Supplemental Data for:

Sheng et al., Cell, *115*, pp. 603–613

Monte Carlo Analysis of Sip1 Genomic Regions

A 4020 bp region of the human Sip1 gene was analyzed to find putative ChCh binding sites. This region includes 952 bp upstream of the transcription initiation site, the first exon, the first intron, and the second exon (where the translation initiation site ATG is located). Since ChCh protein binds to some degree to many NGGGNN sequences, all sites matching this string were initially mapped. Figure S1 shows the density distribution of such sites in the human Sip1 promoter region. In total, 120 putative sites were found, and two conspicuous peaks of density are observed immediately upstream of the 1st and 2nd exons, the latter forming a very strong peak (Figure SA, blue arrow). To test for the significance of these findings, a Monte Carlo simulation experiment was performed. 10,000 sequences were simulated that were of the same length and had the same nucleotide composition as the original Sip1 region analyzed. Strikingly, the number of sites matching the putative ChCh binding consensus found in the simulated sequences was considerably less ($\bar{x} = 75.5 \pm 8.2$ sites per sequence) than the 120 observed in the region analyzed ($p < 0.0002444034$).

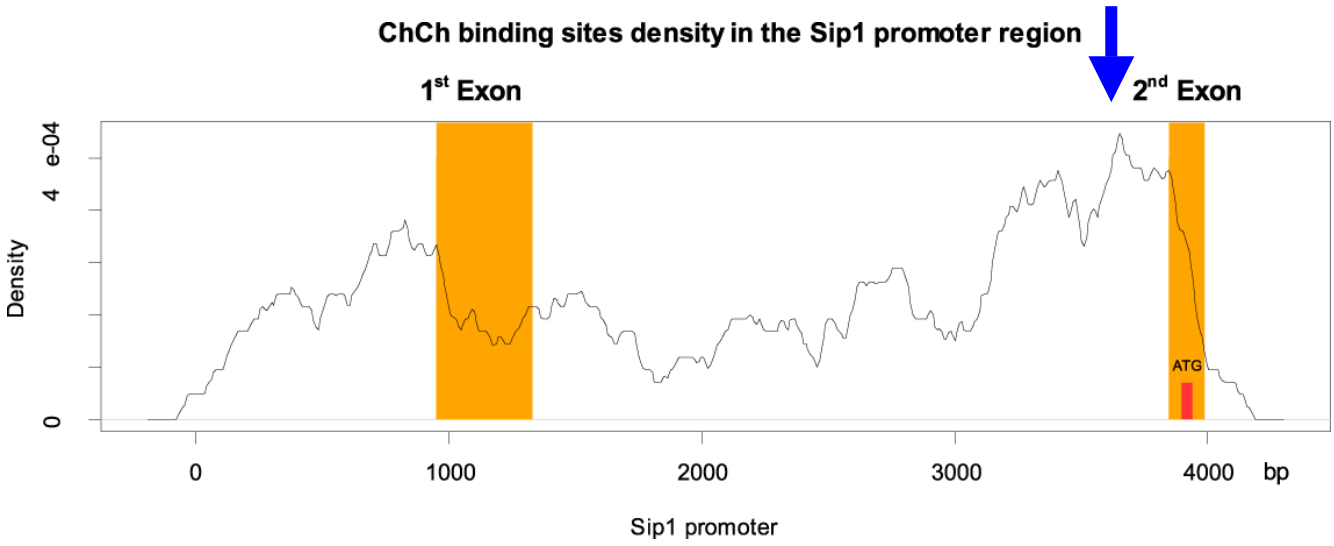

Figure SA. Density Distribution of Putative ChCh Binding Sites.

Although this analysis suggests that there are many more NGGGNN sites than expected by chance, it does give information on the quality of these sites, i.e., how similar they are to the optimal ChCh binding sequence CGGGRR. To solve this problem the following approach was taken. Suppose the site CGGGAG is found in the analyzed region, two questions arise: how likely is this motif in a sequence with the same background nucleotide composition, and how likely is it to bind to ChCh? The first question is easy to answer, the probability of observing the above sequence is

$$P(CGGGAG \mid bg) = p(C).p(G)^3.p(A).p(G)$$

Where p(C), p(G) and p(A) are the frequencies of C, G and A in the background (bg) sequence model. Since the Sip1 region analyzed has the following composition p(A) = 0.25, p(T) = 0.28, p(C) = 0.23, p(G) = 0.24 then

$$P(CGGGAG \mid bg) = 0.23 \times 0.24^3 \times 0.25 \times 0.24 = 0.00020$$

It is known that ChCh has the following preferred motif model (the nucleotide composition of the sequences obtained in the SELEX experiments):

Consensus	C	G	G	G	R	R
A	0.14	0	0	0	0.28	0.31
T	0.17	0	0	0	0.24	0.07
G	0.17	1	1	1	0.38	0.41
C	0.52	0	0	0	0.10	0.21

So the probability of this site under the ChCh motif model would be

$$P(CGGGAG \mid \text{ChCh motif}) = 0.52 \times 1^3 \times 0.28 \times 0.41 = 0.060$$

Finally both probabilities need to be compared and assess if the difference is big enough. The log-likelihood of both probabilities is calculated as:

$$\log L = \frac{P(CGGGAG|motif)}{P(CGGGAG|bg)}$$

In the above example log L = log(0.060/0.00020) = 5.7

Similarly, the log L for multiple matches (i.e., for many ChCh binding sites found) would be sum of the individual log L's. For example

log L (cgggcc cgggag) = log L(cgggcc) + log L(cgggag)

So the higher the value of log L, the more likely is that the observed site is real. But how big does log L must to be in order to be statistically significant? To answer this, log L was calculated for all matches in each of the simulated sequences, and the log L value obtained for the analyzed Sip1 region was compared to the value distribution for the log L values obtained in the simulations. As can be seen in Figure SB, the observed log L value in the Sip1 promoter region is very unlikely to have arisen by chance. Additionally, the motif scores found in the Sip1 promoter region are biased towards higher values than in the simulations (Figure SC)

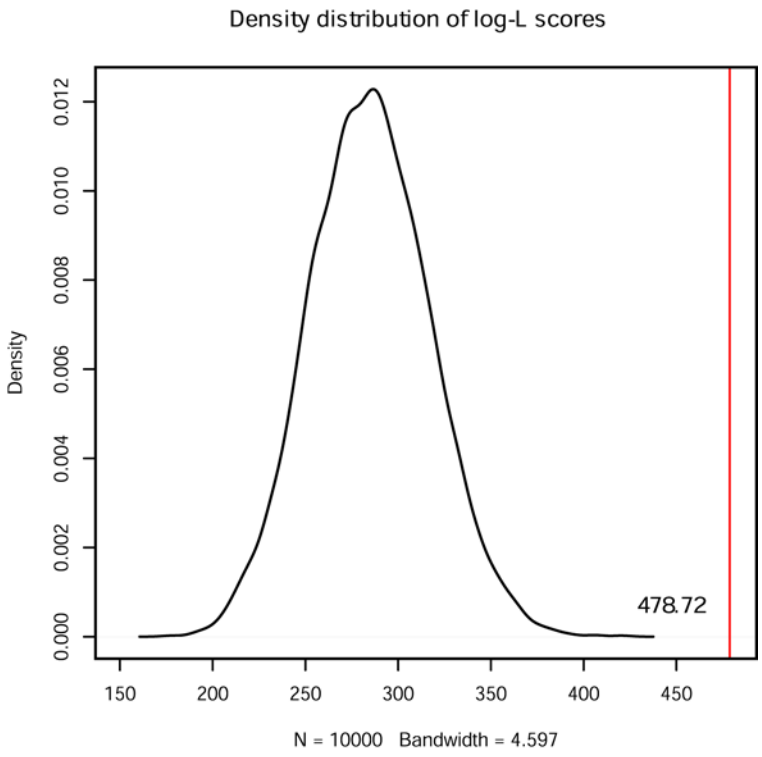

Figure SB. Density Distribution of log-L Scores Obtained from the 10,000 Simulated Sequences ($\bar{x} = 284.9 \pm 32.2$).
The actual value observed in the Sip1 promoter region is depicted as a red vertical line.

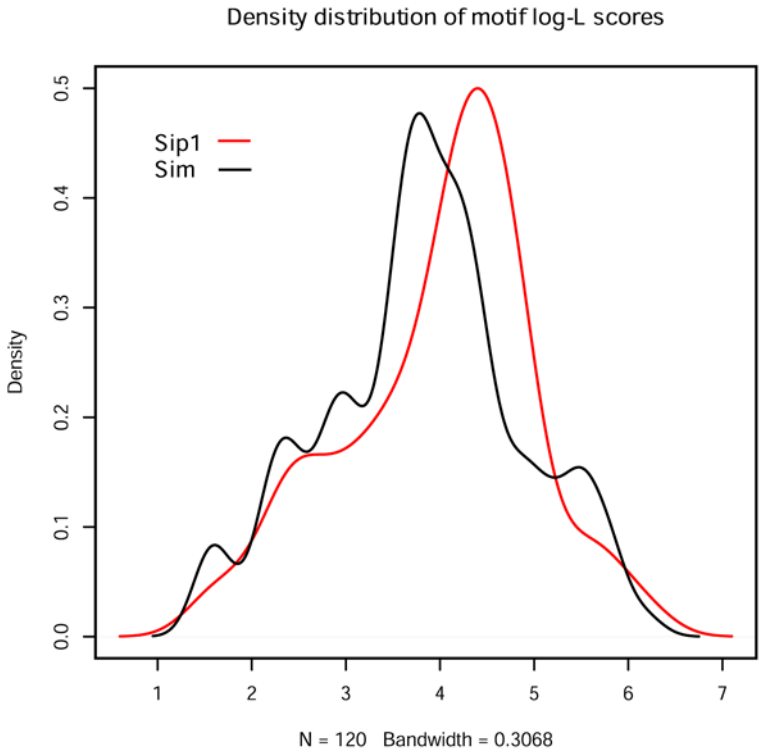

Figure SC. Density Distribution of Motif log-L Scores for the Matches Found in the Sip1 Promoter Region (red) and in the Simulated Sequences (black)

It can be seen clearly that in the real sequence, the observed scores are biased towards 4.5, while in the simulated sequences the mode is close to 4. (i.e., in the Sip1 regions motifs like cgggcg or cggggga are more abundant while in the simulated sequences motifs like agggaa or agggcg are more common).

Additional Information:

(A) Human Sip1 promoter region analyzed:

Blue: repeats; red: exons; yellow: translation start site.

ACCTATGTAACCTGATTGGATTACGACAGAAGCGTCACGTTGGAAGCTTTTGAGTGATTA
TTTAATTAACCATACAGTAGAAAGCTGTATTCTCCAGGAAATCCCTTCCATATTTGCATA
ACCAATCCCTTCAGAGCAAAGGTGGAGTCTTTTCTTTTTAATCTTACTTTAATTGTAATA
TCAAAAAATATATACTCCAAAACTTAGACCCCATGCCGTTTAATATCATGCTCTCCCTTC
CCTGCTAAGTTTCTCTATGGCCTTTCTCGTTTCTCCTCCTGCCTCCCTACACACTCTCCC
TTACTTTGTAGTGAGGTCTCCCCGAGGTGTAGAGAGATTCAGAGATCGGCCAACCGAGTG
TTCTATTTTAATTTACTTAGAGACCCTTTATTAAAATGCCAAACTACTTTTTAATATTGG
GATCCAGTCCAGAAATTCATCATGCACACACCCTAATACACATGCCCTAAGATGCAGCTC
CCATGCAGCATTTTTTTTTCTGGCTCTGGTACCTAAAAAGAAAAAAATAACAATAAGAGA
AAGGGCAGAGAACTTTGTTCCAGAAGCTGTACTGAGATACCTACACAATTTGATGTGCAT
CTCAAATCTGGTCATTAGAGATATCTGTATAAGAAGAGACTATCTGGATTGAGGACCCGG
GATCTTTCCCTTTAACTTTCGCCCCTTGGAGTTCTCCAGTTCTGTGAATGGTGTGCACCG
TTTTCCGCCCTGTACTCTGTAGGATTTAGTGATGAGGATAATGATGCCAAAGGCTTGACG
GGCG**GGGAGGGGGGGGTGGAGGGGGGGAGAAGGGAGGGAGGGGGGAGGGAGGCGAAGGCG**
AAAGGGAGGGAGAGGAGGAAGGGAGGGAGGTGGAATTTCATTTCTTCCACTAAAGCGTTT
GCGGAGACTTCAAGGTATAATCTATCCCAGATCCTTTCCCAGAGAGAAACTT**ggcgatca**
cgttttcacatgatgctcacgctcagggcgcttcaattatccctccccacaaagataggt
ggcgcgtgtttcagggtctctcgtctctctcctacagaaaagaaaaagaaaaaaatgtca
ttagaagaggcgtaacacgtcagtccgtccccaggtttgtgtttcctggagtggccgaaa
gagatcagttctaacctgctctgcaggaataacggtcctgcctcccgacactcttggcga
ggtttttgtacagtttgctccgggagctgtttcttcgcttccacctttttctcccccaca
cttcgcggcttcttcatgctttttcttctcaccatttctggccaaaactacaaacaagac
ttcgcagGTAGGTTTTTTTTCCTCCCCTTTTCTCTCTTTTTATCCCTTTTTGGTGTGCTC
GTCCTCCATCCTCCTTTTCTAATTTTCTCATTTTGAGTGGGGATGTGAGTCTGAAGTGAG
AAGGGGTGTCCGTGGGTAGGAATTTCAGGTGGGTTTAGCTCCTTTATGGCAAGCTTTTCT
CAAGAGTGGCTTCTTCGCTTTCTCTTACACTCACACTCTCTCTCCTTGCTTTTCTCTCCT
TTTTCCCTCCCCCAAATAATTTCCCCTGTACCTGTTCATTCAGGTAACCAAAGGTGTCCT
TTGCTATCCAGCCCAGGAAAAGTTTTATTGAGAATTATATGGCTGGAATAAAAACTCATC
AAGGGGAGAGGTCCTGCTCTATTTTCGCATCTAGTTCCAATTCATTATATTCAGCAATTA
GGACTTGGCAAATACCTGTTAATAATTACAAAACTTGGTCAGTTACAGACCTACAAATTC
GGGGGCAAGTCTTTCTCTTAGACATGGCAGTTCATCCTCTCCCACCGAGAAGGCTGTTTT
CTAAGCAGAAAGTTGTCTGTTTTTGTCTGCTTAGCGAGTTTCATTTCTTTATTTCATTCT
TTGTGTGTGTATGTGTGTGAAGTAACACCTCACTTCTGTGATTATAAATAAAGAAGTGAA
AGTCCTCATGGAACTTGAGTCGTTTACCAAAGTGGAGGGGGGGGCACTACAGCTGGGAAG
AAAACTGCTGTCAGGATATCAGGGTTTTAAATTAGGATAACTTTTAAAATTGTATTTTTA
GTGGCAAAGTAATATAGACATGGATTTAATAGAAGACACATACACACATATATCCCATTC
CATACTTTGGCACTGAAAAGTCGAGAGAGAGATAGCAGGAGTTCCCAGGAACAGTGATGA
GCCCGGCTAACATGAATGTGTTTATCTCCCACCTACTGATATATTTTACATGCTCCCCCT
GTAATGATGAAGGCAAGATTACTAACTCCAGACTTAGGAGTAACAGGCAAAGATAAAGTT
TACAGCACAAATGCCCTGTCTCTCTCTCTCACACACACACATGCCCATAAAACTAGCAAC
TTACTTTATTAAACATGTATTTATATGCAGTCGTGTGTCTCAGGATGAATATAGACAGGC
CAGCTGTTTCATCCCAGACACTTGTGTGACCATTTCTTTATTTCAGGATTTCCCCTAACA
ATTTCAAATCAATAGGACAAGGAAAGAAATGATAAAATATGTCAAACTTAAAAGTCTCCT
ACTATGTATGCCTAGGTAAATTTAAGGTGCTTTCTAAAATATCCACTTCCCCTTCTCTCT
GCCCCCCTCCCTTCACCCCTCTCAGCAAATGTGTGGAAACTGATACTCACTTTCCAGGTT
TTCTGCAAAGCTTCTAGCACAGGAGACAATAGGTGTGGGGCGTGGGGACTGAGTGTGCCC
GGCAGAGAAAGGGTTAATGGCCCTTGTGTTACAGTTTGCATCTGTTACATCTTTTATAGC
CCCCTGTTTAAACTAATCCTCGAGGCACAGCATCTTCCCAACCTCCACCGAGTTTTCAGG

ATTCTCCTCCTCTCCTGCCTGCCTCCGAGGTCCCCCTTCTCCAAGTTTTGACAGTTTCAA
CTGAAAAAATGCAGCGCCTCTCTCGATTTTTAGGTACAAAAGTTTCGAAGCCAGAAAGGA
TACGATCATTTAGTAGAGCTTAAAAGTATTGCCGTAGTGGTTGGACTTTCTCTTTCCTTT
CTTAAGGGAAAAAAAAAATTCTACTGGAACTATTTAAGGGAACCAGTGTGGGGGAGCATG
TGTGTGAGTGAGGTGGGGGGAAATACTTAAGTAAATATTGTCAAAGTGGCGAGGGGGGGT
GGGGTAAAGGACAGTGTCCAAAGAGGCTTATAATTATATTTATCATTGCGGCGTAAAACG
GGAAATCTTAATTTGGGGGCGAGGGTGGGGGGAGGAAGAGACAGTGCCTCGATAAACTCC
AAAAGGGGAGGGGAGGGGGCGCTGGGCGAGTGGGCTTCCTATAATTACTATTATCAATTT
GCCCGAGTGTGCTTCATCCAGGGGCTGTTTCGCGTTGTTTGTTTGTTTGTTTCCTCTGGG
AATTGGGGGGTGGGGGGATGAGAAAAGATGAGAACGAAAAGAAAACCTCGCCTCACCCCA
AGTTCCCGAGCCGGGGGCTGAGGACTTTTCGAAGGAAAGGGTCCGCCTAGCCCTGAGTCA
AGCGGTCTTACTGTACCGCCGTGTGCATTCCCTCATACGGTCAGGAGTTATGACTCATTT
TGAAGATGTAATTCTTGTCTCTCTGATCCCCTCGCGGGTGCAACACACCAAACAGTAACA
AACACAAAGCGCCTCGGGGCCAAGGCTGGGGGTGGGGGGCGGGGAGGGGGCCGCCGAAGT
TTCGCTTTGGCGTTGGGGGGCGCGGATGGGTGCT**CGCCGGGGCCCTGGCCCGGCTCCCCT**
GGGCGGCCCGCGCGCGTTTCAATGGGCGCGGGCGATGCCCACATTGTCGCTGTGTTTGGT
TGCTAG**atcgagcctgcgtgctgccgaagcagggcgccgagtccatgcgaactgccatct**
gatccgctcttatcaatgaagcagccgatcatggcggatggcccccggtgcaagaggcgc
aaacaagccaatcccaggaggaaaaacgGTAAGAAGCAGCCCGAACCAAACTTTTCCGGG

(B) ChCh binding sites found in the Sip1 promoter region:

match	log L	position	strand
agggat	2.38	106	-
agggat	2.38	130	-
ggggtc	3.75	212	-
agggag	4.35	238	-
agggag	4.35	288	-
agggag	4.35	301	-
ggggag	4.64	323	-
agggtc	3.46	387	-
tgggat	2.49	423	+
agggtg	4.11	454	-
agggca	3.13	468	-
tgggag	4.46	483	-
agggca	3.13	547	+
cgggtc	4.9	659	-
cgggat	3.82	663	+
agggaa	3.99	671	-
ggggcg	3.78	685	-
agggcg	3.49	731	-
cgggcg	4.93	784	+
aggggg	4.74	793	+
aggggg	4.74	805	+
agggag	4.35	816	+
aggggg	4.74	824	+
agggag	4.35	831	+
agggag	4.35	848	+
agggag	4.35	865	+
tgggat	2.49	929	-
tgggaa	4.1	941	-
agggcg	3.49	990	+
agggat	2.38	1004	-

agggtc	3.46	1038	+
ggggac	3.99	1112	-
cgggag	5.79	1187	-
cgggag	5.79	1226	+
ggggag	4.64	1256	-
ggggag	4.64	1347	-
agggat	2.38	1367	-
tgggga	4.49	1423	+
aggggt	2.77	1447	+
tgggta	3.86	1458	+
tgggtt	2.25	1475	+
agggaa	3.99	1568	-
tggggg	4.85	1574	-
ggggaa	4.28	1586	-
tgggct	1.63	1635	-
agggga	4.38	1687	+
cggggg	6.18	1805	+
tgggag	4.46	1844	-
aggggg	4.74	2021	+
tgggaa	4.1	2039	+
agggtt	2.14	2066	+
tgggat	2.49	2157	-
tgggaa	4.1	2207	-
cgggct	2.96	2225	-
tgggag	4.46	2251	-
ggggag	4.64	2278	-
agggca	3.13	2357	-
tgggca	3.24	2387	-
tgggat	2.49	2476	-
ggggaa	4.28	2515	-
ggggaa	4.28	2632	-
ggggca	3.42	2645	-
agggag	4.35	2652	-
ggggtg	4.4	2659	-
tggggc	4.2	2741	+
tgggga	4.49	2748	+
cgggca	4.57	2761	-
agggtt	2.14	2775	+
agggcc	2.84	2784	-
ggggct	1.81	2823	-
tgggaa	4.1	2860	-
ggggac	3.99	2915	-
agggaa	3.99	3070	+
agggaa	3.99	3102	+
tggggg	4.85	3114	+
tggggg	4.85	3139	+
aggggg	4.74	3177	+
tggggt	2.88	3185	+
cgggaa	5.43	3244	+
tggggg	4.85	3259	+
agggtg	4.11	3267	+

ggggga	4.67	3273	+
agggga	4.38	3309	+
ggggag	4.64	3315	+
ggggcg	3.78	3321	+
tgggcg	3.6	3328	+
tgggct	1.63	3336	+
cgggca	4.57	3365	-
aggggc	4.09	3385	+
tgggaa	4.1	3422	+
tggggg	4.85	3429	+
tggggg	4.85	3436	+
ggggtg	4.4	3479	-
cgggaa	5.43	3488	-
cggggg	6.18	3497	+
agggtc	3.46	3523	+
agggct	1.52	3534	-
agggaa	3.99	3573	-
ggggat	2.67	3631	-
cgggtg	5.55	3640	+
cggggc	5.53	3680	+
tggggg	4.85	3692	+
tggggg	4.85	3698	+
cgggga	5.82	3705	+
gggggc	4.38	3711	+
tggggg	4.85	3739	+
tgggtg	4.22	3752	+
cggggc	5.53	3763	+
agggcc	2.84	3766	-
cgggcc	4.28	3772	-
ggggag	4.64	3779	-
tgggcg	3.6	3785	+
cgggcc	4.28	3790	-
tgggcg	3.6	3808	+
cgggcg	4.93	3814	+
tgggca	3.24	3821	-
agggcg	3.49	3876	+
ggggcc	3.13	3945	-
tgggat	2.49	3976	-
cgggct	2.96	4003	-