SUPPLEMENTARY DATA
TABLES

Supplementary Table 1 – Introns Selected for Branch Point Mapping

Gene	PPT Score (bits)	Initial SF1 Binding Model Score (bits)	BP A - PPT Distance (bp)	Intron
PPS1	11.528	8.254	15	6 of 6
HMGN3	14.499	8.254	14	5 of 5
MSN	11.734	8.315	10	11 of 12
SNRNPF	6.803	8.472	9	2 of 3
CDC20	7.566	8.399	12	8 of 10
ZMPSTE24	6.182	8.187	14	8 of 9
AP2M1	4.227	7.383	6	11 of 11

Supplementary Table 2 – Sequences Assayed for SF1 Binding

Sequence	Initial SF1 Binding Model Score (bits)	Initial SF1 Binding Model Rank	SF1 Affinity Model Score (bits)	SF1 Affinity Model Rank
Highest Scoring		**Score Rank**		**Score Rank**
UACUAAC	9.4952	1	9.4055	1
UACUGAC	9.2321	2	8.7037	2
UAGUAAC	8.0803	3	5.8523	37
UAGUGAC	7.8172	4	5.1505	69
UACUAAU	7.4953	5	7.4166	9
CACUAAC	7.4953	5	7.4543	8
GACUAAC	7.4953	5	8.2371	3
UACUAAG	7.4953	5	5.8483	38
UGCUAAC	7.3254	6	8.194	4
CACUGAC	7.2322	7	6.7525	16
Very Low Scoring		**Score Rank**		**Score Rank**
ACUACCA	-11.9028	263	-8.7002	13835
AUAACCA	-11.9028	263	-7.9673	13002
ACAGUCA	-11.9028	263	-8.3703	13487
ACGGCGG	-9.3178	257	-10.6111	15167
ACGGUGG	-9.3178	257	-11.2399	15379
CGAACAG	-5.581	222	-4.3369	7275
GUCUUGA	-4.5811	208	-3.109	5299
UGAGUAG	-3.5811	193	-3.3012	5575
High Frequency		**Frequency Rank**		**Frequency Rank**
UGCUGAC	7.0623	1	7.4922	1
UCCUGAC	6.0623	2	6.4198	2
CACUGAC	7.2322	3	6.7525	3
CAGUGAC	5.8173	5	3.1993	61*
UACUAAU	7.4953	6	7.4166	15
UUCUAAC	6.3254	7	6.6091	13
UAUUAAC	6.4953	8	5.7244	17
UACUGAU	7.2322	10	6.7148	18
UGCUAAC	7.3254	11	8.194	19
UAAUAAC	6.4953	12	6.9698	20
UACUAAA	6.4953	13	7.0303	27
GGCUGAC	5.0624	38*	6.3238	6
UGCUGAU	5.0624	54*	5.5033	11
Unique Point Mutants		**Score Rank**		**Score Rank**
UACUUAC	6.9103	9	6.9341	15
UACUCAC	6.9103	9	7.5629	5
AACUAAC	6.4953	10	7.3598	10
UCCUAAC	6.3254	11	7.1216	11
UACUACC	6.1734	13	6.2725	27
UACUAUC	6.1734	13	4.7287	97
UACCAAC	6.1734	13	5.1896	63
UACAAAC	6.1734	13	6.6613	18
UACUAGC	6.1734	13	5.7024	45
UACGAAC	6.1734	13	6.3746	23

Disease BPS		**Score Rank**		**Score Rank**
TH				
GGCUGAU	3.0625	51	4.3349	138
GGCAGAU	-0.2593	119	1.5907	782
LCAT				
CCCUGAC	4.0624	36	4.4686	124
CCCCGAC	0.7406	97	0.2527	1533
ITGB4				
GGCUCAC	2.7406	60	5.183	65
GGCACAC	-0.5811	129	2.4388	482
COL5A1				
GACUGAU	5.2323	22	5.5464	50
GACGGAU	1.9105	74	2.5155	464

Supplementary Table 3 – Intronic regions selected for Tat-hybrid intron context SF1 binding assay

Gene	Intron	SF1 Affinity Model BPS	SF1 Affinity Model Score (bits)	Literature BPS Model BPS	Literature BPS Model Score (bits)	PPT Sequence	PPT Score (bits)	SF1 Affinity Model BP A - PPT Distance
High BPS - High PPT								
CUL4B	10	UACUAAC	14.220	UACUAAC	5.207	UUUUUUUCCUU	14.499	7
TPK1	8	UACUAAU	13.287	UACUAAU	4.943	UUUUUUCCCUU	15.236	5
NUP133	18	UGCUAAC	11.931	UGCUAAC	5.621	UUUUUUCCCAU	13.166	6
PSMD7	2	AACUAAC	10.832	AACUAAC	4.792	UUUUUUUCCUC	13.138	2
DTX2	4	GACUAAC	10.762	GACUAAC	5.207	UUGUUUUCCUU	11.799	4
RNF41	3	UACUGAC	8.704	UACUGAC	5.470	UUUUUUUCCUG	12.330	5
STK11	6	GACUGAC	7.535	GACUGAC	5.470	UUUCUUCCCUC	13.488	8
DOK5	6	UGCUGAC	7.492	UGCUGAC	5.885	UUUCUUCCCUU	14.849	9
HINT3	1	CACUAAC	7.454	CACUAAC	4.792	GUUUUUUCCUU	12.848	6
ANKRD13C	10	UCCUAAC	7.122	UCCUAAC	5.621	UUUUUUUCCUC	13.138	4
Low BPS - Low PPT								
TTLL10	10	GGCUCAG	1.626	UCCUGGU	2.162	GCUCUCUGCAG	-0.717	7
HPCAL4	2	GGCUGUC	1.647	CCCUGAG	2.885	CUGCGCCCCAA	0.784	3
LHX3	6	AGCUCAA	1.931	GCUCAAU	2.773	UCUAAGCCCCU	0.961	5
YL01_HUMAN	3	CGCUGAG	1.984	AGCUGAG	2.885	UCUCGUUUUUG	1.023	9
STCBP2	4	GAGUGAU	1.993	GAGUGAU	3.622	CACCUUCCCCA	0.907	3
Low BPS - High PPT								
CECR1	1	UCUUCAC	1.598	UCUUCAC	3.037	CUUCUUCCCUC	11.029	6
TAP2	6	GCCUCAU	2.122	GCCUCAU	3.037	UUUCUUCUCCU	10.025	5
IQCC	1	UGAUAAG	2.201	CCGUGAU	3.622	CUUCUUUCCUC	10.292	5
MOCS1	10	GACUUAG	2.209	UCUUCAU	2.773	UUUCUUCUCUU	12.417	4
JTB	1	UUUUGAC	2.226	UUUUGAC	5.207	UUUUGUCCCUU	12.362	2
MLC1	9	UGCAACC	2.317	CCCUGGC	2.010	UUUUUUUCCAG	10.260	7
CJ053_HUMAN	2	UGCUGCU	2.370	UCUUGAU	5.358	UUUCUUCUCUG	10.248	1
NFXL1	2	UGCAGAA	2.373	AGGUGAC	3.885	UUUUUUCUUUU	9.997	9
MTA2	7	UUCUCAA	2.391	UUUUGAC	5.207	AUUUCUCCCUU	10.318	1
NP_077018.1	8	CGCUGCC	2.408	GGUUUAC	4.037	CUUCUUCCCUU	12.390	9

Supplementary Table 4 - Affinity data for profile models

A

Sequence	Interim SF1 Affinity
UACUAAU	117.9
CACUAAC	111.2
UAUUAAC	104.7
UAAUAAC	102.7
UACUAAC	100
GGCUGAC	81.2
GGCUGAC	76.3
UACUCAC	75.4
UACUAAA	74.4
UACUUAC	74.1
CACUGAC	71
UAGUAAC	69.5
UGCUGAC	66.8
UACUGAC	58.7
UUCUAAC	57.1
UCCUGAC	52.9
UACCAAC	43.7
UACGAAC	42.4
UACUAUC	41.6
UACUAAG	38.9
CAGUGAC	38.4
UACUACC	36.7
UGCUGAU	15

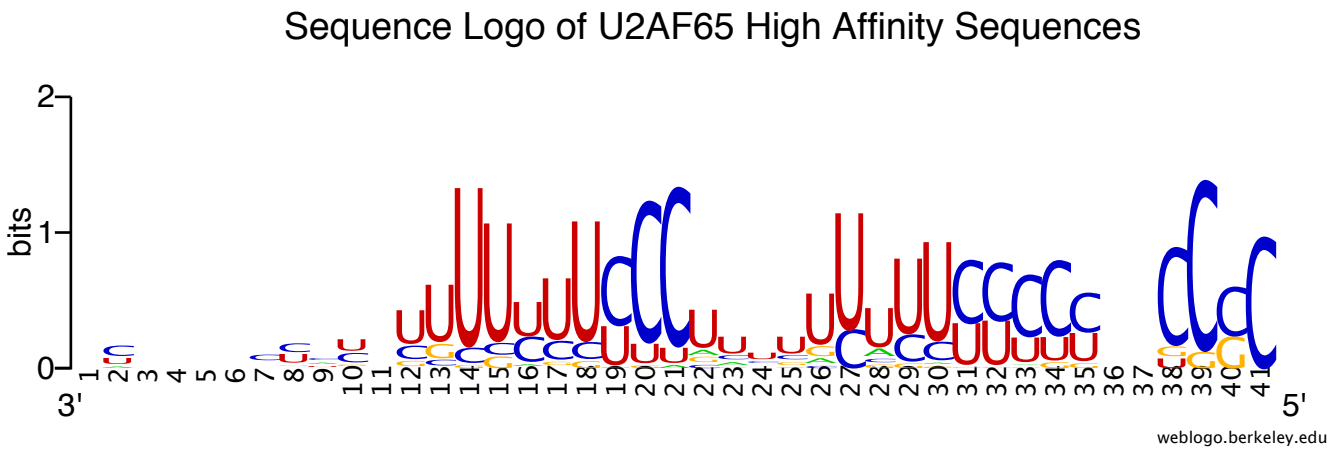
B

Sequence	SF1 Affinity
UAAUAAC	151.61
GGCUGAC	116.09
UGCUAAC	110.56
CACUAAC	105.11
UACUAAC	100
UACUAAU	95.02
AACUAAC	92.17
GACUAAC	83.17
AUAACCA	82.6
UACUAAA	73.22
UAUUAAC	70.45
CACUGAC	68.23
GACGGAU	68.1
UGCUGAU	67.87
UGCUGAC	60.35
GGCUCAC	59.34
GUCUUGA	54.89
CCCUGAC	53.89
UACUGAC	52.83
UACUUAC	49.66
GGCACAC	47.55
UACUCAC	47.49
GGCUGAU	46.64
UCCUAAC	46.52
UAGUAAC	42.85
UACUACC	40.98
GGCAGAU	38.03
ACAGUCA	36.94
UACGAAC	36.29
UACUGAU	31.3
GACUGAU	31.13
UACUAUC	30.91
UACCAAC	29.95
ACUACCA	28.69
UACUAAG	27.32
CAGUGAC	26.33
UACAAAC	26.16
UCCUGAC	25.97
UACUAGC	24.36
CCCCGAC	23.86
UUCUAAC	21.19
CGAACAG	17.74
ACGGUGG	17.45
UGAGUAG	15.72
ACGGCGG	15.32
UAGUGAC	12.27

Supplementary Table 5 – Fractional agreement between predicted and experimentally validated BPSs

	Initial SF1 Binding	SF1 Affinity	YNCURAY	Literature BPS	Pentamer BPS
Intronic region	0.065	0.068	0.286	0.024	0.099
Expected BPS region	0.286	0.357	0.200	0.176	0.286
Gao et al. BPS range	0.235	0.357	0.158	0.182	0.241

Supplementary Figure 1



Supplementary Table 1 – Introns selected for branch point mapping
BPS and PPT criteria for selecting introns for branch point mapping are described in Methods. Gene lists are HUGO gene identifiers (1); bit scores are from the U2AF65 Affinity Model and Initial SF1 Binding Model; BP A-PPT Distance lists the number of nucleotides from the predicted BP A to the start of the predicted PPT; Intron lists the number of each intron selected from each gene.

Supplementary Table 2 – Sequences assayed for SF1 binding
Sequences are grouped according to selection criteria described in Methods. 'Score Rank' is the rank of the bit score for a sequence relative to all unique BPS scores generated by a model. 'Frequency Rank' represents the ranking of a sequence based on how frequently that sequence occurs in an expected BP A region, with or without a score threshold, within the set of matches from 109,455 introns for the Initial SF1 Binding Model or 117,499 introns for the SF1 Affinity Model. Note that two overlapping sets of intronic regions were used in model construction: The first set required 100% identity between the exon downstream of the intron and its genomic locus and <100% sequence identity between every pair of introns, yielding 109,455 non-redundant intronic regions. The second set required ≥99% identity between the exon downstream of the intron and the genomic locus and no sequence uniqueness, yielding 148,643 intronic regions, including some with 3' splice site reuse in different mRNA isoforms. A nonredundant subset of this second set consisting of 117,499 sequences, with <100% sequence identity between every pair of sequences, was used to evaluate models while the full second set was used to construct the annotated dataset of intronic regions. 'Highest Scoring' lists the 10 highest scoring sequences; for the Initial SF1 Binding Model there were 4 sequences with the 7th highest score 7.2322 and only one of these was selected. 'High Frequency' lists frequently observed sequences ranked using all profile model matches in the expected BPS region above a score threshold (or in the absence of a score threshold for ranks marked with '*') (see Methods). Sequences overlapping between the 'Highest Scoring' and 'High Frequency' categories are underlined. 'Low Scoring' are the 8 selected low scoring sequences. 'Unique Point Mutants' are 10 UACUAAC point mutants assayed for binding that do not overlap with any other already selected sequences. 'Disease BPS' lists 4 wild-type and mutant sequences where BPS mutations have been implicated in disease: *TH* – tyrosine hydroxylase; *LCAT* – lecithin cholesterol acetyltransferase; *ITGB4* – integrin beta-4; *COL5A1* – collagen 5A1.

Supplementary Table 3 – Intronic regions selected for Tat-hybrid intron context SF1 binding assays
Intronic regions were selected for Tat-hybrid assays with the SF1 Affinity Model and U2AF65 Affinity Model using criteria described in Methods. Intronic regions are grouped by high BPS and high PPT scores (High BPS - High PPT), low BPS and high PPT scores (Low BPS - High PPT), or low BPS and low PPT scores (Low BPS - Low PPT). Gene and intron definitions are as in Supplementary Table 1. The listed BPS scores are from the SF1 Affinity Model and Literature BPS Model.

Supplementary Table 4 - Affinity data for profile models
The two sets of affinity values from SF1 binding assays as used for profile model construction. Affinity values are expressed as percent activation relative to UACUAAC. This data corresponds to the weights used for constructing the two affinity-weighted BPS models: interim SF1 Affinity Model (A) and SF1 Affinity Model (B).

Supplementary Table 5 - Fractional agreement between predicted and experimentally validated BPSs

The fraction of correct predictions made using BPS profile models using the same thresholds as used for constructing the models, in comparison with BPSs experimentally validated by Gao et al. (2). BPS predictions were made within the entire -199 to -1 'Intronic Region', the -46 to -16 'Expected BPS Region,' as well as the -63 to -3 region in which all reported BPSs were identified by Gao et al. ('Gao et al. BPS Range').

Supplementary Figure 1 - Sequence Logo of U2AF65 High Affinity Sequences

The sequence logo constructed using a manually edited multiple alignment of high affinity U2AF65 sequences identified by SELEX (3). The logo suggested a dimeric motif with each monomer having similar information content. Therefore only the 3' half of this putative dimeric motif, representing positions 13 to 23 of the alignment, was used for constructing the U2AF65 Affinity Model.

SUPPLEMENTARY REFERENCES:

1. Eyre, T.A., Ducluzeau, F., Sneddon, T.P., Povey, S., Bruford, E.A. and Lush, M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res*, **34**, D319-321.
2. Gao, K., Masuda, A., Matsuura, T. and Ohno, K. (2008) Human branch point consensus sequence is yUnAy. *Nucleic Acids Res*, **36**, 2257-2267.
3. Singh, R., Valcarcel, J. and Green, M.R. (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, **268**, 1173-1176.