doi:10.1038/nature03877

LETTERS

A high-resolution map of active promoters in the human genome

Tae Hoon Kim[1]*, Leah O. Barrera[1]*, Ming Zheng[3], Chunxu Qu[1], Michael A. Singer[4], Todd A. Richmond[4], Yingnian Wu[3], Roland D. Green[4] & Bing Ren[1,2]

In eukaryotic cells, transcription of every protein-coding gene begins with the assembly of an RNA polymerase II preinitiation complex (PIC) on the promoter[1]. The promoters, in conjunction with enhancers, silencers and insulators, define the combinatorial codes that specify gene expression patterns[2]. Our ability to analyse the control logic encoded in the human genome is currently limited by a lack of accurate information regarding the promoters for most genes[3]. Here we describe a genome-wide map of active promoters in human fibroblast cells, determined by experimentally locating the sites of PIC binding throughout the human genome. This map defines 10,567 active promoters corresponding to 6,763 known genes and at least 1,196 un-annotated transcriptional units. Features of the map suggest extensive use of multiple promoters by the human genes and widespread clustering of active promoters in the genome. In addition, examination of the genome-wide expression profile reveals four general classes of promoters that define the transcriptome of the cell. These results provide a global view of the functional relationships among transcriptional machinery, chromatin structure and gene expression in human cells.

The PIC consists of the RNA polymerase II (RNAP), the transcription factor IID (TFIID) and other general transcription factors[4]. Our strategy to map the PIC binding sites involves a chromatin immunoprecipitation-coupled DNA microarray analysis (ChIP-on-chip), which combines the immunoprecipitation of PIC-bound chromatin from formaldehyde crosslinked cells with parallel identification of the resulting bound DNA sequences using DNA microarrays[5,6]. We have previously demonstrated the feasibility of this strategy by successfully mapping active promoters in 1% of the human genome corresponding to the 44 genomic loci known as the ENCODE regions[6,7].

To apply this strategy to the entire human genome, we made a series of DNA microarrays[8] containing roughly 14.5 million 50-mer oligonucleotides, designed to represent all the non-repeat DNA throughout the human genome at 100-base pair (bp) resolution. We immunoprecipitated TFIID-bound DNA from primary fibroblast IMR90 cells using a monoclonal antibody that specifically recognizes the TAF1 subunit of this complex (TBP associated factor 1, formerly $TAF_{II}250$, ref. 9; Fig. 1a). We then amplified and fluorescently labelled the resulting DNA, and hybridized it to the above microarrays along with a differentially labelled control DNA (Fig. 1a). We determined 9,966 potential TFIID-binding regions using a simple algorithm that requires a stretch of four neighbouring probes to have a hybridization signal significantly above background. To independently verify these TFIID-binding sequences, we designed a condensed array that contained a total of 379,521 oligonucleotides to represent these sequences, and 29 control genomic loci selected from the 44 ENCODE regions[7] at 100-bp resolution. ChIP-on-chip analysis of two independent samples of IMR90 cells confirmed the binding of TFIID to a total of 8,597 regions, ranging in size from 400 bp to 9.8 kb (Fig. 1b). We further resolved a total of 12,150 TFIID-binding sites within the 8,597 fragments using a peak-finding algorithm that predicts the most likely TFIID-binding sites based on the hybridization intensity of consecutive probes with significant signals (Fig. 1b, see Supplementary Information for details).

Next, we matched these 12,150 TFIID-binding sites to the 5′ end of known transcripts in three public transcript databases (DBTSS[10], RefSeq[11] and GenBank human mRNA collection[12]) and the EnsEMBL gene catalogue[13]. To account for the uncertainty of our knowledge regarding the true 5′ end of transcripts and the uncertainty of predicted TFIID-binding positions due to noise within the microarray data, we chose an arbitrary distance of 2.5 kb as a measure of close proximity. We found that 10,553 (87%) TFIID-binding sites were within 2.5 kb of annotated 5′ ends of known messenger RNA. We resolved common TFIID-binding sites mapping to similar 5′ ends to define a non-redundant set of 9,328 5′-end-matched TFIID-binding sites. Of these TFIID-binding sequences 7,789 (83%) were found within 500 bp of the putative transcription start sites (TSS) (Fig. 1c). As these 9,328 DNA sequences were bound by TFIID *in vivo* and are within close proximity to the 5′ end of known transcripts, we defined them as promoters for the corresponding transcripts (Supplementary Table S1). Of these 9,328 promoters, 8,960 were mapped within 2.5 kb of the 5′ end or within annotated boundaries of 6,763 known genes in the EnsEMBL gene catalogue[13] (Fig. 1d and Supplementary Table S1). The remaining 368 promoters corresponded to transcripts not contained within these boundaries of EnsEMBL genes, and therefore provide support for inclusion of these transcripts to the current gene catalogues. The list of promoters also confirmed 5,118 previously annotated promoters[10], and defined 4,210 new promoters for at least 2,627 genes (Fig. 1e and Supplementary Table S1).

Four independent analyses validated the high specificity and accuracy of the active promoters detected in IMR90 cells. First, ChIP-on-chip analysis using an anti-RNAP antibody (8WG16) confirmed the binding of RNAP to at least 9,050 (97%) of the 9,328 promoters in IMR90 cells (Supplementary Fig. S1). Second, standard chromatin immunoprecipitation (ChIP) experiments performed on 28 promoters randomly selected from the above list confirmed the occupancy of RNAP on all but one promoter (Supplementary Fig. S2). Third, the 9,328 active promoters are enriched for known promoter-associated sequences such as CpG islands and the INR and DPE core promoter elements (Fig. 1f). The percentage

[1]Ludwig Institute for Cancer Research and [2]Department of Cellular and Molecular Medicine and Moores Cancer Center, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, California 92093-0653, USA. [3]8125 Math Sciences Building, UCLA Department of Statistics, Los Angeles, California 90095-1554, USA. [4]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711, USA.
*These authors contributed equally to this work.

of CpG-associated promoters (88%) was significantly higher than the previous estimate (56%, ref. 14), suggesting that CpG islands might play a more general role in gene expression than previously appreciated. Notably, we did not find the TATA box to be significantly enriched in these promoters (Fig. 1f). This might be due to a lack of conservation of the TATA box in human promoters, or it might alternatively indicate that the TATA box is not a general promoter motif for human genes. This observation is consistent with previous reports that the TATA box is only present in a small number of promoters in yeast and in *Drosophila*[15]. Fourth, ChIP-on-chip analysis using antibodies that recognize acetylated histone H3 (AcH3) or dimethylated lysine 4 on histone H3 (MeH3K4) showed that over 97% of the 9,328 promoters were associated with these known epigenetic markers of active genes[16] (Fig. 2a). The localization of MeH3K4 in these promoters was predominantly downstream of the TFIID-binding site (Fig. 2b), but the mechanisms for such chromatin organization at human promoters are currently unknown.

Among the 12,150 mapped TFIID-binding sites, 1,597 are found more than 2.5 kb away from previously defined 5′-ends of mRNA, and might represent promoters for new transcripts or genes (Supplementary Table S2). Of these, 607 non-redundant TFIID-binding sites were matched within 2.5 kb of the 5′ ends of the expressed-sequence-tag (EST)-based gene models, indicating that they may indeed produce mRNA (Supplementary Table S2). The remaining TFIID-binding sites were further filtered to a set of 632 putative promoters by requiring the occupancy of RNAP and presence of AcH3 and MeH3K4 within 1 kb of these sites (Supplementary Fig. S3). To verify that these promoters drive transcription, we analysed mRNA from the IMR90 cells using 50-mer oligonucleotide arrays that represent a 28 kb sequence surrounding 567 of the 632 unmatched putative promoters. At least 35 new transcription units were identified near the putative promoter regions, suggesting that these might represent new transcription units yet to be annotated in the human genome (Supplementary Table S3). The failure to detect mRNA from the other putative promoters might indicate that these transcripts are highly unstable. Indeed, at least one putative promoter is located in the 250 bp upstream of a predicted micro-RNA[17] (Supplementary Fig. S4), suggesting that some putative promoters could transcribe non-coding RNA that might have escaped detection by conventional mRNA-isolation techniques.

In total, we defined a set of 1,239 putative promoters that correspond to previously un-annotated transcription units (Supplementary Table S2). Evolutionarily conserved regions were found

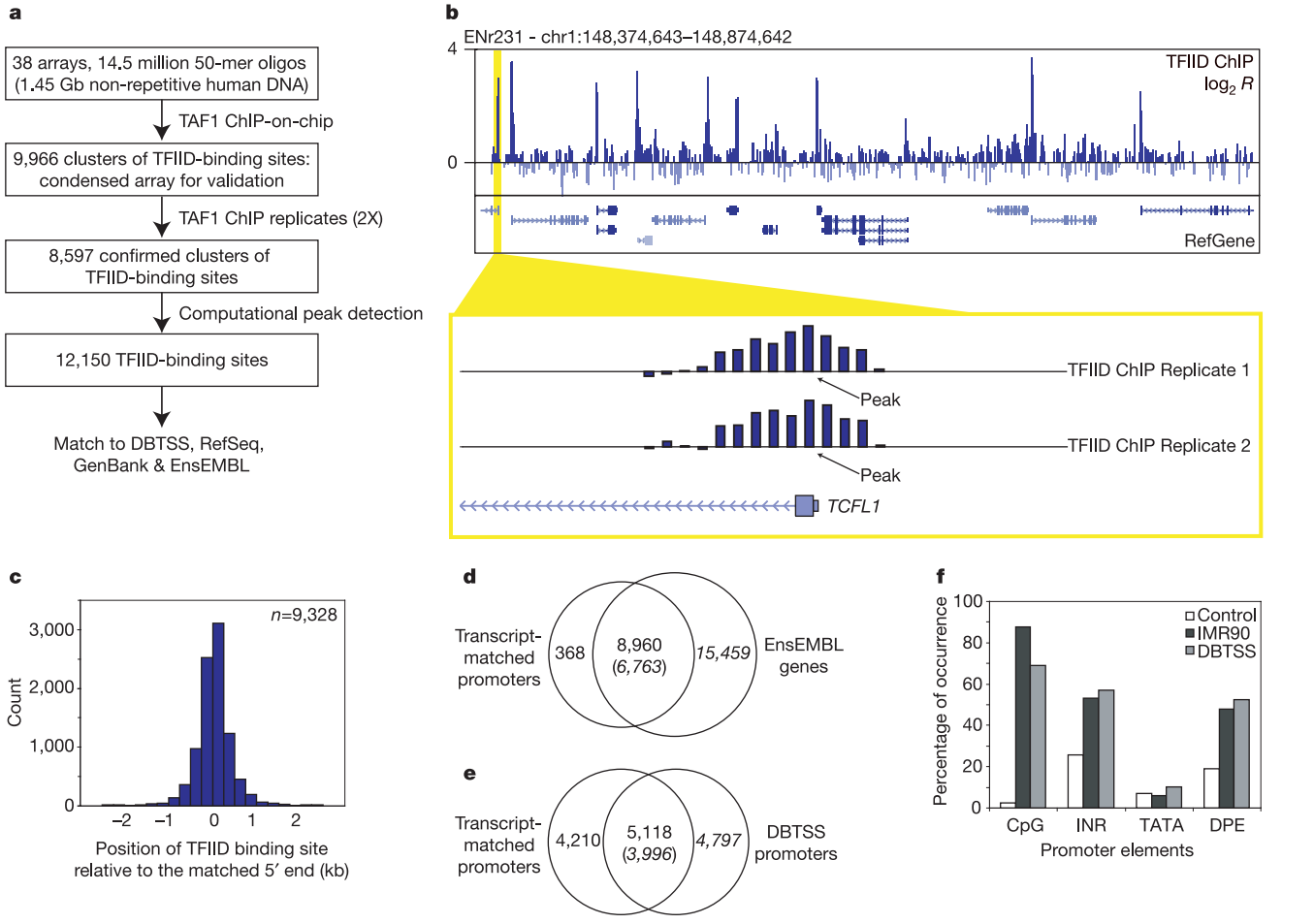

Figure 1 | Identification and characterization of active promoters in the human genome. **a**, Outline of the strategy used to map TFIID-binding sites in the genome. **b**, A representative view of the results from TFIID ChIP-on-chip analysis. Top panel, the logarithmic ratio ($\log_2 R$) of hybridization intensities between TFIID ChIP DNA and a control DNA. Middle panel, RefSeq gene annotation. Bottom panel, a close-up view of two replicate sets of TFIID ChIP-on-chip hybridization signals around the 5′ end of the *TCFL1* gene. Arrows indicate the position of the TFIID-binding site determined by a peak-finding algorithm. **c**, Distribution of TFIID-binding sites relative to the 5′ end of the matched transcripts. **d**, **e**, Venn diagrams showing the number of identified promoters that matched EnsEMBL genes (**d**) or promoters annotated in DBTSS (**e**). **f**, Chart showing the percentages of IMR90 or DBTSS promoters overlapping with CpG islands, or containing conserved TATA box, INR or DPE elements (see Supplementary Information for details).

in a majority of these putative promoters (Supplementary Fig. S5). In addition, they were significantly enriched for core promoter motifs including INR (46%) and DPE (40%), and overlapped with CpG islands (40%, Supplementary Fig. S6). These results indicate that many of the putative promoter sequences that we have defined by TFIID-binding sites may indeed be functional promoters. There are 828 putative promoters located in the intergenic regions. These promoters, together with the 368 promoters that matched to transcripts outside the EnsEMBL genes, suggest the existence of 1,196 new transcription units outside the current gene annotation[18]. This number corresponds to about 13% of the 8,960 promoters that were matched to known genes. We therefore estimate that there are probably an additional 13% of human genes that remain to be annotated in the genome. This number agrees well with a recent estimate of the total number of human genes[18], but is considerably lower than estimates based on the number of transcripts detected by microarrays, serial analysis of gene expression (SAGE) and other methods[19–22]. It is conceivable that promoters for many low-abundance transcripts may be infrequently occupied by TFIID and possibly escaped detection by our assays. Alternatively, it is possible that the new transcripts detected by the other studies are products of a different transcription machinery or process.

Two notable features were apparent in this map of active promoters. First, we observed that large domains of four or more consecutive genes were simultaneously bound by PIC and probably transcribed in the IMR90 cells. At least 256 clusters, consisting of 1,668 EnsEMBL genes, can be classified into such regions, and the number of clustered promoters is highly significant ($P < 0.001$, Supplementary Table S5). The clustering of active promoters is consistent with previous findings that co-regulated genes tend to be organized into coordinately regulated domains[23–26]. Second, a large number of genes contained two or more active promoters (Supplementary Table S4). In general, these multiple promoters correspond to transcripts with either different 5′ UTR sequences or distinct first exons (for example, *PTEN*) but do not affect the open reading frames. In some cases, however, distinct proteins were produced from multiple promoters (for example, *NR2F2* and *WEE1*). In other cases, transcripts undergo differential splicing and polyadenylation (for example, *NFKB2* and *STAT3*). The widespread use of multiple promoters in this single cell type indicates greater complexity of the cellular proteome than previously expected, and also reveals highly coordinated regulation of transcriptional initiation, splicing and polyadenylation throughout the genome[27].

To verify experimentally our observations regarding multiple promoter use in IMR90 cells, we selected the *WEE1* gene for further analysis. Two TFIID-binding sites were mapped within this gene, corresponding to the 5′ ends of two distinct mRNAs, NM_003390 and AK122837 (Fig. 3a). Each mRNA encodes a distinct protein: one encodes a well-characterized full-length version of WEE1 protein, and the other only the kinase domain. We detected both transcripts in a steady-state, asynchronous population of IMR90 cells (Fig. 3b). The shorter transcript appears to be most abundant in the G0 phase, and the longer transcript is highly transcribed in both G0 and S phase (Fig. 3c), suggesting that the two promoters in the *WEE1* gene might have distinct cell-cycle functions.

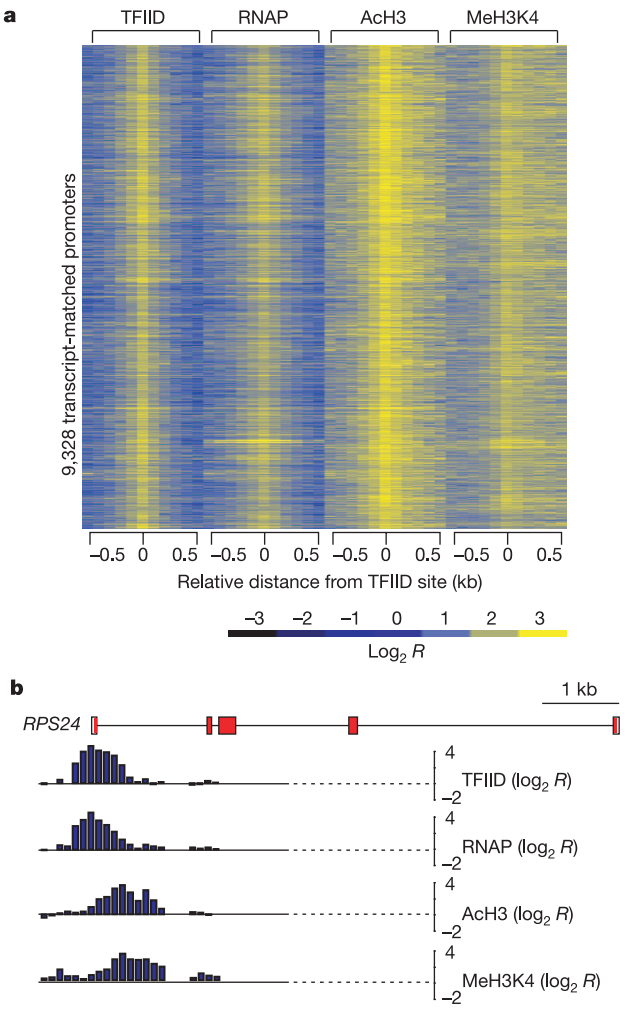

Figure 2 | The chromatin-modification features of the active promoters. **a**, Logarithmic ratios of the ChIP-on-chip hybridization intensities ($\log_2 R$) of probes from 0.5 kb upstream to 0.5 kb downstream of the identified TFIID-binding sites for TFIID, RNAP, AcH3 and MeH3K4 are plotted in a yellow–blue colour scale for 9,328 transcript-matched promoters. The bottom panel shows the colour scale with corresponding $\log_2 R$ values. **b**, A detailed view of TFIID, RNAP, AcH3 and MeH3K4 profiles on the promoter of *RPS24* gene.

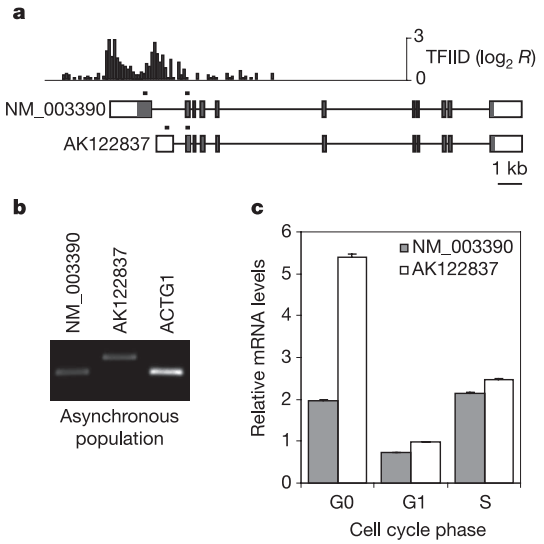

Figure 3 | Use of multiple promoters by human genes. **a**, Annotation of the *WEE1* gene locus and the corresponding TFIID-binding profile. Black bars over the first and second exons in transcripts indicate the positions of the primers used for analysis of each transcript, using real-time quantitative PCR with reverse transcription (RT–PCR). **b**, RT–PCR analysis of NM_003390 and AK122837 transcripts in an asynchronous population of IMR90 cells. **c**, Real-time quantitative RT–PCR analysis of NM_003390 and AK122837 transcripts in cell-cycle synchronized populations of IMR90 cells. Transcript levels observed for each cell-cycle phase were normalized to the level observed in the asynchronous population. Error bars represent standard deviation.

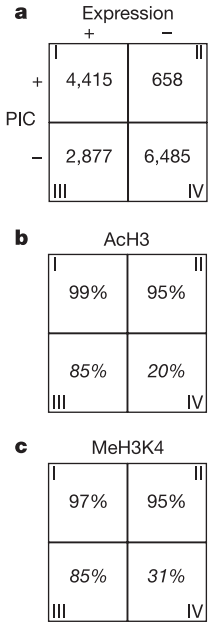

Figure 4 | Four distinct classes of promoters define the transcriptome of IMR90 cells. a, A 2 × 2 matrix describes the distribution of genes defined by expression and PIC occupancy on the promoter. **b**, **c**, Matrices showing the percentages of genes associated with AcH3 (**b**) or MeH3K4 (**c**) modification for each of the four classes of genes. Italicized numbers in some boxes represent extrapolation from the 29 ENCODE regions.

The active promoter map in IMR90 cells allowed us to systematically investigate the functional relationship between the transcription machinery and gene expression. We examined the genome-wide expression profiles of IMR90 cells and correlated the expression status of 14,437 EnsEMBL genes with promoter occupancy by the PIC. This comparison revealed four general classes of genes (Fig. 4 and Supplementary Table S6). Class I consists of 4,415 genes for which promoters were bound by the PIC and transcripts were detected. Class II includes 658 genes for which promoters were bound by the PIC but no transcript was detected. Class III contains 2,879 genes that were transcribed in IMR90 cells but for which the PIC was not detected on their promoters. Class IV contains the remaining 6,485 genes, for which the promoters were not bound by PIC and their corresponding transcripts were not detected.

The genes in class I and class IV, representing over 75% of the genes examined, support the general model that formation of the PIC on the promoters leads to transcription. The class II and III genes, on the other hand, are inconsistent with this model and may indicate that another mechanism is responsible for expression of these genes. We postulate that the discrepancy between PIC formation and transcription on the class II promoters can result from at least two possibilities. The first possibility is that the PIC assembles on these promoters, but that PIC formation is not sufficient to initiate transcription. Additional regulatory steps, such as promoter clearance or elongation, might be rate-limiting in the transcription of these genes[28]. Some notable examples in class II are the immediate early genes *FOS* and *FOSB*, the heat shock protein genes *HSPA6* and *HSPD1*, and the DNA damage repair genes *MSH5* and *ERCC4*. The second possibility is that transcription actually takes place at these promoters but that the resulting mRNAs are post-transcriptionally degraded, as in miRNA-mediated post-transcriptional silencing[29].

In contrast to class II, genes in class III appear to be transcribed, but the PIC binding on their promoters was not detected. This could simply be due to moderate sensitivity of our method[6]. To address this issue, we performed standard ChIP assays to detect binding of TFIID and RNAP on ten randomly selected class III gene promoters. Nearly 60% of the promoters were weakly associated with TFIID and RNAP in these cells, and were marked by enrichment ratios less than twofold but nonetheless above the observed background (Supplementary Fig. S2). Hence, the failure to detect TFIID and RNAP occupancy in roughly 60% of the class III promoters (~1,700) might be due to weak signals that fall below the detection sensitivity of our method. This result indicates that the promoters of a significant fraction of class III genes are open and accessible for transcription, but that PIC assembles on these promoters transiently, weakly or only during the early stage of fibroblast differentiation.

In order to understand the functional relationship between histone modification status and gene expression, we examined the AcH3 and MeH3K4 histone modifications in 29 ENCODE regions[7] (Supplementary Table S7), focusing specifically on the four classes of gene promoters. As expected, these epigenetic markers were associated with virtually all class I and class II genes, and the vast majority of class III genes. However, approximately 20% of the class IV genes were also associated with these markers (Fig. 4). This result indicates that a significant number of genes not actively transcribed are also associated with these epigenetic markers. We speculate that these histone modifications may serve to restrict genome expression potential and define the transcriptome capacity of the cell, and that transcription regulators and machinery collaborate with these epigenetic markers to further restrict the transcriptome to generate a unique pattern of genome expression.

Our results provide an initial framework for analysis of the *cis*-regulatory logic[30] in human cells. The high-resolution map of active promoters in IMR90 cells will enable detailed analysis of transcription factor binding sites within these regions. The promoter map described here can also serve as a reference for investigating gene expression in other cell types. We expect that a survey of additional cell types using the same approach will allow comprehensive mapping of all promoters in the human genome, and help elucidate the control logic that governs gene expression in different cell types in the body.

METHODS

Detailed descriptions of the experimental design and data analysis algorithms can be found in the Supplementary Information.

Briefly, IMR90 cells were obtained from the American Type Culture Collection and maintained under recommended conditions. ChIP-on-chip analysis was performed using commercial antibodies (anti-RNAP, MMS-126R, Covance; anti-TAF1, sc-735, Santa Cruz Biotechnology; anti-AcH3, 06-599, Upstate; anti-diMeH3K4, 07-030, Upstate) following the methods in ref. 6, with modifications. Microarray data from the initial 38 genome scan arrays were normalized, filtered and the TFIID-binding sites were identified as regions with a minimum of 4 probes separated by a maximum of 500 bp, with a logarithmic ratio of the ChIP-on-chip hybridization intensities ($\log_2 R$) greater than 2.5 standard deviations from the mean logarithmic ratio of the probes on each array. ChIP-on-chip hybridization intensities from the condensed arrays were normalized, averaged, and the TFIID-binding sites were identified using a computational peak-finding algorithm. The results were compared to annotated 5′-ends of transcripts from RefSeq, GenBank (downloaded from http://genome.cse.ucsc.edu; HG16, NCBI Build 34), DBTSS (http://dbtss.hgc.jp; Jan. 2004 version) and EnsEMBL (v26). Analysis of the promoter motifs was performed on a 400-bp sequence of each TFIID-binding site (from 200 bp upstream to 200 bp downstream) using matrices defined previously for the TATA box and the Inr and DPE elements. The analysis of CpG islands was carried out on a 1,200-bp sequence of each TFIID-binding site (from 1,000 bp upstream to 200 bp downstream). Standard ChIP assays were performed in duplicate with 0.5 ng of TFIID or RNAP ChIP DNA and the unenriched chromatin DNA from IMR90 cells using quantitative real-time polymerase chain reaction (PCR). Clusters of active promoters were defined by identifying runs of consecutive EnsEMBL genes with active promoters, and the significance of the number of genes found in the identified clusters was empirically determined by performing 1,000 times the same analysis on 6,763 randomly selected EnsEMBL genes. Gene expression analysis was performed in duplicate with total RNA extracted from the IMR90 cells using HU133 Plus 2.0 arrays (Affymetrix), according to the manufacturer's instructions.

Received 5 April; accepted 24 May 2005.
Published online 29 June 2005.

1. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72,** 449–479 (2003).
2. Tjian, R. & Maniatis, T. Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77,** 5–8 (1994).
3. Trinklein, N. D., Aldred, S. J., Saldanha, A. J. & Myers, R. M. Identification and functional analysis of human transcriptional promoters. *Genome Res.* **13,** 308–312 (2003).
4. Reinberg, D. *et al.* The RNA polymerase II general transcription factors: past, present, and future. *Cold Spring Harb. Symp. Quant. Biol.* **63,** 83–103 (1998).
5. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309 (2000).
6. Kim, T. H. *et al.* Direct isolation and identification of promoters in the human genome. *Genome Res.* **15,** 830–839 (2005).
7. The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
8. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnol.* **17,** 974–978 (1999).
9. Ruppert, S., Wang, E. H. & Tjian, R. Cloning and expression of human $TAF_{II}250$: a TBP-associated factor implicated in cell-cycle regulation. *Nature* **362,** 175–179 (1993).
10. Suzuki, Y., Yamashita, R., Sugano, S. & Nakai, K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.* **32** (database issue), D78–81 (2004).
11. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* **31,** 34–37 (2003).
12. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank: update. *Nucleic Acids Res.* **32** (database issue)), D23–26 (2004).
13. Birney, E. *et al.* Ensembl 2004. *Nucleic Acids Res.* **32** (database issue), D468–470 (2004).
14. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90,** 11995–11999 (1993).
15. Ohler, U., Liao, G. C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3,** RESEARCH0087 (2002).
16. Schubeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18,** 1263–1271 (2004).
17. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **32** (database issue), D109–111 (2004).
18. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
19. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306,** 2242–2246 (2004).
20. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14,** 331–342 (2004).
21. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20,** 508–512 (2002).
22. Rinn, J. L. *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* **17,** 529–540 (2003).
23. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA* **99,** 4465–4470 (2002).
24. Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1,** 5 (2002).
25. Roy, P. J., Stuart, J. M., Lund, J. & Kim, S. K. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418,** 975–979 (2002).
26. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289–1292 (2001).
27. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416,** 499–506 (2002).
28. Krumm, A., Hickey, L. B. & Groudine, M. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev.* **9,** 559–572 (1995).
29. Ambros, V. The functions of animal microRNAs. *Nature* **431,** 350–355 (2004).
30. Yuh, C. H., Bolouri, H. & Davidson, E. H. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279,** 1896–1902 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Kadonaga, R. A. Young, R. Kolodner, W. K. Cavenee, S. Van Calcar and C. K. Glass for discussion and comments on the manuscript. This research was supported by a Ruth L. Kirschstein National Research Service Award (T.H.K.) a Ford Foundation Predoctoral Fellowship (L.O.B.); the Ludwig Institute for Cancer Research (B.R.); NIH grants (B.R.) and the NSF (Y.W.).

Author Contributions B.R. and T.H.K. conceived the experimental design; T.H.K. performed the experiments; data analysis was by L.O.B. and C.Q.; microarray fabrication, hybridization and data acquisition were by M.A.S., T.A.R. and R.D.G.; M.Z. and Y.W. worked on the computational peak detection program; writing of the manuscript was primarily by T.H.K. and B.R.

Author Information The microarray data sets are available from GEO (Gene Expression Omnibus) under accession number GSE2672, and from http://licr-renlab.ucsd.edu/download.html. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu).