Supplemental Figures

A

cell type	DLD-1	Ramos	BEAS2B	HEK293	MCF7	TIG3	brain	kidney	heart	fetal brain	fetal kidney	fetal heart	Total
#TSS tag in RefSeq region	7679112	13875727	20106118	8880335	7476053	8883872	7761770	7285131	5379867	9607689	6783071	7982779	111701524
#TSS tag in intergenic region	957301	2355840	1744854	736616	796086	783112	3472374	3458501	3653984	1542402	1592092	1817819	22910981
#TSS tag in antisense of RefSeq region	213119	408592	623077	117363	82131	217908	602851	579412	441998	559169	487960	500645	4834225
Total	8849532	16640159	22474049	9734314	8354270	9884892	11836995	11323044	9475849	11709260	8863123	10301243	139446730

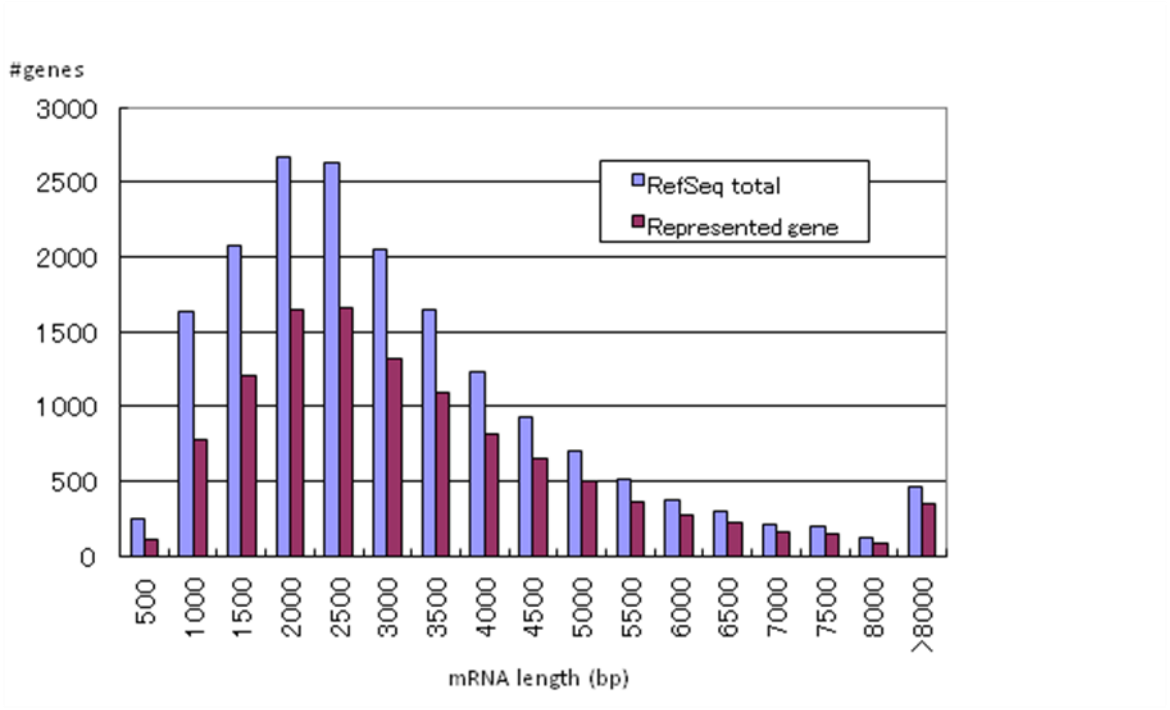
B

cell type	DLD-1	Ramos	BEAS2B	HEK293	MCF7	TIG3	brain	kidney	heart	fetal brain	fetal kidney	fetal heart	Total
# TSC (total)	81181	111437	161974	65183	50810	47816	251453	112418	86112	137721	137721	223919	700371
#TSC (>5ppm)	6265	5690	5955	5605	5635	5934	6287	5687	4299	7959	7233	7199	21030
# Represented NM (>5ppm)	5666	4879	5213	5249	4959	5222	4878	4300	3715	6027	5401	4309	11406
# Represented NM with >=2 TSCs	444	601	562	327	530	498	603	555	298	979	1085	1018	4937

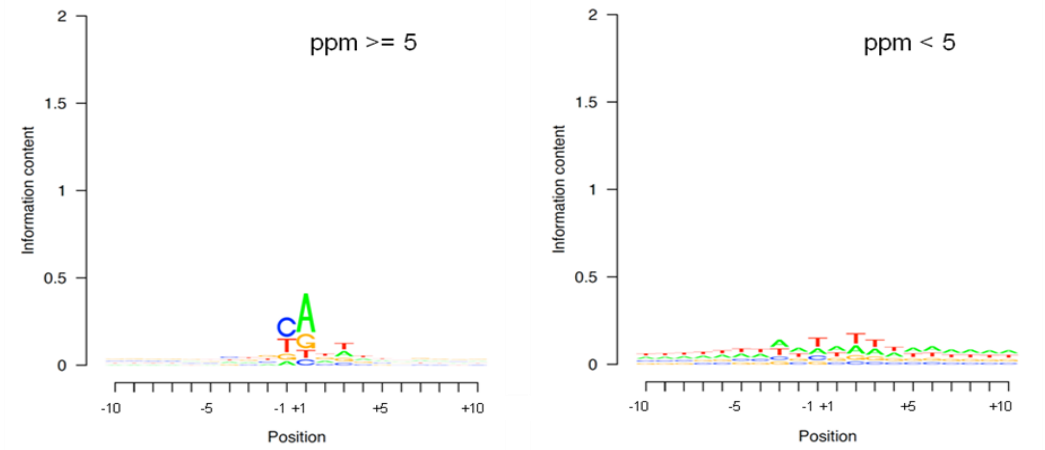
C

	cell type	DLD-1	Ramos	BEAS2B	HEK293	MCF7	TIG3	brain	kidney	heart	fetal brain	fetal kidney	fetal heart	total
outside of the Refseq region		1170420	2764432	2367931	853979	878217	1001020	4075225	4037913	4051630	2101571	1780052	2318464	27400854
within the Refseq region		7679112	13875727	20106118	8880335	7476053	8883872	7761770	7285131	5379867	9607689	6783071	7982779	111701524
	upstream	2144404	4121651	5572919	2687663	2201336	2204737	1882576	1848778	1333475	3077876	1555107	1786696	30417218
	first exon	4507777	7602531	11341708	5112453	3945099	5388107	3608274	3465387	2976629	3910793	2925934	3036124	57820816
	second or later exon	544841	738366	1667229	707253	625280	685944	1060720	958070	595387	965187	775681	1306482	10630440
	intron	482090	1413179	1524262	372966	704338	605084	1210200	1012896	474376	1653833	1226349	1853477	12533050
	%full	0.93	0.95	0.92	0.92	0.92	0.92	0.86	0.87	0.89	0.90	0.88	0.84	0.90

D


#genes
3000
2500
2000
1500
1000
500
0
RefSeq total
Represented gene
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
>8000
mRNA length (bp)

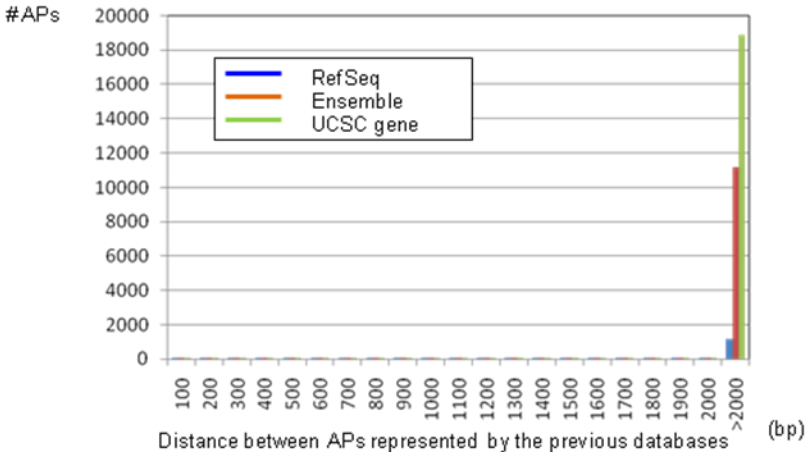
E


ppm >= 5
Information content
2
1.5
1
0.5
0
-10
-5
-1 +1
+5
+10
Position
ppm < 5
Information content
2
1.5
1
0.5
0
-10
-5
-1 +1
+5
+10
Position

F

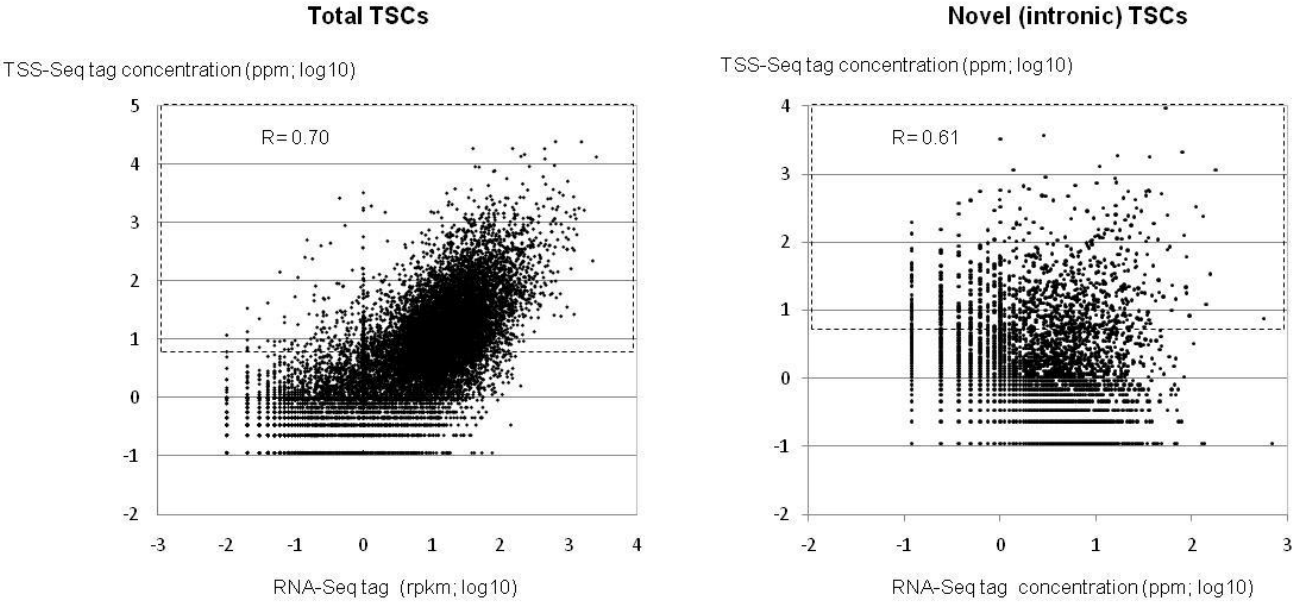
	cluster size			
Expression level of TSCs	**100 bin**	**250 bin**	**500 bin**	**1000 bin**
0-1 ppm	776,136	657,301	548,211	458,822
1-2.5 ppm	123,678	115,609	107,485	103,072
2.5-5 ppm	26,542	25,048	23,645	23,016
5-10 ppm	9,932	9,356	8,836	8,612
>10 ppm	13,113	12,590	12,194	12,092

G



	%APs separated by >500 bp
RefSeq	82%
Ensemble	96%
UCSC gene	98%

H



I

RefSeq	Definition	Genomic position	Expression level (ppm)	Overlap with Ensemble	Overlap with spliced ESTs
NM_001016	ribosomal protein	Chr6,Fw,133177401	24805	2	1505
NM_002046	glyceraldehyde-3-phosphate dehydrogenase	Chr12,Fw,6513944	24576	5	9565
NM_021130	peptidylprolyl isomerase A	Chr7,Fw,44802805	18867	1	3816
NM_000969	ribosomal protein L5	Chr1,Fw,93070184	18606	2	1664
NM_002799	proteasome beta 7 subunit proprotein	Chr9,Rv,126217542	14930	2	489
NM_001469	ATP-dependent DNA helicase II, 70 kDa subunit	Chr22,Fw,40347245	13873	4	2675
NM_021103	thymosin, beta 10	Chr2,Fw,84986291	13556	1	1471
NM_001028	ribosomal protein S25	Chr11,Rv,118394263	12568	1	1935
NM_006088	tubulin, beta, 2	Chr9,Fw,139255560	9679	1	2134
NM_152350	hypothetical protein LOC125144	Chr17,Fw,16283076	9293	2	1150

J

RefSeq	Definition	Genomic position	Expression level (ppm)	Overlap with Ensemble	Overlap with spliced ESTs
NM_001017975	HFM1, ATP-dependent DNA helicase homolog	Chr1,Rv,91625633	3305	1	0
NM_198393	testis expressed sequence 14 isoform a	Chr17,Rv,54064196	2627	1	0
NM_005789	proteasome activator subunit 3 isoform 2	Chr17,Fw,38226346	2582	0	0
NM_021251	calpain 10 isoform g	Chr2,Fw,241193520	2288	2	2
NM_016476	APC11 anaphase promoting complex subunit 11	Chr17,Fw,77397665	2132	0	26
NM_018057	solute carrier family 6, member 15 isoform 1	Chr12,Rv,83811936	1803	1	0
NM_133474	zinc finger protein 721	Chr4,Rv,457939	1528	1	82
NM_006372	synaptotagmin binding, cytoplasmic RNA	Chr6,Rv,86445170	1107	1	182
NM_003217	testis enhanced gene transcript	Chr12,Fw,48421607	916	2	1919
NM_002556	oxysterol binding protein	Chr11,Rv,59151068	694	0	0

K

Genomic position	Expression level (ppm)	Overlap with Ensemble	Overlap with spliced ESTs
Chr1,Fw,58869091	26218	0	0
Chr17,Rv,19290780	4528	1	0
chr5,Fw,167975894	3325	0	0
Chr5,Fw,33198019	1460	1	0
Chr6,Fw,154939112	1006	1	0
ChrX,Rv,39532410	860	1	0
Chr8,Rv,68000330	620	0	0
Chr1,Rv,118959813	531	1	0
Chr21,Rv,121549439	506	0	0
Chr22,Fw,41391194	496	1	0

Supplemental Figure 1 Characterization of TSS Seq

(A) Statistics for the TSS Seq data used in this study. (B) Statistics for the TSCs and AP genes identified in this study. (C) Mapped positions of the TSS Seq tags relative to the RefSeq genes are shown. %full was calculated as (#"upstream"+#"first exon"+#"intron")/(#"total tags") within the RefSeq region. (D) Length distribution of the represented mRNA of the RefSeq genes. Note that TSSs of the long mRNAs are also represented in this dataset because random primers were used for the synthesis of first strand cDNA. Further experimental validation of the TSS Seq analysis using HEK293 cells is provided in our previous paper (Tsuchihara et al. Nucleic Acids Res 37, 2249-63 (2009)). (E) Consensus sequences observed around the TSCs at >5 ppm (left panel) and the TSCs at <5 ppm (right panel). (F) Number of TSCs identified using different bin sizes for clustering. Note that the number of the TSCs at >5 ppm remained almost the same regardless of the clustering bin size. (G) Distribution of the distance between the 5'-ends of RefSeq transcript models (putative APs represented in RefSeqs). Populations of the APs that are separated by >500 bp in the respective databases are shown in the

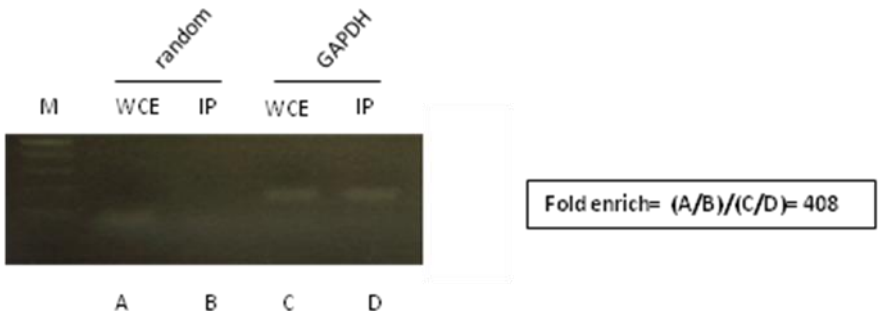
margin. (H) Comparison of expression levels between those observed by TSS-Seq and RNA-Seq. Left panel: Concentration (ppm) of the TSS Seq tags (y-axis) versus rpkm (read per kb mRNA) of RNA Seq (x-axis). Right panel: Concentration of the TSS tags (y-axis) and the RNA-Seq tags for the newly identified (intronic) APs. Correlation coefficients between the TSS-Seq and RNA-Seq data (total RNA) are shown in the margin in each panel. The population of the TSCs at >5 ppm is indicated by a dashed box. Note that the correlation in the right panel was calculated using the concentration (ppm) of the RNA-Seq tags (not normalized to transcript length) that overlapped the intronic TSCs because the entire transcript structures were unknown. We hypothesize that significant correlation between TSS-Seq and RNA-Seq less than the left panel may be accounted for by the lack of a proper normalization method. (I-K) Top ten TSCs with the highest expression levels in DLD-1 cells evaluated by TSS-Seq. I: total TSCs; J: novel (intronic) TSCs; K: iTSCs; Fw: forward strand; and Rv: reverse strand of the genome according to the UCSC information.

Related discussion on the expression levels of TSCs:

Usually, 30 μg total RNA is recovered from $3x10^6$ DLD-1 cells (cultured in a 10 cm dish). If we assume 5% are polyA+ RNA, there are 1.5 μg polyA+ RNA per $3x10^6$ cells. If we further assume that the molecular weight of an mRNA is 2 kb x 330 (molecular weight per base), this corresponds to 0.5 million copies of mRNA per cell. We cannot directly estimate the recovery rate of RNA from a cell, but for DNA, usually 5 μg of genomic DNA is recovered, which is consistent with the amount described in the manufacturer's protocol. This indicates that the recovery rate of the genomic DNA is >50%. Based on these observations, we estimated that there are 0.5-1.0 million copies

of mRNAs per cell, and 1 ppm corresponds to 1 copy per cell.

A



B

	#total peak	#peak within RefSeq region	#peak outside of RefSeq region	estimated false detection rate
fold (IP/WCE in 120 bp bin)= 1	3428818	1871740	1557078	0.5
2	686578	402512	284066	0.2
3	158725	109362	49363	0.09
4	63293	49993	13300	0.03
5	39150	32374	6776	0.007
6	29455	24750	4705	0.002
7	22772	19375	3397	0.0003
8	18597	15997	2600	6E-05
9	15187	13204	1983	9E-06
10	12659	11077	1582	1E-06

C

	#total peak	#peak in RefSeq region	#peak outside of RefSeq region
window size= 60 bp	77591	61135	16456
120 bp	39150	32374	6776
180 bp	18751	15961	2790

D

	#total peak	#peak in RefSeq region	#peak outside of RefSeq region
DLD-1	39150	32374	6776
HEK293	43214	37696	5518
MCF-7	28962	26137	2825
TIG-3	9099	8588	511

E

genic TSCs

# cell types having the expression	# Total TSCs	# TSCs overlapping pol II binding site
1	369533	5600(1.5%)
2-3	221767	5111(2%)
4-5	58866	2599(4%)
6-7	19519	1689(9%)
>=8	9656	2762(29%)

intergenic TSCs

# cell types having the expression	# Total iTSCs	# iTSCs overlapping pol II binding site
1	264642	1821(0.7%)
2-3	78026	3177(4%)
4-5	14670	1329(9%)
6-7	5169	752(15%)
>=8	8585	763(23%)

F

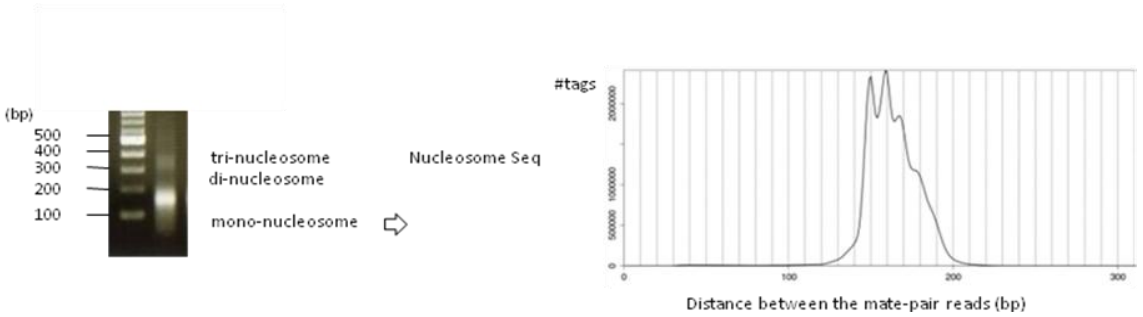
exonic TSCs

expression level	DLD-1 (pol II plus)	HEK293 (pol II plus)	MCF-7 (pol II plus)	TIG-3 (pol II plus)
0-1ppm	21,411 (141; 1%)	23,810 (228; 1%)	1,1360 (113; 1%)	10,426 (12; 0.1%)
1-2.5ppm	5,249 (59; 1%)	4,881 (54; 1%)	4,911 (43; 1%)	6,009 (11; 0.1%)
2.5-5ppm	2,165 (29; 1%)	2,242 (43; 2%)	2,196(31; 1.4%)	2,242 (11; 0.5%)
5-10ppm	1,169 (38; 3%)	1,318 (35; 2%)	1,278 (43; 3%)	1,337 (13; 1%)
>10ppm	1,092 (225; 11%)	1,272 (115; 9%)	1,394 (111; 8%)	1, 241(53; 4%)

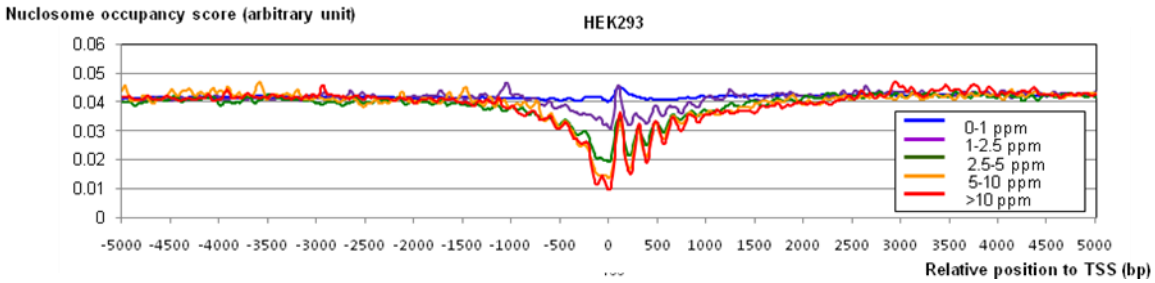
Supplemental Figure 2 Statistics for ChIP Seq analysis of pol II

(A) Enrichment of known pol II binding sites by individual quantitative PCR. The PCR primers used were 5'-CGTGTCCCCCATATCAGAAC-3' and 5'-TCAGCCTCAGTCTCCCTTGT-3' for the random genomic region and 5'-CGTAGCTCAGGCCTCAAGAC-3' and 5'-GTCGAACAGGAGGAGCAGAG-3' for GAPDH (-95 bp to +49 for the reported TSS). The enrichment rate, which is shown in the margin, was calculated based on the Ct value from real time PCR. (B) and (C) Number of peaks (putative pol II binding sites) when different cutoffs were used. (A)-(C) Representative results for DLD-1 cells; similar results were obtained for the other cell lines (data not shown). (D) Number of peaks identified in the indicated cell types, using the parameters described in the main text. (E) Frequency of the "weak" TSCs (whose expression levels are <5 ppm in all cell types) that overlapped the pol II binding sites. First column: number of the cell types in which expressions of the TSCs (>0 ppm) was observed. (F) Frequency of the TSCs that are located in internal exons of RefSeq at the indicated expression levels (exonic TSCs). Populations of the TSCs that overlapped the pol II binding sites is shown in parentheses.

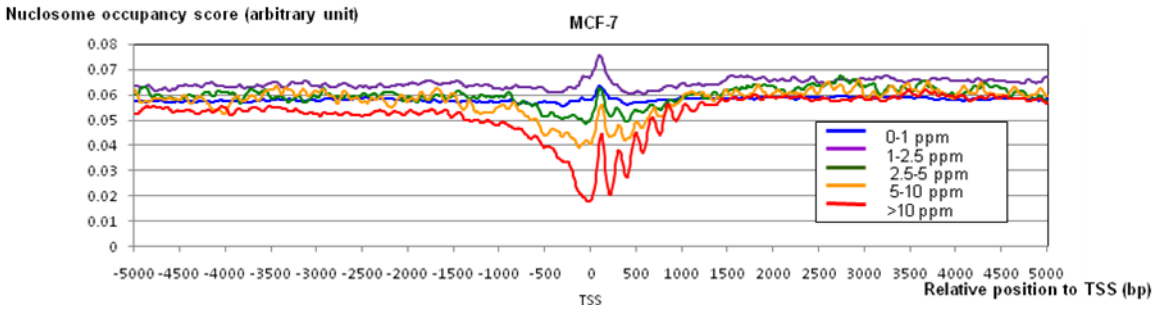
A


(bp)
500
400
300
200
100
tri-nucleosome
di-nucleosome
mono-nucleosome
Nucleosome Seq
#tags
Distance between the mate-pair reads (bp)

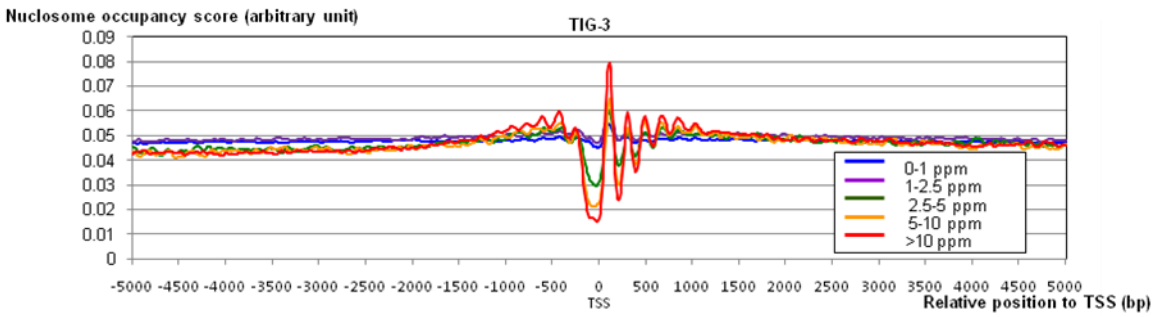
B


Nuclosome occupancy score (arbitrary unit)
HEK293
0-1 ppm
1-2.5 ppm
2.5-5 ppm
5-10 ppm
>10 ppm
Relative position to TSS (bp)

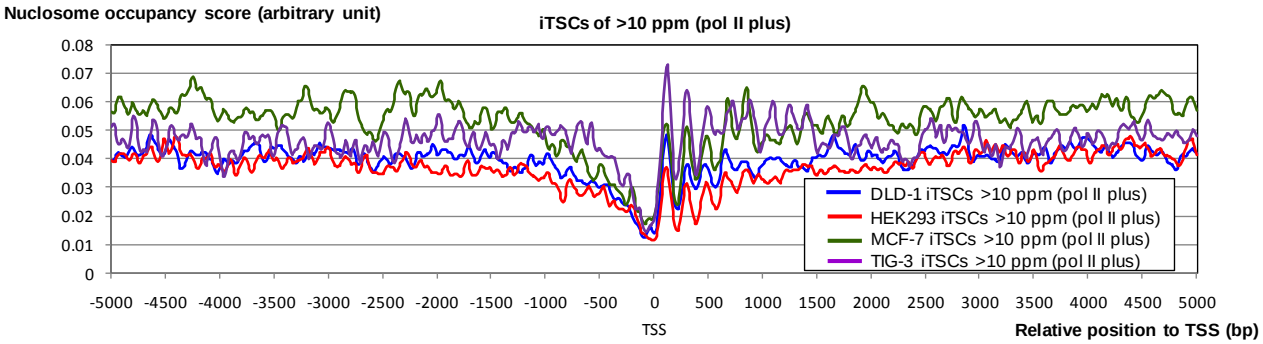
C


Nuclosome occupancy score (arbitrary unit)
MCF-7
0-1 ppm
1-2.5 ppm
2.5-5 ppm
5-10 ppm
>10 ppm
TSS
Relative position to TSS (bp)

D


Nuclosome occupancy score (arbitrary unit)
TIG-3
0-1 ppm
1-2.5 ppm
2.5-5 ppm
5-10 ppm
>10 ppm
TSS
Relative position to TSS (bp)

E



Supplemental Figure 3 Nucleosome structure around the TSCs

(A) Characterization of the micrococcal nuclease-digested DNA used for the Nucleosome Seq analysis by agarose gel electrophoresis and the mate-pair Illumina reads (left and right panels, respectively). In the right panel, the distribution of the distance between the mate-pair reads calculated using 2,143,342 paired-end sequence tags is shown. The figure shows data from DLD-1 cells, although similar results were obtained for the other cell lines. (B)-(D) Nucleosome occupancy score (y-axis) around the TSCs (x-axis) in HEK293 (B), MCF-7 (C) and TIG-3 (D) cells. The TSC populations are as indicated in the inset. (E) Nucleosome occupancy score of the iTSCs (>10 ppm) that overlapped the pol II binding signals in the indicated cell types.

Related discussion on statistical analysis for differences in nucleosome patterns:

To evaluate the statistical significance for the nucleosome patterns in the surrounding region of the TSCs compared to the background noise, we calculated the average and standard deviation of the nucleosome occupancy score at distal regions (i.e., from -5 kb to -2.5 kb and from +2.5 kb to +5 kb). Deviations of nucleosome occupancy scores in the proximal regions of TSCs were calculated assuming a normal distribution of scores

in the distal regions. The calculated statistical significances (p-values) are as follows:

	P-value
Fig.3A blue line	1.0E-05
Fig.3A purple line	6.2E-62
Fig.3A green line	1.2E-72
Fig.3A yellow line	1.4E-102
Fig.3A red line	8.4E-127
Fig.3B blue line	7.5E-06
Fig.3B red line	1.3E-285
Fig.3C blue line	4.1E-16
Fig.3C green line	5.1E-24
Fig.3C yellow line	2.8E-22
Fig.3C red line	5.8E-72
Fig.5B blue line	1.7E-04
Fig.5B yellow line	1.6E-09
Fig.5B red line	2.5E-28
Fig.5B green line	3.6E-27

	P-value
Sup Fig.3B blue line	8.6E-06
Sup Fig.3B purple line	1.8E-27
Sup Fig.3B green line	1.4E-81
Sup Fig.3B yellow line	2.1E-119
Sup Fig.3B red line	3.8E-133
Sup Fig.3C blue line	5.6E-05
Sup Fig.3C purple line	5.1E-04
Sup Fig.3C green line	2.1E-08
Sup Fig.3C yellow line	1.8E-18
Sup Fig.3C red line	1.4E-41
Sup Fig.3D blue line	6.7E-15
Sup Fig.3D purple line	1.5E-03
Sup Fig.3D green line	1.1E-24
Sup Fig.3D yellow line	3.0E-42
Sup Fig.3D red line	3.6E-51
Sup Fig.3E blue line	1.5E-29
Sup Fig.3E red line	4.5E-26
Sup Fig.3E green line	8.7E-28
Sup Fig.3E purple line	4.4E-18
Sup Fig.3E blue line	6.3E-03
Sup Fig.9A blue line	1.3E-27
Sup Fig.9A red line	1.6E-03
Sup Fig.9A green line	4.3E-22
Sup Fig.9A purple line	1.0E-03

A

	EPD	RefSeq	DBTSS
#total putative promoters	196	31554	32037
#represented putative promoters in this study	161 (82%)	19,879 (63%)	29,474 (92%)
#total AP genes	64	745	7894
#represented AP genes in this study	59 (92%)	582 (78%)	5055(64%)

B

expression levels of TSCs	#TSC	CAGE	GENCODE
0-1 ppm	544,168	208568(38%)	8,051(1%)
1-2.5 ppm	105,043	43725(42%)	1,779(2%)
2.5-5 ppm	21,612	10019(46%)	418(2%)
5-10 ppm	6,774	3549(52%)	139(2%)
>10 ppm	5,093	3079(60%)	83(2%)
<5 ppm, expressing cell types >=8, pol II plus	1018	904(89%)	10(1%)

C

expression levels of TSCs	>500 bp from the 5'-ends of GENCODE	>500 bp from the 5'-ends of GECOMDE (pol II plus)	#TSCs of intergenic to GENCODE genes	#TSCs of intergenic to GENCODE genes (pol II plus)
0-1 ppm	547672	10801	540078	10755
1-2.5 ppm	107297	4253	105658	4277
2.5-5 ppm	23548	2543	23197	2574
5-10 ppm	8772	2224	8662	2229
>10 ppm	11998	6897	11971	6928
<5 ppm, expressing cell types >=8, pol II plus	-	2747	-	2754

Supplemental Figure 4 Overlap of TSCs with pre-existing databases

(A) Number of records is shown for each of the indicated categories. EPD (Eukaryotic Promoter Database), RefSeq and DBTSS (DataBase of Transcription Start Sites) data were current as of June 1, 2009. AP: alternative promoter; AP gene: genes having multiple promoters. (B) Population of the TSCs covered by the CAGE

(http://fantom.gsc.riken.jp/4/) or GENCODE (http://genome.crg.es/gencode/) databases. The data were current as of March 1, 2010. (C) Number of the putative novel promoters against the GENCODE database. Second and third columns: TSCs that overlapped genic regions, as defined by GENCODE, but are separated by >500 bp from the 5'-ends of the transcript models; fourth and fifth columns: TSCs located >50 kb outside of the genic regions, as defined by GENCODE. Populations of the TSCs that overlapped the pol II binding sites in at least one of the four examined cell types (DLD-1, HEK293, MCF-7 and TIG-3) are shown in the third and fourth columns.

A

expression levels of TSCs	total	FLJ	MGC
>5 ppm	7206	3792	3414
<5 ppm	8874	6566	2308
total	16080	10358	5722

B

expression levels of TSCs	total	possible NMD	ORF <100 aa	UTR >750 bp	either of the translation caveats
0-1 ppm	4882	723	1309	1542	2588(53%)
1-2.5 ppm	2376	372	505	432	953(40%)
2.5-5 ppm	1616	246	204	232	511(32%)
5-10 ppm	1533	202	159	142	462(30%)
>10 ppm	5673	677	483	320	879(15%)
<5 ppm; expressing cell types >=8; pol II plus	1258	194	86	114	293(23%)

C

expression levels of TSCs	#total shotgun sequenced cDNA	#tag	average coverage	assembled length (bp)
>5 ppm	398	170,625,240	303	1515
<5 ppm	448	207,385,452	333	1579
total	846	378,010,692	319	1549

D

expression levels of TSCs	total	possible NMD	ORF <100 aa	UTR >750 bp	either of the translation caveats
0-1 ppm	213	35	85	32	108(51%)
1-2.5 ppm	139	23	62	17	73(53%)
2.5-5 ppm	96	20	35	7	44(46%)
5-10 ppm	108	23	42	10	49(45%)
>10 ppm	290	62	91	24	112(39%)
<5 ppm; expressing cell types >=8; pol II plus	65	19	23	5	30(46%)

Supplemental Figure 5 Statistics for complete cDNA sequencing

(A) Statistics calculated for independent populations. For the dataset used in this study, non-redundant cDNAs were first selected from the FLJ and then from the MGC collections. However, essentially the same results were obtained from the analysis, even with a dataset for which the selection order was reversed (data not shown). (B) Number of the TSCs that exhibit the indicated possible translation caveat. TSCs were associated with the MGC or FLJ cDNAs when the 5'-ends of the cDNAs were located within 500 bp of the TSCs. The indicated expression levels represent the maximum expression levels of the TSCs over twelve cell types. aa: amino acids; NMD: non-sense-mediated mRNA decay; and UTR: untranslated region. (C) Statistics for the shotgun sequencing of the cDNAs. (D) Results of an analysis similar to (B) that was used for the newly sequenced cDNAs.

A

target	fold (N/C)	PCR primer	PCR primer
SCRNA 17	46.3	CGGAGCCTGAGAGGATTATG	TCACCTGGTTCCCAGAACTC
SCARNA 2	49.4	ACGCGTGAGTGTGTGAGTGT	GCAGGAGGAGAGCTTTTCATT
SCARNA 12	76.5	TGATGAGACTAAGGCGAATGC	GCACCAGAAATGAAGGCAAG
SCARNA 10	47.2	AATCTTGGTGGGCGATACAG	CCCTGATACCCTGAACATGC
SCARNA 13	25.0	GTAGTCTTGGAGCCGCACAG	GTGGCAACAGTGACCAGAAA
SNORD3AT	89.6	CGTGTAGAGCACCGAAAACC	CACTCCCCAATACGGAGAGA
SNORNA 73A	103.5	CTCTGTCCAAGTGGCGTAGG	GACAGGACTCTGGGAAGCTG

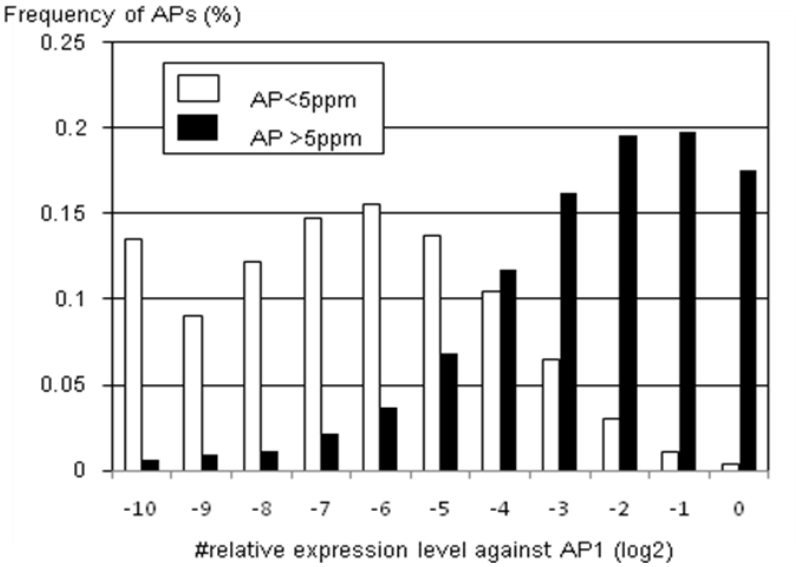
B

category	expression level in DLD-1	total	>=3 tag support in polysome	polysome enriched (p<0.01)	nucleus enriched (p<0.01)
TSCs	**0-1 ppm**	66880	10125(15%)	1564(2%)	9165(14%)
	1-2.5 ppm	5644	2730(48%)	461(8%)	943(17%)
	2.5-5 ppm	2392	1701(71%)	311(13%)	289(12%)
	5-10 ppm	1890	1486(79%)	297(16%)	189(10%)
	>10 ppm	4375	3802(87%)	975(22%)	378(9%)
<5 ppm, expressing cell types >=8, pol II plus	**<5 ppm**	2762	1921(70%)	296(11%)	174(6%)
assembled cDNAs	**>5 ppm**	178	154(87%)	36(20%)	18(10%)
	<5 ppm	668	276(41%)	43(6%)	137(21%)
CCDS	**>5 ppm**	5853	5627 (96%)	2999(51%)	1367 (23%)
	<5 ppm	8951	4291 (48%)	1546(17%)	305 (3%)

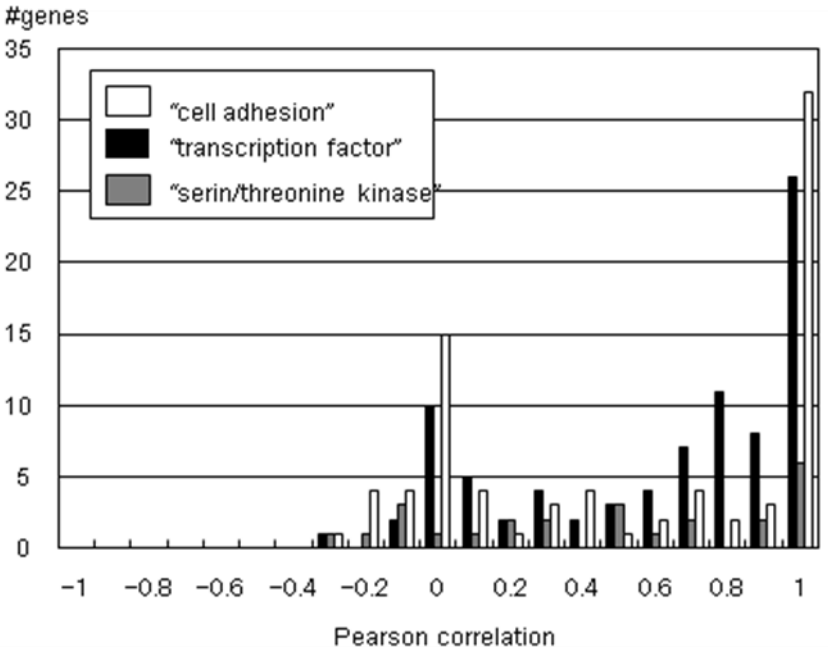
Supplemental Figure 6 Characterization of the RNA Seq library for the subcellular fractions

(A) Evaluation of the enrichment for the nuclear fraction from the cytoplasmic fraction using several nuclear RNAs. Fold differences, which were calculated based on the results of real time PCR, are shown. The PCR primers are also shown. (B) Summary of the statistics calculated for the individual populations. The computational procedures used to calculate the statistical significance were as described in the Methods. Consensus coding sequences (CCDSs) were obtained from NCBI as of June 1, 2009.

A


Frequency of APs (%)
0.25
0.2
0.15
0.1
0.05
0
AP<5ppm
AP >5ppm
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
#relative expression level against AP1 (log2)

B


#genes
35
30
25
20
15
10
5
0
"cell adhesion"
"transcription factor"
"serin/threonine kinase"
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
Pearson correlation

Supplemental Figure 7 Differential expression patterns of the APs belonging to different GO categories

(A) Distribution of relative expression levels of AP2 (second or later AP) against AP1 (the AP that had the highest expression levels in the twelve cell types). Black and gray bars represent the populations of the AP2 at >5 ppm and the AP2 at <5 ppm, respectively. (B) Distribution of Pearson correlation between the APs for the genes annotated with the GO annotation terms “cell adhesion” (white bars), “serine/threonine kinase” (gray bars) and “transcription factor” (black bars).

A

	#total iTSCs	average G+C%	%iTSCs in CpG island	evolutional conservation score	relative position within the ORF (N terminal: 0%; C terminal: 100%)	%iTSCs overlapping pol II
0-1 ppm	328148	41%	1%	0.10	52%	2%
1-2.5 ppm	30540	42%^{*1,2}	4%^{*a,b}	0.12^{*3,4}	37%^{*c,d}	6%^{*e}
2.5-5 ppm	7122	43%	7%	0.15	29%	12%
5-10 ppm	3005	45%^{*1}	11%^{*a}	0.20^{*3}	22%^{*c}	16%^{*e}
>10 ppm	3034	47%	15%	0.22	19%	23%
<5 ppm; expressed in >8 cell types and overlapping pol II	763	64%^{*2}	70%^{*b}	0.20^{*4}	7%^{*d}	-

*1:p=1e-57,*2:p<1e-200,*3:p=1e-72,*4:p=1e-51,
*a:p=1e-79,*b:p<1e-200,*c:p=1e-54,*d:p=1e-62,*e:p=1e-89

B

expression levels of iTSCs	#total cDNAs	FLJ	MGC	possible NMD	ORF<100 aa	UTR >750 bp	either of the translation caveats
>5 ppm	395	357	38	156	225	98	325(82%)
<5 ppm	1617	1539	78	502	1003	555	1348(83%)
total	2012	1896	116	658	1228	653	1673(83%)

C

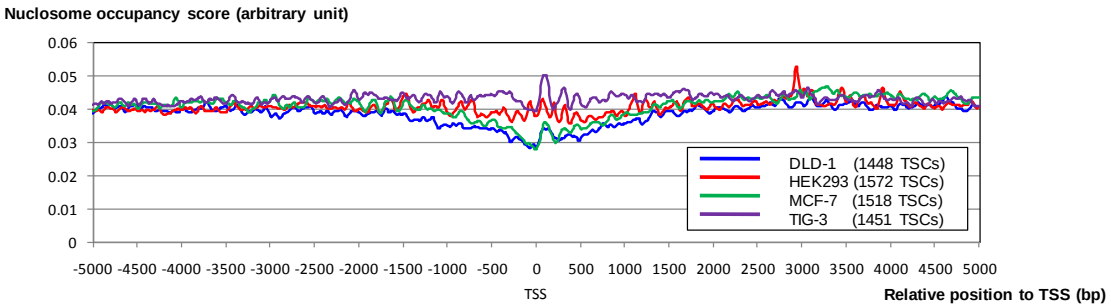
expression levels of iTSCs	#total shotgun sequenced cDNAs	#tag	average coverage	assembled length	possible NMD	ORF<100 aa	UTR >750 bp	either of the translation caveats
>5 ppm	361	2964285	281	1198	79	293	85	316(88%)
<5 ppm	103	418616	285	1259	39	78	44	91(88%)
total	464	3382901	283	1212	118	371	129	407(88%)

Supplemental Figure 8 Statistics for the iTSC

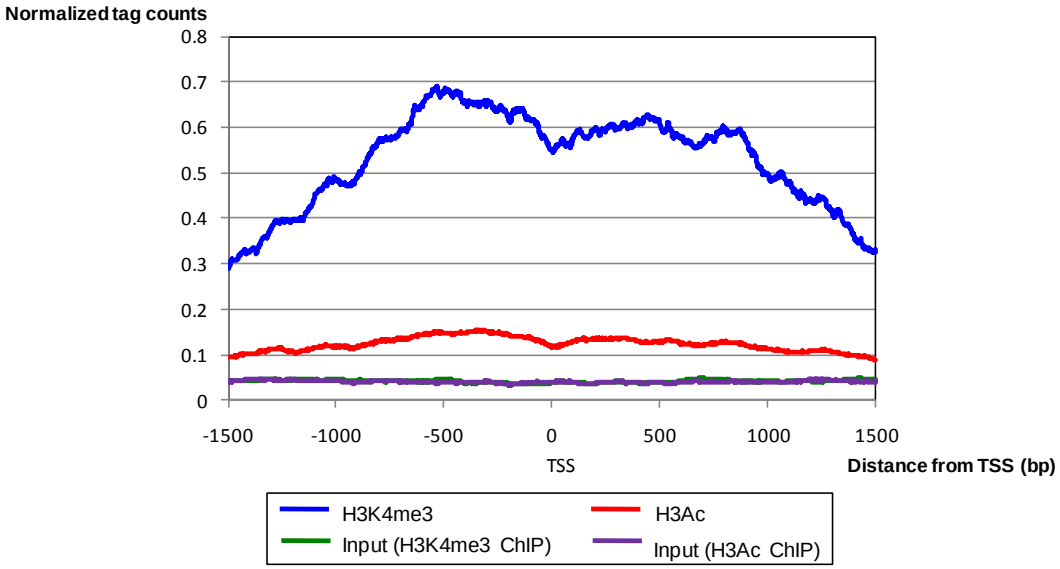
(A) Results of an analysis similar to that shown in Table 2. Statistical significances for differences between populations indicated by asterisks were evaluated by the Wilcoxon signed rank test (*1-4) or the proportion test (*a-e) and are shown in the margin. The legends for the Table are as described in Table 2. (B) Sequence features characteristic of the intergenic TSCs for the indicated categories. Legends for the Table are as described in Table 2. Further details of the characterization of the iTSCs will be published

elsewhere. (C) Results of a similar analysis that used the newly shotgun sequenced cDNAs.

A



B



Supplemental Figure 9 Chromatin structure in the absence of transcription

(A) The nucleosome occupancy score (y-axis) around the putative TSCs (x-axis) in the cell types indicated in the inset. Those TSCs had expression levels of >5 ppm in other cell types but were not expressed in the indicated cell type. Number of the analyzed TSCs in each cell type is shown in parentheses. (B) Histone modification patterns in DLD-1 cells around the putative TSCs that are not expressed in this cell type (population represented by blue line in (A)). Also, note that H3K4me3 and H3Ac show different distribution patterns around these TSCs.

TSSseq Sample	Accession Number
DLD-1 / HEK293 / MCF-7 / TIG-3	SRA003625
Beas2B	SRA008162
Ramos	SRA008164
Fetal Brain	DRA000023
Fetal Heart	DRA000024
Fetal Kidney	DRA000025
Brain	DRA000026
Heart	DRA000027
Kidney	DRA000028

Nucleosome Sample	Accession Number
DLD-1	DRA000003
HEK293	DRA000038
MCF-7	DRA000077
TIG-3	DRA000074

ChIPseq Sample	Accession Number
DLD-1 pol II	DRA000007/DRA000008
HEK293 pol II	DRA000036/DRA000037
MCF-7 pol II	DRA000192/DRA000193
TIG-3 pol II	DRA000190/DRA000191
DLD-1 H3K4me3	DRA000278/DRA000280
DLD-1 H3Ac	DRA000286/DRA000288

RNAseq Sample	Accession Number
DLD-1 Cytoplasma	DRA000005
DLD-1 Nuclear	DRA000006
DLD-1 Polysome	DRA000003
DLD-1 TotalRNA	DRA000308

Supplemental Figure 10 Accession numbers for sequence data used in this study