190–201 *Nucleic Acids Research, 2011, Vol. 39, No. 1* *Published online 14 September 2010*
doi:10.1093/nar/gkq775

Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq

Hao Sun[1], Jiejun Wu[2], Priyankara Wickramasinghe[1], Sharmistha Pal[1], Ravi Gupta[1], Anirban Bhattacharyya[1], Francisco J. Agosto-Perez[1], Louise C. Showe[1], Tim H.-M. Huang[2] and Ramana V. Davuluri[1,*]

[1]Center for Systems and Computational Biology, Molecular and Cellular Oncogenesis Program, The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104 and [2]Human Cancer Genetics, Comprehensive Cancer Center, Dept. of Mol. Virology, Immunology & Med. Genetics, Ohio State University, 460 W 12th Avenue, BRT, Columbus, OH 43210, USA

Received May 28, 2010; Revised August 13, 2010; Accepted August 17, 2010

ABSTRACT

Alternative promoters that are differentially used in various cellular contexts and tissue types add to the transcriptional complexity in mammalian genome. Identification of alternative promoters and the annotation of their activity in different tissues is one of the major challenges in understanding the transcriptional regulation of the mammalian genes and their isoforms. To determine the use of alternative promoters in different tissues, we performed ChIP-seq experiments using antibody against RNA Pol-II, in five adult mouse tissues (brain, liver, lung, spleen and kidney). Our analysis identified 38 639 Pol-II promoters, including 12 270 novel promoters, for both protein coding and non-coding mouse genes. Of these, 6384 promoters are tissue specific which are CpG poor and we find that only 34% of the novel promoters are located in CpG-rich regions, suggesting that novel promoters are mostly tissue specific. By identifying the Pol-II bound promoter(s) of each annotated gene in a given tissue, we found that 37% of the protein coding genes use alternative promoters in the five mouse tissues. The promoter annotations and ChIP-seq data presented here will aid ongoing efforts of characterizing gene regulatory regions in mammalian genomes.

INTRODUCTION

Recent analyses of mammalian genomes and microarray data suggest that the majority of mammalian genes generate multiple transcripts and protein isoforms with distinct functional roles. This transcript diversity is generated, in part, through the use of alternative promoters (1) and alternative splicing (2), which produce pre-mRNA and mRNA isoforms respectively. The use of alternative promoters plays a fundamental role in regulating different gene isoforms, e.g. *LEF1*, *TP73*, *RUNX1* and *MYC* in various mammalian tissues and at different developmental stages. For example, in case of *LEF1,* the protein isoforms generated from the two promoters perform opposing biological functions. While the full-length LEF1, transcribed from upstream promoter, interacts with β-catenin and regulates Wnt target genes, the shorter isoform is incapable of binding β-catenin and suppresses the regulation of Wnt targets through β-catenin (3). Moreover, activation of upstream promoter and silencing of the internal promoter is observed in most colon cancers (4). Therefore, identifying primary and alternative gene promoters in various normal tissues is critical to understanding a diversity of physiological processes associated with normal and diseased states in different tissues. The advent of high-throughput molecular technologies and computational methods to support this technology has significantly improved our ability to annotate mammalian gene regulatory regions. High-throughput technologies, such as cap analysis gene expression (CAGE); chromatin immunoprecipitation (ChIP) followed by microarray analysis (ChIP–chip); ChIP coupled with pair-end ditag sequencing analysis (ChIP-PET) (5,6); and, more recently, ChIP coupled with sequencing (ChIP-seq) (7) are enabling the genome-wide identification of alternative promoters and their patterns of use. This information will help us to

*To whom correspondence should be addressed. Tel: +215 495 6903; Fax: +215 495 6848; Email: rdavuluri@wistar.org
Present address:
Hao Sun, Li Ka Shing Institute of Health Sciences and Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China.

The authors wish it to be known that, in their opinion, the first five authors should be regarded as joint First Authors.

understand the use of alternative promoters in a wide variety of cell/tissue types, different developmental stages and their misuse in disease conditions.

Growing evidence suggests that about half of the mammalian genes have multiple alternative promoters that can span up to thousands of bases (8–12). For example, a comprehensive analyses of 1% of the human genome in 16 diverse human cell lines, using transient transfection reporter assays demonstrated the presence of functional alternative promoters in >20% of genes (12). Similarly, it has been reported that 35% of 100 human erythroid genes examined have alternative promoters and that 24% of active genes in human fibroblast cells possess multiple promoters (13). This is quite a high percentage of genes showing multiple promoter usage in a single biological process or cell type suggesting extensive use of multiple promoters by mammalian genes. The knowledge of alternative promoter usage in different mammalian tissues is very limited and cannot be addressed without high-resolution genome-wide mapping of the promoter regions. However, the high-throughput approaches, such as CAGE (14), deepCAGE (15), ChIP–chip (16,17) or ChIP-seq (7), to annotate promoters at genome level need to be applied with caution because of the inherent problems with each method. For example, cytoplasmic enzyme complexes can add caps to 5′-monophosphate RNA molecules generated by ribonuclease cleavage (18), and hence CAGE tags could represent 5′ ends of RNAs generated by cleavage and subsequent re-capping (19). CAGE analysis can also capture some non-capped transcripts that may represent cleaved decaying mRNA (20). Furthermore, a large number of CAGE tags are distributed throughout the gene transcripts rendering it inefficient as a sole source of promoter identifier. Previously, we (16) and others (17) have performed ChIP–chip analyses to identify the activity of mammalian promoters across different cell and tissue types. However, ChIP–chip requires design of genome-wide microarray to probe the ChIP-bound DNA sequences. Additionally, with either ChIP-chip or ChIP-seq technology promoters cannot be identified solely on the presence of Pol-II enrichment on a genomic location because of its enrichment throughout the transcribed genomic region and lack of highly specific antibodies that can distinguish promoter bound Pol-II from elongating Pol-II. In order to overcome these limitations of previous studies, we pursued a combined Pol-II ChIP-seq and bioinformatics promoter prediction approach to identify promoter regions and their activity in five different mouse tissues. We provide a genome-wide catalog of active promoters in five tissues of adult mouse along with tissue-specific promoters that will help future studies of transcriptional regulation in mammalian genomes.

MATERIALS AND METHODS

Chromatin immunoprecipitation, massive parallel sequencing and real-time polymerase chain reaction

About 1 g of freshly dissected brain, kidney, liver, lung or spleen tissue from 2-month-old FBV mice was minced finely and cross-linked with 1% formaldehyde for 10 min at room temperature. To stop cross-linking, glycine was added to a final concentration of 0.125 M. Next, the tissue sample was treated to isolate individual cells and cross-linked chromatin was fragmented to a size range of 0.2–0.8 kb as previously described (21). ChIP was performed using 10 μg of Pol-II antibody bound Dynal magnetic beads. The antibodies against Pol-II were purchased from Abcam Inc (ab5408) and Santa Cruz Biotechnology (sc-899X). Following immunoprecipitation, the bound nucleoprotein complexes were extensively washed six times with wash buffer 1 and once with wash buffer 2 [Wash buffer 1: 50 mM HEPES-KOH (pH 7.55), 500 mM LiCl, 1 mM EDTA, 1.0% NP-40 and 0.7% Na-deoxycholate; wash buffer 2: TE containing 50 mM NaCl] and the ChIP enriched DNA was eluted and purified by phenol:chloroform:isoamyl alcohol extraction. This purified DNA (10 ng) was further processed according to the Illumina Inc. instructions to prepare the library for sequencing ChIP enriched DNA. For ChIP-qPCR, Pol-II ChIP was conducted the same way as described above and same amount of either input or Pol-II enriched DNA was PCR amplified in the presence of specific primers using the SYBR green-based detection (Applied Biosystems, Foster City, CA, USA) as described previously (22). The primers used were as follows: Promoter forward: 5′ gacggttggagaagaaggtg 3′, Promoter reverse: 5′ aggagaggaggaggttttgg 3′ and Control region forward: 5′ gtaacctctgccgttcagga 3′, Control region reverse: 5′ tttctccctttccggagatt 3′.

ChIP-seq data processing

We have adapted a similar approach used by previous published studies (7,23) for ChIP-seq data analysis. Briefly, the analysis involves the following steps: (i) Identification of statistically significant sequence read-enriched genomic regions (of length 1 kb). A region will be considered statistically significant if the number of reads within that region is higher than the number expected due to random background. We used Poisson distribution to estimate the background read count at a given significant level P ($P \leq 0.01$). (ii) Creating the read overlapping profile for each identified region from step 1, by extending the sequence reads from the 5′ end to the 3′ end of the reads up to 400 bp (the average length of the ChIP–DNA fragment sequenced from the Solexa GA with Illumina standard ChIP-seq protocol). (iii) Peak identification—by counting the number of overlapped reads at each nucleotide position and defining the genomic position with the highest number as the peak position within the 1 kb significant region.

Identification and annotation of Pol-II promoter peaks

To identify the Pol-II bound promoters from the ChIP-seq data we used our recently developed program to discriminate Pol-II enrichment peak associated with promoter region from peaks associated with non-promoter region (24). The method uses DNA sequence composition, physico-chemical-structural property of DNA sequences, CAGE tags, Pol-II and H3K4me3 enrichment profiles

from ChIP-seq data sets as features for discrimination. In total, 39 features were calculated for each peak and as described in ref. 24. Classification model using the aforementioned features was constructed with three different state-of-the-art ensemble and meta classifiers: Random forest (25), Bagging (26) and LogitBoost (27–30). The performance of the model was evaluated based on the promoter prediction metrics suggested by (31): sensitivity (SN), positive predictive value (PPV), correlation coefficient (CC) and true-positive cost (TPC). The performance measures were calculated for 10-fold cross-validation and independent test set.

For annotating the predicted promoters, we referred to gene information tracks from various sources available at UCSC genome browser. The tracks include protein coding and non-coding genes from Refseq gene, UCSC gene, Ensembl gene, Vega gene and miRNA. We also downloaded recently discovered large intervening non-coding RNAs (lincRNA) (32) information for annotating promoters related to non-coding genes. The other non-coding RNAs gene information including snoRNA and snRNA are part of the Refseq gene, UCSC gene, Ensembl gene and Vega gene models. A non-redundant set of coding and non-coding transcripts was generated after combing the transcript information from various gene models stated above. A total of 42 924 and 38 159 coding and non-coding transcripts, respectively, were obtained for annotation of Pol-II peaks (Supplementary Table S1A in additional data file 2). The known protein-coding and non-coding genes for organism other than mouse were also considered for annotation. This was done to identify those promoters that are evolutionarily conserved and are known in other organisms but still unknown in mouse. The non-mouse gene track was also downloaded from UCSC genome browser and is referred as XenoRef Gene. A non-redundant set of promoters for XenoRef Gene track genes (Supplementary Table S1B in additional data file 2) was also generated for annotation.

We divided the result of promoter annotation into three categories: (i) known promoters, (ii) novel promoters and (iii) novel promoters-unassigned. All those promoters that overlapped with first exon of known mouse transcripts are categorized as 'known promoters'. The rest of the promoters are categorized as 'novel promoters'. Further, the novel promoters are assigned to known genes if those fall inside a transcript of known mouse genes (or orthologous non-mouse genes) or within −10 kb of the 5′end of known mouse genes (or orthologous non-mouse genes). The rest of the novel promoters are left as 'Novel promoters –unassigned'. Those promoter peaks that overlap with the 5′ends of both protein coding and non-coding genes are assigned to both.

Cloning novel promoters and luciferase assay

PCR on mouse genomic DNA was performed with specific primers to amplify 0.5–1.0 kb of the randomly selected novel promoter (0.5–1.0 kb) and non-promoter (~0.9 kb) regions. The genomic coordinates of each cloned promoter/non-promoter region is provided in Supplementary Table S2. As a positive control for the luciferase assay, the promoter of *DLL1* gene was also amplified. Amplified PCR products were cloned in pCRII vector (Invitrogen) and the clones were confirmed by sequencing. The confirmed clones were subcloned in the promoter less luciferase vector pGL3basic (Promega Inc.). DNA for the pGL3 basic constructs (1.8 μg for calcium chloride method, 0.9 μg for Lipofectamine 2000 or Fugene) along with pGL4-renilla-luciferase (0.2 μg for calcium chloride method, 0.1 μg for Lipofectamine 2000 or Fugene) were individually transfected in HEK293 (calcium chloride-based transfection), A549, HepG2 (Lipofectamine 2000, Invitrogen Inc.), NIH3T3 and DAOY (Fugene, Roche Inc.) cell lines in triplicates in six-well plate for about 48 h. After 48 h, cells were washed and lyzed in 200 μl of passive lysis buffer provided in the dual luciferase assay kit (Promega Inc.). The lysates were cleared by centrifugation and luciferase assay was performed with 5–20 μl of the lysate as per manufacturer's instructions (Promega Inc.). Renilla luciferase activity was used to normalize for transfection efficiencies and fold enrichment of luciferase activity was calculated relative to the vector backbone (pGL3 basic alone).

Core promoter identification and analysis

We searched for core-promoter elements for each identified promoter, by scanning a sequence of length 200 bp (–100 to +100 around the Pol-II peak position). The sequences were analyzed by MATCH program (33) for the five known core-promoter elements (INR, TATA, MTE, BRE and DPE) using the position weight matrices published earlier (34). We used the default parameters for the MATCH search with the following cutoffs for each element (INR-0.85 and 0.8; TATA-0.73 and 0.58; MTE-0.79 and 0.53; BRE-0.70 and 0.65; DPE-0.92 and 0.92). In this process, search was done first for the INR element because it is the most abundant core promoter element, and if found, that position plus 3 was considered as the true TSS for the corresponding promoter. If INR was not found, the rest of the elements (TATA, MTE, BRE and DPE) were searched in that order of importance, and then the TSS was assigned relative to the first element found, by adjusting the relative distance between the TSS and the corresponding element (34). The next priority was given to TATA because though MTE is the second most abundant core promoter element it shows high co-occurrence with INR and the co-occurrence tendencies of TATA element with others is least. If there are more than one element identified in a sequence, priority is given to the one with highest score. Once this assignment is done, we looked for the presence of the remaining core promoter elements in that promoter. If none of the elements were present, the original peak position was considered as the true TSS.

RESULTS

Pol-II ChIP-sequencing data quality

To identify the active promoter regions in the adult mouse genome, we used the ChIP-seq approach to find genome-wide binding regions of Pol-II in five mouse

Table 1. Summary of Pol II Chip-seq data processing for five tissues

Number of	Brain	Kidney	Liver	Lung	Spleen
Read obtained from Solexa sequencer	1 79 59 062	2 05 12 406	2 49 55 539	1 76 62 905	2 10 81 878
Read aligned back to mouse genome (mm9)	1 02 92 266	1 37 86 546	1 49 64 945	1 08 46 984	1 32 83 139
Read in identified enriched region	54 82 173	67 44 056	68 23 960	53 09 423	54 24 148
Peaks identified	91 712	81 375	57 816	72 971	31 594
Peaks predicted as promoters	18 900	17 548	14 145	19 415	10 589

After alignment to the mouse genome, the enriched regions were first identified and, using a statistical cutoff, the significantly enriched peaks were determined. The promoter prediction algorithm identified the significantly enriched peaks that reside in promoter regions.

tissues (brain, kidney, liver, lung, and spleen). Through mapping of the Pol-II binding regions, we seek to investigate the usage of alternative promoters and uncover novel promoters in the mouse genome. Previous studies have indicated that performing two biological replicates for ChIP-seq studies is enough to achieve the sequencing depth required for robust identification of target binding sites (35,36), and hence we performed two replicates of Pol-II ChIP-seq experiment on each tissue and analyzed their overlap. Following the ENCODE consortium standards (36), we analyzed the agreement between our two biological replicates and since our results indicate a good correlation we combined the two datasets for further analysis (Supplementary Figure S1 in additional data file 1). Sequencing of Pol-II enriched DNA from two biological replicates in the five tissues yielded a total of over 102 million sequence reads of 36 bp length. Using the ELAND (Illumina, Inc., San Diego, CA, USA) and Bowtie (37) programs and allowing a maximum of two mis-matches, ~62% (63.2 million) of the reads were uniquely mapped back to the mouse reference genome (version mm9). The aligned reads were processed to identify significantly enriched Pol-II binding regions and significant peaks as described in 'Materials and Methods' section. Following a three-phase peak identification approach, we identified a total of 335 468 significantly Pol-II-enriched peaks across the five tissues with brain showing the maximum number of peaks followed by kidney, lung, liver and spleen (Table 1).

As Pol-II binding is expected to be highly enriched in promoter regions compared to intragenic locations, we looked at the enrichment profile of reads relative to known transcription start sites (TSS). The distribution of read counts per million mapped reads for each tissue indicates an increased enrichment of Pol-II near TSS as compared to intra and inter-genic regions (Figure 1A). To further verify the quality of the ChIP-seq experimental data, we performed ChIP-qPCR experiment on the ubiquitous *Polr2a* locus. We analyzed the enrichment of Pol-II at the promoter region as well as at a downstream region of *Polr2a* gene as indicated in Figure 1B. Consistent with our ChIP-seq data, we observed significant Pol-II enrichment only at the promoter region (Figure 1B and C).

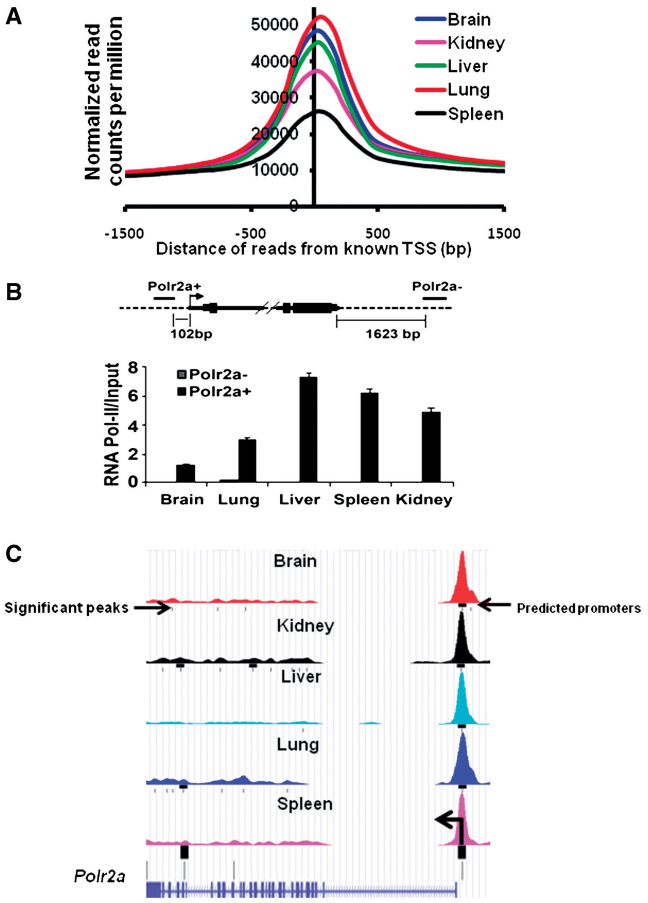

Figure 1. Pol-II ChIP-seq data quality. (**A**) Plot to show the distribution of ChIP-seq reads as normalized read counts per million around known transcription start site (TSS) of known genes in the five tissues. High Pol-II enrichment is observed around known TSS. (**B**) Real-time-PCR results show the enrichment of *Polr2a* promoter in Pol-II ChIP-enriched DNA from five mouse tissues (brain, lung, liver, spleen and kidney). The primers (sequences provided in the 'Materials and Methods' section) used in this ChIP-PCR cover both the promoter region of *Polr2a* and the downstream control region. The promoter region of *Polr2a* gene shows high recruitment of Pol-II while no significant enrichment is observed in downstream *Polr2a* region. (**C**) Enrichment of ChIP-seq reads around *Polr2a* sgene. The black box under the wiggle profile shows the position of the predicted promoters and the line below identifies the peaks that are statistically significant. The *y*-axis represents the normalized read counts per million mapped reads. All five tissues show huge enrichment of read around promoter region of *Polr2a* gene. The enrichment profile is low inside gene for all tissues.

Identification and annotation of active promoters in the mouse tissues

The major challenge in identifying promoters based on Pol-II enriched regions/peaks is the presence of the

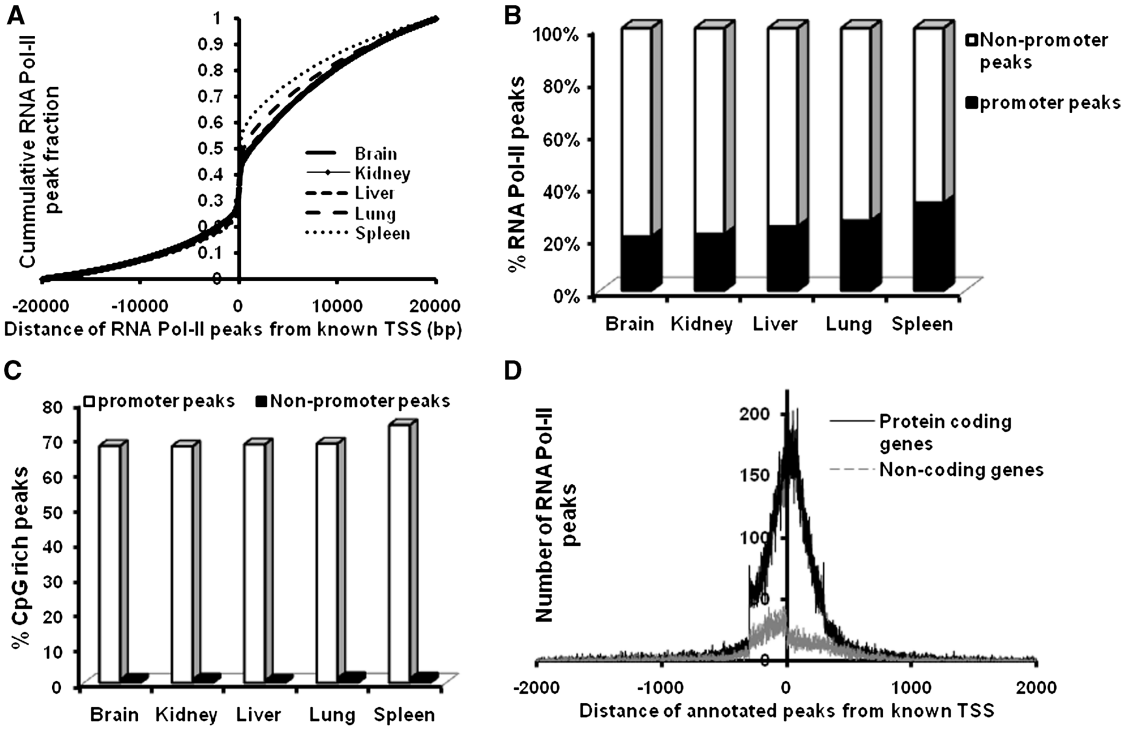

Figure 2. Identification of Pol-II peaks related to promoters. (**A**) Cumulative distribution of Pol-II peaks around known transcription start site (TSS) of known genes in five tissues. The distribution shows that many Pol-II bound peaks are either upstream or downstream of known TSS. (**B**) Percentage of significant Pol-II peaks predicted as promoter and non-promoter peaks by our program in five tissues. The graph indicates that majority of the Pol-II peaks do not correspond to promoters (**C**) Percentage of CpG-rich promoter and non-promoter peaks in five tissues. (**D**) Distribution of annotated peak promoters relative to known TSS for both protein coding and non-coding genes.

transcribing polymerase throughout the gene and, as a result, all genomic regions bound by Pol-II are enriched in the ChIP-seq experiments, producing significantly large number of enriched peaks after the initial statistical analysis. This is clear from the cumulative distribution of Pol-II peaks around known TSS, which indicates that a significant percentage of Pol-II bound loci are also present outside known TSS/promoter regions (Figure 2A). We have recently developed a computational method that discriminates the Pol-II bound promoter regions from Pol-II associations at non-promoter regions (24). Moreover, in our study, we do not use an IgG control as we have determined that our promoter identification analysis is not influenced by the use of IgG background subtraction (Supplementary Table S3 in additional data file 2). We have performed promoter prediction on Pol-II-enriched DNA from liver tissue with and without the subtraction of IgG bound DNA background of liver tissue. As shown in Supplementary Table S3, there is no major change upon inclusion of IgG control but rather we achieve slightly better results without background subtraction in terms of the number of promoters identified and the overlap of identified promoters with the known first exons from various databases. Using this program, we predicted ~24% (80 597) of the significant Pol-II peaks from the five tissues as promoter associated. As shown in Figure 2B, ~20–26% of the Pol-II peaks in brain, kidney, liver and lung are in promoter regions and in spleen almost 34% peaks correspond to promoters. Further analysis suggested that nearly 85% of the non-promoter Pol-II bound peaks were localized in intragenic regions in liver and kidney compared to about 71% of non-promoter-enriched peaks in spleen intragenic locations. Next, we compared the coverage of Pol-II enrichment in the entire transcript with the sequence enrichment at the corresponding promoter (Supplementary Figure S2 in additional data file 1). For ~5–7.5% of the promoters in these five adult tissues, we found 4-fold or more enrichment of the read around the promoter region than the rest of the corresponding transcript, and we speculate that these represent the paused promoters in the adult tissues (38).

We analyzed the CpG richness of promoter and non-promoter peaks based on the previously defined criteria (39). As expected, we found that 69% of the predicted promoter peaks were CpG rich, while only 1.1% of non-promoter peaks were localized in CpG-rich regions (Figure 2C). Further, the proportion of CpG rich promoters did not vary significantly across the tissues, with a maximum of 73.8% and a minimum of 67.7% CpG-rich promoters found in spleen and kidney respectively. Next, we analyzed the enrichment of Pol-II on the CpG rich and CpG poor (non CpG) promoters and found higher binding of Pol-II on CpG rich than non-CpG promoters in all tissues except spleen where the reverse is observed (Supplementary Figure S3A in additional data file 1). In order to provide genome-wide annotation of active promoters in the mouse genome, we combined the promoter

Table 2. Summary of identified promoters across five tissues

Tissue	Known promoters			Novel promoters				Total
	Protein-coding	Non-coding	Both	Protein-coding	Non-coding	Both	Unassigned	
Brain	11 611	3873	207	6274	152	27	4241	21 926
Kidney	11 141	3520	217	5729	153	25	3520	20 301
Liver	9441	2421	114	3889	104	21	2322	15 720
Lung	12 020	3259	187	6375	169	37	4033	21 599
Spleen	7163	1492	63	2742	70	3	1638	11 401
Total number of non-redundant promoters identified								38 639

The breakup of Pol-II bound promoters in each tissue is provided as known and novel promoters and their assignment to either protein coding or/and non-coding genes is indicated.

regions identified in all the five tissues into a master table (additional data file 3). We merged any two consecutive promoters into a single promoter region if the distance between corresponding Pol-II peaks in those promoter regions was <300 bp.

Next, we annotated the identified promoters using a non-redundant set of 42 924 known protein-coding transcripts and 38 159 known non-coding transcripts as described in 'Materials and Methods' section (Supplementary Table S1A in additional data file 2). A schematic representation of the step-wise pipeline that was followed for promoter annotation is shown in Supplementary Figure S4 in additional data file 1. We identified 21 926, 20 301, 15 720, 21 599 and 11 401 promoters that are active in brain, kidney, liver, lung and spleen, respectively (Table 2). About 8173 (21%) of the promoters were left unassigned to any gene based on our annotation strategy. In order to account for the false-negative predictions of the program (known promoters that were not predicted by the program despite the presence of significantly enriched Pol-II peak), all the Pol-II significantly enriched peaks that were predicted as non-promoters but overlap with the first exons of known transcripts were added into the final promoter annotations presented in Table 2. Using this strategy we had further annotated 5356 (2.1%), and 8374 (3.2%) of Pol-II-enriched peaks to known protein-coding and non-coding genes, respectively (Supplementary Table S4 in additional data file 2). Eventually, we have identified a total of 38 639 promoters and annotated 21 739 promoters to only protein-coding genes (15 503), another 7406 promoters to only non-coding genes (5354) and 1321 promoters were assigned to both protein coding and non-coding genes. The list of all annotated promoters along with the annotation is provided in additional data file 4. Many of these promoters were tissue specific, while others were shared between two or more tissues as shown in the Venn diagram (Supplementary Figure S5 in additional data file 1). In particular, our analysis has identified 8727 promoters for Pol-II transcribed non-coding genes with 856, 184, 31 and nine promoters assigned to lincRNA, miRNA, snoRNA and snRNA genes, respectively (Table 3). As promoters are localized around TSS, we analyzed the positioning of the identified promoters relative to the known TSS and found that for both protein coding and non-coding genes, promoters were mostly upstream of the known TSS (Figure 2D). Next, we examined the presence of bidirectional promoters among the identified promoters. We consider a promoter as bidirectional if it is shared between two genes which are in opposite orientation and the promoter region either overlaps with the first exon of the transcripts or is within −1 kb of known TSS as previously described (40). We identified 1093, 1029, 989, 1125 and 852 bidirectional promoters in brain, kidney, liver, lung and spleen, respectively. Interestingly, we found that more than 93% of the bidirectional promoters were CpG rich (Supplementary Table S5 in additional data file 2).

Table 3. Summary of identified promoters assigned to non-coding RNA class across five tissues

Tissue	lincRNA	miRNA	snoRNA	snRNA	Others	All
Brain	405	113	21	9	4058	4547
Kidney	419	70	19	8	3729	4186
Liver	315	55	17	9	2701	3048
Lung	460	64	16	8	3489	3974
Spleen	249	32	13	6	1752	2011
Overall	856	184	31	9	7740	8727

This table shows the number of promoters assigned to each category of non-coding RNA in each tissue.

Novel promoter identification and experimental validation

One of the major goals of this study was the identification of novel promoters in the mouse genome. We have identified a total of 12 270 novel promoters, which represents 32% of all the identified promoters. This suggests that a large number of promoters that are active in at least one of the five tissues were unknown in any of the current genome-wide annotations. The tissue-wise distribution of the novel promoters is presented in Table 4. We observed higher enrichment of Pol-II on the novel promoters than known promoters in spleen, while in brain and lung the reverse was true. In contrast, in kidney and liver the binding of Pol-II near TSS was similar for known and novel promoters (Supplementary Figure S3B in additional data file 1). Additionally, we found that about 34% of the novel promoters were CpG rich and these novel promoter regions show similar level of conservation as known promoters across 30 vertebrate species (Supplementary Figure S6 in additional data file 1). When we analyzed the novel promoters with the known

Table 4. Novel promoters and relationship with existing information

Tissue	Number of novel promoters	Percentage of CpG-rich promoters	Number of promoters			
			Homologous to other organisms (overlap with 5′ends of XenoRef mRNAs)	Supported by CAGE cluster	Found in Bing Ren's study (17)	Overlap with 5′ of ESTs
Brain	6649	41.21	405	6479	184	4320
Kidney	5857	42.38	372	5722	177	3728
Liver	3972	36.61	250	3887	131	2670
Lung	6507	42.98	392	6318	208	4015
Spleen	2809	42.11	160	2724	95	1775
Overall	12 270	33.85	671	11 902	371	7663

This table shows the support of the identified novel promoters from other experimental sources in every tissue.

promoters of homologous genes, we found that 671 of these promoters had corresponding known promoters in other organisms. Next, we looked for the overlap of novel promoters from our study with CAGE tag clusters generated by the FANTOM4 project (41). CAGE tag clusters are found at the 5′ end of transcript as well as in other regions including the internal exons, introns, 3′UTR and intergenic regions. Because CAGE analysis relies on 5′ Cap trapper techniques, thus capturing even post-transcriptionally re-capped mRNA, it has inherent deficiencies as a sole tool to identify promoters (20). We have observed that for our promoter peaks, 95–97% are supported by CAGE clusters and surprisingly even 53–59% of non-promoter peaks also show the presence of CAGE clusters (Supplementary Figure S7 and Supplementary Table S6). When we focused on the overlap of novel promoters with CAGE clusters, as expected, we observed a 97% correlation. It is worth noting that only 1.4% of all CAGE tag based predicted promoters are actually identified as active promoter by our approach in the five adult tissues (Supplementary Figure S7A). Additionally, we compared non-redundant EST sequences with the novel promoters and found that about 62.4% of the novel promoters overlapped with the 5′ ends of the ESTs. Furthermore, using the published mRNA-seq data from mouse brain and liver, we detected mRNA-seq reads for 68% and 60% of novel promoters identified in brain and liver, respectively (example shown in Supplementary Figure S8 in additional data file 1) (42). Thus, the novel promoters identified by our approach of promoter prediction on Pol-II-enriched ChIP-seq data are supported by other independent experimental methods. To further validate the activity of these novel promoters, we cloned 10 of the randomly selected novel promoters (NP1-10) and two non-promoter regions (Ctrl1, 2) upstream of a promoter-less luciferase gene and measured promoter activity (Supplementary Table S2 in additional data file 2). As shown by the wiggle tracks in Figure 3A, we have identified a new promoter in all five mouse tissues that lies ~16 kb upstream of the known *mKIAA1632* gene promoter (NP1). The homologous region in humans represents the promoter for the human *KIAA1632* gene. Similarly, we have identified a new promoter in three of the mouse tissues for *GPM6A* that lies ~175 kb upstream of the known *GPM6A* gene promoter in mouse (NP8). The homologous region in humans represents the promoter for the human *GPM6A* gene (Figure 3B). These promoters either represent the unidentified promoters for *KIAA1632* and *GPM6A* or drive the expression of unknown genes. Figure 3C shows the Pol-II binding in one of the regions that was not predicted as a promoter in our analysis and is considered as non-promoter. Using transient transfection experiments, we introduced these constructs in five different cell lines (HEK293, DAOY, A549, HepG2 and NIH3T3) and measured the expression of luciferase gene which is controlled by the novel promoters or non-promoter regions. We observed significant luciferase expression (7-fold for NP5 to 304-fold for NP1) from nine of the 10 selected promoters in at least one of the cell lines (Figure 3D). We did not observe any promoter activity for the novel promoter NP7 which was identified in spleen tissue and it is possible that this is due to the absence of the proper cell system in our luciferase assays or NP7 is a false promoter prediction. Based on our results, we conclude that non-CpG promoters are underrepresented in the current promoter inventory and advances in high-throughput sequencing technology coupled with bioinformatics analysis can help to identify these.

Alternative promoter usage in the mouse tissues

It is well established that many mammalian genes have multiple promoters and that these are differentially used in different cellular context. In agreement, we found 23 060 promoters for 16 330 protein-coding and 8727 promoters for 6314 (24%) non-coding genes in the five mouse tissues that were analyzed. To identify the genes that use alternative promoters in these different tissues, we adopted a two-step procedure. The first step involved identification of the promoter(s) with bound Pol-II for each gene in each tissue individually. In the second step, we compared the identified promoter for each gene across the five tissues. In case the promoters from two or more tissues overlap with each other by at least 300 bp then they are considered as the same promoter, otherwise they are defined as distinct promoters (additional data file 4). Examples of alternative promoter genes identified by

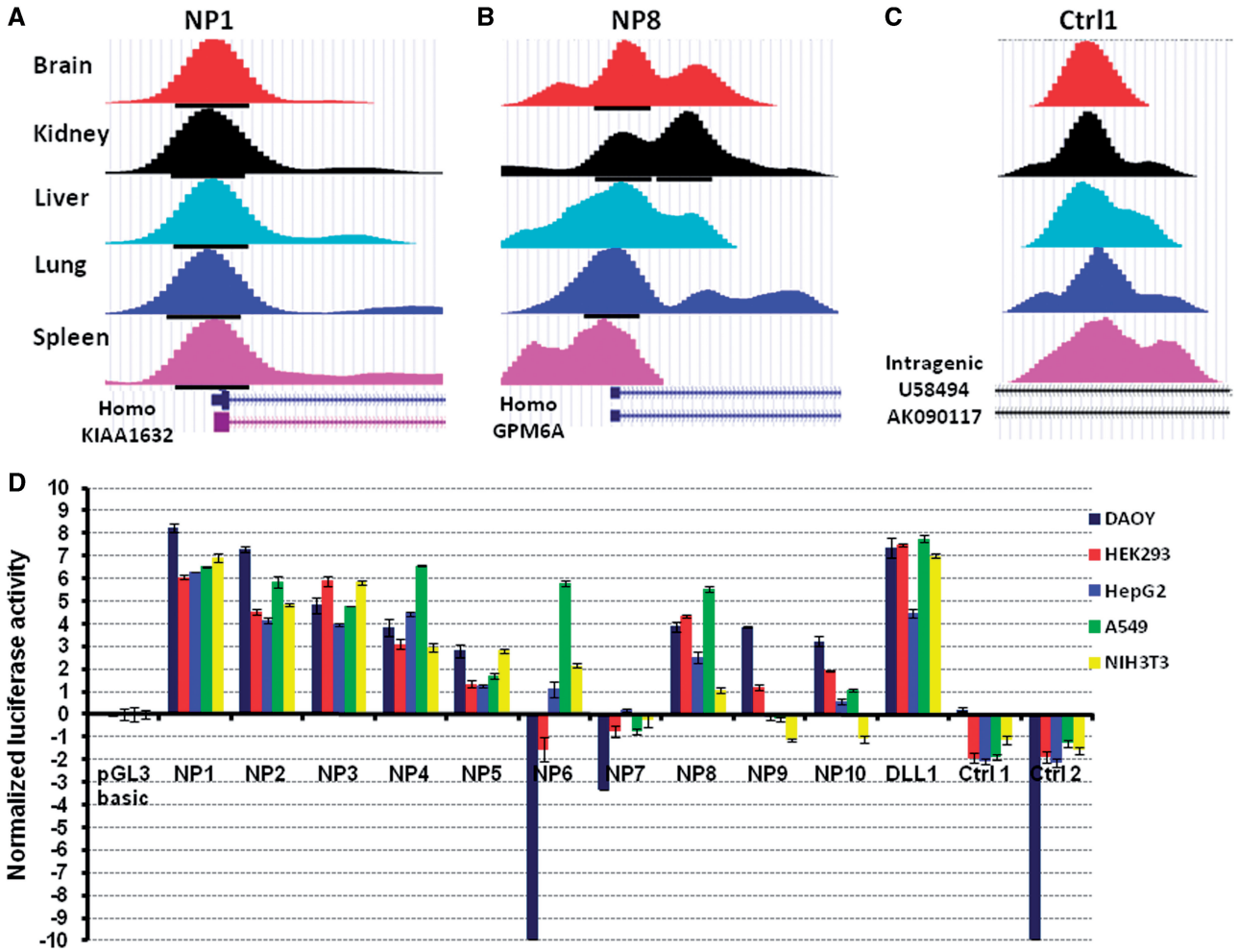

Figure 3. Identification of novel promoters and experimental verification. (**A–C**) The wiggle profile shows the enrichment of Pol-II and prediction of novel promoters for mouse *KIAA1632* (A) and *GPM6* (B) gene in brain, kidney, liver, lung, and spleen tissues and (C) shows the Pol-II profile on a non-promoter region that lies within the transcripts (AK090117 and U58494). *Y*-axis shows the normalized read counts/million mapped reads and the black boxes below the Pol-II-binding profile represent the identified novel promoters by our program. The two novel promoters shown above have a corresponding conserved promoter in human genome. (**D**) Luciferase activity of the novel promoters in five distinct cell lines. The *x*-axis represents either the vector alone background (pGL3 basic) or the activity of novel promoters (NP1-10) and non-promoter region (ctrl1 and 2). On the *y*-axis, the normalized luciferase activity has been plotted in a logarithmic scale to the base of 2. Thus, the value of 1on the *y*-axis represents a 2-fold activity of the promoter. Only if a promoter shows an activity that is more than 2-fold it is considered as active in that cell line. NP7 does not show any luciferase activity in any of the five cell lines. Luciferase activity is expressed as log2 of the fold change over the vector alone (pGL3basic) after normalization with Renilla-luciferase for transfection efficiency. Dll1 promoter has been included as a positive control and Ctrl1, 2 represents the negative controls (non-promoter regions).

our approach are shown for *Adar1*and *Hdgf* gene in Figure 4A. In case of *Adar1*, there are two active promoters that are differentially used in the five tissues. The upstream promoter P1 is used in brain, kidney and lung, while downstream P2 promoter has been identified as active in kidney, liver, lung and spleen. Similarly for *Hdgf,* four distinct active promoters have been identified that are differentially enriched with Pol-II. Based on this analysis, we have found the distribution of multi-promoter usage in five mouse tissues (Figure 4B and C). We observed that 37% of the annotated protein-coding genes and 31% of the non-coding genes use alternative promoters in the five mouse tissues. Furthermore, we found that the use of alternative promoters changes the coding protein in 34.5% of the alternative promoter genes. As our analysis is based on only five tissues, it suggests that a significant number of mouse genes use alternative promoters.

Identification of tissue-specific promoters

Having identified the promoters that are active in one or more of the five tissues, we further investigated the usage of promoters in a tissue-specific manner. Two different parameters were used for identifying tissue-specific promoters in our study. The first parameter is based on Shannon entropy that was previously employed for identifying tissue-specific promoters from ChIP–chip (17), gene expression and EST data (43). As the second parameter, we define fold change for each promoter (p) as $fp = max1/max2$, where $max1$ and $max2$ are the first and second highest normalized read counts for promoter p

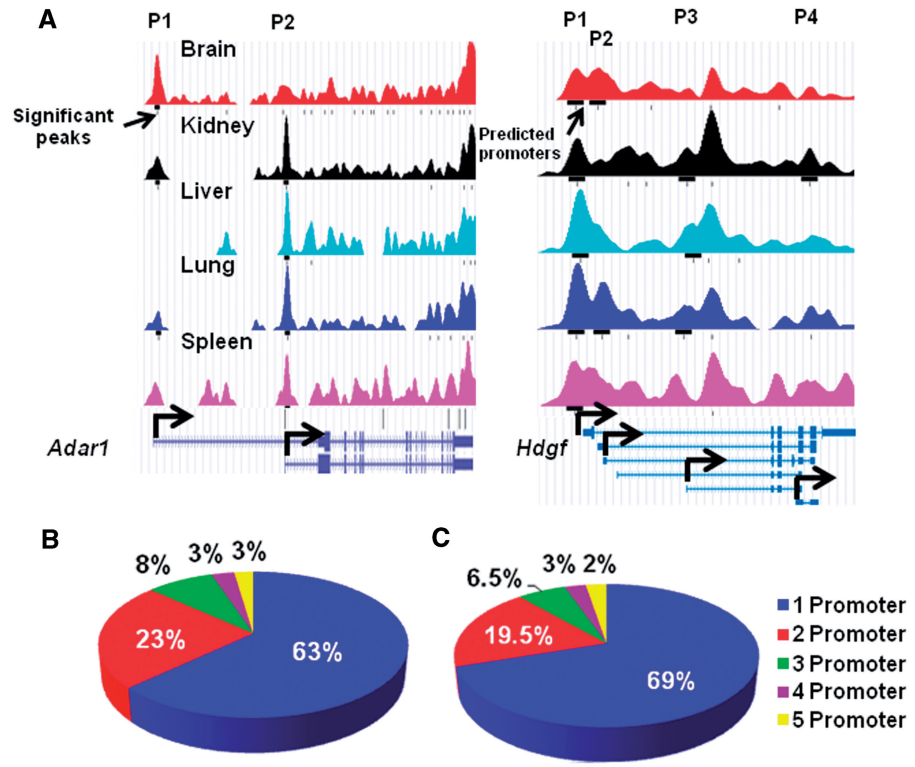

Figure 4. Alternative promoter usage in five mouse tissues (**A**) Wiggle profile showing examples of alternative promoter usage identified by our analysis in the five mouse tissues. For *Adar1* and *Hdgf* our approach has identified the use of two and four known promoters among the five tissues, which are indicated by arrows on the transcripts at the bottom of the figure. *Y*-axis shows the normalized read count per million reads, the black box below the Pol-II enrichment shows the position of the identified promoters and the black line under it indicates the significantly enriched peaks of each track. It is evident that only few of the enriched peaks reside in the promoter region. (**B** and **C**) Pie chart shows distribution of alternative promoter usage for protein coding (B) and non-coding genes (C) in the five tissues analyzed from mouse.

among the five tissues, respectively. To define tissue specificity, we have set the maximum cutoff as 1.25 for Shannon entropy and a minimum cutoff of 2.0 for the fold change determinant. Note that while Shannon entropy is inversely correlated, the fold-change parameter is directly correlated with tissue specificity. Using the above-defined parameters, we have identified 6384 tissue-specific promoters across the five mouse tissues (Supplementary Table S7 in additional data file 2 and additional data file 5). These results are further supported by the box plot, which shows as an example for the brain-specific promoters a significantly higher read count in brain compared to the other four tissues in the distribution of normalized read counts (Figure 5A). Similar results were obtained for other tissue-specific promoters (Supplementary Figure S9 in additional data file 1). The highest number of tissue-specific promoters was identified in brain while spleen has the least number of tissue-specific Pol-II-associated promoters. Further analysis revealed that overall only 29% of the tissue-specific promoters were CpG rich, with brain and lung exhibiting highest CpG richness (~41%), while in spleen only 9% of promoters were CpG rich. We further studied the relationship of tissue-specific parameters: Shannon entropy and normalized read fold change with CpG richness in promoters of genes (Figure 5B and C). A direct relationship is observed between CpG richness and Shannon entropy and an inverse relationship is observed between CpG richness and normalized read fold-change for promoters, suggesting that globally the tissue-specific promoters are CpG poor compared to ubiquitous promoters. Furthermore, detailed analysis of core promoter elements in the tissue-specific promoters versus ubiquitous promoters show that TATA (p = 1.75e-14) and INR (p = 1.69e-12) elements are more enriched in tissue-specific promoters, while BRE (p = 4.56e-29) and MTE (p = 6.82e-18) elements are significantly enriched in ubiquitous genes (Supplementary Table S8A and B in additional data file 2, statistical significance was calculated using proportion test). DPE element did not show any preference for either class of promoters. Thus, our data suggests hat CpG-poor tissue-specific promoters and CpG-rich ubiquitous promoters tend to possess different core promoter composition (44).

Correlation of Pol-II binding at promoter and the corresponding transcript expression

As binding of Pol-II precedes transcription, we expected that promoters with high occupancy of Pol-II will be transcribed at a higher rate than others. To address this issue, we studied the correlation of Pol-II recruitment to the promoter and the consequential transcript expression in the mouse tissues. We performed this analysis for only

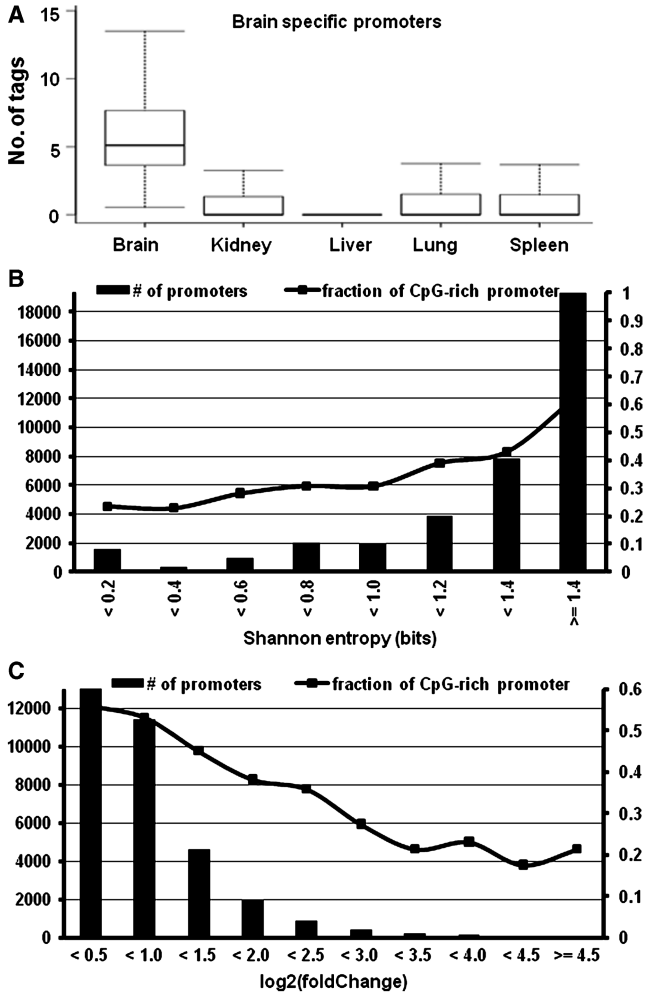

Figure 5. Identification of tissue specific promoters and their relationship to CpG islands. (**A**) Box plot shows the normalized read counts of promoters that have been assigned to be brain specific in all five mouse tissues. The plot shows that the brain specific promoters indeed show specific and increased Pol-II binding in brain relative to other tissues. (**B** and **C**) Tissue specificity of promoters and its association with CpG richness. An inverse relationship is observed between tissue-specific promoters and CpG richness when we look at either the Shannon entropy (B) or normalized read count fold change (C) measures of tissue specificity across all five tissues. The left-hand side of *y*-axis represents the total promoter count. The right side of *y*-axis represents fraction of CpG-rich promoters. The *x*-axis in (B) represents Shannon entropy and in (C) it represents log2 (fold change) in read count value around annotated promoters.

brain and liver using the publicly available mRNA-seq data (42). We used Cufflinks software (45) to estimate the transcript expression from the mRNA-seq data sets using the default parameters. For each tissue, the expression scores from promoters were broken up into four quartiles: high, medium, low and very low. Next, we calculated the average Pol-II ChIP-seq read count around annotated TSS at base pair resolution for promoters found in the four quartiles and plotted the Pol-II-enriched profile around annotated TSS (Figure 6). Because there is no mRNA-seq data available for kidney, lung and spleen, we performed similar analysis at the gene level using microarray gene expression data (Supplementary Figure S10). The gene expression data used in our present study for brain, kidney, liver, lung and spleen were downloaded from NCBI (GEO ID: GDS592) (46). As these data contained profiles for eight different brain tissues, (frontal cortex, cerebral cortex, substantia nigra, cerebellum, amygdala, hypothalamus, hippocampus and dorsal striatum), the average of these scores for each gene was taken as the expression of the corresponding gene in brain. Altogether, we observe that the promoters driving higher mRNA expression exhibit increased Pol-II recruitment, suggesting that the binding of Pol-II at the promoter is a good signature for global expression from a promoter.

DISCUSSION

Identification and annotation of all human and mouse gene promoters that are differentially used in different cell/tissue types, during development, or aberrantly activated in disease conditions are still incomplete and are essential for defining the transcriptome and proteome of the mammalian genome. It is well known that differential gene expression is a characteristic of different tissues; however, not much work has been done to characterize the global isoform specific expression of genes in various tissues (47,48). One of the important aspects to understand the regulation of gene expression is the study of all the promoters of a gene. Currently, our promoter knowledge is partial and our goal in this study was to expand our promoter inventory and to determine the tissues where each promoter is active. To provide a catalog of active promoters in various tissues and identify tissue-specific promoter usage, we used a combination of ChIP-seq and bioinformatics approaches. In this study, we focused on five adult tissues—brain, kidney, liver, lung and spleen, and have successfully identified 38 639 promoters for both protein coding and non-coding genes. Our approach has identified 12 270 novel promoters including promoters for genes such as *Dnmt1*, *Bmp4*, *Jmjd3, Cyclin E1* and *D1*, *MeCP2*, which have been associated with tumorigenesis. We have been able to annotate a large number of the newly discovered promoters to known genes like 60% in case of brain, and we anticipate that the remaining 40% un-annotated promoters might mostly represent the promoters of unknown non-coding genes. Our results also show that about 37% of the protein coding genes possess alternative promoters. This is lower than the expected 50–60% genes from other genome-wide analysis due to the small number of tissues assayed as well as the sole use of adult tissue in this study (11,44). This is supported by our analysis where we compared the alternative promoter use in two, three, four, five tissues and observed an increase in alternative promoter usage from 27% (two-tissue) to 37% (five-tissue). The use of alternative promoters results in different proteins in about 34% of multi-promoter genes as seen for *Adar1*, *Hdgf* (Figure 4A). In case of *Adar1*, the upstream promoter P1 is responsive to interferon and produces a 150-kDa protein compared to the constitutive promoter P2 that produces an N-terminally truncated 110-kDa protein (49). We have found that 5–8% of the

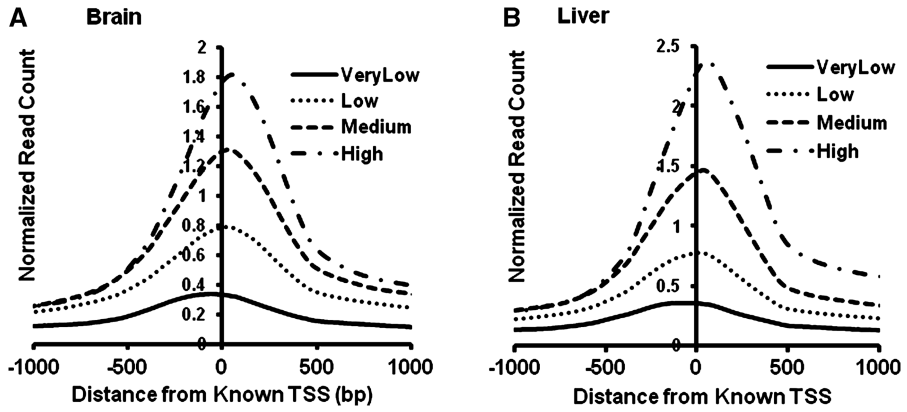

Figure 6. Correlation of Pol-II enrichment on promoters with the expression of corresponding transcripts. (**A** and **B**) Based on the mRNA expression estimated by Cufflinks software the promoters were divided into four groups for brain and liver (high, medium, low and very low). The normalized read/tag count of Pol-II bound DNA for each group of promoters at base pair resolution from –1 kb to +1 kb relative to TSS was calculated and plotted as a function of distance from TSS. Total reads of gene promoter were counted based on our annotation for each group in each tissue. The *x*-axis represents distance of reads from annotated TSS and the *y*-axis represents total number of reads per million mapped reads.

promoters in these tissues are bidirectional and 4–6% of the promoters are shared by protein coding and non-coding genes in each tissue. Our analysis suggests that nearly 17% of the promoters in the mouse genome are used in a tissue-specific manner and these tissue restrictive promoters tend to be CpG poor. We found that almost 70% of the known promoters are CpG rich, while 66% of the novel promoters are CpG poor suggesting that many of these new promoters are tissue-specific and not easily identifiable without high-throughput genome-wide analysis. This is supported by the finding that, while 1/4th of the highly tissue restrictive (active in only one tissue) novel promoters are CpG rich, 2/3rd of the novel promoters active in the five tissues show CpG richness. In conclusion, we have identified the active promoters in five mouse tissues and we plan to expand our study to identify the differential and overlapping use of promoters in normal human tissues and diseased tissue counterparts.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. ChIP-seq data has been deposited in GEO under accession number GSE 21773.

ACKNOWLEDGEMENTS

The use of Genomics core facility and computational resources in the Centre for Systems and Computational Biology and Bioinformatics Facility of Wistar Cancer Centre (supported by grant # P30CA010815) are gratefully acknowledged.

FUNDING

National Human Genome Research Institute (grant # R01HG003362 to R.D.). R.D. holds a Philadelphia Healthcare Trust Endowed Chair Position and this work is also supported by Philadelphia Healthcare Trust. Funding for open access charge: National Institutes of Health (grant # R01HG003362 to R.D.).

Conflict of interest statement. None declared.

REFERENCES

1. Davuluri,R.V., Suzuki,Y., Sugano,S., Plass,C. and Huang,T.H. (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet*., **24**, 167–177.
2. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
3. Van de Wetering,M., Castrop,J., Korinek,V. and Clevers,H. (1996) Extensive alternative splicing and dual promoter usage generate Tcf-1 protein isoforms with differential transcription control properties. *Mol. Cell. Biol*., **16**, 745–752.
4. Hovanes,K., Li,T.W., Munguia,J.E., Truong,T., Milovanovic,T., Lawrence,M.J., Holcombe,R.F. and Waterman,M.L. (2001) Beta-catenin-sensitive isoforms of lymphoid enhancer factor-1 are selectively expressed in colon cancer. *Nat. Genet*., **28**, 53–57.
5. Sandelin,A., Carninci,P., Lenhard,B., Ponjavic,J., Hayashizaki,Y. and Hume,D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet*., **8**, 424–436.
6. Kapranov,P., Willingham,A.T. and Gingeras,T.R. (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet*., **8**, 413–423.
7. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
8. Baek,D., Davis,C., Ewing,B., Gordon,D. and Green,P. (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res*., **17**, 145–155.
9. Sun,H., Palaniswamy,S.K., Pohar,T.T., Jin,V.X., Huang,T.H. and Davuluri,R.V. (2006) MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. *Nucleic Acids Res*., **34**, D98–D103.
10. Takeda,J., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res*., **35**, D104–D109.

11. Kimura,K., Wakamatsu,A., Suzuki,Y., Ota,T., Nishikawa,T., Yamashita,R., Yamamoto,J., Sekine,M., Tsuritani,K., Wakaguri,H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55–65.
12. Cooper,S.J., Trinklein,N.D., Anton,E.D., Nguyen,L. and Myers,R.M. (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, **16**, 1–10.
13. Bajic,V.B., Tan,S.L., Christoffels,A., Schonbach,C., Lipovich,L., Yang,L., Hofmann,O., Kruger,A., Hide,W., Kai,C. *et al.* (2006) Mice and men: their promoter properties. *PLoS Genet.*, **2**, e54.
14. Maeda,N., Nishiyori,H., Nakamura,M., Kawazu,C., Murata,M., Sano,H., Hayashida,K., Fukuda,S., Tagami,M., Hasegawa,A. *et al.* (2008) Development of a DNA barcode tagging method for monitoring dynamic changes in gene expression by using an ultra high-throughput sequencer. *Biotechniques*, **45**, 95–97.
15. Balwierz,P.J., Carninci,P., Daub,C.O., Kawai,J., Hayashizaki,Y., Van Belle,W., Beisel,C. and van Nimwegen,E. (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.*, **10**, R79.
16. Singer,G.A., Wu,J., Yan,P., Plass,C., Huang,T.H. and Davuluri,R.V. (2008) Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. *BMC Genomics*, **9**, 349.
17. Barrera,L.O., Li,Z., Smith,A.D., Arden,K.C., Cavenee,W.K., Zhang,M.Q., Green,R.D. and Ren,B. (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.*, **18**, 46–59.
18. Otsuka,Y., Kedersha,N.L. and Schoenberg,D.R. (2009) Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA. *Mol. Cell. Biol.*, **29**, 2155–2167.
19. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project *et al.* (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature*, **457**, 1028–1032.
20. Schoenberg,D.R. and Maquat,L.E. (2009) Re-capping the message. *Trends Biochem. Sci.*, **34**, 435–442.
21. Lee,T.I., Johnstone,S.E. and Young,R.A. (2006) Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.*, **1**, 729–748.
22. Cheng,A.S., Jin,V.X., Fan,M., Smith,L.T., Liyanarachchi,S., Yan,P.S., Leu,Y.W., Chan,M.W., Plass,C., Nephew,K.P. *et al.* (2006) Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Mol. Cell*, **21**, 393–404.
23. Zhang,Z.D., Rozowsky,J., Snyder,M., Chang,J. and Gerstein,M. (2008) Modeling ChIP sequencing in silico with applications. *PLoS Comput. Biol.*, **4**, e1000158.
24. Gupta,R., Wikramasinghe,P., Bhattacharyya,A., Perez,F.A., Pal,S. and Davuluri,R.V. Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics*, **11(Suppl. 1)**, S65.
25. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
26. Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.
27. Friedman,J., Hastie,T. and Tibshirani,R. (1998) Additive logistic regression: a statistical view of boosting. *Ann. Stat.*, **28**, 337–407.
28. Freund,Y. and Schapire,R.E. (1996) *Thirteenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, pp. 148–156.
29. Freund,Y. and Schapire,R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.
30. Rätsch,G., Onoda,T. and Müller,K.R. (2001) Soft margins for AdaBoost. *Mach. Learn*, **42**, 287–320.
31. Bajic,V.B., Tan,S.L., Suzuki,Y. and Sugano,S. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.*, **22**, 1467–1473.
32. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
33. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
34. Jin,V.X., Singer,G.A., Agosto-Perez,F.J., Liyanarachchi,S. and Davuluri,R.V. (2006) Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics*, **7**, 114.
35. Blahnik,K.R., Dou,L., O'Geen,H., McPhillips,T., Xu,X., Cao,A.R., Iyengar,S., Nicolet,C.M., Ludascher,B., Korf,I. *et al.* (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.*, **38**, e13.
36. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
37. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
38. Wu,J.Q. and Snyder,M. (2008) RNA polymerase II stalling: loading at the start prepares genes for a sprint. *Genome Biol.*, **9**, 220.
39. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.
40. Koyanagi,K.O., Hagiwara,M., Itoh,T., Gojobori,T. and Imanishi,T. (2005) Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene*, **353**, 169–176.
41. Kawaji,H., Severin,J., Lizio,M., Waterhouse,A., Katayama,S., Irvine,K.M., Hume,D.A., Forrest,A.R., Suzuki,H., Carninci,P. *et al.* (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol.*, **10**, R40.
42. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
43. Schug,J., Schuller,W.P., Kappen,C., Salbaum,J.M., Bucan,M. and Stoeckert,C.J. Jr (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.
44. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
45. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
46. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
47. Zhang,W., Morris,Q.D., Chang,R., Shai,O., Bakowski,M.A., Mitsakakis,N., Mohammad,N., Robinson,M.D., Zirngibl,R., Somogyi,E. *et al.* (2004) The functional landscape of mouse gene expression. *J. Biol.*, **3**, 21.
48. Naef,F. and Huelsken,J. (2005) Cell-type-specific transcriptomics in chimeric models using transcriptome-based masks. *Nucleic Acids Res.*, **33**, e111.
49. Patterson,J.B. and Samuel,C.E. (1995) Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol. Cell. Biol.*, **15**, 5376–5388.