Chromatin Tandem Affinity Purification Sequencing (ChTAP-Seq)

Vahab D Soleimani, Gareth A Palidwor, Theodore J Perkins and Michael A Rudnicki.

SUPPLEMENTARY INFORMATION

Caption to figures

Figure S1. Read distribution in ChIP-seq, control and input sample. (**a**) H3K4me3 ChIP-seq (GSM798328), control (GSM798324) and input reads (GSM923574) were obtained from UCSC ENCODE DCC, available from the NCBI database. The genome was divided into 500 bp windows and, for each data set; the number of reads per window was calculated as described in Figure 3a. (**b-c**) Correlation analysis between read distribution of input DNA with H3K4me3 ChIP-seq and IgG control ChIP (mock ChIP) shows that IgG ChIP-seq signal is more correlated to the H3K4me3 ChIP-seq than the input DNA. Correlation analysis was performed as described in Fig. 3-4.

Figure S2. Representative screenshot of flagged genomic regions enriched in input, control and ChIP-seq datasets. Many of these genomic regions are enriched in all datasets including those generated by ENCODE. See table S1 for a complete list of these genomic regions.

Supplementary Discussion

Inferring the composition of ChIP-seq signal. Discrimination of putative transcription factor-binding sites (true peaks) from a significant amount of noise inherent in ChIP-seq data is a critical initial step that can affect the outcome of all downstream data analysis. In recent years, various peak calling algorithms have been developed [1-5]. To adjust for the inherent noise in ChIP-seq data, many studies have used a generic null model of the read distribution for control such as a Poisson or negative binomial distribution, to calculate p-values and false discovery rates (FDR) [1,2,6]. Analysis of ChTAP data, which uses identical affinity reagents in the ChIP and the control experiment, challenges the notion that input DNA is an appropriate control to adjust for background noise. We have shown that using input DNA as a means to subtract background results in a significantly larger set of peaks compared to using a matching control ChIP. Importantly, the large set of extra peaks that are called when using input DNA for background subtraction show hallmarks of false positive peaks including low peak scores and no significant enrichment for the DNA motif to which the transcription factor of interest binds within the boundaries of these peaks (Fig. 6).

To adjust for background noise emanating from ChIP-seq data we have used reads from an empty vector – matching control experiment for background subtraction [7,8]. Consistent with previous observations [9] that read distributions in control (mock IP) experiments bear significant resemblance to genome wide read distributions of the actual ChIP, rather than input DNA libraries.

Comparative analysis using a linear regression model also indicates that the ChIP-seq signal contains a significant amount of control-like signal that cannot be captured by using input signal as control. To provide a quantitative estimate of the extent of noise in ChIP-seq data we: (1) analyzed the distribution of reads in 500 bp windows for a Poisson model and various data sets, and (2) correlation/linear regression analysis of ChTAP-seq of ChIP-seq reads using control and input densities as independent variables. We first divided the mouse genome (mm9) into 500 bp windows (500 bp approximates the average length of the peaks identified in our ChTAP-seq datasets). We analyzed three datasets; transcription factor (experimental) ChTAP-seq, empty vector (control) ChTAP-seq and a sequenced input DNA library. We performed these analyses on a MyoD ChIP-seq dataset in myoblasts, a MyoD ChIP-seq dataset in myotubes, a Myf5 ChIP-seq dataset in myoblasts, and a Pax7 ChIP-seq dataset in myoblasts. For comparative purposes we extended this analysis to other ChIP-seq datasets generated by an independent group. We obtained ChIP-seq (GSM798328), control (GSM798324) and input reads (GSM923574) from UCSC ENCODE DCC, available from the NCBI database. For each dataset we counted the frequency of windows with 0, 1, 2, …, 29 reads. In addition, we also combined all windows with ≥30 reads into one bin. To adjust for the differences in tag number among different dataset we normalized, the window counts to 10 million tags. We also computed the expected number of windows with 0, 1, 2, ..., ≥30 reads under a Poisson model (Fig. 3, Fig. S1).

Our data shows that the Poisson model is a poor fit to any of the real libraries (i.e., ChIP, Control, and the input DNA). The Poisson model attributes very low probability to even modestly enriched windows. The input DNA distribution is much broader than the Poisson distribution, with a peak around 4 reads per window, and an inflection point at 1 read per window (Fig. 3, Fig. S1). This is most likely due to the occurrence of windows with 0 reads, which are fairly common, at least in part because large regions of the genome are unmappable. The input DNA also includes some highly enriched windows, some of which may be mapping artifacts or bias in sequencing the input library, as described below in more detail. The control and the ChIP data appear quite similar for low numbers of reads (≤5 per window) which are the most common kinds of windows (Fig. 3, Fig. S1). Interestingly, many genomic regions enriched in the input DNA library were equally enriched in all controls and ChIP-seq datasets. Furthermore, these regions were also enriched in the ENCODE datasets available from the UCSC genome browser (Fig. S2, Table S1). Consequently, we have flagged these artificially enriched regions and excluded them from all our downstream analysis.

To study how well the signals are correlated on a per-window basis we performed correlation analysis (Fig. 4, Fig. S1b). Our analysis shows that on a per-window basis the empty vector ChTAP-seq control data is better correlated with the actual transcription factor ChTAP-seq data than is the input DNA (Fig. 4, Fig. S1b). The input DNA library also shows non-trivial but lower correlations with the actual transcription factor ChTAP-seq (Fig. 4). Thus, the total signal of a transcription factor (ChIP-seq signal) contains a substantial amount of non-specific noise.

To disassemble the transcription factor (in this case Pax7, MyoD and Myf5) ChTAP-seq signal into true signal versus noise we formulated the following linear model. TF_Tags_Normalized = w1 × InputDNA_Tags_Normalized + w2 × ChTAP-seq_Control_Tags_Normalized, where w1 denotes the weight of input and w2 denotes the weight of empty vector ChTAP-seq (control). We normalized each dataset by subtracting the mean count from each window and divided it by the standard deviation. Therefore, the normalized tag counts per window have mean zero and standard deviation one. By applying this model to MyoD dataset we obtain

w1	0.2296	(0.2288 , 0.2303)
w2	0.5091	(0.5083 , 0.5098)

where numbers inside the parenthesis indicate 95% confidence intervals. Importantly, this model shows that input DNA library as well as the empty vector ChTAP-seq control are both useful in predicting the total signal in the ChIP-seq data. However, the empty vector ChTAP-seq (w2 = 0.5091) has a larger weight and much more predictive power than the input DNA (w1 = 0.2296). By extending this analysis to publicly available ChIP-seq datasets from an independent laboratory we observed very similar pattern. Namely, that ChIP signal is more closely related to a mock ChIP (IgG ChIP) than the input (Fig. S1). Therefore, these analyses indicate that an affinity based matching control is more superior for the normalization of background noise in ChIP-seq data.

The alternative control commonly used to eliminate background noise in ChIP-seq is sonicated input DNA. Although such input DNA libraries have been shown to have underlying signals corresponding to gene expression, mappability and chromatin structure [10]; the reads tend to be rather smoothly distributed with far less localized enrichment compared to treatment ChIP-seq data or an actual control ChIP-seq.

Whilst the signal present in the input DNA library is not trivial, the empty vector control shows a much greater similarity to the treatment ChTAP-seq libraries (Fig. 3-4). However, tag distribution in sonicated DNA library (input) is also non-uniformly distributed and showed some bias towards transcription start sites (TSS) (Fig. 5). This kind of bias in high throughput DNA sequencing experiments has been reported previously [11,12] and may be related to GC compositional bias around TSS. Previous studies have shown that eukaryotic transcription start sites (TSSs) are associated with elevated GC content [13]. The GC content of the TSS may also be due to the large overlap of TSSs with CpG islands, as most mouse and human TSSs have been shown to be associated with CpG islands [14]. A direct correlation between enrichment of reads in the input DNA library and the GC content around the TSS of ENSEMBL genes (known protein coding genes) (Fig. 5) confirms the above observations. Therefore, some enrichment around TSS and other genomic features may be independent of immunoprecipitation steps and may be due to structural differences in the genome, resulting in biased shearing of DNA, or due to PCR bias during library construction during high throughput sequencing.

In conclusion, ChIP-seq signals contain a significant amount of noise. While input DNA is widely used for the subtraction of background noise, and can eliminate some noise emanating from GC compositional bias, a significant amount of non-specific signal in the ChIP-seq data is not amenable to normalization by using input DNA. By using ChTAP-seq we first provide an efficient alternative to the traditional ChIP, and second, by using identical affinity reagents, we dissect the ChIP-seq signal into real signal versus background noise. Our methodology of dealing with background noise in ChIP-seq data is broadly applicable to the analysis of all ChIP-seq data.

References Cited:

1 Tuteja, G., White, P., Schug, J. & Kaestner, K. H. Extracting transcription factor targets from ChIP-Seq data. *Nucleic acids research* **37**, e113, doi:10.1093/nar/gkp536 (2009).

2 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

3 Feng, X., Grossman, R. & Stein, L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC bioinformatics* **12**, 139, doi:10.1186/1471-2105-12-139 (2011).

4 Xu, J. & Zhang, Y. A generalized linear model for peak calling in ChIP-Seq data. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 826-838, doi:10.1089/cmb.2012.0023 (2012).

5 Spyrou, C., Stark, R., Lynch, A. G. & Tavare, S. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC bioinformatics* **10**, 299, doi:10.1186/1471-2105-10-299 (2009).

6 Diaz, A., Park, K., Lim, D. A. & Song, J. S. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical applications in genetics and molecular biology* **11**, Article 9, doi:10.1515/1544-6115.1750 (2012).

7 Soleimani, V. D. *et al.* Transcriptional dominance of Pax7 in adult myogenesis is due to high-affinity recognition of homeodomain motifs. *Developmental cell* **22**, 1208-1220, doi:10.1016/j.devcel.2012.03.014 (2012).

8 Soleimani, V. D. *et al.* Snail regulates MyoD binding-site occupancy to direct enhancer switching and differentiation-specific transcription in myogenesis. *Molecular cell* **47**, 457-468, doi:10.1016/j.molcel.2012.05.046 (2012).

9 Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813-1831, doi:10.1101/gr.136184.111 (2012).

10 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).

11 Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36**, e105, doi:10.1093/nar/gkn425 (2008).

12 Vega, V. B., Cheung, E., Palanisamy, N. & Sung, W. K. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PloS one* **4**, e5241, doi:10.1371/journal.pone.0005241 (2009).

13 Fujimori, S., Washio, T. & Tomita, M. GC-compositional strand bias around transcription start sites in plants and fungi. *BMC genomics* **6**, 26, doi:10.1186/1471-2164-6-26 (2005).

14 Bajic, V. B. *et al.* Mice and men: their promoter properties. *PLoS genetics* **2**, e54, doi:10.1371/journal.pgen.0020054 (2006).

Figure S1

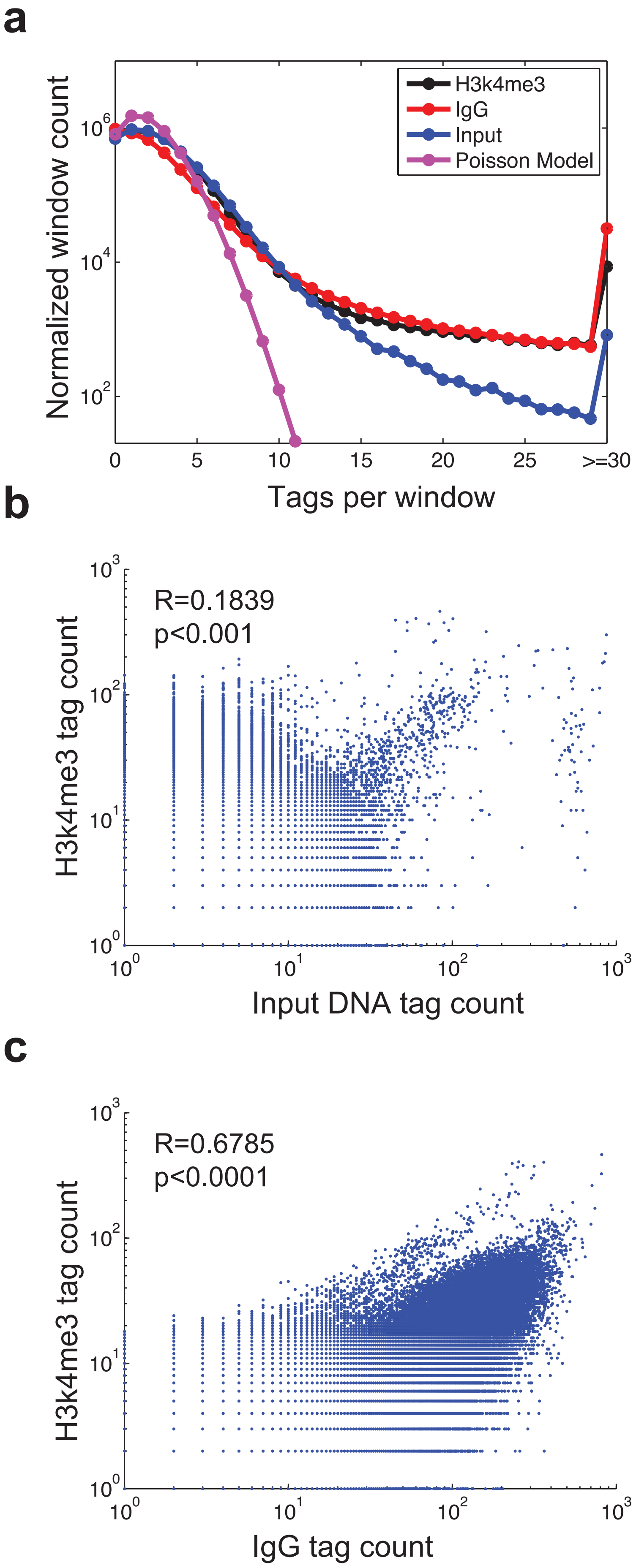
a
Normalized window count
10^6
10^4
10^2
0
5
10
15
20
25
>=30
Tags per window
H3k4me3
IgG
Input
Poisson Model
b
R=0.1839
p<0.001
H3k4me3 tag count
10^0
10^1
10^2
10^3
Input DNA tag count
c
R=0.6785
p<0.0001
H3k4me3 tag count
10^0
10^1
10^2
10^3
IgG tag count

Figure S2

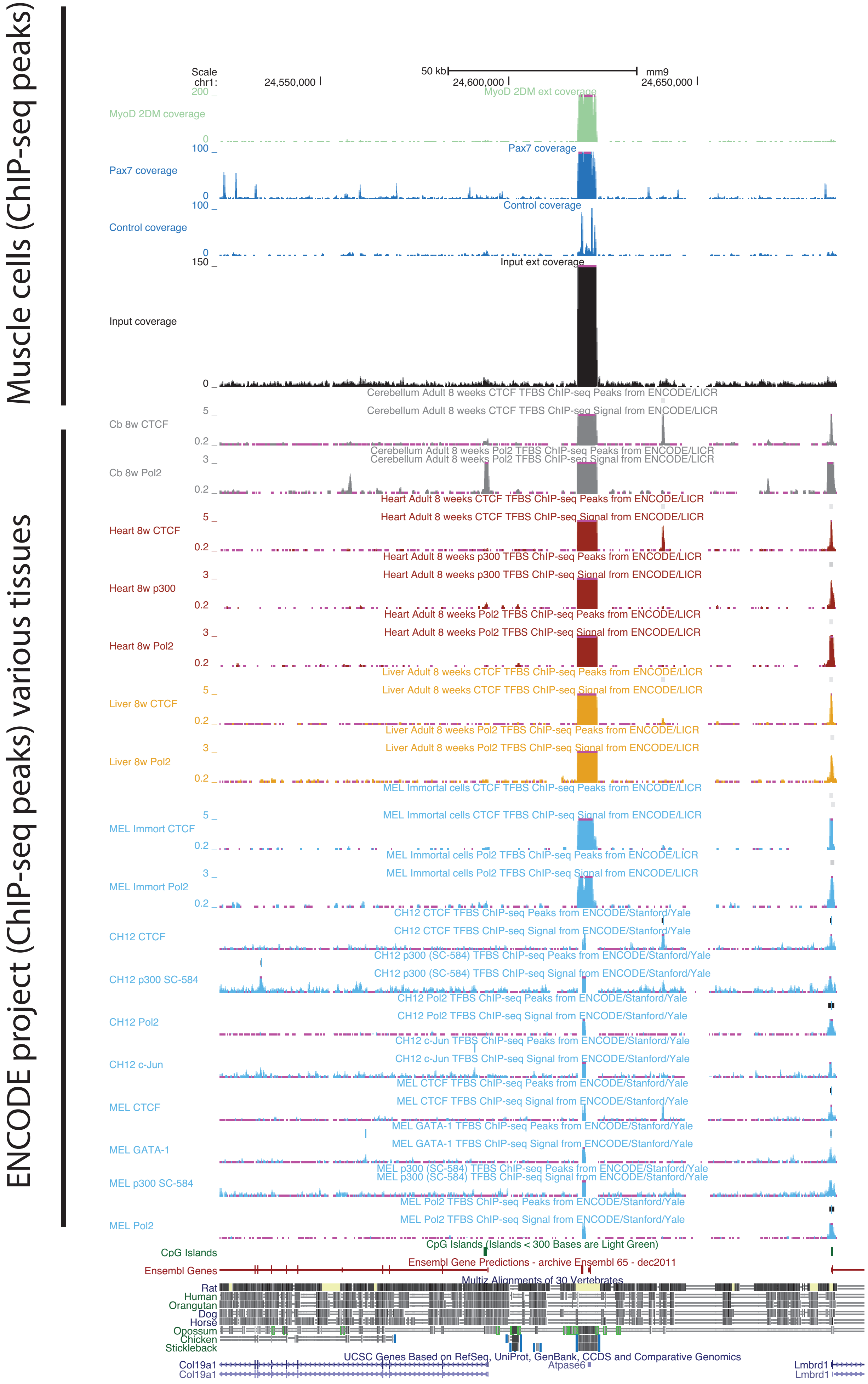
Muscle cells (ChIP-seq peaks)
ENCODE project (ChIP-seq peaks) various tissues
Scale
chr1:
50 kb
mm9
24,550,000
24,600,000
24,650,000
200
MyoD 2DM ext coverage
MyoD 2DM coverage
0
100
Pax7 coverage
Pax7 coverage
0
100
Control coverage
Control coverage
0
150
Input ext coverage
Input coverage
0
Cerebellum Adult 8 weeks CTCF TFBS ChIP-seq Peaks from ENCODE/LICR
Cerebellum Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR
5
Cb 8w CTCF
0.2
Cerebellum Adult 8 weeks Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR
Cerebellum Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
3
Cb 8w Pol2
0.2
Heart Adult 8 weeks CTCF TFBS ChIP-seq Peaks from ENCODE/LICR
Heart Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR
5
Heart 8w CTCF
0.2
Heart Adult 8 weeks p300 TFBS ChIP-seq Peaks from ENCODE/LICR
Heart Adult 8 weeks p300 TFBS ChIP-seq Signal from ENCODE/LICR
3
Heart 8w p300
0.2
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR
Heart Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
3
Heart 8w Pol2
0.2
Liver Adult 8 weeks CTCF TFBS ChIP-seq Peaks from ENCODE/LICR
Liver Adult 8 weeks CTCF TFBS ChIP-seq Signal from ENCODE/LICR
5
Liver 8w CTCF
0.2
Liver Adult 8 weeks Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR
Liver Adult 8 weeks Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
3
Liver 8w Pol2
0.2
MEL Immortal cells CTCF TFBS ChIP-seq Peaks from ENCODE/LICR
MEL Immortal cells CTCF TFBS ChIP-seq Signal from ENCODE/LICR
5
MEL Immort CTCF
0.2
MEL Immortal cells Pol2 TFBS ChIP-seq Peaks from ENCODE/LICR
MEL Immortal cells Pol2 TFBS ChIP-seq Signal from ENCODE/LICR
3
MEL Immort Pol2
0.2
CH12 CTCF TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
CH12 CTCF TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
CH12 CTCF
CH12 p300 (SC-584) TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
CH12 p300 (SC-584) TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
CH12 p300 SC-584
CH12 Pol2 TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
CH12 Pol2 TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
CH12 Pol2
CH12 c-Jun TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
CH12 c-Jun TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
CH12 c-Jun
MEL CTCF TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
MEL CTCF TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
MEL CTCF
MEL GATA-1 TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
MEL GATA-1 TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
MEL GATA-1
MEL p300 (SC-584) TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
MEL p300 (SC-584) TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
MEL p300 SC-584
MEL Pol2 TFBS ChIP-seq Peaks from ENCODE/Stanford/Yale
MEL Pol2 TFBS ChIP-seq Signal from ENCODE/Stanford/Yale
MEL Pol2
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands
Ensembl Gene Predictions - archive Ensembl 65 - dec2011
Ensembl Genes
Multiz Alignments of 30 Vertebrates
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Col19a1
Col19a1
Atpase6
Lmbrd1
Lmbrd1

Table S1. Regions of mouse genome (flagged region) that are enriched in input, affinity control and experimental ChIP.

Chromosome	Peak start	Peak end	Peak width (bp)	In input
chr1	197067705	197068115	411	TRUE
chr10	21862065	21862874	810	TRUE
chr10	95540115	95540577	463	TRUE
chr10	95540115	95540577	463	TRUE
chr11	53953491	53954238	748	TRUE
chr11	53953491	53954238	748	TRUE
chr11	53953491	53954238	748	TRUE
chr11	53953491	53954238	748	TRUE
chr11	108872898	108873446	549	TRUE
chr12	3109866	3110177	312	TRUE
chr12	56701255	56701450	196	FALSE
chr12	68160101	68160929	829	TRUE
chr12	75904825	75905560	736	TRUE
chr12	98299607	98300069	463	TRUE
chr12	98299607	98300069	463	TRUE
chr13	3372078	3372970	893	TRUE
chr13	3372078	3372970	893	TRUE
chr13	3372078	3372970	893	TRUE
chr13	77578107	77578378	272	TRUE
chr13	77578107	77578378	272	TRUE
chr13	85266203	85267181	979	TRUE
chr13	97960438	97960646	209	TRUE
chr13	100560813	100561130	318	TRUE
chr13	100560813	100561130	318	TRUE
chr13	100560813	100561130	318	TRUE
chr15	72399062	72399239	178	FALSE
chr15	74916570	74917534	965	TRUE
chr15	74916570	74917534	965	TRUE
chr15	74916570	74917534	965	TRUE
chr15	74916570	74917534	965	TRUE
chr15	74916570	74917534	965	TRUE
chr15	74916570	74917534	965	TRUE
chr15	74916570	74917534	965	TRUE
chr15	77950232	77950459	228	FALSE
chr16	10975106	10975426	321	TRUE
chr16	35981645	35981897	253	TRUE
chr16	35981645	35981897	253	TRUE
chr16	57391552	57391833	282	TRUE
chr16	57391552	57391833	282	TRUE

Chromosome	Peak start	Peak end	Peak width (bp)	In input
chr16	98244219	98244413	195	TRUE
chr16	98244219	98244413	195	TRUE
chr16	98244219	98244413	195	TRUE
chr17	36368168	36368596	429	TRUE
chr17	36368168	36368596	429	TRUE
chr17	39979904	39985817	5914	TRUE
chr17	39979904	39985817	5914	TRUE
chr17	39979904	39985817	5914	TRUE
chr17	39979904	39985817	5914	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3005011	3006236	1226	TRUE
chr18	3638812	3639187	376	TRUE
chr18	40467658	40468068	411	TRUE
chr18	56099081	56099335	255	TRUE
chr18	82390528	82390698	171	TRUE
chr19	4989120	4989292	173	FALSE
chr19	42330454	42330685	232	FALSE
chr19	45724406	45724768	363	TRUE
chr19	45724406	45724768	363	TRUE
chr19	61275555	61275792	238	TRUE
chr2	22444012	22446095	2084	TRUE
chr2	22600333	22600516	184	TRUE
chr2	69193555	69193763	209	TRUE
chr2	79294386	79294591	206	FALSE
chr2	98502333	98503197	865	TRUE
chr2	98502333	98503197	865	TRUE
chr2	98502333	98503197	865	TRUE
chr2	98502333	98503197	865	TRUE
chr2	98503753	98504307	555	TRUE
chr2	98506340	98507525	1186	TRUE
chr2	98506340	98507525	1186	TRUE
chr2	152858042	152858252	211	FALSE
chr2	164794060	164794160	101	FALSE
chr2	181652014	181652818	805	TRUE
chr2	181652014	181652818	805	TRUE
chr2	181652014	181652818	805	TRUE
chr2	181665130	181665729	600	TRUE

Chromosome	Peak start	Peak end	Peak width (bp)	In input
chr2	181665130	181665729	600	TRUE
chr3	5860569	5860868	300	TRUE
chr3	5860569	5860868	300	TRUE
chr3	8245754	8246603	850	TRUE
chr3	8245754	8246603	850	TRUE
chr3	8245754	8246603	850	TRUE
chr3	56379314	56379744	431	TRUE
chr3	56379314	56379744	431	TRUE
chr3	56379314	56379744	431	TRUE
chr3	56379314	56379744	431	TRUE
chr3	99785033	99785316	284	TRUE
chr4	3016622	3018387	1766	TRUE
chr4	3016622	3018387	1766	TRUE
chr4	3016622	3018387	1766	TRUE
chr4	21639564	21639635	72	FALSE
chr4	28104023	28104197	175	TRUE
chr4	64158840	64159150	311	TRUE
chr4	64158840	64159150	311	TRUE
chr4	70039090	70039358	269	TRUE
chr4	79648396	79651153	2758	TRUE
chr5	7276770	7277009	240	TRUE
chr5	115372382	115372654	273	TRUE
chr5	115372382	115372654	273	TRUE
chr5	115372382	115372654	273	TRUE
chr5	147072509	147072968	460	TRUE
chr5	147072509	147072968	460	TRUE
chr6	3151399	3151612	214	TRUE
chr6	9840054	9840216	163	TRUE
chr6	48071227	48071390	164	TRUE
chr6	79768041	79768360	320	TRUE
chr6	79768041	79768360	320	TRUE
chr6	103598999	103599287	289	TRUE
chr6	121423177	121423379	203	FALSE
chr6	136752829	136753076	248	TRUE
chr7	7230792	7230996	205	TRUE
chr7	88638379	88638443	65	FALSE
chr7	89920145	89920590	446	TRUE
chr7	118940974	118941162	189	TRUE
chr8	14072623	14072825	203	FALSE
chr8	19784504	19784991	488	TRUE
chr8	20020262	20020528	267	TRUE
chr8	24205893	24206328	436	TRUE

Chromosome	Peak start	Peak end	Peak width (bp)	In input
chr8	24205893	24206328	436	TRUE
chr8	127317183	127317380	198	FALSE
chr9	2999946	3001152	1207	TRUE
chr9	3002908	3003075	168	TRUE
chr9	3005825	3006009	185	TRUE
chr9	3015212	3015406	195	TRUE
chr9	3015644	3015808	165	TRUE
chr9	3016025	3016407	383	TRUE
chr9	3018714	3019517	804	TRUE
chr9	3025313	3026171	859	FALSE
chr9	3026262	3027857	1596	TRUE
chr9	3030097	3030318	222	TRUE
chr9	3031825	3032769	945	TRUE
chr9	3035658	3036643	986	TRUE
chr9	15123942	15124316	375	TRUE
chr9	24346388	24346665	278	TRUE
chr9	24346388	24346665	278	TRUE
chr9	24346388	24346665	278	TRUE
chr9	24549507	24549705	199	TRUE
chr9	35112718	35113179	462	TRUE
chr9	86916747	86916747	1	FALSE
chr9	108729231	108729531	301	FALSE
chr9	123371015	123371328	314	TRUE
chr9	123782164	123782369	206	TRUE
chrX	73843981	73844520	540	TRUE
chrX	73843981	73844520	540	TRUE
chrX	109484283	109484882	600	TRUE
chrX	109484283	109484882	600	TRUE
chrX	139917495	139917705	211	TRUE
chrX	139917495	139917705	211	TRUE
chrY	2765733	2765870	138	FALSE

Most of these genomic sites hereafter designated as flagged regions are also enriched in other publically available control and experimental ChIP datasets. True indicates whether the given genomic locus is equally enriched in input sample. False indicates that the region is enriched in affinity based control and experimental ChIP but not in the input genomic DNA.