**Zhang et. al. Supporting Information**

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

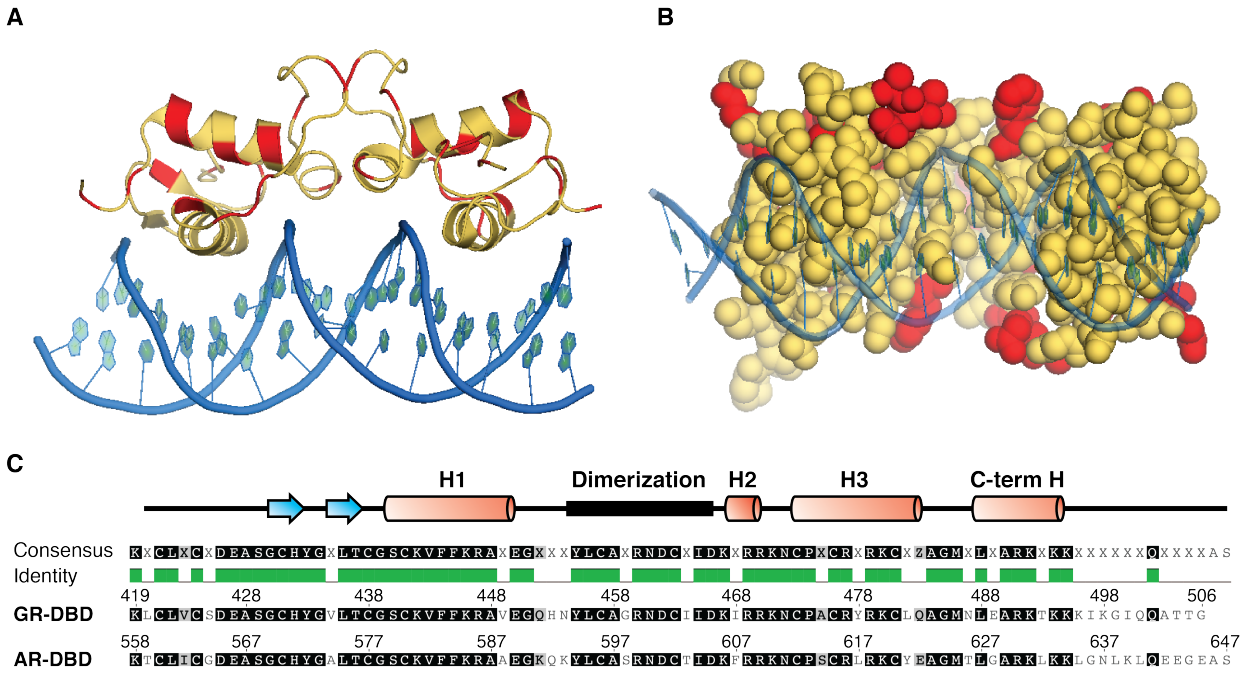


**Figure S1. AR- and GR-DBD share 100% identity in the amino acids that contact DNA. (A, B)** Crystal structure of AR-DBD:DNA complex. Non-identical residues between AR- and GR-DBD were colored by red. Strikingly, all residues that are located within 6Å of DNA are completely conserved between AR and GR. **(C)** Primary sequence alignment of AR- and GR-DBDs that were used in this study. The two proteins are 71% identical.

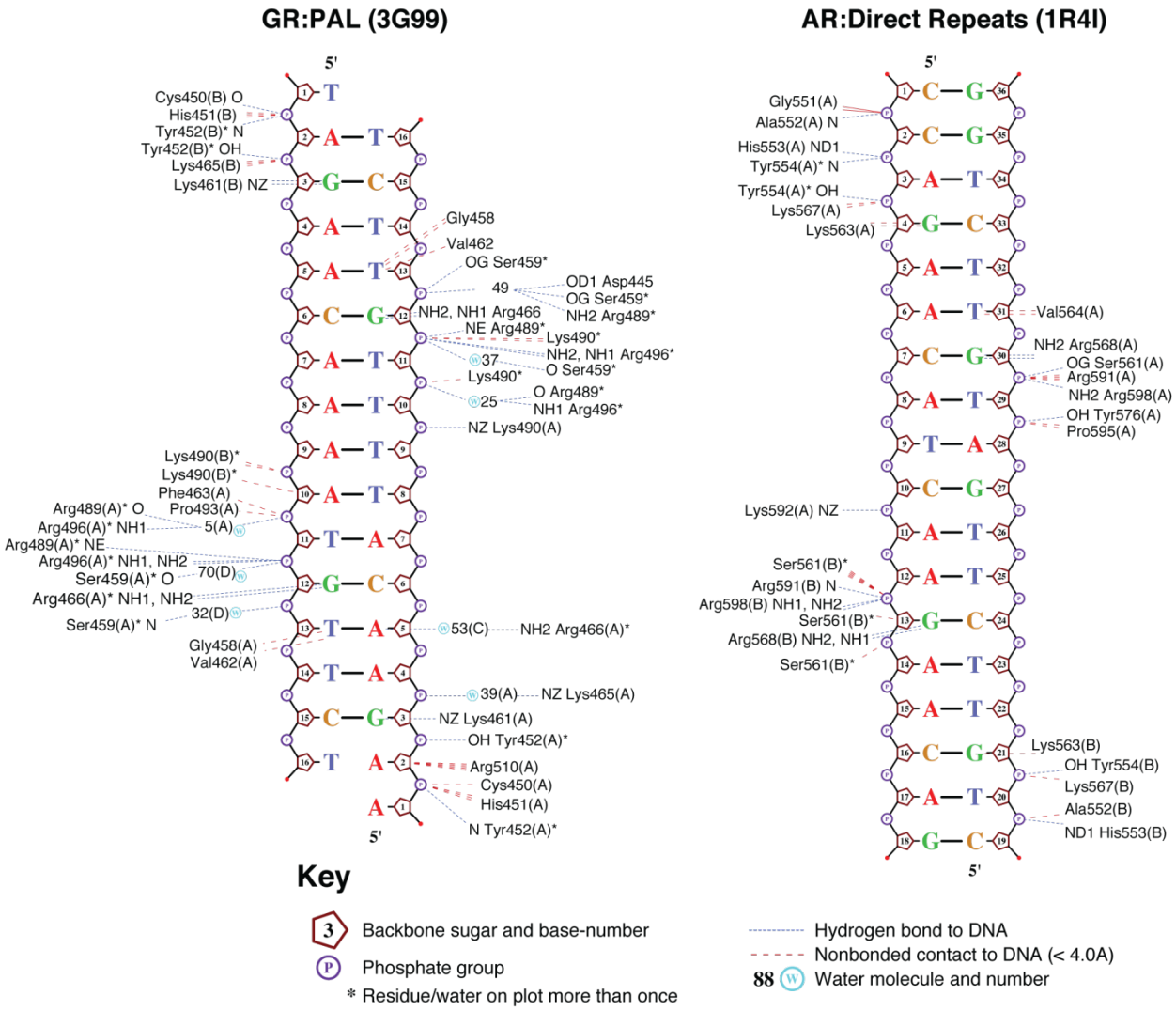


**Figure S2. Molecular interactions between AR- or GR-DBD and DNA identified within the respective crystal structures (calculated using NUCPLOT).** The interactions between AR- and GR-DBDs and DNA are conserved on an atomic level. The presence of water-mediated contacts for GR are likely not due to novel contacts, but instead because they are resolvable in the higher resolution (1.8 Å) GR structure but not in the lower resolution (3.1 Å) AR structure.

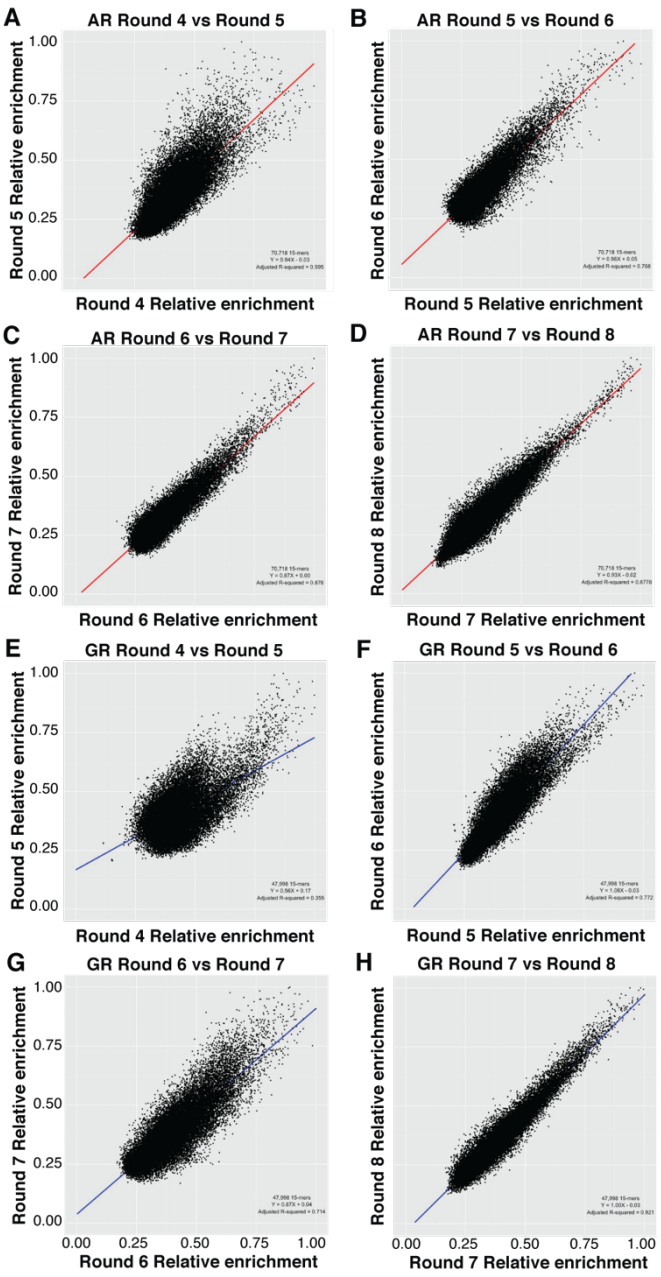


**Figure S3. Round-to-round comparison of 15-mer relative enrichment in AR/GR SELEX. (A-D)** Round to round comparison of the relative enrichment of 70,718 15-mers with at least 100 counts in R3-R8 AR-DBD SELEX libraries. **(E-H)** Round to round comparison of the relative enrichment of 47,908 15-mers with at least 100 counts in R3-R8 GR-DBD SELEX libraries. The relative enrichment became stable in the late rounds of SELEX (Since R6 for AR-DBD, and R7 for GR-DBD), indicating that protein is limiting, at which point the relative enrichment for each sequence is proportional to affinity after these rounds.

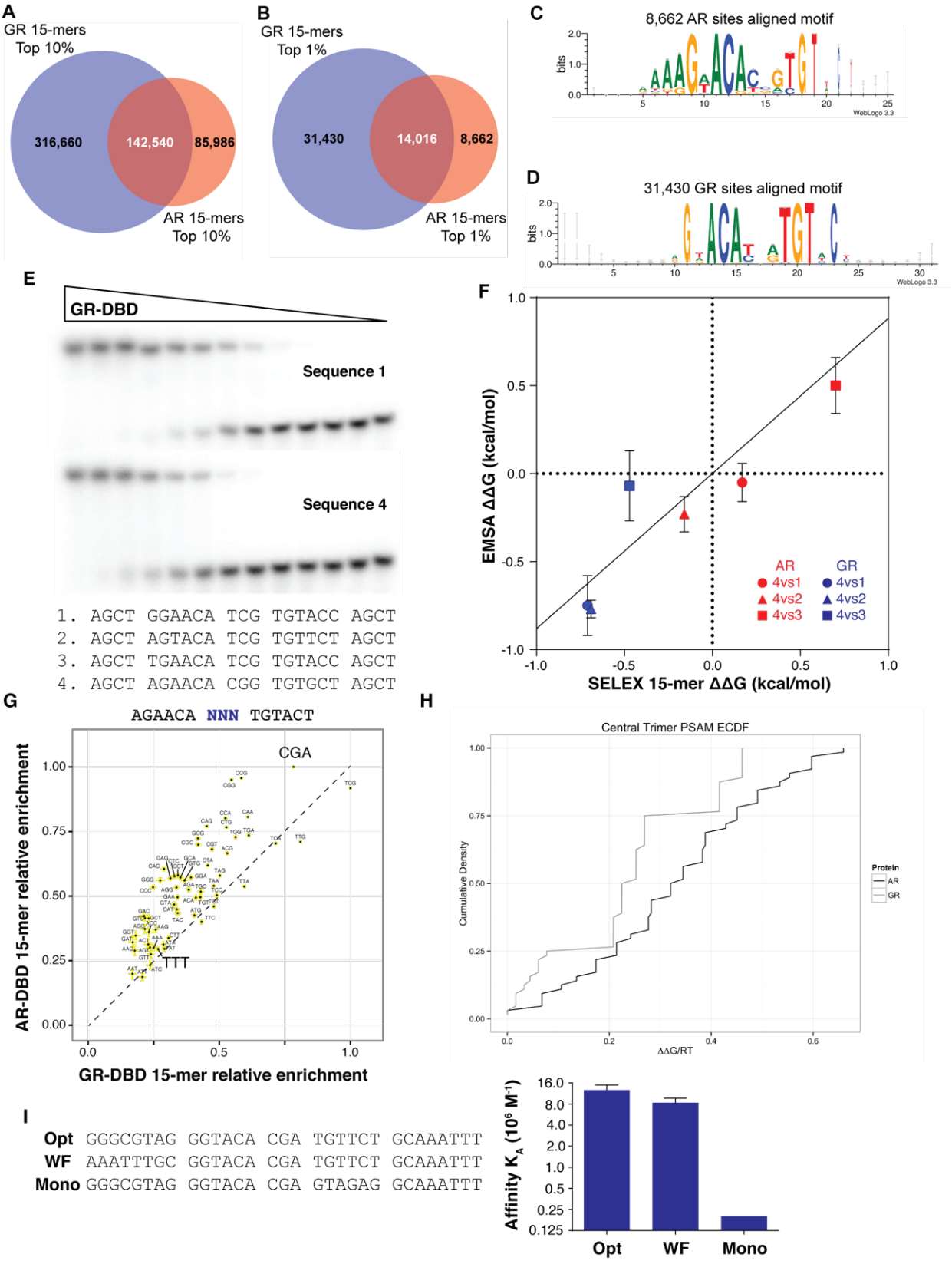
A
GR 15-mers
Top 10%
316,660
142,540
85,986
AR 15-mers
Top 10%
B
GR 15-mers
Top 1%
31,430
14,016
8,662
AR 15-mers
Top 1%
C
8,662 AR sites aligned motif
bits
WebLogo 3.3
D
31,430 GR sites aligned motif
bits
WebLogo 3.3
E
GR-DBD
Sequence 1
Sequence 4
1. AGCT GGAACA TCG TGTACC AGCT
2. AGCT AGTACA TCG TGTTCT AGCT
3. AGCT TGAACA TCG TGTACC AGCT
4. AGCT AGAACA CGG TGTGCT AGCT
F
EMSA ΔΔG (kcal/mol)
SELEX 15-mer ΔΔG (kcal/mol)
AR
4vs1
4vs2
4vs3
GR
4vs1
4vs2
4vs3
G
AGAACA NNN TGTACT
CGA
TTT
AR-DBD 15-mer relative enrichment
GR-DBD 15-mer relative enrichment
H
Central Trimer PSAM ECDF
Cumulative Density
ΔΔG/RT
Protein
AR
GR
I
Opt GGGCGTAG GGTACA CGA TGTTCT GCAAATTT
WF AAATTTGC GGTACA CGA TGTTCT GCAAATTT
Mono GGGCGTAG GGTACA CGA GTAGAG GCAAATTT
Affinity K_A (10^6 M^-1)
Opt
WF
Mono

**Figure S4. Flanking and spacer sequence preference for AR- and GR-DBD revealed by 15-mer analysis. (A)** Venn diagram of the top 10% of 15-mers enriched from R7 to R8 with at for AR-DBD (red) and GR-DBD (blue). **(B)** Venn diagram of the top 1% of 15-mers enriched from R7 to R8 with at for AR-DBD (red) and GR-DBD (blue). Although there are many more sequences in the top 10%, the ratio of 15-mers that are shared or specific to AR and GR remain constant. The reduced number of AR compared to GR sequences indicate that it more specific tan GR. **(C, D)** Motif logos generated by *k*-mer analysis of AR and GR specific sequences from the top 1% (**Figure S4B**). A visual comparison of two motifs indicates difference in preferences within the flanking and spacer sequences. **(E)** Validation of 15-mer tables for AR- and GR-DBD *SELEX-seq* data performed by EMSA. An example GR gel is shown. Sequences 1-4 were chosen from 15-mer tables by rank ordered GR affinity, yet having the order 2,1,3,4 for AR-DBD. **(F)** The affinity of each individual sequence was then measured by quantitative EMSA. The rank ordered affinity is consistent with the relative preference for binding for both AR and GR. Each sequence was measured a minimum of 3 times, with error bars representing the standard deviation. **(G)** Analysis of the effect of spacer on AR-DBD (y-axis) versus GR-DBD (x-axis) affinity. Examining the effect of spacer on a base sequence (AGAACA NNN TGTACT), we found that spacers could cause a large difference in affinity (see TTT vs. CGA). **(H)** The difference in spacer contribution to affinity is apparent from the increase in cumulative distribution function for AR compared to GR (zoom of **Figure 2E**). **(I)** The effect of changing flanking sequence (WF) and half-site sequences (Mono) on the affinity of GR-DBDs is compared to the optimal GR sequence (Opt) quantitative EMSA. The contribution of flanking sequence to GR affinity is significant, though less than the effect for AR (**Figure 2F**), and much less than ablating a half-site.

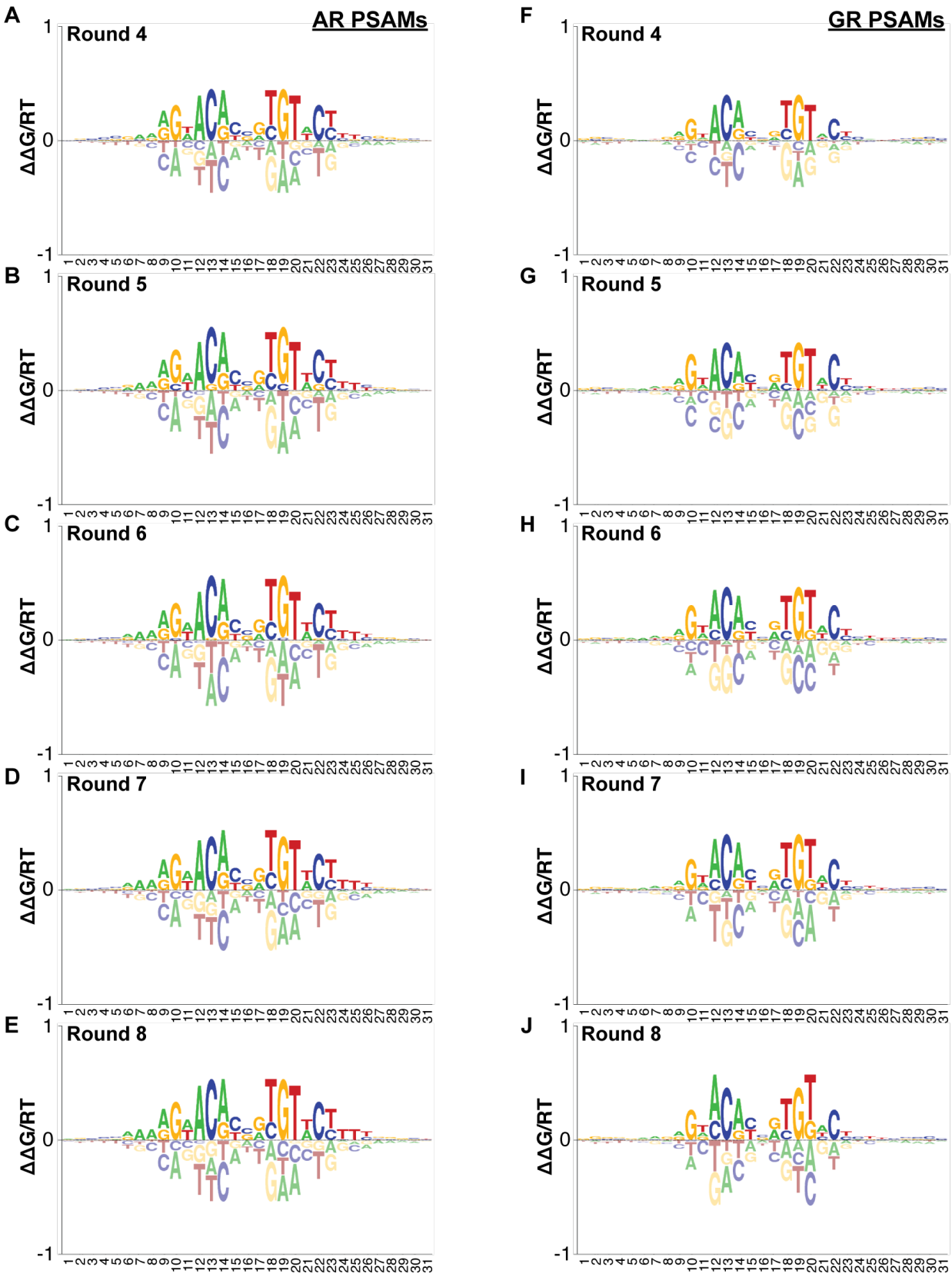
A
AR PSAMs
Round 4
B
Round 5
C
Round 6
D
Round 7
E
Round 8
F
GR PSAMs
Round 4
G
Round 5
H
Round 6
I
Round 7
J
Round 8
ΔΔG/RT
1
0
-1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

**Figure S5. *SelexGLM* is able to derive highly similar PSAMs from earlier rounds of SELEX-seq data.** SELEX-seq was performed using AR-DBD and GR-DBD using 8 rounds of selection, after which linear amplification of sequences was observed. Although later rounds of selection produced less noisy PSAMs (See **Figure S6**), highly similar PSAMs were able to be generated form earlier rounds using *SelexGLM*. **(A-E)** Logos generated from PSAMs for AR-DBD using libraries from rounds 4-8. The preference of AR for flanking As is clearly visible even in round 4 SELEX result. **(F-J)** Logos generated from PSAMs for AR-DBD using libraries from rounds 4-8. It should be noted that, although the preferred sequence is stable over these rounds for each protein, the most detrimental base pairs are refined by the later rounds of selection, reflecting a more accurate description of the free energy of binding.

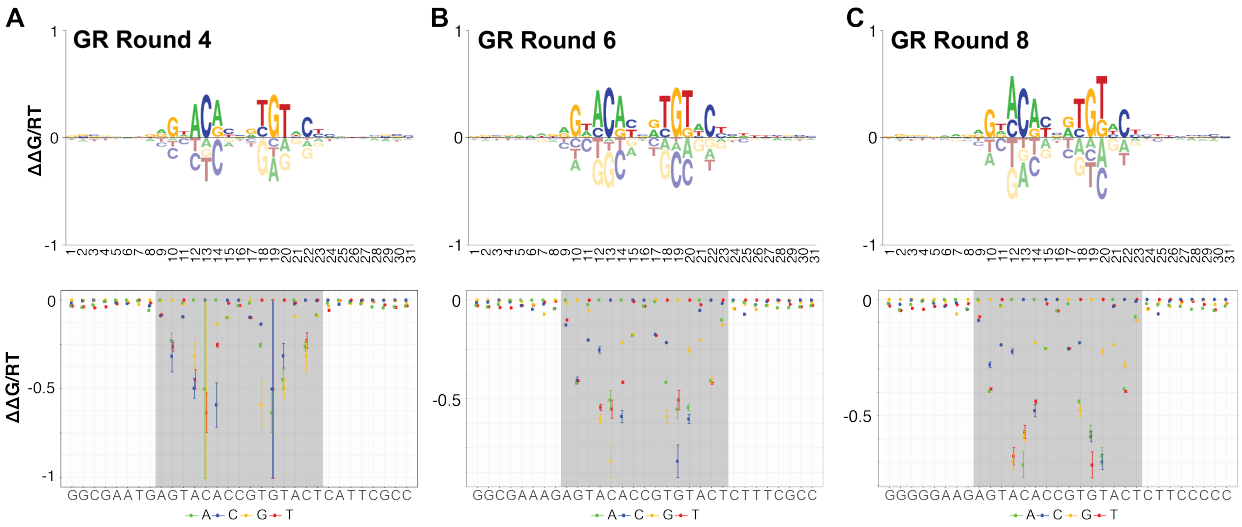


**Figure S6. More rounds of selection improve the precision and accuracy of PSAM model.** *SelexGLM* was used to generate a PSAM from each round of selection for both AR-DBD and GR-DBD (**Figure S5**). **(A-C top)** Logos representing the PSAMs from round 4, 6, and 8 for GR-DBD are shown here. As stated in **Figure S5**, the contribution of the preferred sequences are very stable from round to round, but the estimation for how much non-preferred base pairs are disfavor binding become more stable in later rounds. This can be more clearly visualized using the raw nucleotide affinity coefficients estimated by GLM **(A-C bottom)**. The most preferred sequence (ΔΔG = 0) remains constant from round to round, but estimates for most disfavored have large error bars in round 4, and have greater dispersion in later rounds. This noise in early rounds is the result of few sequences of low affinity that are available to estimate the deleterious effects of substitutions. As sequences are further selection, greater resolution is evident among low affinity sequences, without a cost from over-selection **(Figure 1E)**. This shows that *SelexGLM* can identify the most preferred sequence accurately from early rounds of selection, but the greater resolution provided by further selection can help distinguish preferences among low affinity sequences.

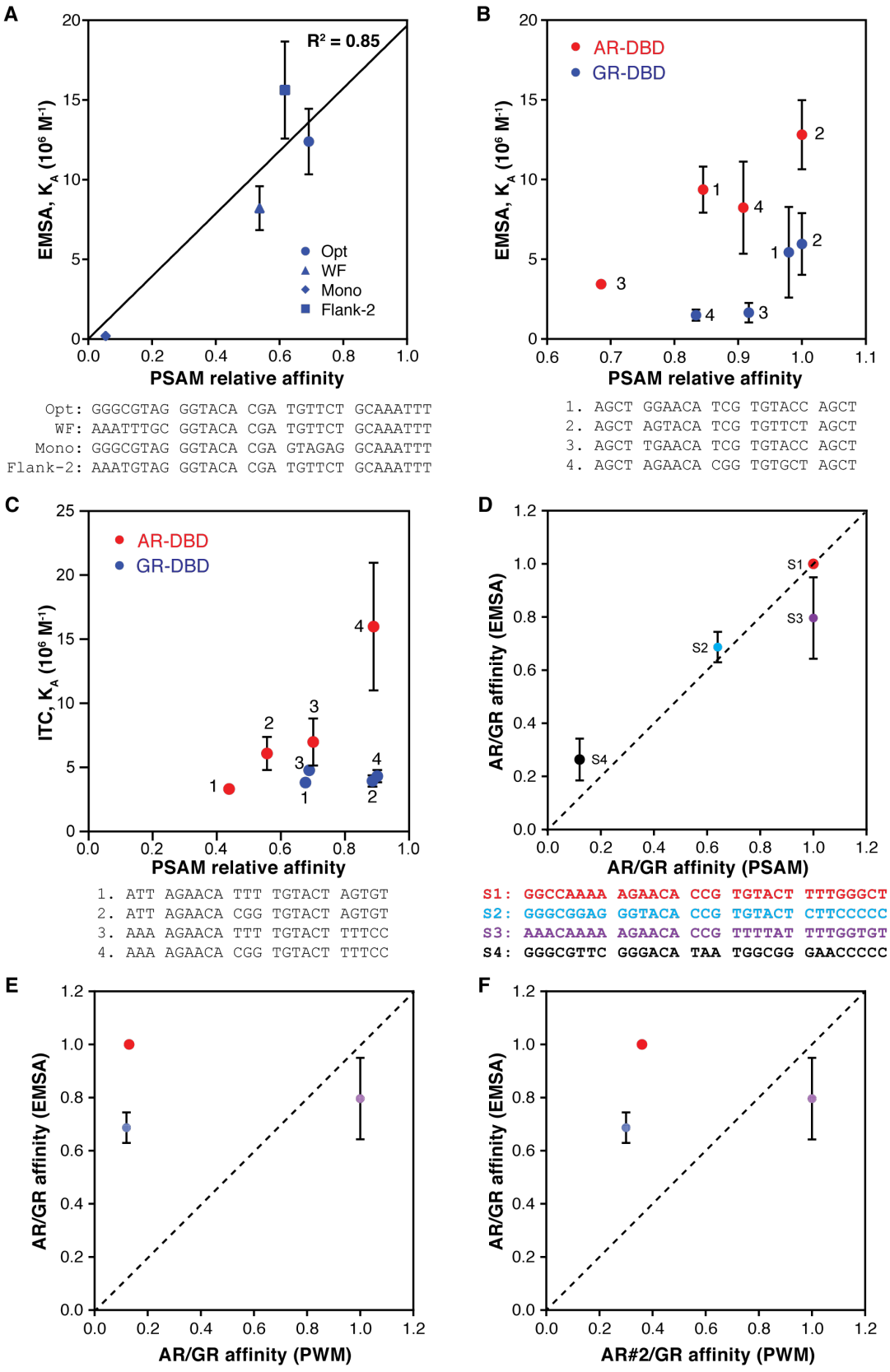
A
R² = 0.85
EMSA, K_A (10⁶ M⁻¹)
PSAM relative affinity
Opt
WF
Mono
Flank-2
Opt: GGGCGTAG GGTACA CGA TGTTCT GCAAATTT
WF: AAATTTGC GGTACA CGA TGTTCT GCAAATTT
Mono: GGGCGTAG GGTACA CGA GTAGAG GCAAATTT
Flank-2: AAATGTAG GGTACA CGA TGTTCT GCAAATTT
B
AR-DBD
GR-DBD
EMSA, K_A (10⁶ M⁻¹)
PSAM relative affinity
1. AGCT GGAACA TCG TGTACC AGCT
2. AGCT AGTACA TCG TGTTCT AGCT
3. AGCT TGAACA TCG TGTACC AGCT
4. AGCT AGAACA CGG TGTGCT AGCT
C
AR-DBD
GR-DBD
ITC, K_A (10⁶ M⁻¹)
PSAM relative affinity
1. ATT AGAACA TTT TGTACT AGTGT
2. ATT AGAACA CGG TGTACT AGTGT
3. AAA AGAACA TTT TGTACT TTTCC
4. AAA AGAACA CGG TGTACT TTTCC
D
AR/GR affinity (EMSA)
AR/GR affinity (PSAM)
S1: GGCCAAAA AGAACA CCG TGTACT TTTGGGCT
S2: GGGCGGAG GGTACA CCG TGTACT CTTCCCCC
S3: AAACAAAA AGAACA CCG TTTTAT TTTGGTGT
S4: GGGCGTTC GGGACA TAA TGGCGG GAACCCCC
E
AR/GR affinity (EMSA)
AR/GR affinity (PWM)
F
AR/GR affinity (EMSA)
AR#2/GR affinity (PWM)

**Figure S7. Position-specific affinity matrix accurately captures binding specificity within and outside the core motif. (A)** Plot of the PSAM predicted affinity (x-axis) versus the qEMSA measured affinity for 31-bp sequences: Optimal (Opt), Opt with suboptimal flanking sequences (WF), predicted monomeric site (Mono), and Opt with an AAA vs GGG leader (Flank-2). The measured affinities match well with the expected relative affinities calculated by *SelexGLM*. **(B)** Plot of the PSAM predicted affinity for 23-bp sequences (See **Figure S4B**) identified by 15-mer enrichment (x-axis) versus affinity measured by quantitative EMSA. The measured affinities match well with the expected affinities. Sequences are listed by relative affinity for GR. **(C)** Plot of the PSAM predicted affinity for GR- and AR-DBDs for DNA sequences vs measured affinities by isothermal titration calorimetry (ITC). Measured affinities are in good agreement with predicted. Sequences are listed by relative affinity for AR. **(D)** Plot of the ration of AR to GR affinity calculated from PSAMs (x-axis) for four sequences (See **Figure 3B**) with distinct shape features vs measured selectivity by EMSA (y-axis). To normalize the scores, the AR or GR relative affinity of each sequence was first calculated using Shape-1 (S1) as reference, then computed as the ratio of relative affinity of AR and GR for each sequence. Lower AR/GR affinity scores represent a higher preference for GR. This relative preference between AR and GR measured by EMSA is in good agreement with predicted on quantitative level. All points are based on at least three experiments, with error bars representing the standard error of mean. **(E, F)** To compare the performance of our calculated PSAMs to existing motifs, we performed the same analysis described for **(D)**, but used position weight matrices calculated from HT-SELEX (data found in cisBP). Human AR PWMs AR-1 (Cis-BP ID: M5288_1.02) and AR-2 (Cis-BP ID: M5289_1.02) were used in the AR affinity calculation of Figure S7E and S7F respectively. The only Human GR PWM (Cis-BP ID: M5683_1.02) from HT-SELEX was used to calculate GR affinity in both cases. Sequence S4 was excluded because its PWM-based affinity for GR is 0. The AR/GR affinity calculated from these PWMs does not correspond well to the measured relative affinities.

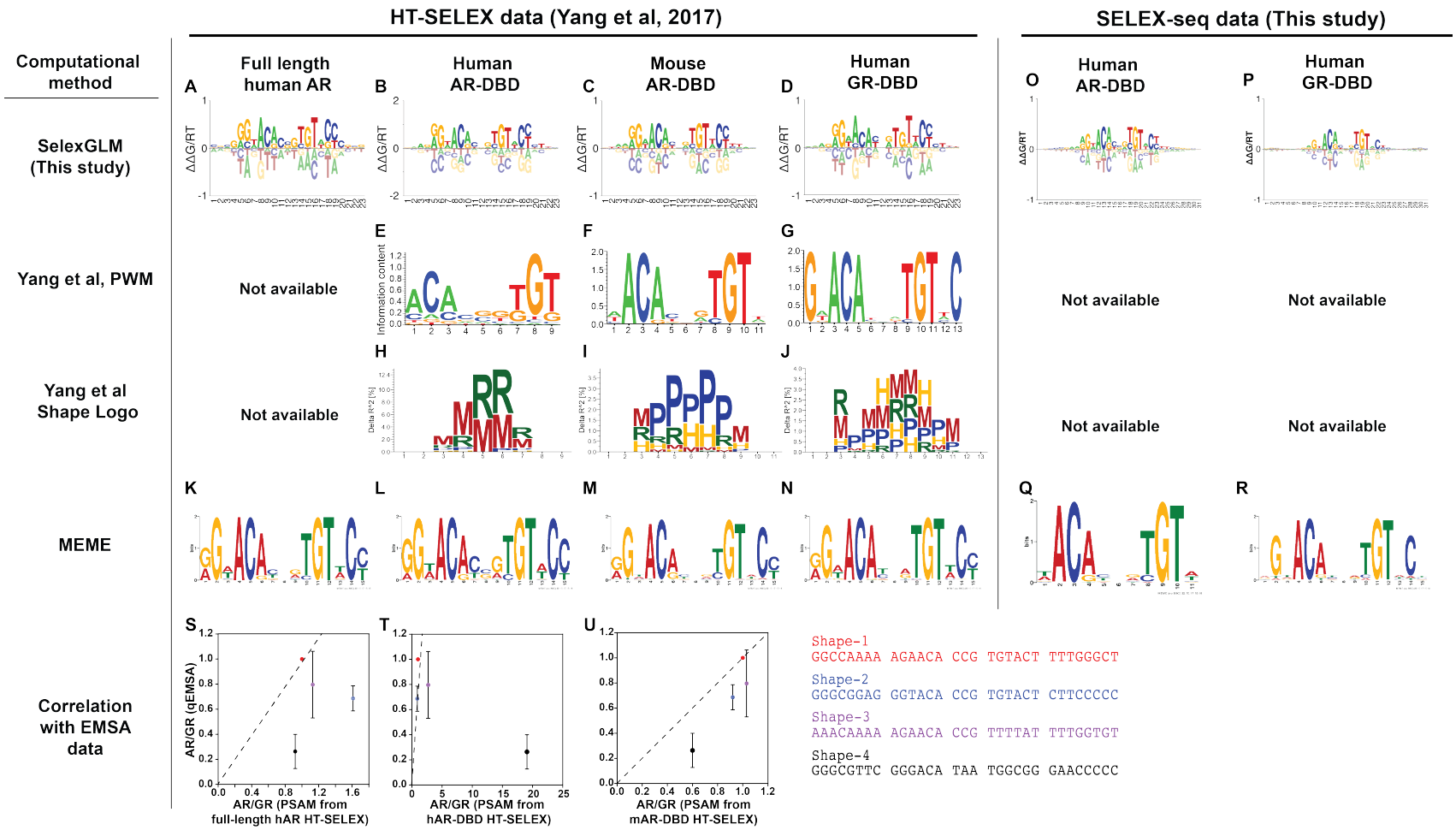
HT-SELEX data (Yang et al, 2017)
SELEX-seq data (This study)
Computational method
SelexGLM (This study)
Yang et al, PWM
Yang et al Shape Logo
MEME
Correlation with EMSA data
Full length human AR
Human AR-DBD
Mouse AR-DBD
Human GR-DBD
Human AR-DBD
Human GR-DBD
A
B
C
D
O
P
E
F
G
H
I
J
K
L
M
N
Q
R
S
T
U
Not available
Not available
Not available
Not available
Not available
Not available
ΔΔG/RT
Information content
Delta R^2 [%]
AR/GR (qEMSA)
AR/GR (PSAM from full-length hAR HT-SELEX)
AR/GR (PSAM from hAR-DBD HT-SELEX)
AR/GR (PSAM from mAR-DBD HT-SELEX)
Shape-1
GGCCAAAA AGAACA CCG TGTACT TTTGGGCT
Shape-2
GGGCGGAG GGTACA CCG TGTACT CTTCCCCC
Shape-3
AAACAAAA AGAACA CCG TTTTAT TTTGGTGT
Shape-4
GGGCGTTC GGGACA TAA TGGCGG GAACCCCC

**Figure S8: *SelexGLM* is able to discern a large, but inconsistent, binding site from HT-SELEX data.** *SelexGLM* was used to calculate a PSAM from recently re-sequenced HT-SELEX data (Yang et al. 2017). Three data sets were available for AR: full length human AR **(A)**, human AR-DBD **(B)**, and mouse AR-DBD **(C)**; one data set was available for GR: human GR-DBD **(D)**. Consistent with our own SELEX-seq data, *SelexGLM* identified contributions to affinity over a 23-bp binding site **(A-D)**. A preference for poly-A flanking sequences is evident in the mouse AR HT-SELEX PSAM logo (C), but not in the other two AR logos (A,B). The relative contribution of the flanking poly-A sequences to affinity of the mouse AR PSAM is less pronounced than in our SELEX-seq data **(O-P).** Oligomer table analysis performed by Yang *et al*. was able to discern preferences over a more limited binding site, with some differences at individual sites **(E-G)**. **(H-J)** Previous shape analysis performed on the HT-SELEX data identifies differences in shape preference within the 15-bp core motif**,** but the shape preference for AR is not consistent, and shape data was not available for flanking positions outside this core. **(K-P)** MEME analysis of the HT-SELEX data **(K-N)** shows that probabilistic modeling yields a motif for the 15-bp core similar to that inferred by *SelexGLM* **(O-P)**. However, some differences are evident within the core, and no preference is discerned in the flanking regions. **(Q,R)** MEME analysis of 1,000 sequences randomly samples sequences from our SELEX-seq data is able to identify important core base pairs, but did not provide information for the full 15-bp core motif or flanking regions. **(S-U)** Plots of the relative affinity of AR to GR for the 'Shape' sequences validated individually by EMSA (y-axis) versus the HT-SELEX PSAMs (x-axis). The correlation between validated sequences is substantially worse for the HT-SELEX data than for our SELEX-seq data **(**cf. **Figure S7D)**.

**Table S1. DNA sequences of SELEX library, primers and EMSA/ITC validation sites**

| Oligo Name | Purpose | Sequence |
|---|---|---|
| SELEX library | **Library amplification / sequencing** | 5′ GTTCAGAGTTCTACAGTCCGACGATC (N23)TGGAATTCTCGGGTGCCAAGG 3′ |
| TSSR0 | | 5′ GTTCAGAGTTCTACAGTCCGACG 3′ |
| TSSR1 | | 5′ CCTTGGCACCCGAGAATTCCA 3′ |
| Cy5-TSSR1 | | 5′ Cy5-CCTTGGCACCCGAGAATTCCA 3′ |
| TSSR2 | | 5′ AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA 3′ |
| TSSR-RPIX* | | 5′ CAAGCAGAAGACGGCATACGAGAT**XXXXXX**GTGACTGGAGTTCCTTGGCACCCGAGAATTCCA 3′ |
| FS-23bp-1 | **EMSA** Figure 2C | 5′ AAA AGAACA CGA TGTACT TTTCC 3′ |
| FS-23bp-2 | | 5′ ATT AGAACA CGA TGTACT AGTGT 3′ |
| FS-23bp-3 | | 5′ AAA AGAACA TTT TGTACT TTTCC 3′ |
| FS-23bp-4 | | 5′ ATT AGAACA TTT TGTACT AGTGT 3′ |
| Core-23bp-1 | **EMSA** Figure S4D Figure S7B | 5′ AGCT GGAACA TCG TGTACC AGCT 3′ |
| Core-23bp-2 | | 5′ AGCT AGTACA TCG TGTTCT AGCT 3′ |
| Core-23bp-3 | | 5′ AGCT TGAACA TCG TGTACC AGCT 3′ |
| Core-23bp-4 | | 5′ AGCT AGAACA CGG TGTGCT AGCT 3′ |
| Opt-31bp | **EMSA** Figure S4 Figure S7A | 5′ GGGCGTAG GGTACA CGA TGTTCT GCAAATTT 3′ |
| WF-31bp | | 5′ AAATTTGC GGTACA CGA TGTTCT GCAAATTT 3′ |
| Mono-31bp | | 5′ GGGCGTAG GGTACA CGA GTAGAG GCAAATTT 3′ |
| Flank-2-31bp | | 5′ AAATGTAG GGTACA CGA TGTTCT GCAAATTT 3′ |
| Shape-1 | **EMSA** Figure 3B Figure S7D | 5′ GGCCAAAA AGAACA CCG TGTACT TTTGGGCT 3′ |
| Shape-2 | | 5′ GGGCGGAG GGTACA CCG TGTACT CTTCCCCC 3′ |
| Shape-3 | | 5′ AAACAAAA AGAACA CCG TTTTAT TTTGGTGT 3′ |
| Shape-4 | | 5′ GGGCGTTC GGGACA TAA TGGCGG GAACCCCC 3′ |
| ITC-23bp-1 | **ITC** Figure 4 Figure S7C | 5′ ATT AGAACA TTT TGTACT AGTGT 3′ |
| ITC-23bp-2 | | 5′ ATT AGAACA CGG TGTACT AGTGT 3′ |
| ITC-23bp-3 | | 5′ AAA AGAACA TTT TGTACT TTTCC 3′ |
| ITC-23bp-4 | | 5′ AAA AGAACA CGG TGTACT TTTCC 3′ |

*Available TSSR-RPIX with different barcodes (RPI1 – RPI48) can be found in *Illumina Custom Sequence letter, 2012, Sep 7th*. XXXXXX: 6bp barcode for Illumina sequencing.

## REFERENCES CITED

Yang L, Orenstein Y, Jolma A, Yin Y, Taipale J, Shamir R, Rohs R. 2017. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol* **13**: 910–14.