Supporting Information:

Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding

Daniel D. Le[a,1], Tyler C. Shimko[a,1], Arjun K. Aditham[b,c], Allison M. Keys[c,d], Scott A. Longwell[b], Yaron Orenstein[e], and Polly M. Fordyce[a,b,c,f,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305
[b]Department of Bioengineering, Stanford University, Stanford, CA 94305
[c]Stanford ChEM-H, Stanford University, Stanford, CA 94305
[d]Department of Chemistry, Stanford University, Stanford, CA 94305
[e]Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, POB 653
[f]Chan Zuckerberg Biohub, San Francisco, CA 94158

December 29, 2017

1 Supplemental Methods

1.1 Binding energy calculations

Using the model proposed by Djordjevic, *et al.* (1) that was refined by Stormo and colleagues (2, 3), we consider the binding of DNA sequences S_i among many S_j competitor substrates at equilibrium with a transcription factor. The probability of being bound to the transcription factor as a function of sequence identity, represented by $P(bound \mid S_n)$, is given by:

$$P(bound \mid S_n) = \frac{e^{-\Delta G_n}}{e^{-\mu} + e^{-\Delta G_n}}$$

Where ΔG_n is the binding free energy and μ is the chemical potential, equal to the natural log of the transcription factor concentration. Both terms are in units of RT.

Sequencing of TF-bound substrates yields $P(S_i \mid bound)$, which is the inclusion probability of S_i among the bound substrate distribution. From the partition probability of substrates between bound and unbound, $P(S_i \mid bound)$, application of Bayes' theorem and the law of total probability yields:

$$P(S_i \mid bound) = \frac{P(bound \mid S_i)P(S_i)}{\sum_j P(bound \mid S_j)P(S_j)}$$

Where $P(S_n)$ describes the probability of a given species within the input substrate distribution. The combination of these two equations returns $P(S_i \mid bound)$ as a function of binding energies and input probabilities.

Next, we make the following assumptions:

- The transcription factor concentration is substantially lower than the total DNA concentration.
- As the TF concentration is minimized, μ approaches negative infinity. In effect, this causes the denominator in equation 1 to be dominated by the $e^{-\mu}$ term, which we can approximate with a constant $C \simeq e^{-\mu} + e^{-\Delta G}$.
- The sum of $P(S_j)$ in a large population is ~ equal to one.

Applying these assumptions to equation 2, it becomes possible to isolate ΔG_i as a function of $P(S_i \mid bound)$ and $P(S_i)$. This equality is given in equation 3, in units of RT:

$$\Delta G_i = -ln\Big(\frac{P(S_i \mid bound)}{P(S_i)}\Big) + \Delta G_j$$

Where ΔG_j equals the total binding energy excluding contribution from S_i. Setting this value to zero yields the relative binding affinity of S_i, represented by $\Delta\Delta G_i$ in units of RT, which is given in equation 4:

$$\Delta\Delta G_i = -ln\Big(\frac{P(S_i \mid bound)}{P(S_i)}\Big)$$

1.2 Monte Carlo assay simulations

The distribution of theoretical reference binding energies is assumed to be normal and symmetric about $\Delta\Delta G = 0$ kcal/mol. Representation of input substrate species is assumed to follow a uniform distribution. We tested all combinations of the following conditions: affinity range = 0.25-5 kcal/mol, library size = 10^2–10^6 species, and total read counts = 10^3–10^8 reads equally allocated between bound and input fractions. From these conditions, the bound species frequency distribution was derived and randomly sampled. The simulated bound and input count ratio for each substrate was used to calculate putative individual $\Delta\Delta G$ values. Each combination of parameters was used to simulate ten replicate datasets, which were compared to the theoretical $\Delta\Delta G$ by Pearson's correlation. Correlation coefficients (r) with Bonferroni-corrected p-values above 0.05 were set to zero. Accuracy was defined as the product of r^2 and the fraction of observed species.

For datasets with large energetic ranges of 3–5 kcal/mol (160- to 4590-fold change in bound over input), excellent correlation (Pearson's r^2) was observed despite comparatively low overall read depths (Figure S1). However, under these conditions, only a narrow fraction of species, primarily high affinity substrates, were observed relative to the input population (Figure S2). We reasoned that sequencing such a population disproportionately samples the minority of strong binding species, resulting in high measurement accuracy. For the majority of weaker binders that are under-sampled, few reads are allocated to these substrates, yielding inaccurate measurements. While the simulation explicitly assumes a normal energy distribution among species with uniform input frequency, these assumptions are potentially violated in real systems, which would further contribute to sampling imbalance.

For the simulations probing the effect of $\Delta\Delta G$ ranges and library size on equilibrium concentrations, The following assumptions were made:

1. Each substrate in the library is present in the same initial concentration.
2. Since simulation of 1,000,000+ sequences is computationally intractable, we instead uniformly sample 100 substrates across the energetic range and use one high concentration sequence as a stand in for the remaining substrates present in the highest density portion of the distribution.
3. The concentration of the total input material is 1 μM and the concentration of protein in the assay is 30 nM.
4. The concentration of any individual species queried is $\frac{1\mu M}{n}$ where n is the total number of species in the library (not just the number simulated).

Based on these simulations, we note that large energetic spreads and large libraries can cause depletion of the unbound fraction (Figure S3), which causes systematic overestimation of the value of $\Delta\Delta G$ for tightly bound species (Figure S4). When combined with the stochastic sampling error simulations discussed above, we note that the input library serves as a good approximation for unbound the unbound fraction in situations where the energetic range is relatively small (0–4 kcal/mol). Beyond this range, ligand depletion can cause inaccurate estimates for tightly bound species (Figure S5).

The results from these simulations have broad implications for other sequencing-based binding assays: (1) For *de novo* core motif discovery, in which the fraction of strong binders is small and the difference in energy relative to weak binders is large, low read depth is sufficient to identify the sub-population of very strong binders; and, (2) when the energetic range is narrow, as in the flanking context library or for other motif refinement approaches, comprehensive sequence coverage is attained at the expense of overall accuracy. While current NGS instruments have the output necessary for high sequence diversity applications (*e.g.* flanking library), operational costs scale with library size, which limits the libraries that can be probed. These general guidelines for measuring accurate binding energies from sequencing-based assays can facilitate broader use of thermodynamic metrics for assessing TF binding specificity.

1.3 DNA substrate preparation (Compatible with Illumina NGS)

Synthesis of ssDNA oligos by IDT or Sigma:

- DNA substrate template:
  - GATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTNNGTATCACGAG NCAATACACTGTTATCNNNNNCACGTGNNNNNCTACTCGTTCGGTTANCAG GAGAGCTNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
- Illumina Read 2 primer:
  - GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Polymerase extension:

- 50 μL of Q5 Hot Start High-Fidelity 2x Master Mix (M0494S)
- 40 μL of 10 μM DNA substrate template
- 5 μL of 100 μM Illumina Read 2 primer
- 5 μL of DNase-/RNase-free water
- Zymo Clean and Concentrate - 25 column purification kit (D4005)

All components were allowed to equilibrate to 4°C on ice. To a PCR strip tube, the following were added in order: 50 μL of Q5 Hot Start High-Fidelity 2x Master Mix, 40 μL of 10 μM IDT synthetic DNA substrate template, 5 μL of 100 μM Illumina Read 2 primer, and 5 μL of DNase-/RNase-free water. The solution was incubated at the following thermocycler settings:

1. Initial melt: 98°C for 30 sec
2. 4.0°C/s ramp rate
3. 1 cycle:
   (a) Cycle melt: 98°C for 10 sec
   (b) 0.1°C/s ramp rate
   (c) Cycle anneal: 30°C for 1 min 15 sec
   (d) 0.1°C/s ramp rate
4. Final extension: 65°C for 5 min
5. Store at 4°C

The extended product was purified using a Zymo Clean and Concentrate - 25 column purification kit using a 5:1 ratio of binding buffer to extended product solution as specified in the user manual. The purified eluent was quantified using a low-volume UV/Vis spectrometry and adjusted to 1 μM in elution buffer (EB). This solution was stored for several weeks at -20°C.

1.4 Protein expression using *n vitro* transcription and translation (IVTT)

PHO4 and CBF1 open reading frames were sub-cloned into a pTNT vector (Promega/Genscript) with a C-terminal monomeric enhanced green fluorescent protein (meGFP) tag using Golden Gate Assembly, as previously reported (4, 5).

Promega Wheat Germ Extract IVTT kit:

- 50 μL of Promega TNT Wheat Germ Extract (L411A)
- 4 μL Promega TNT Reaction Buffer (L462A)
- 0.66 μL of Promega Amino Acid Mixture Minus Methionine, 1 mM (L996A)
- 0.66 μL of Promega Amino Acid Mixture Minus Leucine, 1 mM (L995A)
- 0.66 μL of Promega Amino Acid Mixture Minus Cysteine, 1 mM (L447C)
- 2 μL of Promega Recombinant RNasin Ribonuclease Inhibitor (N251A)
- 2 μL of Promega TNT T7 Wheat Germ Polymerase (L516A)
- 1-2 μg of expression plasmid

All components were allowed to equilibrate to 4°C on ice. To a PCR strip tube, the following were added in order: 50 μL of Promega TNT Wheat Germ Extract, 4 μL of Promega TNT Reaction Buffer, 0.66 μL of 1 μM Promega Amino Acid Mixture Minus Methionine, 0.66 μL of 1 μM Promega Amino Acid Mixture Minus Leucine, 0.66 μL of 1 μM Promega Amino Acid Mixture Minus Cysteine, 2 μL of Promega Recombinant RNasin Ribonuclease Inhibitor, 2 μL of Promega TNT T7 Wheat Germ Polymerase, and 50 μL of DNase-/RNase-free water. Then, 1–2 μg of expression plasmid was added and incubated at 30 °C for 3 hours. After incubation, the solution was clarified by benchtop centrifugation set to 15,000 RPM for 10 minutes at room temperature. The supernatant was collected and loaded directly onto a MITOMI device.

1.5 MITOMI-seq assay

1.5.1 Microfluidic device operation

MITOMI PDMS devices were attached to a custom pneumatic control system to pressurize valves and control fluid flows (6). Input flow lines were pressurized to 4.25 PSI and the control lines were pressurized to 32 PSI. The device was degassed by introducing a 1x phosphate-buffered saline (PBS) into the main channels with the output valve closed. Using MATLAB scripts, the following automation steps were performed:

Automation step A - Surface chemistry

- 200 μL of 2 mg/mL biotinylated bovine serum albumin in 140 nM citrate, pH 6.8 + 50 μL of 250 μg/mL of poly(deoxyinosinic-deoxycytidylic) acid in EB (bBSA),
- 100 μL of 1 mg/mL neutravidin in 1x PBS (NA),
- 50 μL of 0.04 mg/mL biotin-conjugated antibody specific to green fluorescent protein (αGFP),
- 500 μL of PBS

1. For each input, 15 sec purge to waste followed by 15 sec PBS
2. 20 min bBSA
3. 6 min 40 sec PBS
4. 20 min NA
5. 6 min 40 sec PBS wash
6. Close BUTTON valves
7. 5 min PBS wash
8. 20 min bBSA
9. 6 min 40 sec PBS
10. 1 min 20 sec αGFP
11. Open BUTTON valves
12. 11 min 40 sec αGFP
13. 6 min 40 sec PBS
14. Increase flow line pressure to 8.50 PSI
15. Increase control line pressure to 35 PSI

Automation step B - TF-DNA equilibration

- 40 μL of 1 μM DNA substrate in 1x Tris-EDTA buffer (DNA)
- 100 μL of IVTT protein expression solution (PROT)

1. For each input, 15 sec purge to waste followed by 15 sec PBS
2. 20 min PROT
3. 6 min 40 sec PBS
4. 4 min DNA
5. Close OUT valve
6. Equilibrate for 1 hr
7. Close BUTTON valves
8. 5 min PBS

Automation step C - Device wash

- 250 μL of freshly-prepared 2 mg/mL bovine trypsin in 1x PBS (TRYP)

1. For each input, 15 sec purge to waste followed by 15 sec PBS
2. 30 min TRYP
3. 5 min PBS
4. 10 min bBSA
5. 10 in PBS

Automation step D - Elution

1. 300 x 3 sec BUTTON cycle under constant PBS
2. Collect bound fraction eluent in pipette tip attached to OUT port

1.5.2 Sequencing library preparation

- 35 μL of Q5 Hot Start High-Fidelity 2x Master Mix (M0494S)
- 5 μL of 10 μM P5 indexed adapter primer [Table S7]
- 5 μL of 10 μM P7 indexed adapter primer [Table S8]
- ThemoFisher GeneJET Cleanup Kit (K0851)

To respective PCR strip tubes, approximately 25 μL of bound fraction eluent and 25 μL of 100 pM input was added. The following reagents were equilibrated to 4°C and added in order: 35 μL of Q5 Hot Start High-Fidelity 2x Master Mix, 5 μL of 10 μM P5 indexed adapter primer and 5 μL of 10 μM P7 indexed adapter primer. The solution was incubated at the following thermocycler settings:

1. Initial melt: 98°C for 30 sec
2. 4.0°C/s ramp rate
3. 11-14 cycles:
   (a) Cycle melt: 98°C for 10 sec
   (b) 4.0°C/s ramp rate
   (c) Cycle anneal and extend: 65°C for 1 min 15 sec
   (d) 4.0°C/s ramp rate
4. Final extension: 65°C for 5 min
5. Store at 4°C

The PCR product was purified using a ThemoFisher GeneJET Cleanup Kit following user manual Protocol B for adapter removal. The purified library DNA quality was analyzed using a fluorogenic assay (*e.g.* Qubit) to determine concentration and using a electrophoretic migration assay (*e.g.* Bioanalyzer) to determine size. Libraries were sequenced at Stanford Protein and Nucleic Acid Facility, Stanford Functional Genomics Facility or Chan-Zuckerberg Biohub Core using Illumina MiSeq and NextSeq instruments with 2 x 75 cycle reagent kits.

1.5.3 Sequencing read count analysis

Sequencing reads were demulitplexed using unique paired index sequences. For each indexed set, paired-end reads were merged using the Paired-End reAd mergeR (PEAR) algorithm (7). Resultant merged sequences were filtered as follows: (1) average PHRED scores greater than 30 (*i.e.* $>$ 99.9% base call accuracy), (2) PHRED score greater than 30 at functional positions (UMI, variable flank, and core motif), and (3) matched constant region sequence identity adjacent to variable flank regions (*i.e.* ATCNNNNNCACGTGNNNNNCTA). Total counts per flank sequence were determined by the de-duplicated frequency of associated unique molecular identifier (UMI) sequences.

1.6 MITOMI Titration Binding

1.6.1 Library Preparation

DNA sequence libraries were normalized to a concentration of 100 μM in water [Tables S3, S6]. A "universal" AlexaFluor647-functionalized primer was diluted to 100 μM in water ("Universal primer" sequence: 5'- /5Alexa647/ GTC ATA CCG CCG GA-3') [Table S6]. DNA oligos were provided by IDT. Prior to use, plates were defrosted overnight at 4°C and centrifuged. To generate the dNTP solution, 20 μL of dNTP mixture (25 mM of each dNTP) was combined with 480 μL of Milli-Q water and kept on ice.

To generate libraries of fluorescently-labeled double-stranded DNA, the primer and library DNA were first annealed and then extended with recombinantly expressed Klenow exo$^-$ (New England Biolabs). The annealing reaction recipe was as follows:

- 6 μL library DNA substrate (100 μM stock)
- 2 μL NEBuffer 2
- 6 μL dNTPs
- 6 μL "universal" Alexa647-labeled primer (100 μM stock)

The library DNA substrate was added to a 96-well PCR plate using a Liquidator 96-channel pipette (Rainin). NEBuffer2, dNTPs, and the "universal" oligo were combined into a master mix scaled for 100 reactions, and 14 μL of mastermix was transferred into wells of the 96-well plate using a multichannel pipette. The plate was sealed and placed in a thermocycler for annealing using the following protocol:

1. 95°C for 3 min
2. Anneal to 37°C over 45 min

The Klenow mix was prepared as a master mix scaled to 100 reactions as follows and stored on ice until use. Per reaction:

- 1 μL Klenow exo$^-$
- 1 μL NEBuffer 2 (B7002S)
- 8 μL Milli-Q water

After annealing the primer to the DNA, the PCR plate was centrifuged at 4°C and placed back into the thermocycler at 37°C. To this plate, 10 μL of the Klenow mastermix was added using a multichannel pipette. The extension protocol was as follows:

1. 37°C for 60 min
2. 72°C for 20 min
3. Anneal to 10°C over 45 min

The prepared library was then serially diluted into a sterile-filtered print solution formulated as follows:

- 12.5 mg/mL D-(+)-trehalose dihydrate (Sigma Life Science T9531-25G)
- 1% bovine serum albumin (Sigma Life Science B4287-25G)
- 3x saline-sodium-citrate (SSC) Buffer

The serially diluted library was then transferred into 384-well plates for printing onto 2" x 3" SuperChip epoxysilane coated glass slides using with a custom-built robotic microarrayer. MITOMI devices were aligned to these printed libraries and bonded for 12 hours at 95°C.

1.6.2 Equilibrium binding protocol

Briefly, MITOMI experiments were run as previously described (4) with minor modifications. After the initial biotinylated-BSA passivation, printed DNA was solubilized using wheat germ extract (formulated as 30 μL extract and 30 μL Milli-Q water) before the experiment continued. Secondly, meGFP-tagged Pho4 or Cbf1 expression reactions were prepared as described above and incubated at 30°C on an orbital shaker (300 RPM) for 3 hours before clarification by centrifugation and introduction into the device. Finally, after deposition of protein onto the device, button valves and neck valves were opened to allow DNA to diffuse throughout the protein chamber for 20 minutes. Afterwards, button valves were opened and neck valves were shut to allow the protein-DNA interaction to equilibrate for 60 minutes.

1.6.3 Equilibrium binding analysis

Prior to analysis, all images were flat-field corrected (8) and stitched using the Grid/Collection Stitching plugin (9) available on FIJI. Protein and DNA binding were quantified from images using a MATLAB script that automates feature detection and data extraction. The concentration of DNA in the chamber was quantified using the background subtracted median intensity of Cy5 signal in "prewash" images in which solubilized DNA is in the device protein chambers. The amount of DNA bound to protein was quantified using the background-subtracted median Cy5 signal underneath the buttons of the device in "postwash" images. In order to calculate the DNA/Protein ratio, the intensity of the Cy5 signal was normalized by the median meGFP signal underneath the button. K_d values from binding isotherms were estimated using a global nonlinear fit (4, 10).

2 Supplemental Tables

Protein	Replicate	Bound reads	Input reads
Pho4	#1	6,203,398	7,190,687
Pho4	#2	8,704,634	8,330,396
Pho4	#3	54,959,850	24,275,485
Pho4	#4	50,430,034	24,275,485
Cbf1	#1	6,059,502	4,501,445
Cbf1	#2	14,277,898	24,275,485
Cbf1	#3	5,725,019	24,275,485

Table S1: Sequencing read depths.

Replicate set A	Replicate set B	Pearson's r^2
Pho4 #1	Pho4 #2	0.00
Pho4 #1	Pho4 #3	0.01
Pho4 #1	Pho4 #4	0.01
Pho4 #2	Pho4 #3	0.07
Pho4 #2	Pho4 #4	0.08
Pho4 #3	Pho4 #4	0.67
Cbf1 #1	Cbf1 #2	0.02
Cbf1 #1	Cbf1 #3	0.02
Cbf1 #2	Cbf1 #3	0.18

Table S2: Unprocessed binding energy Pearson's correlation.

Sequence	Pho4 K_d, nM	Pho4 std. error, nM	Cbf1 K_d, nM	Cbf1 std. error, nM
AGACC_TCGAG	61	1	139	7
AGACG_TCGAG	66	2	109	5
AGACA_CCGAG	74	2	104	5
AGACA_GCGAG	91	2	193	9
AGAGA_TCGAG	92	2	104	5
AGACA_TCTAG	107	3	168	8
AGACA_TCCAG	117	3	111	5
AGACA_TCAAG	117	3	145	7
AGAAA_TCGAG	120	3	110	5
AGACA_TCGAA	121	3	153	7
AGACA_TCGAC	121	3	155	7
CGACA_TCGAG	121	3	166	8
AGGCA_TCGAG	135	3	182	10
AGACA_TCGTG	135	3	220	10
AGACA_TTGAG	138	4	208	10
AGACA_TCGGG	138	4	213	12
AGACA_TCGAT	139	4	195	9
GGACA_TCGAG	139	3	217	12
AGACA_TCGCG	143	4	184	10
AAACA_TCGAG	150	3	204	13
AGACA_TCGAG	151	3	236	16
ACACA_TCGAG	152	3	222	17
TGACA_TCGAG	153	4	212	13
AGCCA_TCGAG	154	4	250	15
AGACA_ACGAG	162	4	30	2
AGATA_TCGAG	165	4	185	9
AGACA_TGGAG	166	4	387	22
ATACA_TCGAG	170	5	237	15
AGTCA_TCGAG	196	5	217	13
AGACA_TAGAG	197	8	364	24
AGACT_TCGAG	336	8	82	4
negative control	1230	112	2970	825

Table S3: Titration binding results. Standard error on the fit provided. The "_" represents the E-box motif. The negative control sequence contains a weak binding CAGCTG mutated consensus motif.

Protein	Linear model	Pearson's r^2
Pho4	mononucleotide	0.92
Pho4	nearest neighbor	0.98
Pho4	dinucleotide	0.99
Cbf1	mononucleotide	0.70
Cbf1	nearest neighbor	0.94
Cbf1	dinucleotide	0.99

Table S4: Interpretation of NN model features. Linear models trained on NN-derived predictions were used squared Pearson's correlation coefficient to determine the proportion of variance explained by additive combinations of sequence features.

Replicate set A	Replicate set B	Pearson's r^2
Pho4 #1	Pho4 #3	0.97
Pho4 #1	Pho4 #2	0.97
Pho4 #3	Pho4 #4	0.96
Pho4 #2	Pho4 #4	0.96
Pho4 #2	Pho4 #3	0.96
Pho4 #1	Pho4 #4	0.95
Cbf1 #1	Cbf1 #2	0.87
Cbf1 #2	Cbf1 #3	0.85
Cbf1 #1	Cbf1 #3	0.79
Cbf1 #2	Pho4 #4	0.03
Cbf1 #1	Pho4 #3	0.01
Pho4 #2	Cbf1 #3	0.01
Pho4 #1	Cbf1 #3	0.01
Cbf1 #1	Pho4 #4	0.00
Cbf1 #2	Pho4 #3	0.00
Cbf1 #3	Pho4 #4	0.00
Pho4 #1	Cbf1 #2	0.00
Cbf1 #1	Pho4 #2	0.00

Table S5: Cross-correlation of mononucleotide model coefficients from unprocessed replicates. Pair-wise Pearson's r^2 from replicate mononucleotide model coefficient cross-correlation.

Sequence
AGACA CACGTG TCGAG
GGACA CACGTG TCGAG
TGACA CACGTG TCGAG
CGACA CACGTG TCGAG
AAACA CACGTG TCGAG
ATACA CACGTG TCGAG
ACACA CACGTG TCGAG
AGGCA CACGTG TCGAG
AGTCA CACGTG TCGAG
AGCCA CACGTG TCGAG
AGAAA CACGTG TCGAG
AGAGA CACGTG TCGAG
AGATA CACGTG TCGAG
AGACG CACGTG TCGAG
AGACT CACGTG TCGAG
AGACC CACGTG TCGAG
AGACA CACGTG ACGAG
AGACA CACGTG CCGAG
AGACA CACGTG GCGAG
AGACA CACGTG TAGAG
AGACA CACGTG TGGAG
AGACA CACGTG TTGAG
AGACA CACGTG TCAAG
AGACA CACGTG TCTAG
AGACA CACGTG TCCAG
AGACA CACGTG TCGGG
AGACA CACGTG TCGTG
AGACA CACGTG TCGCG
AGACA CACGTG TCGAA
AGACA CACGTG TCGAT
AGACA CACGTG TCGAC
AGACA CAGCTG TCGAG

Table S6: Single nucleotide variant titration binding substrates. 5' constant region = CAATACACTGTTATC. 3' constant region = CTACTCGTTCGGTTATCCGGCGGTATGAC.

Name	Index (5' to 3')
i501	TATAGCCT
i502	ATAGAGGC
i503	CCTATCCT
i504	GGCTCTGA
i505	AGGCGAAG
i506	TAATCTTA
i507	CAGGACGT
i508	GTACTGAC

Table S7: P5 indexed primers. 5' constant region = AATGATACGGCGACCACCGAGATCTACAC. 3' constant region = ACACTCTTTCCCTACACGACGCTCTTCCGATCT.

Name	Index (5' to 3')
i701	CGAGTAAT
i702	TCTCCGGA
i703	AATGAGCG
i704	GGAATCTC
i705	TTCTGAAT
i706	ACGAATTC
i707	AGCTTCAG
i708	GCGCATTA

Table S8: P7 indexed primers. 5' constant region = CAAGCAGAAGACGGCATACGAGAT. 3' constant region = GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT.

3 Supplemental Figures

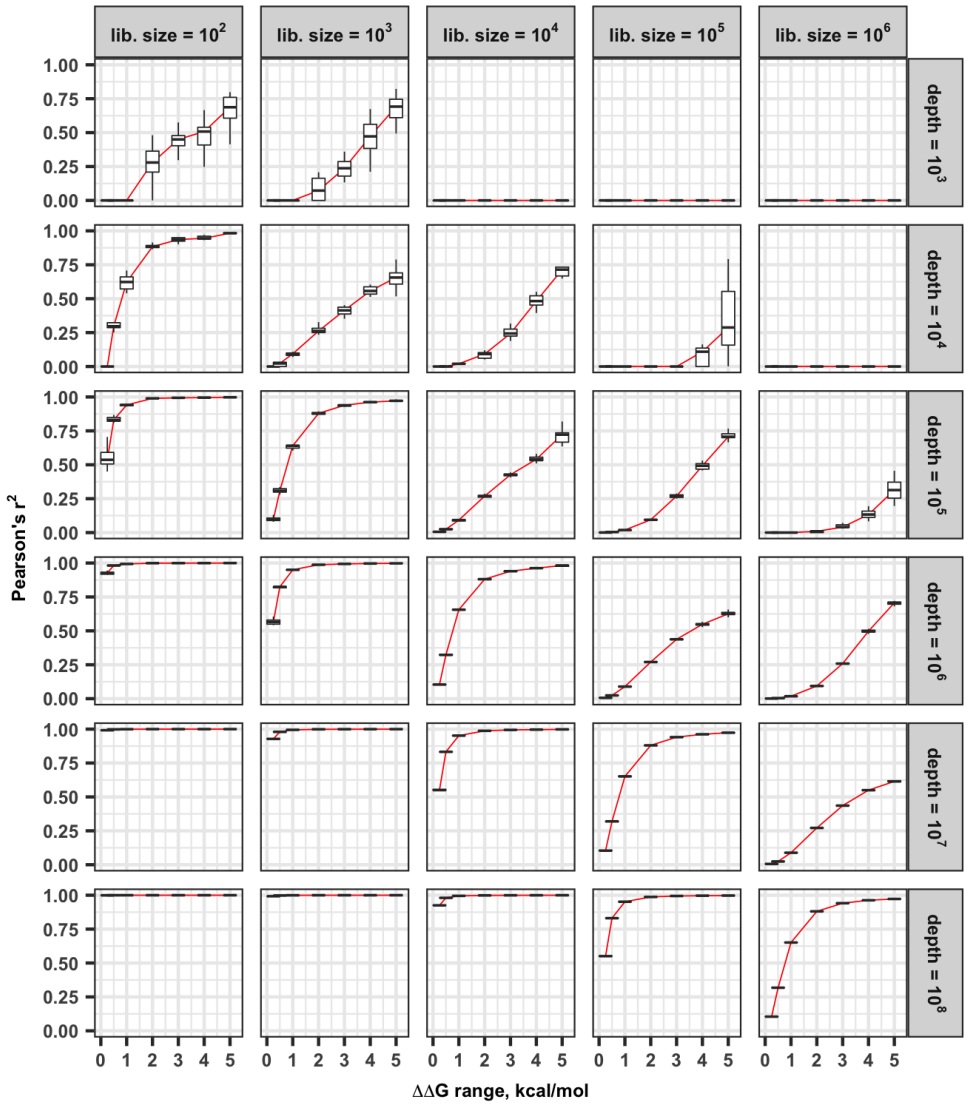

Figure S1: Monte Carlo simulations. Pearson's r^2, which describes the ability to recover theoretical relative binding energies, as a function of binding energy range. Ten simulation replicates shown as box plots [inner line = median, box edges (hinges) = bounds of interquartile range (IQR), upper whisker = extends from upper hinge to largest value less than (1.5 × IQR), lower whisker = extends from lower hinge to smallest value greater than (1.5 × IQR)]. Each set of simulation replicates was derived from a unique combination of assay parameters: library size (x-axis facets) and total sequencing depth (y-axis facets).

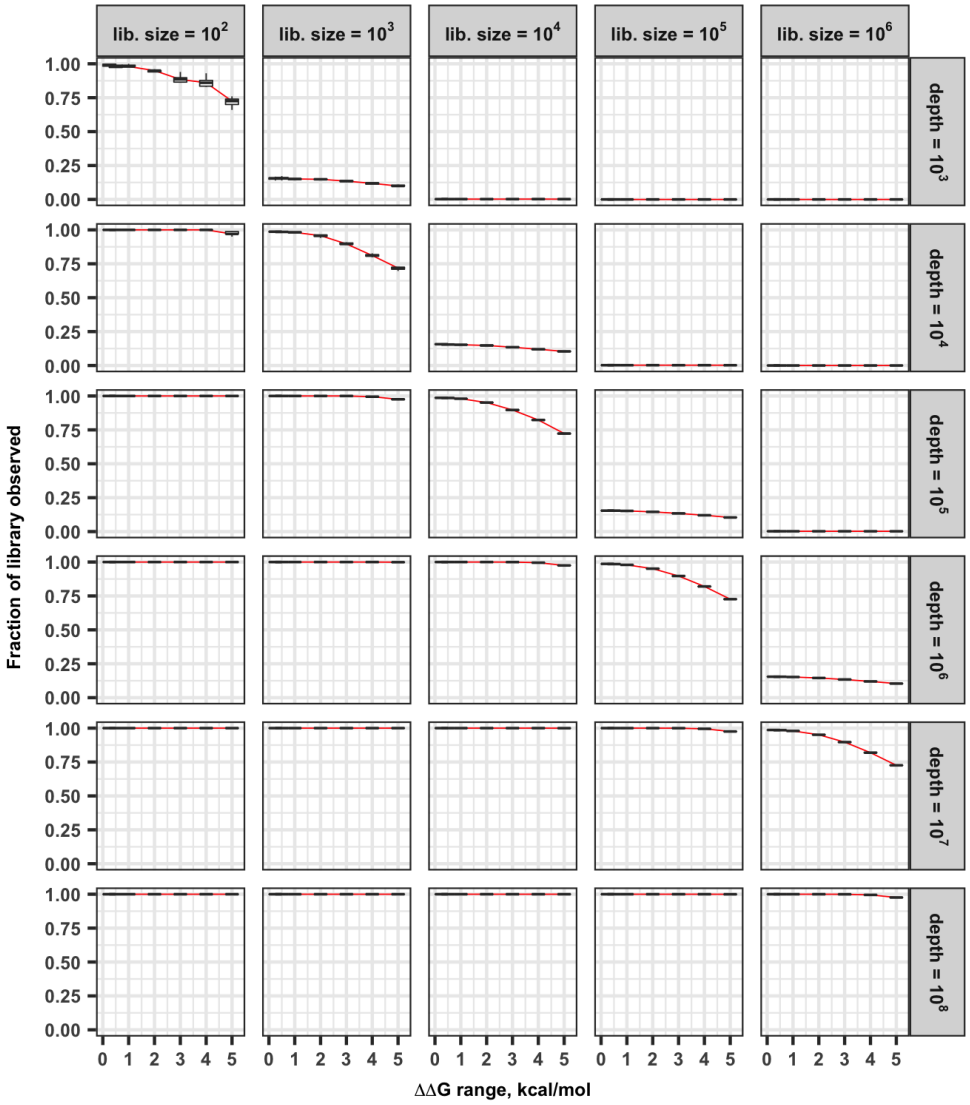

Figure S2: Monte Carlo simulations. Fraction of observed species as a function of binding energy range. Ten simulation replicates shown as box plots [inner line = median, box edges (hinges) = bounds of interquartile range (IQR), upper whisker = extends from upper hinge to largest value less than (1.5 × IQR), lower whisker = extends from lower hinge to smallest value greater than (1.5 × IQR)]. Each set of simulation replicates was derived from a unique combination of assay parameters: library size (x-axis facets) and total sequencing depth (y-axis facets).

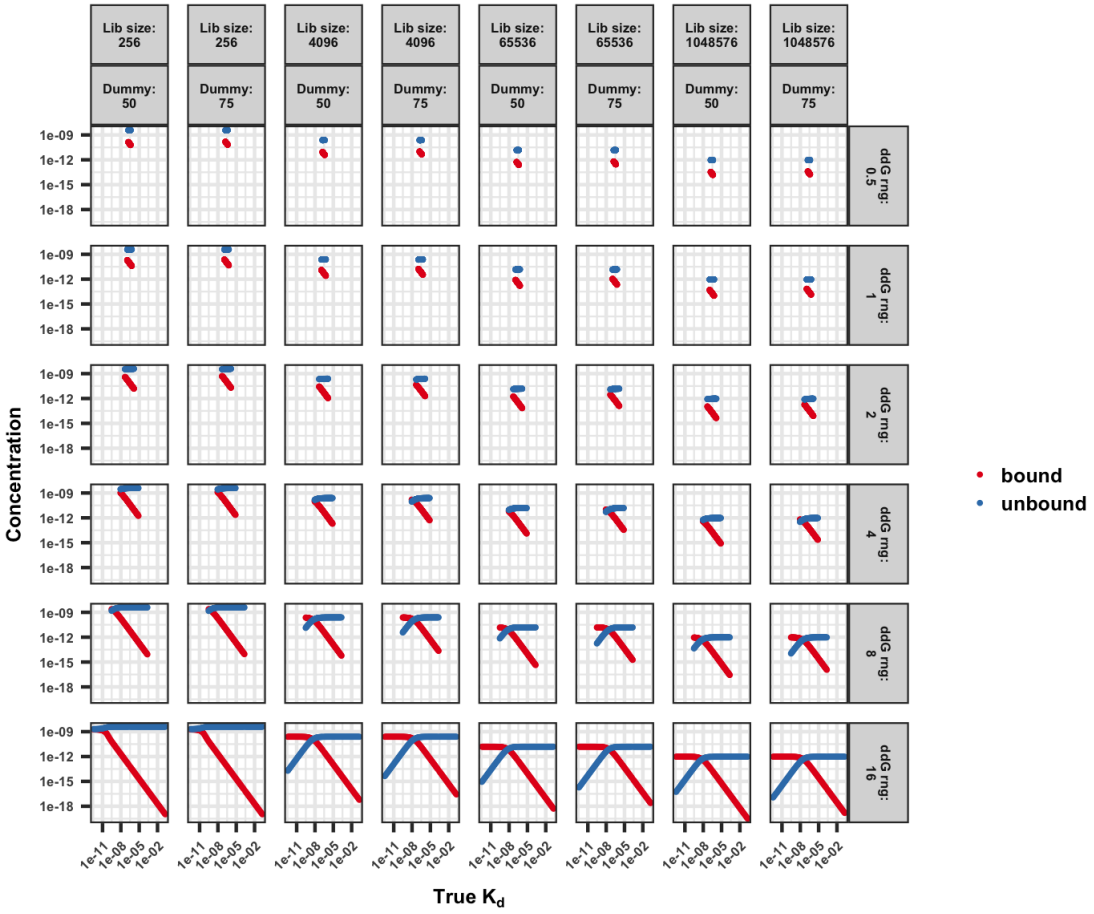

Figure S3: Equilibrium binding simulations. Equilibrium concentrations in both the bound and unbound fraction for each of the simulated species as a function of that species' K_d. As the affinity of the interaction becomes higher (K_d becomes lower), ligands are depleted from the unbound fraction, causing the relative concentrations of the unbound fraction to differ from those of the input library.

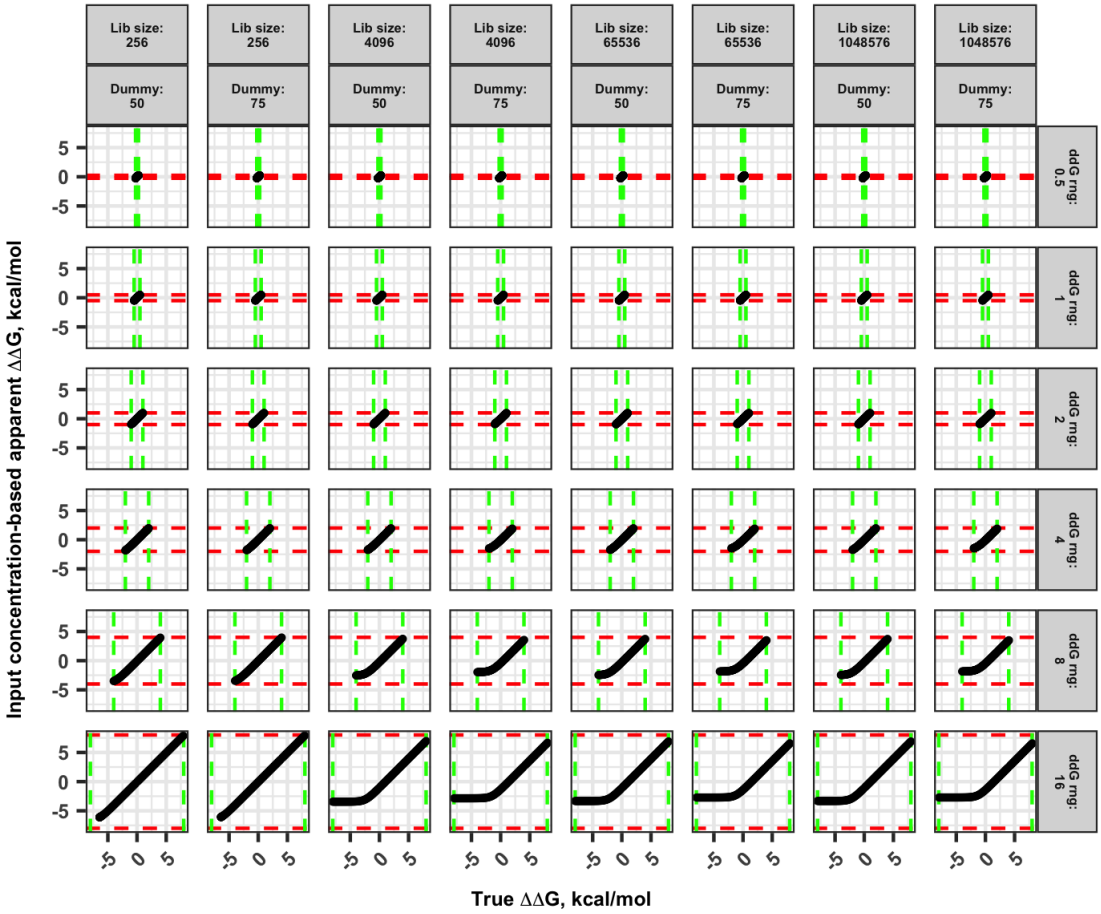

Figure S4: Agreement between true ΔΔG and apparent ΔΔG calculated using $\frac{[bound]}{[input]}$. Agreement between true ΔΔG and apparent ΔΔG values for different library sizes, ΔΔG ranges, and high density positions within the distribution are shown when using $[unbound]$ or $[input]$ in place of the reference concentration in the following equation: $\Delta\Delta G = -RT \ln\left(\frac{[bound]}{[reference]}\right)$. The actual simulated ΔΔG range is shown by the green lines on the x-axis and the red lines on the y-axis.

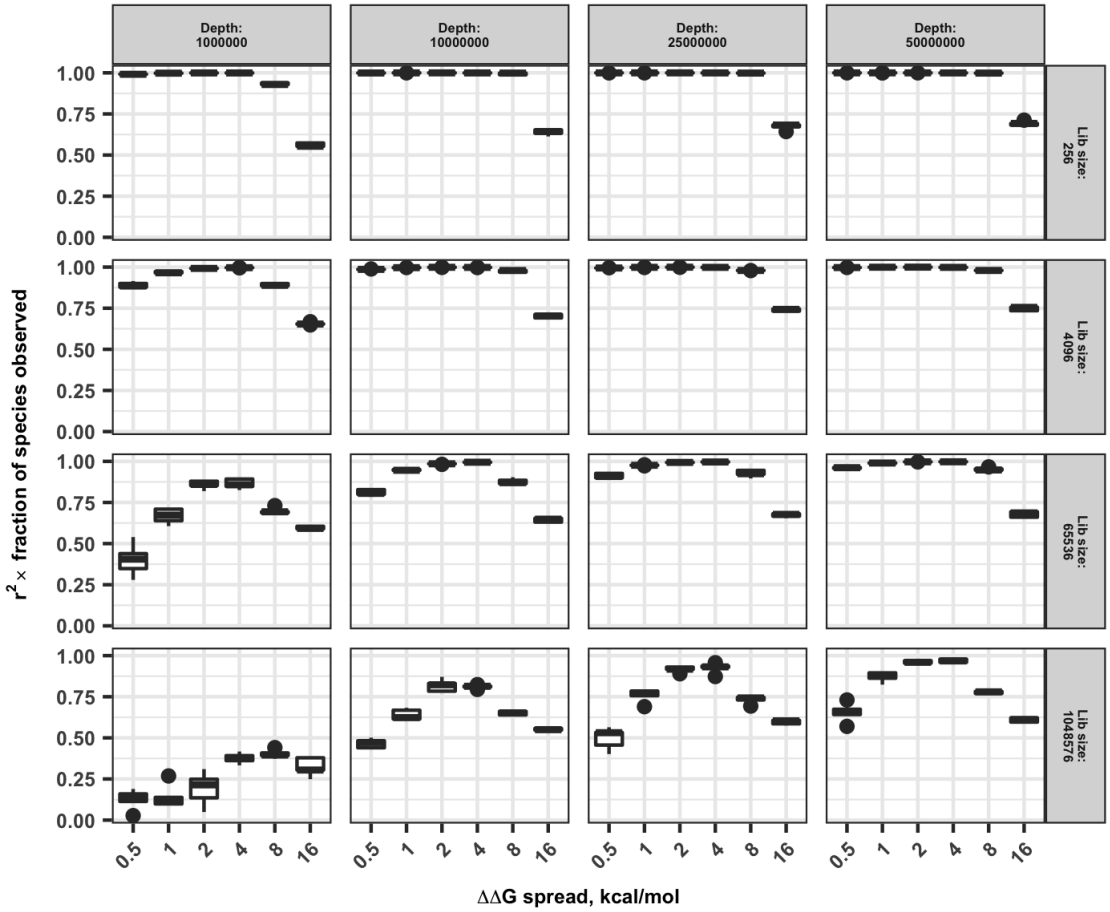

Figure S5: Accuracy of estimating true ΔΔG when approximating unbound concentrations by sequencing the input library as a function of ΔΔG spread, library size, and sequencing depth for 5 replicate simulations. Accuracy ($r^2 \times fraction\ of\ species\ observed$) is shown for different parameterizations of the assay. Generally, assays with highest read depth to library size ratio perform the best. We also note that energetic resolution is highest when the energetic spread is between 2–4 kcal/mol.

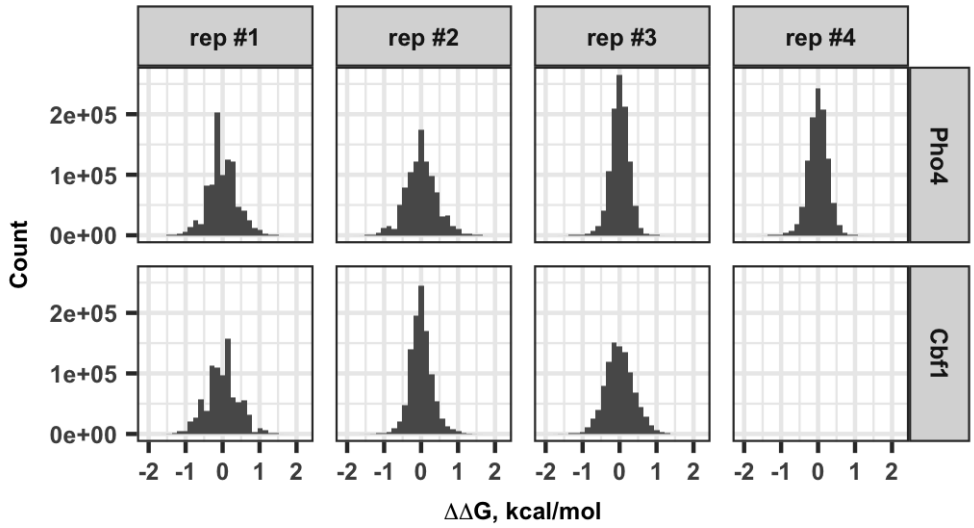

Figure S6: Unprocessed ΔΔG distributions across replicates.

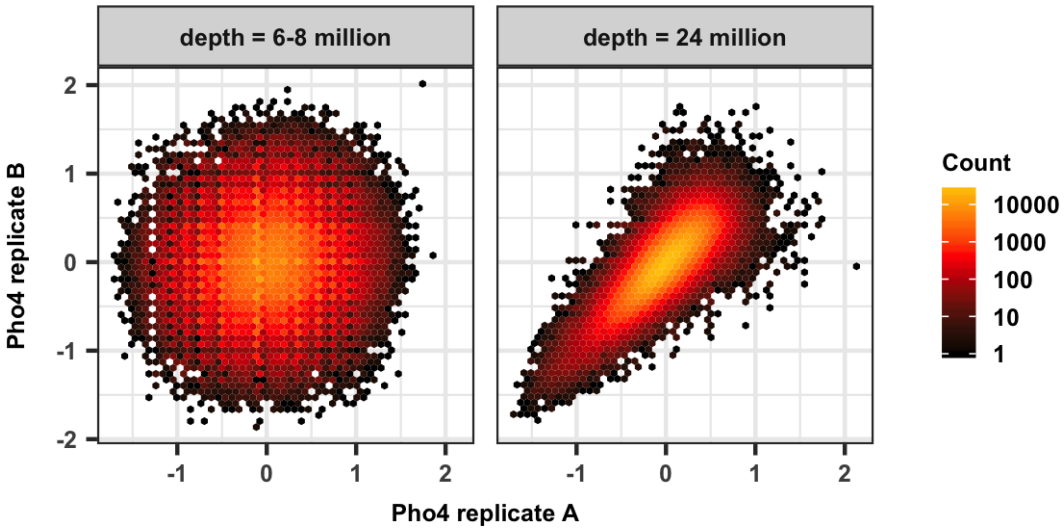

Figure S7: Correspondence of unprocessed ΔΔG measurements as a function of sequencing read depth. (left) Pho4 replicates #1 and #2. (right) Pho4 replicates #3 and #4.

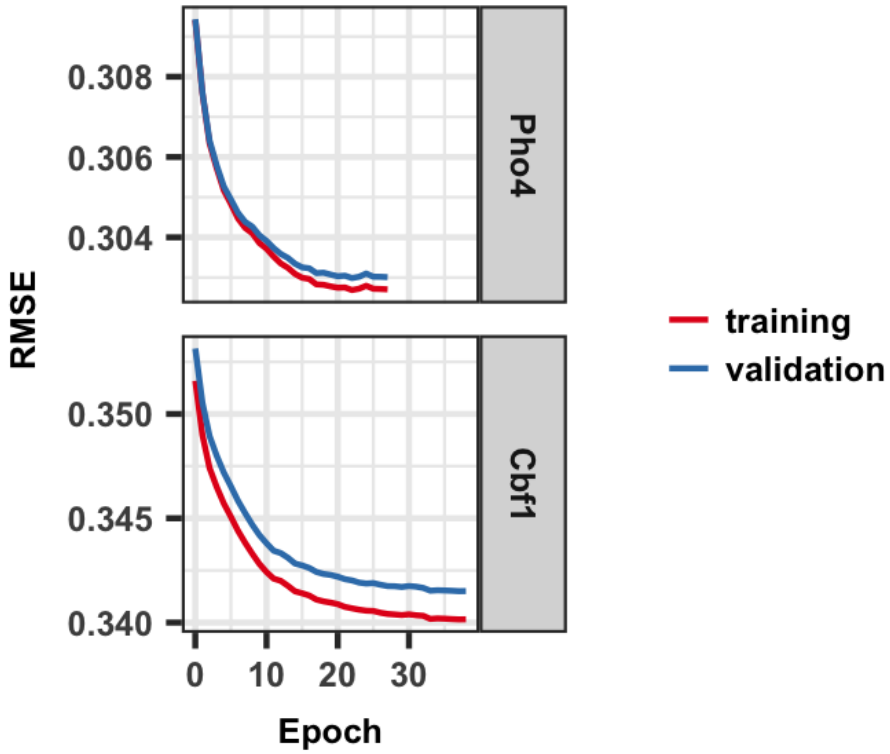

Figure S8: Training and validation dataset error curves for neural network training. Root mean squared error curves are shown for both the neural network training on composite Pho4 data and Cbf1 data. Red line represents error on the training (observed) dataset and blue line represents error on the validation (unobserved) dataset.

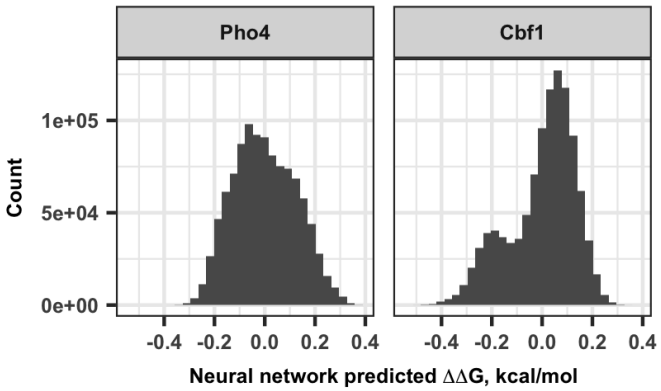

Figure S9: Neural network $\Delta\Delta$G distributions.

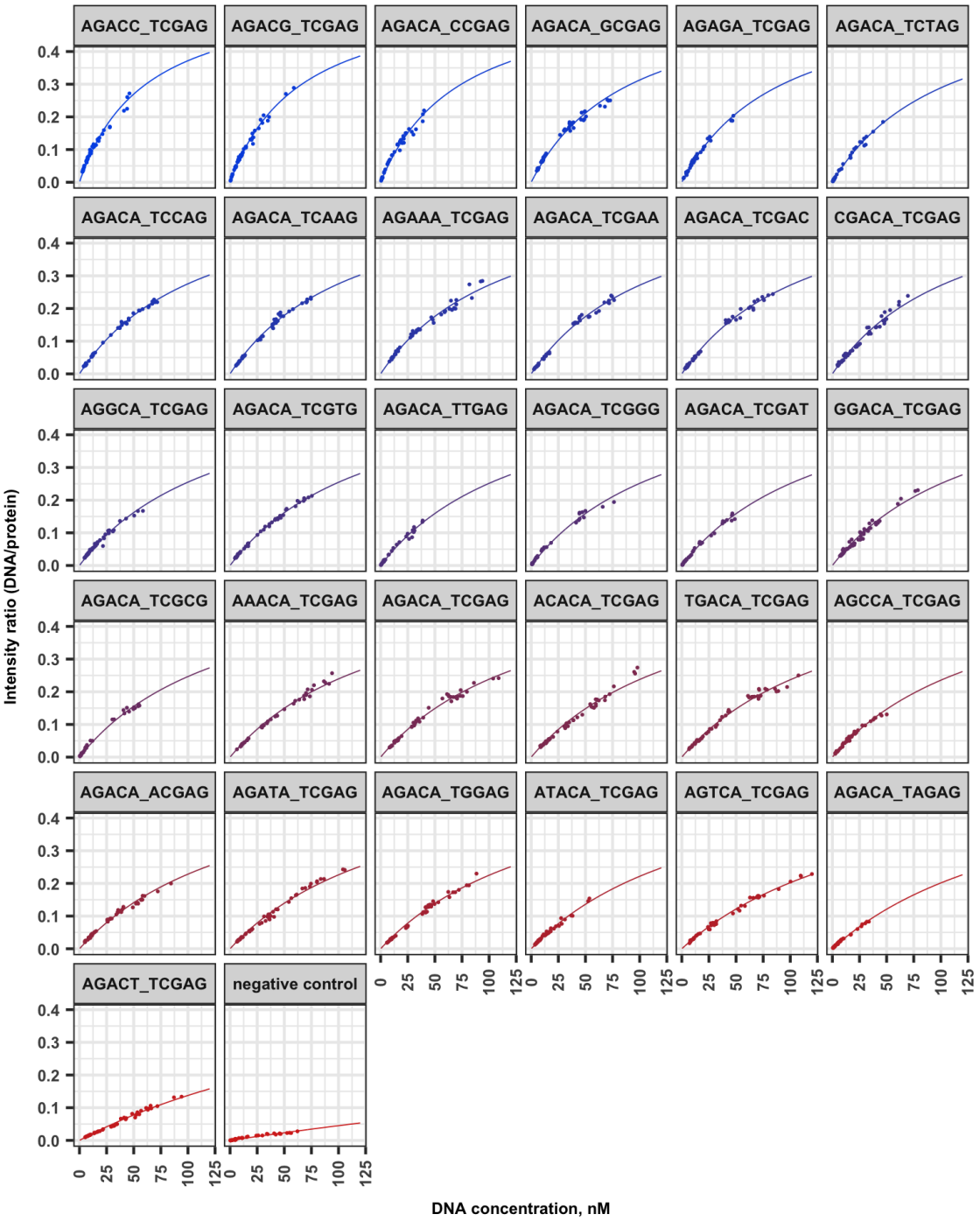

Figure S10: Pho4 titration binding curves. Each point represents binding response as a function of DNA concentration. Global nonlinear fitted line shown, representing binding isotherms. The CACGTG E-box motif is represented by "_". The negative control sequence contains a weak binding CAGCTG mutated consensus motif.

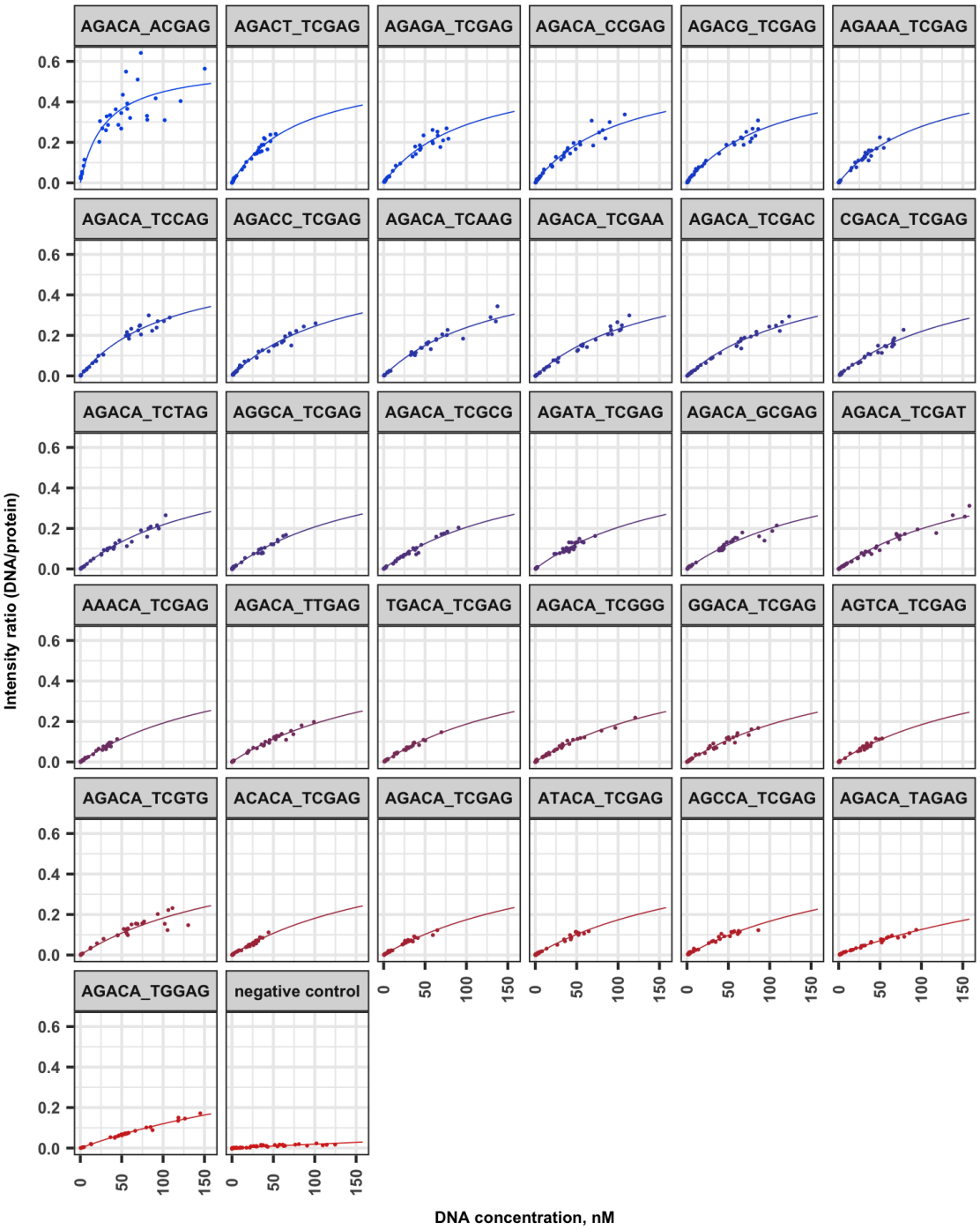

Figure S11: Cbf1 titration binding curves. Each point represents binding response as a function of DNA concentration. Global nonlinear fitted line shown, representing binding isotherms. The CACGTG E-box motif is represented by "_". The negative control sequence contains a weak binding CAGCTG mutated consensus motif.

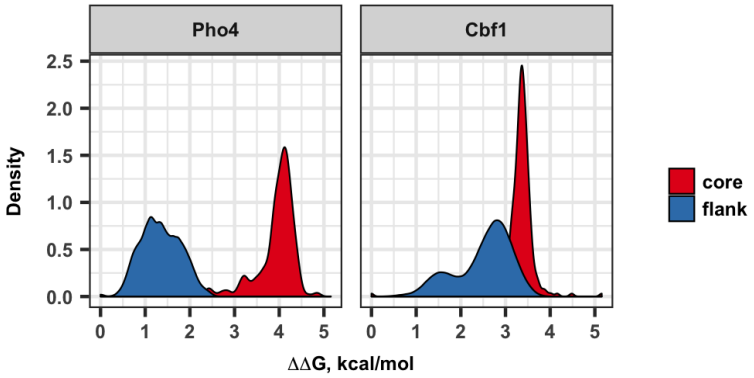

Figure S12: Distributions of $\Delta\Delta G$ measurements between flanking and core sequences. Core binding energies measurements previously reported by Maerkl *et al.* (10).

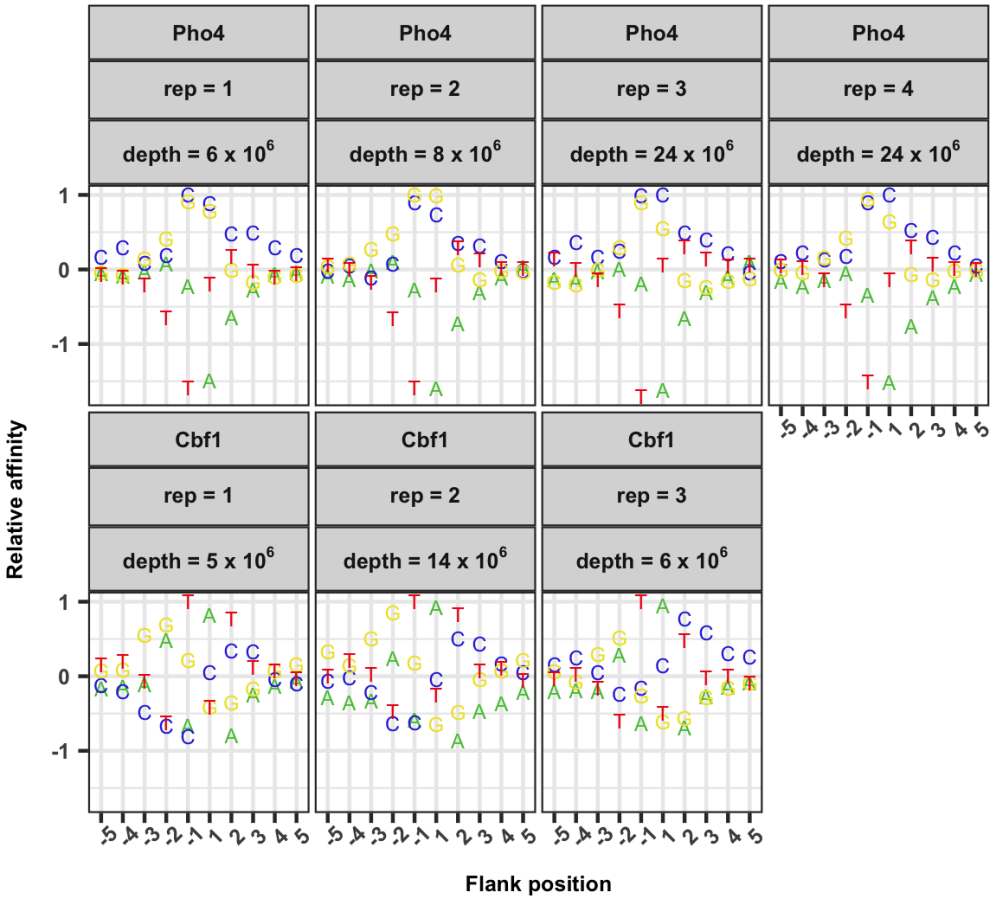

Figure S13: Individual replicate binding preferences. DNA base letters corresponding relative affinity, equal to normalized mean $\Delta\Delta G$, as a function of flanking sequence position. Binding preferences arranged by TF identity, replicate number, and sequencing read depth. The CACGTG motif is located between flank positions -1 and 1.

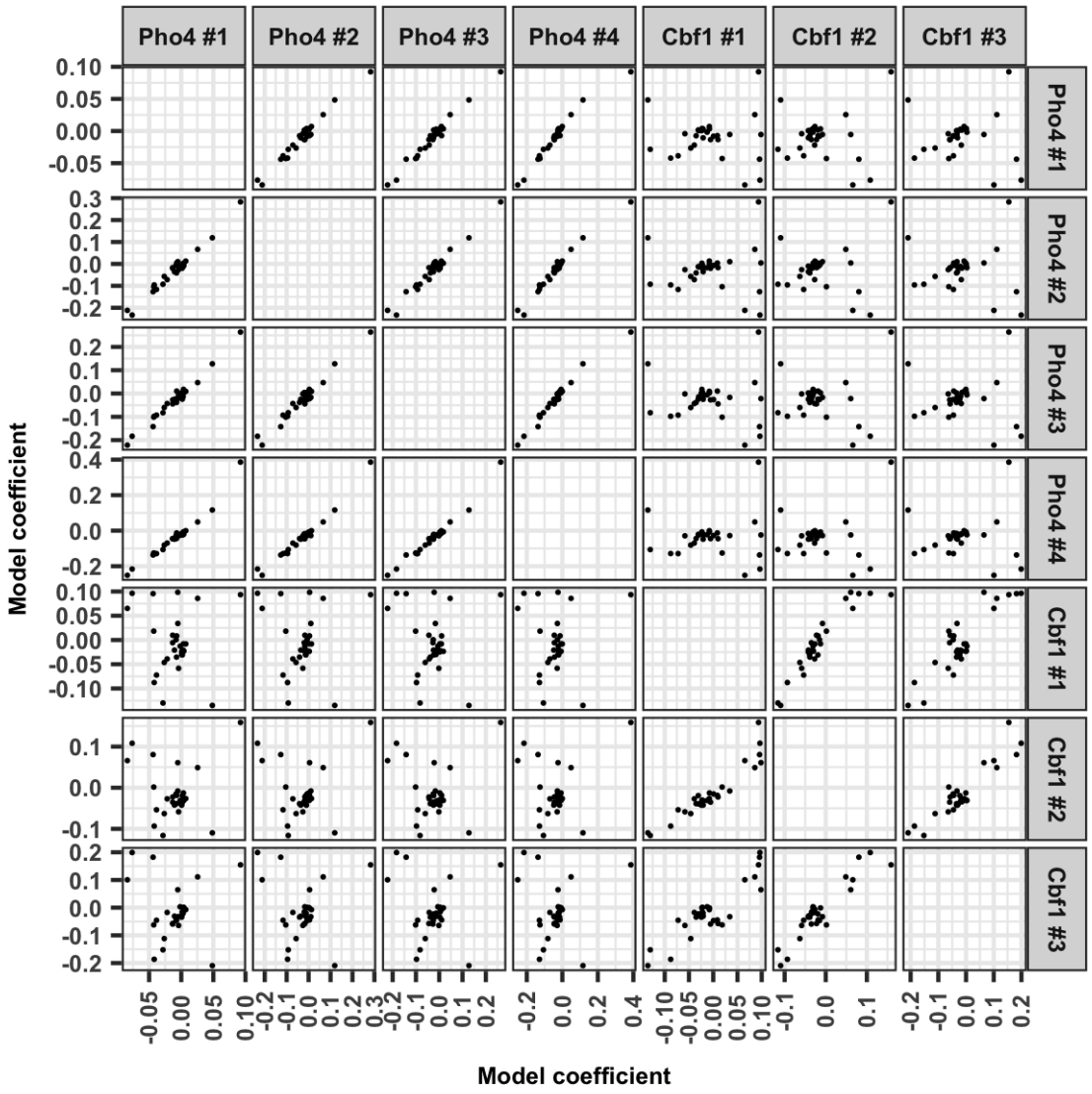

Figure S14: Cross-correlation of replicate mononucleotide model coefficients. Pair-wise comparison of TF replicate mononucleotide model coefficients shown as points.

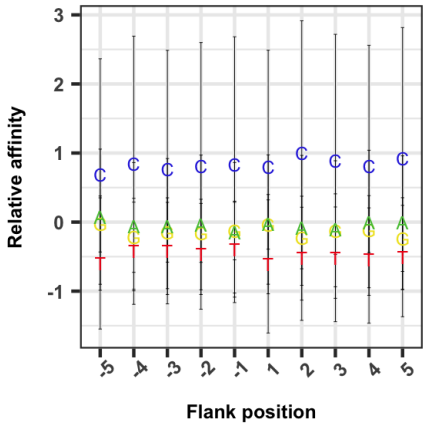

Figure S15: TF-meGFP fusion negative control. DNA base letters corresponding relative affinity, equal to normalized mean $\Delta\Delta G$, as a function of flanking sequence position. The meGFP tag alone (*i.e.* not fused to a corresponding TF) displayed overlapping mononucleotide binding preferences with high variance. The CACGTG motif is located between flank positions -1 and 1. Error bars are standard error on the mean (SEM) among replicates.

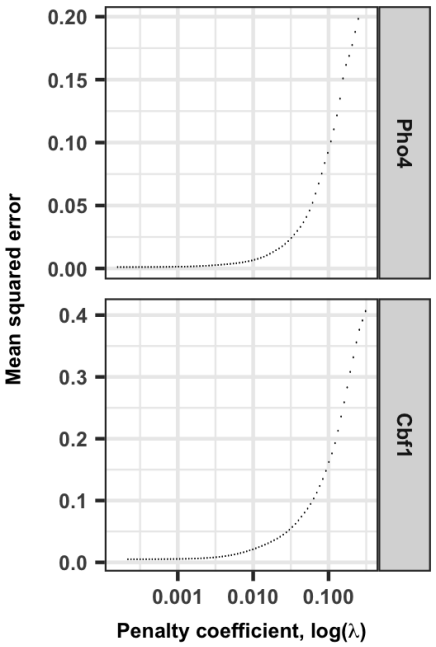

Figure S16: Ten-fold cross-validation of LASSO regression models. The range of mean squared error represented by bracketed vertical lines as a function of penalty coefficient.

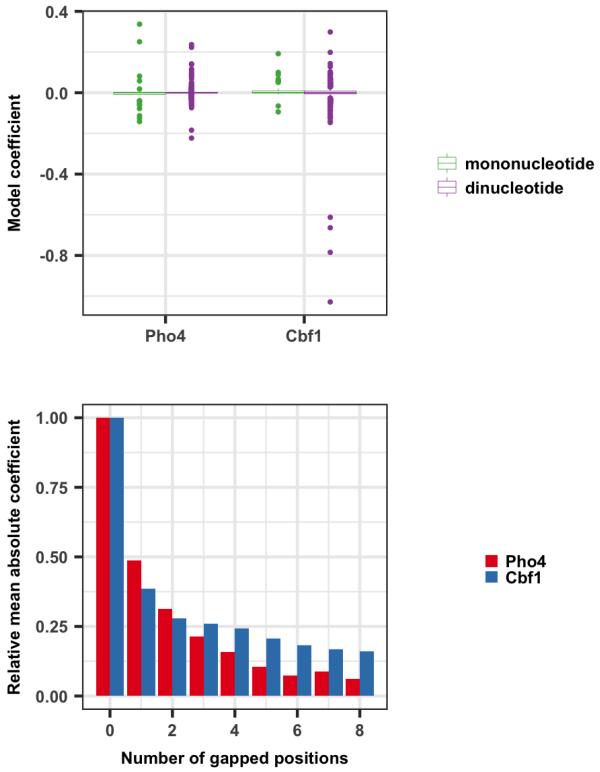

Figure S17: LASSO regression model coefficient magnitudes. (Top) Boxplots of mononucleotide and dinucleotide model coefficients. (Bottom) Bars represent mean absolute model coefficient values as a function of number of gapped positions in a given dinucleotide feature.

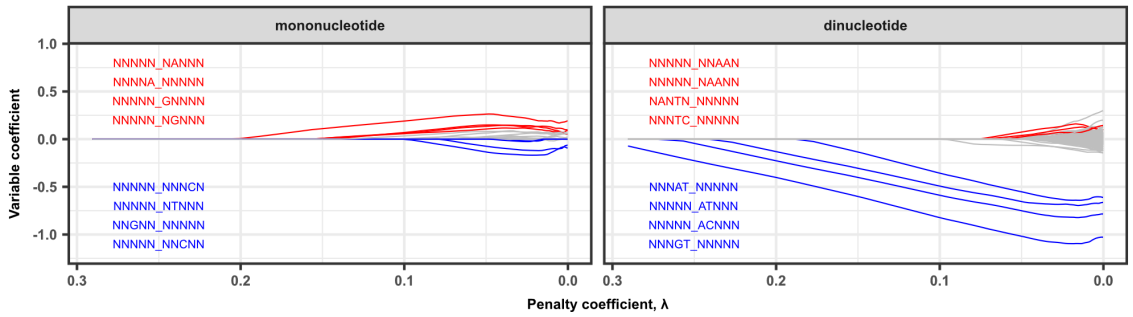

Figure S18: Cbf1 magnitude of individual sequence feature coefficients as a function of penalty coefficient λ during LASSO regression for different flanking sequence features; blue and red indicate negative and positive energetic contributions, respectively.

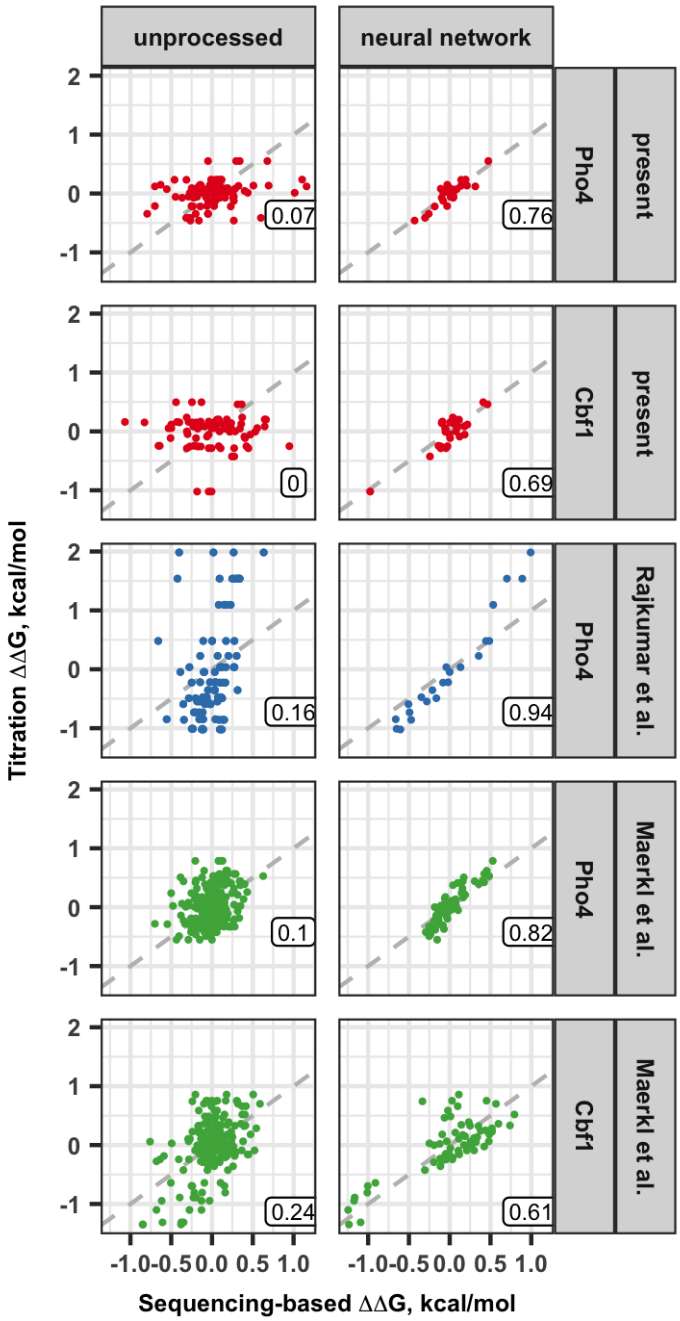

Figure S19: Correlation of sequencing-based results compared to titration measurements. Shown are both unprocessed and NN-predicted ΔΔG estimate compared to titration measurements: substrates presented in this present work and previously reported values (10, 11). All values have been recentered to account for reduced library size within the neural network ΔΔG estimates.

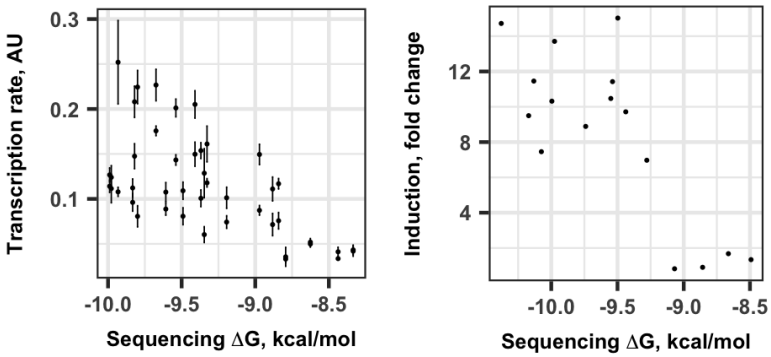

Figure S20: Correlation of sequencing-based neural network model results compared to *in vivo* activity. (Left) Points represent transcription rate with associated standard error on the fit as a function of sequencing-based results (11). (Right) Points represent reporter induction as a function of sequencing-based results (12). The standard deviation of sequencing-based ΔG values for matched substrate species shown as horizontal lines.

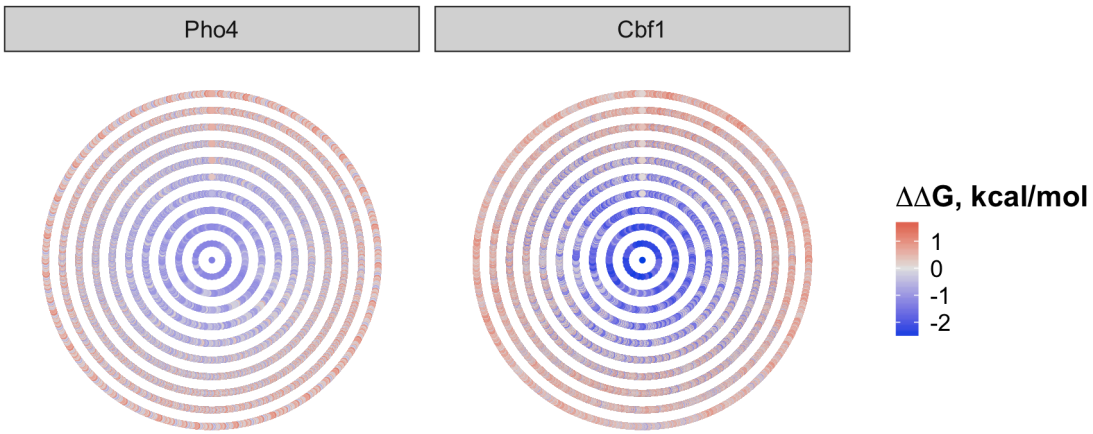

Figure S21: Sequencing-based neural network ΔΔG landscapes. Points represent individual ΔΔG values as a function of Hamming distance from the highest affinity sequence. Sequences are arranged alphabetically and clockwise from the top. Color code: blue = high affinity, red = low affinity.

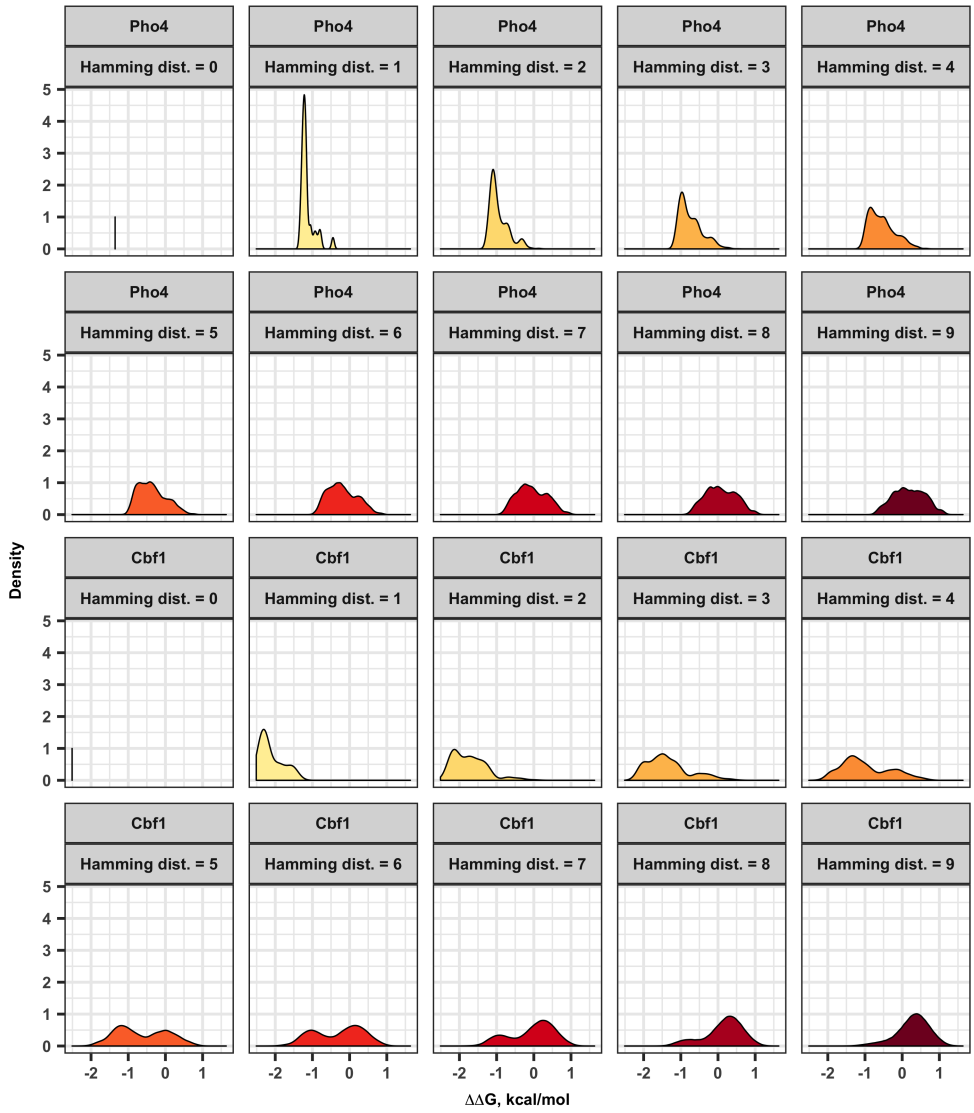

Figure S22: Sequencing-based neural network ΔΔG landscapes summary. Distributions of individual ΔΔG values for each Hamming distance population of sequences, relative to the highest affinity sequence.

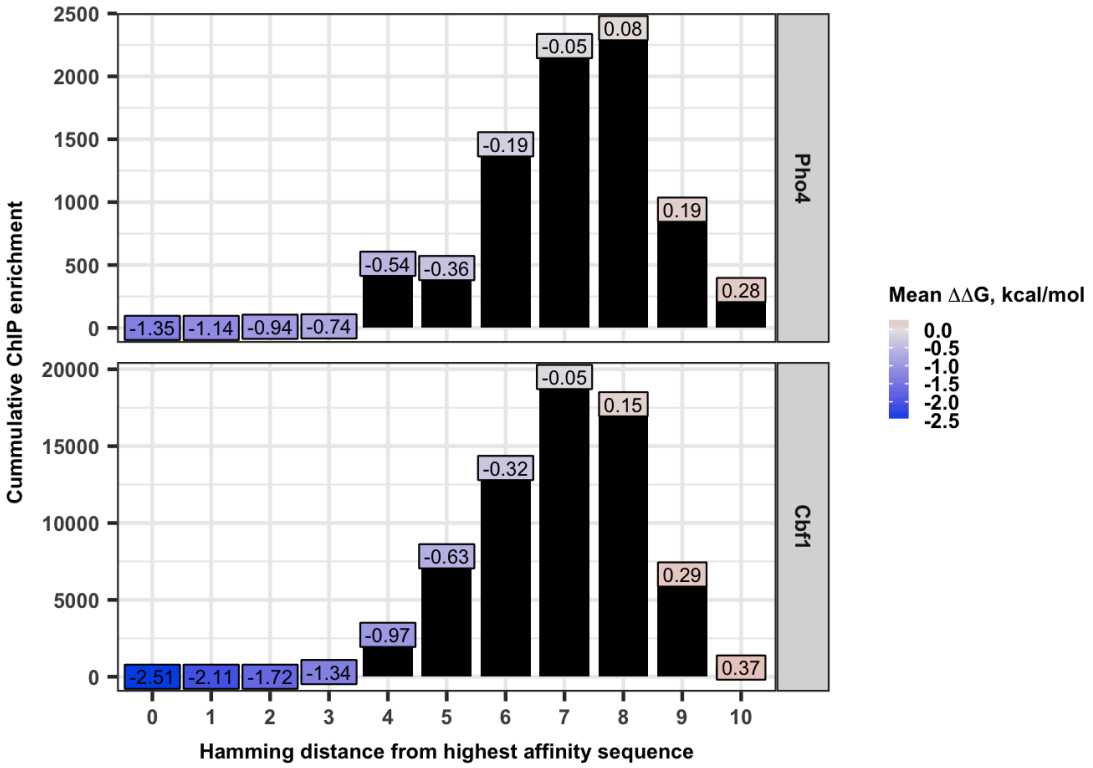

Figure S23: Summary of sequencing-based neural network model results compared to ChIP-seq enrichment. Bars represent the sum of ChIP enrichment as a function of Hamming distance from highest affinity sequence (13). The corresponding mean $\Delta\Delta G$ values shown in color-coded labels: blue = high affinity, red = low affinity.

References

[1] Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. *Genome research* 13(11):2381–2390.

[2] Zhao Y, Granas D, Stormo GD (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5(12):e1000590.

[3] Zuo Z, Stormo GD (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics* 198(3):1329–1343.

[4] Fordyce PM, et al. (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature biotechnology* 28(9):970–975.

[5] Zacharias DA, Violin JD, Newton AC, Tsien RY (2002) Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. *Science* 296(5569):913–916.

[6] Brower K, Puccinelli R, Markin C, Shimko T, Longwell S (2017) An Open-Source, Programmable Pneumatic Setup for Operation and Automated Control of Single-and Multi-Layer Microfluidic Devices. *bioRxiv*.

[7] Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)* 30(5):614–620.

[8] Model MA, Burkhardt JK (2001) A standard for calibration and shading correction of a fluorescence microscope. *Cytometry* 44(4):309–316.

[9] Preibisch S, Saalfeld S, Tomancak P (2009) Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics (Oxford, England)* 25(11):1463–1465.

[10] Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315(5809):233–237.

[11] Rajkumar AS, Dénervaud N, Maerkl SJ (2013) Mapping the fine structure of a eukaryotic promoter input-output function. *Nature Genetics* 45(10):1207–1215.

[12] Aow JSZ, et al. (2013) Differential binding of the related transcription factors Pho4 and Cbf1 can tune the sensitivity of promoters to different levels of an induction signal. *Nucleic Acids Research* 41(9):4877–4887.

[13] Zhou X, O'Shea EK (2011) Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Molecular Cell* 42(6):826–836.