iScience, Volume *24*

Supplemental information

Activation of gene expression by detergent-like protein domains

Bradley K. Broyles, Andrew T. Gutierrez, Theodore P. Maris, Daniel A. Coil, Thomas M. Wagner, Xiao Wang, Daisuke Kihara, Caleb A. Class, and Alexandre M. Erkine

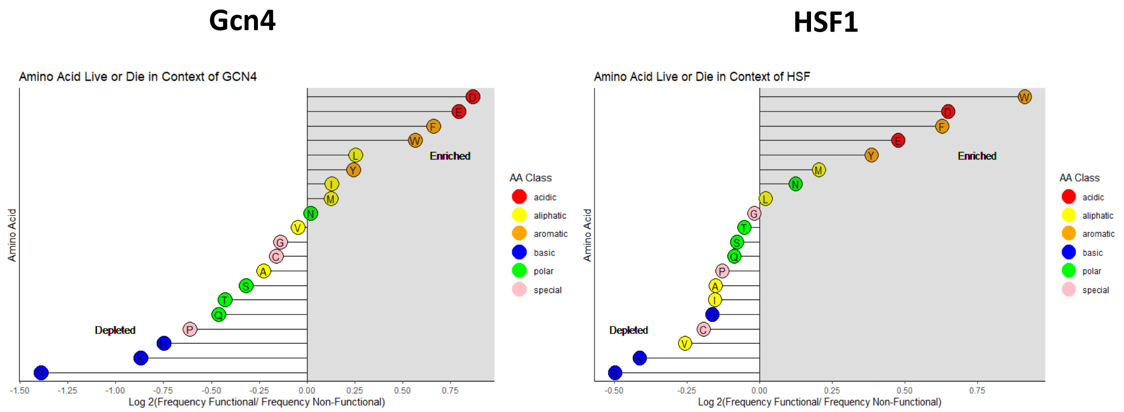


Figure. S1. The profile of frequencies for individual amino acids is looking strikingly similar for both Gcn4 and HSF1 contexts. Related to Figure 2. X-axis: Log2 of frequency in functional to frequency in non-functional sub-pools. Y-axis: individual amino acids.

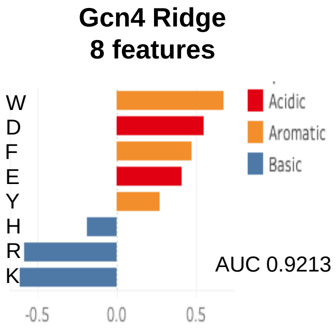
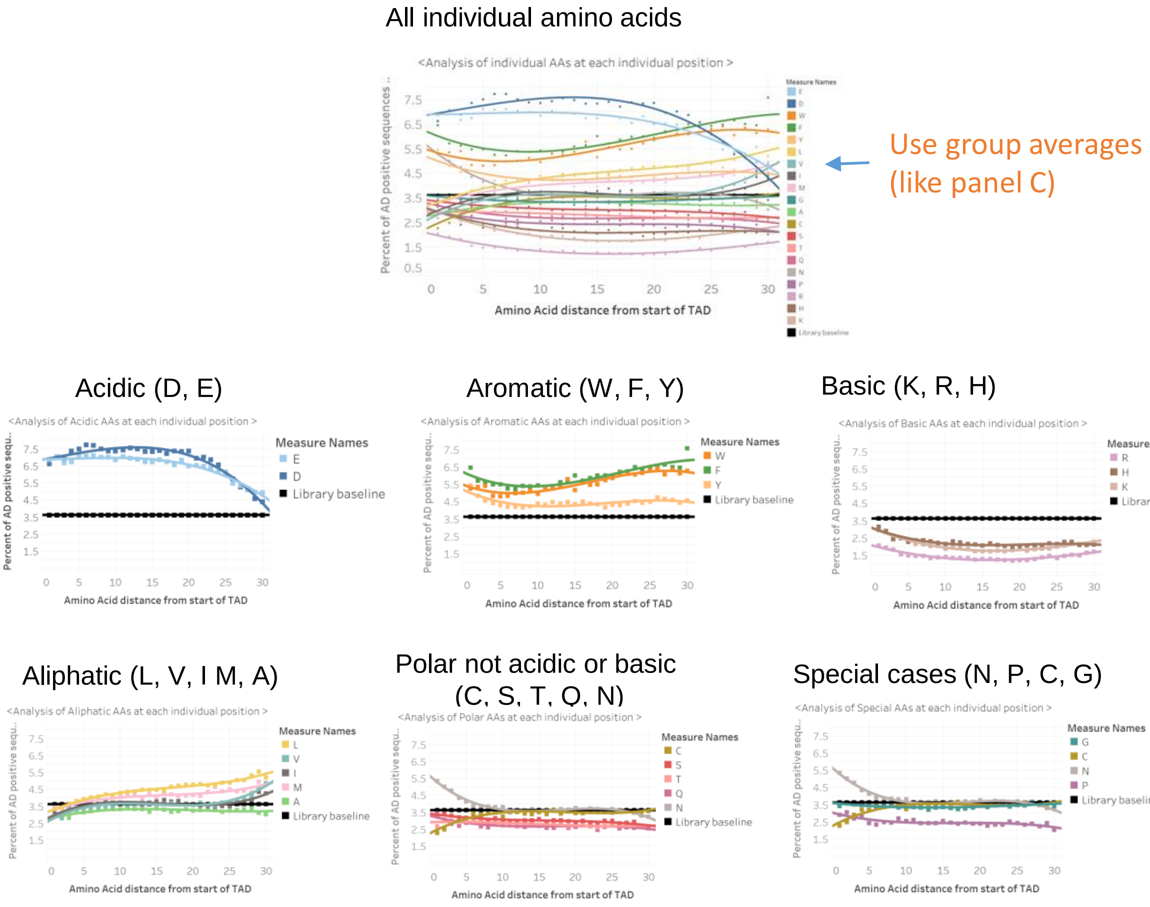


Figure S2. ML coefficients of eight features determining the precision of the Ridge regression model. Related to figure 2. Y-axis: amino acids as ML features. X-axis: ML feature coefficients values.



All individual amino acids

Use group averages (like panel C)

Acidic (D, E)

Aromatic (W, F, Y)

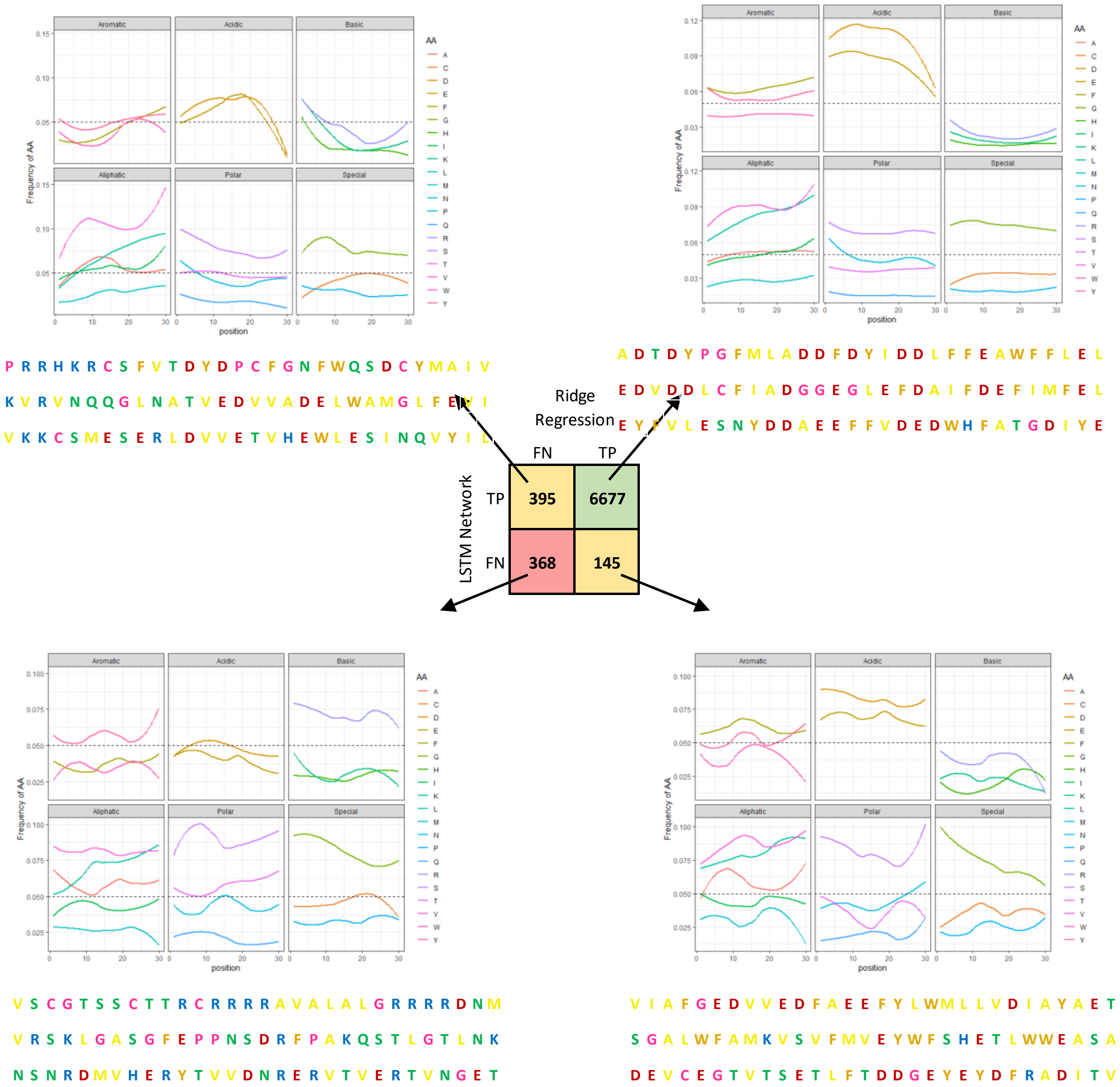
Basic (K, R, H)

Aliphatic (L, V, I M, A)

Polar not acidic or basic (C, S, T, Q, N)

Special cases (N, P, C, G)

Figure S3. Probability of contribution to functionality for individual amino acids as a function of position in tAD. Related to Figure 6. Data for indicated amino acids and amino acid groups are shown. Y-axis: Percent of functional sequences within the pool. X-axis: position of the amino acid with the 30 amino acid stretch tested (C-terminus is position 30).



P R R H K R C S F V T D Y D P C F G N F W Q S D C Y M A I V

K V R V N Q Q G L N A T V E D V V A D E L W A M G L F E V I

V K K C S M E S E R L D V V E T V H E W L E S I N Q V Y I L

A D T D Y P G F M L A D D F D Y I D D L F F E A W F F L E L

E D V D D L C F I A D G G E G L E F D A I F D E F I M F E L

E Y F V L E S N Y D D A E E F F V D E D W H F A T G D I Y E

	FN	TP
TP	395	6677
FN	368	145

V S C G T S S C T T R C R R R R A V A L A L G R R R R D N M

V R S K L G A S G F E P P N S D R F P A K Q S T L G T L N K

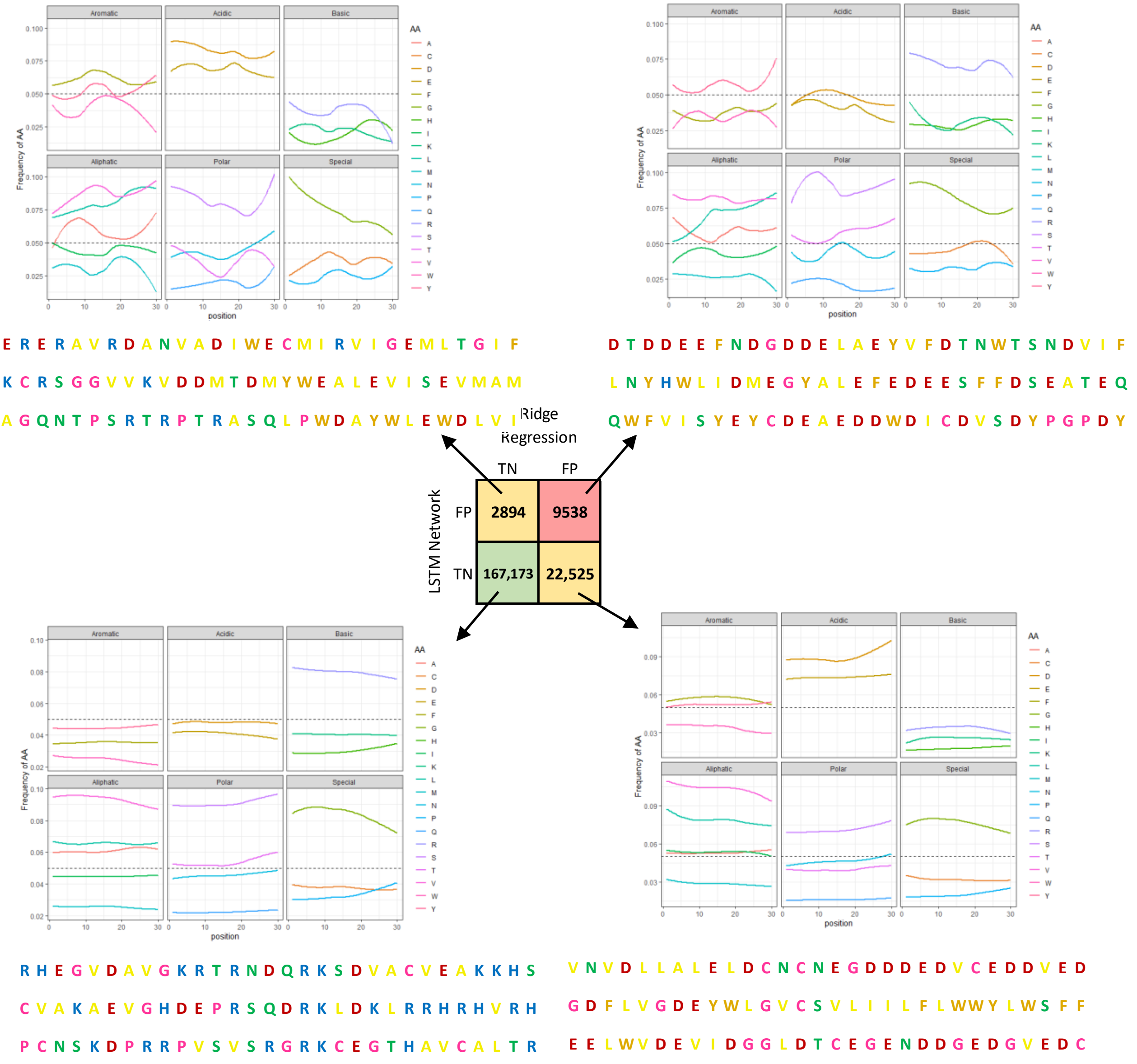
N S N R D M V H E R Y T V V D N R E R V T V E R T V N G E T

V I A F G E D V V E D F A E E F Y L W M L L V D I A Y A E T

S G A L W F A M K V S V F M V E Y W F S H E T L W W E A S A

D E V C E G T V T S E T L F T D D G E Y E Y D F R A D I T V

Figure S4. Comparisons of Ridge Regression and LSTM Network predictions for True Functional sequences. Related to Figure 7. Matrix in center: Number of true functional sequences identified correctly (TP = True Positive) or incorrectly (FN = False Negative) by LSTM network and ridge regression. A consensus plot is provided for the sequences falling in each of the four boxes, along with three example sequences. LSTM network analysis performs better in identifying functional sequences, by identifying more complicated factors (such as locations of relevant amino acids) that contribute to functionality. Sequences mis-identified by both methods could demonstrate unique patterns of functional sequences, or simply experimental error.



	TN	FP
FP	2894	9538
TN	167,173	22,525

Figure S5. Comparisons of Ridge Regression and LSTM Network predictions for True Non-Functional sequences. Related to Figure 7. Matrix in center: Number of true non-functional sequences identified correctly (TN = True Negative) or incorrectly (FP = False Positive) by LSTM network and ridge regression. A consensus plot is provided for the sequences falling in each of the four boxes, along with three example sequences. LSTM network analysis performs better in identifying non-functional sequences, by identifying more complicated factors (such as combinations of aromatic & acidic amino acids) that contribute to functionality.