent changes in expression are due to altered transcription factor specificity.

**METHODS:**

**Expression and purification of GR/AR-DBD.** DNA sequences encoding the DNA binding domains (DBDs) of human Glucocorticoid receptor (GRα: 418-506) and Androgen receptor (AR-B: 557-647) were cloned into an N-terminal his\(_6\)-tagged vector (pET28a, Novagen). Vectors were transformed into BL21DE3 Gold *E.coli* (Agilent) cells, and grown to an OD600 of between 0.2 and 0.4. The temperature was reduced to 27°C and 10 \( \mu \)M of ZnCl\(_2\) was added to the culture. Expression of recombinant protein was induced with 0.5 mM IPTG for 4 hours when
OD600 reaches 0.6 to 1. Cells were then spun down at 6000g for 15 minutes, then resuspended in Ni^{2+} column loading buffer (25 mM TrisHCl, pH 7.5, 500 mM NaCl, 15 mM Imidazole, 1 mM DTT, 1 mM PMSF), snap frozen, and stored at -80°C until purification. Cell suspensions were lysed with an EmulsiFlex C3 homogenizer running at 15,000 psi. After three passes or more through the EmulsiFlex, soluble protein was isolated by ultracentrifugation at 40,000 rpm using Beckman Ti-75 rotor for 1 hour at 4°C then collecting the supernatant. The supernatant was loaded onto nickel affinity column (GE Healthcare Life Sciences) pre-equilibrated with 25 mM TrisHCl, pH 7.5, 500 mM NaCl, 15 mM Imidazole. Unbound protein was washed off the column with equilibration buffer, followed by a low imidazole bump (25 mM TrisHCl, pH 7.5, 500 mM NaCl, 30 mM Imidazole) to remove non-specifically bound protein. A linear gradient of imidazole from 30 mM to 350 mM was used to elute DBD. Fractions containing DBDs were pooled and dialyzed overnight at 4°C in 20 mM TrisHCl, pH 7.5, 50 mM NaCl, 2.5 mM CaCl\(_2\) and 1 mM DTT. Thrombin was used to cleave the his\(_6\)-tag during the dialysis. Dialyzed protein was ultracentrifuged (40,000 rpm, 1 hour, 4°C), and loaded onto a cation exchange column (HiTrap SP HP, GE Healthcare Life Sciences) pre-eliqulibrated with 20 mM TrisHCl, pH 7.5, 50 mM NaCl and 1 mM DTT. The DBD was eluted in a linear gradient of NaCl from 50 to 350 mM over 20 CVs. Fractions containing DBD were pooled, concentrated (Amicon Ultra – 3K, Millipore), filtered (Ultrafree®-CL), and monomers isolated by gel filtration (16/600 Superdex 200 PG, GE LifeSciences) in 20 mM HEPES, pH 7.7, 100 mM NaCl, 1 mM DTT. DBDs were collected and dialyzed against storage buffer (20 mM HEPES, pH 7.7, 100 mM NaCl, 1 mM DTT, 50% Glycerol), quantified (280 nm, \(\varepsilon = 5095\)M\(^{-1}\)cm\(^{-1}\) for both AR- and GR-DBDs).

**SELEX Library design and synthesis.** We designed our SELEX library to contain a 23bp random region flanked by primer binding regions conforming to the Illumina TruSeq Small RNA format
for a total of 70bp. The library was ordered in 1 mmole format from IDT as single strand with handmix option over the randomized region. The complementary strand was synthesized by Klenow extension using 5'-Cy5 labeled TSSR1 primer (Table S2) on a PCR machine. Briefly, a reaction containing 2.5 μM ssDNA library, 5 μM 5'-Cy5-TSSR1, 150 μM dNTPs in NEB buffer 2 was incubated at 94°C for 3 minutes, and then gradually cooled down to 37°C over 45 minutes. Six units of Klenow were added to every 25μl of reaction, incubated at 37°C for 60 minutes, 72°C for 20 minutes, and gradually cooled down to 10°C over 45 minutes. dsDNA was purified using MinElute PCR column (Qiagen) and quantified by absorbance at 260 nm.

**SELEX-seq. Selection:** To ensure at least 1X coverage of all possible 23mers, SELEX was carried out in a 120 µl binding reaction containing 0.63 μM purified DBD and 1 μM DNA library. At this size, there are ~ 7.2×10¹³ DNA molecules in the reaction, representing ~1.02X coverage of all possible 23mers (7.03×10¹³). Binding was carried out in a buffer that approximates the salt and crowding of the nucleus for 1 hour at 4°C (20 mM TrisHCl, pH 8.0, 150 mM KCl, 5% Glycerol, 1 mM EDTA, 5 mM MgCl₂, 40 ng/µl PolydIdC (Sigma: P4925), 200 ng/µl BSA, 1 mM DTT, 200 mg/ml Ficoll PM400 (Sigma: F4375)). The reaction was then run out in multiple wells of a 10% native polyacrylamide gel (19:1 Acrylamide/Bis-acrylamide) in 0.5X TB containing 150 μM MgCl₂ (89 mM Tris-Boric acid, 150 μM MgCl₂, pH 8.3) at 4°C. In order to ensure that SHR-DBD:DNA complexes were trapped, the sample was loaded while running at 200V to minimize the dissociation before entering the gel. DBD:DNA complexes were visualized using Cy5 fluorescence (GE ImageQuant LAS4010), isolated by excision, and bound DNA isolated by electroelution (Novagen D-tube dialyzer, 3.5 kDa) into native PAGE running buffer described above. Resulting DNA sequences were then purified using (Qiagen MinElute PCR clean-up), eluted (10 mM TrisHCl, pH 8.0) to a final volume of 180µl, and amplified to generate next round
of library as described below. Because of the relatively low affinity of the SHRs for DNA, we were unable to shift enough DNA using a limiting amount of protein (<1:5) to allow PCR generation of pools for subsequent rounds without generating high molecular weight artifacts. We therefore incubated the library with a high protein:DNA ratio (0.63 μM:1 μM) to select all potential binding sequences in early rounds. Please note that this protein:DNA ratio requires more rounds of selection to begin linear enrichment of sequences. Since submission of this manuscript we have begun using a 1:10 protein:DNA ratio. It is critical to perform a size-selection (8% 1X TG gel) from the recovered bound DNA to remove the high molecular weight DNA that co-migrates with complexes prior to re-amplification. Together with controlled amplification cycle by qPCR (see below), this improved SELEX protocol is artifact-free, saves a few rounds of selection and is less saturated at early rounds. Typically, 4~5 rounds of selection are sufficient for factors with long binding footprints (~25-30 bp).

Library regeneration: To generate enough DNA (~10^{13} molecules) to perform each round of SELEX and sequencing library prep, the recovered DNA had to be carefully amplified. As these libraries were susceptible to amplification artifacts caused by over-amplification (different type of artifact than those co-migrating with the complex), the optimal number of PCR cycles was first determined by qPCR. Briefly, 1 μl of recovered dsDNA was analyzed in 50 μl qPCR reaction (0.5 μM TSSR0 primer (Table S2), 0.5 μM unlabeled TSSR1 primer, 200 μM dNTPs, 0.1X SybrGreen (Invitrogen: S-7563), 0.5 unit of Phusion polymerase in 1X NEB Phusion HF buffer). The amplification curve was then analyzed to determine the maximal number of rounds within the linear amplification range (Typically less than 16 PCR cycles, depending on the amount of the template). Subsequently, 170 μl recovered library was divided into 85 reactions at 100 μl, and amplified with the determined cycle numbers (with 5'-Cy5-TSSR1). The amplification
reactions were then combined, purified, and concentrated (Qiagen MinElute), and eluted with 45 µl EB. The resulting libraries were then quantified and mixed into a new binding reaction as described above. Additional agarose gel electrophoresis is required to remove polydIdC if it is present in the recovered dsDNA, as the polydIdC interfered with PCR.

**Library sequencing:** To sequence the resulting libraries on the Illumina HiSeq platform, additional adapter sequences were added by limited-cycle PCR. Briefly, 400 ng of each SELEX library was amplified using the 5’ adaptor primer (TSSR2, Table S2) and the 3’ adaptor and barcoding primer (TSSR-RPIX, Table S2) in a 1 ml PCR reaction (300 µM dNTPs, 0.8 µM TSSR2, 0.8 µM TSSR-RPIX, 10 U Phusion polymerase in 1X Phusion HF buffer) for 2 cycles. The added 71bp allowed separation of the sequencing library from the adapter-less library on a 16% 0.5X TBE native polyacrylamide gel (19:1 Acrylamide/Bis-acrylamide) run at 200V. The 141bp band was excised, and the DNA recovered by electroelution as described above. The purity and concentration of the library was determined by Bioanalyzer (High-sensitivity dsDNA chip, Agilent). Multiple sequencing libraries with compatible barcodes were pooled in equimolar concentrations and sequenced on a HiSeq2000 with 10 million reads per library, using single-end, 50bp sequencing mode.

**Motif discovery.** Raw sequencing data processing:** The sequenced libraries were processed using the R package SELEX (http://bioconductor.org/packages/SELEX). We required sequencing reads to match the sequence TGGAA at positions 24-28. The package was also used to construct Markov models of R0 and compute the information gain (KL divergence) after affinity-based selection.

**Model-based analysis (SelexGLM):** To avoid bias in our estimates, we split the reads into two equal-sized random subsets. One half was used to define the “universe” of unique variable
regions (which we refer to as “probes”). Read counts across this universe were defined based on the other half of the reads, and a count of zero was registered for probes that were only seen in the first half. A Markov model of order 5 was constructed from the R0 probes using the selex.mm () function from the SELEX package, and an affinity table for k=15 was constructed using selex.affinities (). An initial position specific affinity matrix (PSAM) was constructed from the relative affinity of all 15x3 single-base mutations of the optimal 15-mer, expanded to the desired size by adding eight neutral columns on each side, and used as a seed. The subsequent iterative procedure alternated between two steps. First, the current PSAM was used to find the position/direction of highest affinity on either strand, the optimal “view” on the probe. If that optimal affinity was larger than 95% of the sum over all positions (including the top position), the probe was used in the analysis; otherwise, it was ignored. The set of optimal positions in each of the accepted probes was used to define a design matrix containing the base identity at each position relative to the start of the optimal view. Using the probe counts as independent variables, the logarithm of the expected probe frequency in R0 according to the Markov model as offset, and a logarithmic link function, a fit was performed using the glm () function. The regression coefficients were interpreted as free-energy differences ddG. All our analyses were implemented in the form of an R software package, which will be submitted to Bioconductor as soon as the manuscript has been accepted for publication. All computational figure panels in this paper were produced fully automatically from the raw sequencing data using R scripts that use the SELEX and SelexGLM packages.

Isothermal Titration Calorimetry (ITC): ITC was performed using a Microcal VP-ITC (GE) at 25°C. Protein and DNA samples were dialyzed into binding buffer (20 mM HEPES-KOH, pH 7.7, 250 mM KCl and 0.5 mM TCEP) at 4°C for 36~48 hours before use. DNA samples were loaded into the syringe and titrated into the protein in the reaction cell. Each ITC experiment consisted
of an initial 2µl injection, followed by 20 x 14.3 µl injections, with 240 seconds between injections. For AR-DBD, 50 µM DNA and 10 µM protein sample were used. For GR-DBD, 100 µM DNA and 20 µM protein sample were used to increase signal. The raw isotherm was analyzed using NITPIC (Brautigam et al. 2016), followed by fitting to a one-site binding model in SEDPHAT (Brautigam et al. 2016). The mean and standard deviation of thermodynamic parameters were calculated based on at least three experimental replicates.

**ChIP-seq data analysis:** Raw reads for AR (GSM759657 and GSM759658), GR (GSM759669), and IgG control (GSM759671) data from ChIP-seq experiments using LNCaP-1F5 cells were downloaded from Gene Expression Omnibus and aligned to the hg19 assembly using bowtie2 with settings ‘--sensitive --score-min L,-1.5,-0.3’. Peaks were called using macs2 width default settings and the “fold enrichment” (column 7) was used to quantify peak strength. The narrowPeak bed-file was filtered retain one entry per unique peak interval and the intervals were standardized to cover 200bp. The PSAM score were then calculated for each offset orientation of the peak and the largest value was recorded.

**Statistical analysis:** Figure 1: To minimize the influence of sequencing count on the calculation of relative enrichment, we only used 15-mers with ≥100 counts in both R7 and R8 libraries. We consider the sequencing count of each k-mer as a random variable of poisson distribution, where the k-mer count (n) is the best estimation of the mean (λ). Therefore, the variance is approximated by the k-mer count, with an absolute error of Sqrt (n). A sequencing count ≥100 therefore restricts the absolute error no more than 10% of the mean (λ).

Figure 2C, 3B, 4, Figure S4, S5: To compare DNA binding affinities of GR and AR for different sequence by EMSA or ITC, we performed at least 3 independent experiments under the same
conditions. The variance of measured affinity is assumed to be normally distributed, and the T-test and P-values are appropriate.

DATA ACCESS: The raw SELEX-seq reads from this study have been submitted to the Sequence Read Archive (SRA; www.ncbi.nlm.nih.gov/sra) (Leinonen et al. 2011) under accession number SRP101815.

ACKNOWLEDGMENTS: This research was supported by the National Institutes of Health (NIH) K99/R00 CA149088, the Roy J. Carver Charitable Trust 01-224, and the National Science Foundation CAREER grant 1552862 (MAP), and grant R01HG003008 (HJB). We thank Todd Riley and members of the Pufall, Bussemaker, Mann, and Rohs labs for useful discussions.

AUTHOR CONTRIBUTIONS: M.A.P and H.J.B conceived and oversaw the project. L.Z designed and performed all experiments. G.D.M, J.F.K, C.R. and H.J.B conceived and developed the GLM analysis approach. G.D.M implemented the SelexGLM software and analyzed SELEX-seq data. H.T.R analyzed ChIP-seq data. L.Z and V.D.F performed preliminary analysis of SELEX-seq and ChIP-seq data. M.A.P analyzed the crystallography data. J.C.H provided the ITC hardware and consultation on interpretation. L.Z, G.D.M, H.T.R, H.J.B and M.A.P wrote the manuscript.

DISCLOSURE DECLARATION: The authors have no competing financial of personal interests to declare.

FIGURE LEGENDS:
Figure 1: SELEX-seq reveals differences in AR- and GR-DBD DNA binding specificity. (A) SELEX-seq. A 70bp ssDNA library with 23bp randomized region was incubated with the DNA binding domains (DBD) of AR or GR and separated into monomer and dimer species by EMSA. Dimer bound DNA was recovered, quantified by qPCR, amplified as the library for the next round, and repeated for eight rounds. Each round of library, including the initial dsDNA library, was sequenced. (B) EMSA gel showing the enrichment of dimer bound sequences after each round of selection for GR-DBD. The intensity of the shifted band plateaus after round 7. A high-affinity palindromic sequence served as a control to locate the dimer band. (*) marks an artifact during the synthesis of control sequence, but not observed in the SELEX library. (C, D) Information gain, or Kullback-Leibler divergence, from R0 to R8, as a function of oligonucleotide length. (E) Boxplot showing the multiplicity of unique 23-mers in each of the last three rounds of SELEX-seq selection for AR and GR. Even for the most highly selected library (AR R8) fewer than 10% of all reads have 10 copies or more, indicating that the libraries are not over-selected. (F) Venn diagram showing the overlap of sequences for AR- and GR-DBD with at least 100 sequencing counts. (G) Scatterplot of sequences that were commonly bound (yellow from panel (F)) by AR- and GR-DBD.
Figure 2: *SelexGLM* shows differences in DNA recognition between AR and GR throughout their binding sites. (A-D) Energy logos for AR-DBD (top) and GR-DBD (bottom), obtained by fitting biophysical models for protein-DNA interaction to the *SELEX* read counts using an iterative generalized linear modeling approach based on Poisson regression, implemented as *SelexGLM*. Highly similar logos were obtained using two separate rounds of data. See Fig. S5 for logos generated using round 4 to 8 data. (B) Cumulative distribution functions for the contribution of half site (squares), spacer (triangles), and flanking (circles) sequences on AR-DBD (red) and GR-DBD (blue) binding energy. (C) Validation of the contribution of flanking A tracts and spacer to AR- and GR-DBD binding performed by quantitative electrophoretic mobility shift assay. Loss of flanking A tracts is more detrimental to AR- than GR-DBD (1 vs. 2), whereas changing spacer can have detrimental effects on the binding of both (1 vs. 3). Error bars are the standard error of the mean based on at least 3 repeats of each experiment. (* P-Value ≤ 0.05, ** P-Value ≤ 0.01, t-test)
Figure 3: Difference in DNA shape readout between AR and GR. (A) Difference in ΔΔG/RT values between AR and GR at each nucleotide position, normalized by their mean across all four bases. (B) Quantitative EMSA was used to measure the affinities of AR- and GR-DBDs for four sequences that maximally favor AR (Shape-1) and GR (Shape-4) and test the importance of half-site minor groove width (Shape-3 and -4). Each measurement was performed at least three times. Error bars represent SEM. (C) Average minor groove width profile for the top/middle/bottom 5% of sequences in terms of affinity for AR and (D) for GR.
Figure 4: ITC analysis reveals distinct DNA binding thermodynamics between AR- and GR-DBD.

(A) The raw heat titration signals (top) and normalized (NITPIC) heat of injection profiles (bottom) of AR- and GR-DBD bound to a given DNA sequence. Standard errors are estimated by NITPIC. (B) The $K_D$ of AR-DBD and GR-DBD for four sets of sequences fit from the ITC data. AR-DBD affinity is increased with flanking As and an optimal spacer, whereas GR is insensitive. (C) Enthalpy, $\Delta H$, is calculated from the heat of binding for each DNA sequence. Flanking sequences decrease $\Delta H$ for AR-DBD, enhancing affinity. Smaller indicates a greater contribution to affinity. (D) Entropy, $\Delta S$, is calculated from the $K_D$ (thus $\Delta G$) and $\Delta H$. GR-DBD affinity is more entropically driven. Larger is indicates a greater contribution to affinity. (* P-Value ≤ 0.05, ** P-Value ≤ 0.01, *** P-Value ≤ 0.001, t-test) Error bars represent the standard deviation from at least 3 experiments.
Figure 5: Differences of intrinsic specificity between AR- and GR-DBD are reflected in the respective cellular genomic binding profiles. **(A)** Ability of the AR (red) and GR (blue) PSAMs to identify true ChIP-seq peaks as measured by the receiver operator characteristic (ROC) of binary classifiers identifying peaks vs. adjacent regions. **(B)** Comparison of the median enrichment over background from ChIP-seq (y-axis) to the relative affinity for the strongest sequences within the ChIP-seq peak calculated from the position specific affinity matrix for AR (red) and GR (blue). In vitro affinity is binned by decile (e.g. 10 is the sequence space representing the top 10% highest affinity sites) **(C)** Multiple linear regression coefficients in models that use AR (red) and GR (blue) PSAM scores to predict AR (left) and GR (right) ChIP-seq peak enrichments. *p*-values were calculated using *t*-tests.

References:

Abe N, Dror I, Yang L, Slattery M, Zhou T, Bussemaker HJ, Rohs R, Mann RS. 2015. Deconvolving the Recognition of DNA Shape from Sequence. *Cell* 161: 307–318.

Arbuckle ND, Luisi B. 1995. A recipe for specificity. *Nature Structural Biology*, May.

Arora VK, Schenkein E, Murali R, Subudhi SK, Wongvipat J, Balbas MD, Shah N, Cai L, Efstathiou E, Logothetis C, et al. 2013. Glucocorticoid Receptor Confers Resistance to Antiandrogens by Bypassing Androgen Receptor Blockade. *Cell* 155: 1309–1322.

Belikov S, Berg OG, Wrange Ö. 2016. Quantification of transcription factor-DNA binding affinity in a living cell. *Nucleic Acids Research* 44: 3045–3058.

Berger MF, Bulyk ML. 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* 4: 393–
Brautigam CA, Zhao H, Vargas C, Keller S, Schuck P. 2016. Integration and global analysis of isothermal titration calorimetry data for studying macromolecular interactions. *Nat Protoc* **11**: 882–894.

Buurma NJ, Haq I. 2007. Advances in the analysis of isothermal titration calorimetry data for ligand-DNA interactions. *Methods* **42**: 162–172.

Chodankar R, Wu D-Y, Schiller BJ, Yamamoto KR, Stallcup MR. 2014. Hic-5 is a transcription coregulator that acts before and/or after glucocorticoid receptor genome occupancy in a gene-selective manner. *Proceedings of the National Academy of Sciences* **111**: 4007–4012.

Crocker J, Abe N, Rinaldi L, McGregor AP, Frankel N, Wang S, Alsawadi A, Valenti P, Plaza S, Payre F, et al. 2015. Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* **160**: 191–203.

Djordjevic M. 2010. Inferring Protein-DNA Interaction Parameters from SELEX Experiments. In *DNA Recombination* (ed. H. Tsubouchi), Vol. 674 of *Methods in Molecular Biology*, pp. 195–211, Humana Press, Totowa, NJ.

Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. 2015. Suboptimization of developmental enhancers. *Science* **350**: 325–328.

Feldman BJ, Feldman D. 2001. The development of androgen-independent prostate cancer. *Nat Rev Cancer* **1**: 34–45.

Fisher F, Goding CR. 1992. Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *The EMBO Journal* **11**: 4103–4109.

Foat BC, Morozov AV, Bussemaker HJ. 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**: e141–9.

Frederick KK, Marlow MS, Valentine KG, Wand AJ. 2007. Conformational entropy in molecular recognition by proteins. *Nature* **448**: 325–329.

Gordân R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML. 2013. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *CellReports* **3**: 1093–1104.

He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. 2012. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Research* **22**: 1015–1025.

Hughes TR. 2011. A Handbook of Transcription Factors. 1–310.

Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA...
interactions. *Nucleic Acids Research* **43**: D117–D122.

Jin H-J, Zhao JC, Wu L, Kim J, Yu J. 2014. Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program. *Nature Communications* **5**: 3972.

Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ, et al. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**: 861–873.

Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327–339.

Kumar R, Calhoun WJ. 2008. Differential regulation of the transcriptional activity of the glucocorticoid receptor through site-specific phosphorylation. *Biologics* **2**: 845–854.

La Baer J, Yamamoto KR. 1994. Analysis of the DNA-binding affinity, sequence specificity and context dependence of the glucocorticoid receptor zinc finger region. *J Mol Biol* **239**: 664–688.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**: 208–214.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. 2011. The sequence read archive. *Nucleic Acids Research* **39**: D19–D21.

Luisi B, Xu W, Otwinowski Z, Freedman L. 1991. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA.

Maerkl SJ, Quake SR. 2007. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**: 233–237.

Mathelier A, Fornes O, Arenillas DJ, Chen C-Y, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al. 2015. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* gkv1176.

Meijsing SH, Pufall MA, So AY, Bates DL, Chen L, Yamamoto KR. 2009. DNA Binding Site Sequence Directs Glucocorticoid Receptor Structure and Activity. *Science* **324**: 407–410.

Meng X, Brodsky MH, Wolfe SA. 2005. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature Biotechnology* **23**: 988–994.

Nelson CC, Hendy SC, Shukin RJ, Cheng H, Bruchovsky N, Ben F Koop, Rennie PS. 1999. Determinants of DNA Sequence Specificity of the Androgen, Progesterone, and Glucocorticoid Receptors: Evidence for Differential Steroid Receptor Response Elements. *Molecular Endocrinology* **13**: 2090–2107.
Ogawa N, Biggin MD. 2011. High-Throughput SELEX Determination of DNA Sequences Bound by Transcription Factors In Vitro. In DNA Recombination (ed. H. Tsubouchi), Vol. 786 of Methods in Molecular Biology, pp. 51–63, Humana Press, Totowa, NJ.

Orenstein Y, Shamir R. 2014. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. Nucleic Acids Research 42: e63–e63.

Persikov AV, Wetzel JL, Rowland EF, Oakes BL, Xu DJ, Singh M, Noyes MB. 2015. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. Nucleic Acids Research 43: 1965–1984.

Pihlajamaa P, Sahu B, Lyly L, Aittomäki V, Hautaniemi S, Jänne OA. 2014. Tissue-specific pioneer factors associate with androgen receptor cistromes and transcription programs. The EMBO Journal n/a–n/a.

Privalov PL, Dragan AI, Crane-Robinson C. 2011. Interpreting protein/DNA interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components. Nucleic Acids Research 39: 2483–2491.

Privalov PL, Dragan AI, Crane-Robinson C, Breslauer KJ, Remeta DP, Minetti CASA. 2007. What Drives Proteins into the Major or Minor Grooves of DNA? J Mol Biol 365: 1–9.

Pufall MA, Lee GM, Nelson ML, Kang H-S, Velyvis A, Kay LE, McIntosh LP, Graves BJ. 2005. Variable control of Ets-1 DNA binding by multiple phosphates in an unstructured region. Science 309: 142–145.

Riley TR, Lazarovici A, Mann RS, Bussemaker HJ. 2015. Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. eLife 4: 307.

Riley TR, Slattery M, Abe N, Rastogi C, Liu D, Mann RS, Bussemaker HJ. 2014. SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes. In DNA Recombination (ed. H. Tsubouchi), Vol. 1196 of Methods in Molecular Biology, pp. 255–278, Springer New York, New York, NY.

Rohs R, West SM, Liu P, Honig B. 2009. Nuance in the double-helix and its role in protein–DNA recognition. Current Opinion in Structural Biology 19: 171–177.

Sahu B, Laakso M, Pihlajamaa P, Ovaska K, Sinielnikov I, Hautaniemi S, Janne OA. 2013. FoxA1 Specifies Unique Androgen and Glucocorticoid Receptor Binding Events in Prostate Cancer Cells. Cancer Research 73: 1570–1580.

Sahu B, Pihlajamaa P, Dubois V, Kerkhofs S, Claessens F, Jänne OA. 2014. Androgen receptor uses relaxed response element stringency for selective chromatin binding and transcriptional regulation in vivo. Nucleic Acids Research 42: 4230–4240.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1985. Information Content of Binding Sites on Nucleotide Sequences. J Mol Biol 188: 415–443.
Shaffer PL, Jivan A, Dollins DE, Claessens F, Gewirth DT. 2004. Structural basis of androgen receptor binding to selective androgen response elements. Proc Natl Acad Sci USA 101: 4758–4763.

Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, et al. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell 147: 1270–1282.

Stormo GD. 2000. DNA binding sites: representation and discovery. Bioinformatics 16: 16–23.

Tanay A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. Genome Research 16: 962–972.

Watson LC, Kuchenbecker KM, Schiller BJ, Gross JD, Pufall MA, Yamamoto KR. 2013. The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. Nat Struct Mol Biol 20: 876–883.

Watson PA, Arora VK, Sawyers CL. 2015. Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. Nat Rev Cancer 15: 701–711.

Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. The EMBO Journal 29: 2147–2160.

Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S, et al. 2013. Evaluation of methods for modeling transcription factor sequence specificity. Nature Biotechnology 31: 126–134.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell 158: 1431–1443.

Yan J, Liu Y, Lukasik SM, Speck NA, Bushweller JH. 2004. CBFβ allosterically regulates the Runx1 Runt domain via a dynamic conformational equilibrium. Nat Struct Mol Biol 11: 901–906.

Yoon C, Privé GG, Goodsell DS, Dickerson RE. 1988. Structure of an alternating-B DNA helix and its relationship to A-tract DNA. Proc Natl Acad Sci USA 85: 6332–6336.

Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. ed. C. Workman. PLoS Comp Biol 5: e1000590.

Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature Publishing Group 29: 480–483.

Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Research 41: W56–62.
Zhou X, O'Shea EK. 2011. Integrated Approaches Reveal Determinants of Genome-wide Binding and Function of the Transcription Factor Pho4. *Molecular Cell* 42: 826–836.

1994. Coupling of local folding to site-specific binding of proteins to DNA. http://www.sciencemag.org/content/263/5148/777.short.