Genomic organization of human transcription initiation complexes

Bryan J. Venters† & B. Franklin Pugh†

The human genome is pervasively transcribed, yet only a small fraction is coding. Here we address whether this non-coding transcription arises at promoters, and detail the interactions of initiation factors TATA box binding protein (TBP), transcription factor IIB (TFIIB) and RNA polymerase (Pol) II. Using ChIP-exo (chromatin immuno-precipitation with lambda exonuclease digestion followed by high-throughput sequencing), we identify approximately 160,000 transcription initiation complexes across the human K562 genome, and more in other cancer genomes. Only about 5% associate with messenger RNA genes. The remainder associates with non-polyadenylated non-coding transcription. Regardless, Pol II moves into a transcriptionally paused state, and TBP and TFIIB remain at the promoter. Remarkably, the vast majority of locations contain the four core promoter elements—upstream TFIIB recognition element (BREu), TATA, downstream TFIIB recognition element (BREd), and initiator element (INR)—in constrained positions. All but the INR also reside at Pol III promoters, where TBP makes similar contacts. This comprehensive and high-resolution genome-wide detection of the initiation machinery produces a consolidated view of transcription initiation events from yeast to humans at Pol II/III/TATA-containing/TATA-less coding and non-coding genes.

The classic model for assembling the minimal core transcription machinery at mRNA promoters starts with the recruitment of TBP to the TATA box core promoter element (CPE). Next is the docking of TFIIB, which straddles TBP and locks onto flanking TFIIB-recognition elements (BRE, and BRE2,3). Together with TFIIF, TFIIB then engages Pol II in its active site to help set the transcription start site (TSS) at INR4–6. The recruitment of the transcription machinery has long been thought to be an important rate-limiting step in gene expression2. Concepts in transcription initiation by all three RNA polymerases (I, II and III) have been guided by this basic theme.

Clashing with this seemingly simplified view is the TATA box has been identified at only ~10% of human promoters9,10, with most genes ostensibly being classified as ‘TATA-less’ in all three RNA polymerase systems. The other CPEs are apparently equally rare. A second complication of the classic view, particular to multicellular eukaryotes, is that the general transcription factors may be largely pre-assembled at promoters. There, Pol II is in a transcriptionally engaged but paused state, approximately 30–50 base pairs (bp) downstream from the TSS11–13. A third complication is that transcription of genomes is not restricted to coding genes, but seems to be quite pervasive, without clear evidence of being coupled to definable promoters4–8. These complications, together, paint a seemingly complex picture of eukaryotic transcription initiation.

Towards reconciling simplistic models against complex data, we recently developed the ChIP-exo assay to map sites of protein–DNA interactions at near single-base resolution14. We discovered in yeast that so-called TATA-less promoters actually possess degenerate versions of the TATA box, and that most yeast promoters assemble the transcription machinery fundamentally in accord with the classic model, although a deep dichotomy between the TATA/SAGA/stress-induced genes and TATA-less/TFIID/housekeeping genes remains. This led us to consider whether similar simplicity was true in humans, albeit with the additional complications of paused polymerase and pervasive non-coding transcription.

TBP/TFIIB separation from paused Pol II

Using ChIP-exo, we detected 159,117 TFIIB locations (peak pairs) in K562 cells, of which 36% were associated with ENCODE-defined transcriptional domains17 (Extended Data Fig. 1a). Remarkably, half were associated with heterochromatic regions, which are generally thought to be devoid of stable RNA production. However, heterochromatic transcription may be more pervasive than previously thought.

We assigned a TBP/TFIIB location to more than 50% of all annotated protein-coding K562-expressed genes (Extended Data Fig. 1b), thereby providing independent validation. Seemingly expressed genes that lacked a TBP–TFIIB location may have arisen from several sources including rare but stable mRNAs, detection noise, and antisense transcription arising from a more distal promoter. TBP/TFIIB/Pol II occupancy and mRNA levels were correlated (Extended Data Fig. 1c), as expected of recruitment being at least partially rate-limiting in gene expression.

We initially focused on all 8,364 K562 TFIIB locations near the TSS of 6,511 coding RNAs as defined by RefSeq18. Figure 1a provides one example of the raw tag distribution and the identified CPEs concentrated ~25 bp upstream of the RPS12 ribosomal protein gene TSS. When individual genes were examined (Fig. 1b), or averaged across all 6,511 genes (Fig. 1c), two regions of high TFIIB/TBP/Pol II occupancy were observed. The major rightward peaks corresponded to primary promoter transcription initiated complexes (Fig. 1c, top panel). Those in the leftward direction matched divergent TSSs19–22, although the resulting RNA was less abundant than expected from TFIIB/TBP/Pol II occupancy levels (Fig. 1c, bottom versus top panel; note that secondary TSS represents only 24% of the total TSS signal). This may result from RNA instability, as seen in yeast. The clear spatial separation of complexes indicates that divergent transcripts arise from distinct initiation complexes, most (78%) of which were in CpG islands. On average, two complexes were detected per CpG island23, regardless of island length, with the centre of the island being enriched ~100 bp downstream of the primary TSS (Extended Data Fig. 2a, b). Complexes...
RETRACTED

**Figure 1** Transcription machinery organization at human mRNA promoters. **a.** Smoothed distribution of strand-separated ChIP-exo tag 5′ ends at the RPS12 gene. Core promoter elements (CPEs) are shown with lower casing denoting mismatches to the consensus. **b.** Peak-pair distribution of RNA at RefSeq genes (rows). Rows are linked, and sorted by TFIIB occupancy. **c.** Top, averaged ChIP-exo patterns around the closest (primary) RefSeq TSS. The ‘spikes’ of TBP and TFIIB are indiscernible (vertically offset in inset). Bottom, distribution of secondary polyadenylated RNA, with traces separated by sense (blue) and antisense (red, inverted trace) orientations relative to the corresponding mRNA TSS.

We looked for CPEs (illustrated in Fig. 2a) within the narrow intervals defined by 8,364 mRNA TSS-proximal TFIIB locations. Remarkably, and consistent with yeast4, nearly 85% of them had a sequence with 0–3 mismatches to the TATA-box consensus (TATAWAWR) (Fig. 2b, c). Less than 3% had a perfect match to the consensus. Deviations from the TATA-box consensus inversely correlated with TFIIB and TBP occupancy levels (Extended Data Fig. 3a), indicating that the TATA element sequence quality contributes to their occupancy level, consistent with previous observations20 on their in vivo functionality.

Several controls put the false positive rate for TATA elements at ~20% (Fig. 2c). First, 10,000 randomly generated sequences having the same human genome sequence bias found that only 16% were called by chance. Second, a scrambled version of the motif (having 0–3 mismatches) was identified only 20% of the time, and had no positional relationship with TFIIB/TBP (not shown). Third, coordinates having a single isolated tag were used to generate an essentially random set of false-positive locations, and the analysis repeated. TATA elements (0–3 mismatches) were identified only 20% of the time. Fourth, whereas control sequences were distributed randomly across the query space, the distribution of TATA elements was not random. Instead it displayed a tight peak 20 bp upstream of the TFIIB and TBP locations (Fig. 2d and data not shown).

TFIIB in complex with TBP makes sequence-specific contacts with BREu and BREd, which flank the TATA box2,3 and are upstream of the INR. Instead it displayed a tight peak 20 bp upstream of the TFIIB and TBP locations (Fig. 2d and data not shown).

**Figure 2** TATA elements at most mRNA genes. **a.** Core promoter schematic. **b.** Nucleotide distribution for TATA elements with 0–3 mismatches (panels) to the consensus, and sorted by ascending Pvalue. Colours are reflected in the logo colour. **c.** Cumulative percentage of TFIIB locations having a TATAWAWR sequence with 0–3 mismatches (solid line). Controls include a randomized sequence (60% GC, dashed black line), a scrambled consensus (dashed red line), and 8,364 locations represented by a single background tag (dashed grey line). **d.** Distance of strand-specific TFIIB peaks (exonuclease stop sites) from TATA element midpoints. Opposite-strand peaks are in red and inverted.

tended to be separated by 70–180 bp (Extended Data Fig. 2c, red), and had largely uncorrelated occupancies (Extended Data Fig. 2c, black), which suggests that they are generally regulated independently.

For the vast majority of transcription units, Pol II crosslinked 50 bp downstream of the primary TSS (Fig. 1b, c), where it is expected to pause after initiating transcription13. Pol II was most depleted over the core promoter, indicating that it does not stably reside there in proliferating K562 cells. Therefore, when Pol II enters the core promoter, it rapidly initiates transcription and then moves into a paused state ~50 bp downstream, thereby preventing any new polymerase from detectably engaging the core promoter.

The crosslinking pattern of human TFIIB was of particular interest because TFIIB in budding yeast crosslinks broadly across the relatively stable single-stranded DNA region within the Pol II active site at core promoters34, in accord with crystallographic models of ‘open’ complexes24. Remarkably, human TFIIB maintained its contact within this region, despite the absence of polymerase (Fig. 1c, top). Mechanistically, this might occur via TFIIB contacts with BREd (ref. 3 and see below), which are absent in budding yeast. The coincidence of TBP and TFIIB crosslinking at the BREu suggests that TBP may be predominantly crosslinking to TFIIB there, rather than directly to the TATA element.

**TBP, TATA, BREu and INR are common**

Several controls put the false positive rate for TATA elements at ~20% (Fig. 2c). First, 10,000 randomly generated sequences having the same human genome sequence bias found that only 16% were called by chance. Second, a scrambled version of the motif (having 0–3 mismatches) was identified only 20% of the time, and had no positional relationship with TFIIB/TBP (not shown). Third, coordinates having a single isolated tag were used to generate an essentially random set of false-positive locations, and the analysis repeated. TATA elements (0–3 mismatches) were identified only 20% of the time. Fourth, whereas control sequences were distributed randomly across the query space, the distribution of TATA elements was not random. Instead it displayed a tight peak 20 bp upstream of the TFIIB and TBP locations (Fig. 2d and data not shown).

TFIIB in complex with TBP makes sequence-specific contacts with BREu and BREd, which flank the TATA box2,3 and are upstream of the INR (Fig. 2a). However, these elements are essentially non-existent in yeast, and ill-defined across mammalian genomes. Using the identified TATA elements as a reference point, we searched upstream for BREu, and downstream for BREd and INR. Notably, in nearly every
instance a sequence with three or less mismatches to the literature-derived consensus for BRE\textsubscript{d} (SSRCGCC\textsuperscript{2}), BRE\textsubscript{u} (RTDKKKKK\textsuperscript{3}) and INR (YYANWYY\textsuperscript{27}) was found (Fig. 3a–c). Remarkably, sequences within BRE\textsubscript{d} and INR seemed to co-vary. For example, the BRE\textsubscript{d} consensus tended towards either GTKGGGG or ATKTTTT, rather than an equal mixture of all possible combinations (Fig. 3b), making them less degenerate than the consensus would suggest. Similarly, the INR consensus tended towards either CCANWCC or TTANWTT (Fig. 3c).

Sequence bifurcation was not observed with TATA or BRE\textsubscript{u} elements. Given the strong bias towards either strong (G/C) or weak (A/T) base pairing, this sequence dimorphism may reflect selection for distinct thermodynamic stabilities towards helix melting, which is an essential first step in initiation at these elements. Consistent with this, A/T-rich BRE\textsubscript{d} and INR elements had substantially higher crosslinking levels of TFIIB than their G/C-rich counterparts (not shown). However, this may not explain the strand bias of the sequences.

Similar to our TATA analysis, the TFIIB peak density was tightly focused at a fixed distance from each CPE (Fig. 3d), and were rarely found in randomized controls (Fig. 3e), thereby validating them. TFIIB peak pairs were centred over BRE\textsubscript{d}, suggesting that the primary crosslinking point is at or near the BRE\textsubscript{d}. Unlike the TATA element, the BRE and INR elements deviated relatively little from their consensus (compare Figs 2c and 3e), and such deviations did not correlate with TBP and TFIIB occupancy levels (not shown). Thus, BRE and INR sequence variability may regulate occupancy of the basal initiation complex to a lesser extent than TATA. Within their search space, the locations of each CPE peaked at previously defined canonical positions (Fig. 3f and Extended Data Fig. 3b), thereby providing cross-validation. These results suggest that although TFIIB occupancy levels varied from gene to gene, most were relatively constant at individual genes across cell lines.

150,000 non-coding initiation complexes

We next examined the remaining 150,753 putative TFIIB locations that were far (>500 bp) from a protein-coding gene (Supplementary Data 1). At a 20% false discovery rate per element, we identified at least three of the four CPEs at 97% of all non-mRNA TFIIB locations (Extended Data Fig. 4a). Deviations from the consensus were no more than at mRNA genes (average of 5 deviations across 28 positions within the four CPEs). TBP, TFIIB and Pol II peaked at the same canonical distances from each motif as found at mRNA promoters (Extended Data Fig. 4b, c). They were also embedded in a similar chromatin environment as mRNA promoters (Fig. 4a, b), but displayed comparatively lower TFIIB occupancy (Extended Data Fig. 4d).

Restricted motif spacing in promoters

We searched for an overall core promoter consensus (SSRCGGCTAT AWAWRNRTDKKKK(N)\textsubscript{13}YYANWYY) and CPE-spacing variants within 60 bp of all TFIIB locations, and plotted their distribution relative to TFIIB (Extended Data Fig. 6). Remarkably, the consensus spacing defined in Fig. 3f displayed the strongest positional relationship with TFIIB (Fig. 5a). For example, a consensus having the spacing between BRE\textsubscript{d} and TATA reduced by 1 bp displayed almost no positional relationship with TFIIB, as would be expected of a random/non-functional sequence.

There was very little or no tolerance for variable spacing between CPEs, which reflects structural constraints of the initiation complex. Surprisingly, proper spacing was accompanied by greater sequence deviations within individual CPEs (Fig. 5a), whereas small spacing deviations were accompanied by stronger elements (Fig. 5b). In short, core promoters may be weak by design, through a compensatory balance of sequence and spacing deviations from the consensus. This allows for greater dependence on transcriptional activators, but also provides for a specified basal output.

We conducted ChiP-exo mapping of TFIIB locations across four ENCODE cancer cell lines: HeLa S3, HepG2 and MCF7 in addition to K562 (cervical, liver, breast and blood, respectively). We detected TFIIB at 9,074 mRNA genes in at least one cell line, and at 1,691 genes in all lines (group 1 in Extended Data Fig. 7). Cluster analysis suggested that although TFIIB occupancy levels varied from gene to gene, most were relatively constant at individual genes across cell lines.
About one-third displayed noticeable cell-type specificity (for example, group 3 in Extended Data Fig. 7). For non-coding initiation complexes, we focused on those present in two or more cell types, and found 100,349 such locations (376,074 locations were found in at least one cell type). Non-coding complexes seemed to have more cell-type specificity and were bimodally distributed at high and low occupancy levels. This heterogeneity may reflect more numerous and diverse roles for the resulting non-coding transcription and/or RNA in cell-type specific physiology compared to proteins.

**tRNA genes have TATA and BRE**

With some exception\(^29\), transfer RNA genes have been classically defined as TATA-less, in which TFIIIC recognizes specific sequences downstream of the TSS, then recruits TFIIIB to a region immediately upstream of the TSS that lacks apparent sequence specificity\(^30,31\). Pol III then binds to form an initiation complex. TFIIIB contains TBP (and BRF, a factor related to TFIIIB), and thus it has been enigmatic as to how TBP in TFIIIB engages the upstream region without a TATA box.

Remarkably, TBP crosslinked \(~21\) bp upstream of 386 tRNA genes (Fig. 6a, left), as seen at Pol II promoters. In nearly every instance we found a TATA element (Fig. 6a, middle) that was \(~18\) bp further upstream (Fig. 6b). Similar to TBP crosslinking through TFIIIB, we suspect that TBP crosslinks through BRF. Indeed, the peaks of BRF upstream (Fig. 6a, right panel) and a BRE\(_d\) immediately upstream of TATA (not shown). Enrichment of these elements, but not the Pol-II-specific INR\(^33\), were statistically significant (Fig. 6c). Thus, TBP in complex

![Figure 4](image1.png) **Figure 4** | Non-coding TFIIIB locations have chromatin marks and non-polyadenylated RNA. **a**, Distribution of chromatin marks around TFIIIB at RefSeq genes (left) and ncRNA (right). **b**, TFIIIB locations that overlap with chromatin marks and epigenetic regulators\(^39\). **c**, Distribution of polyadenylated\(^38\) and non-polyadenylated\(^40\) RNA-seq tags around TFIIIB >500 bp from a RefSeq TSS. Percentages reflect TFIIIB having an RNA tag.

![Figure 5](image2.png) **Figure 5** | Restricted spacing of CPEs. **a**, Candidate core promoter enrichment at varying distances from all 159,117 TFIIIB locations, for spacing variants having motifs with weak and strong \(P\) values. **b**, Moving average of mismatches to the indicated motif consensus as a function of distance from TATA.
with a TFIIB family member engages a set of BREu–TATA–BREd CPEs similarly in Pol II and III systems.

**Consolidated genomic view of initiation**

Genome-wide mapping of the general transcription machinery at near single-base resolution offers a consolidated model of certain transcription initiation events from yeast to humans, Pol II to Pol III, TATA-containing to TATA-less, and mRNA to non-coding RNA (ncRNA). In general, a TFIIB/BRF family member is recruited to all TATA-containing to TATA-less, and mRNA to non-coding RNA transcription initiation events from yeast to humans, Pol II to Pol III, Genome-wide mapping of the general transcription machinery at Consolidated genomic view of initiation CPEs similarly in Pol II and III systems.

**Figure 6** | **TATA and BRE elements at most tRNA genes.** a. Left, TBP peak density separated by forward and reverse strand orientation (blue and red colours, respectively) relative to each tRNA TSS. Corresponding sequences are shown in the right two panels (provided in Supplementary Data 4). b. Average distribution of TBP peaks around all identified tRNA TATA elements. c. Cumulative percentage of tRNA genes with the indicated promoter element having 0–3 mismatches to the consensus. Dashed lines represent calculations for an equivalent number of randomized sequences for the colour-linked solid traces.

We detected promoter transcription initiation complexes at 25% of all ∼24,000 human coding genes, and found that there were 18-fold more non-coding complexes than coding. We therefore estimate that the human genome potentially contains as many as 500,000 promoter initiation complexes, corresponding to an average of about one every 3 kilobases (kb) in the non-repetitive portion of the human genome. This number may vary more or less depending on what constitutes a meaningful transcription initiation event. The finding that these initiation complexes are largely limited to locations having well-defined core promoters and measured TSSs indicates that they are functional and specific, but it remains to be determined to what end. Their massive numbers would seem to provide an origin for the so-called dark matter RNA of the genome34, and could house a substantial portion of the missing heritability35.

**METHODS SUMMARY**

Human cells were grown, treated with formaldehyde, and processed through the ChIP-exo assay as described in the Methods and elsewhere36. Sequence tags were normalized to input, peaks were called and paired, and pairs with >4 tags retained. FIMO37 was used to find literature-defined motifs within pre-defined distances from TFIIB peak-pairs or TATA elements. Illumina sequencing statistics, TFIIB locations, and Position-Specific Percentage Matrix (PSPM) tables are presented in Extended Data Table 1 and Supplementary Data 1 and 2, respectively.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 March; accepted 9 August 2013.
Published online 18 September; corrected online 2 October 2013 (see full-text HTML version for details).

1. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. Cell 56, 549–561 (1989).
2. Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Eibrigt, R. H. New core promoter element in RNA polymerase II-dependent transcription:
sequence-specific DNA binding by transcription factor IIIB. Genes Dev. 12, 34–44 (1998).
3. Deng, W. & Roberts, S. G. A core promoter element downstream of the TATA box that is recognized by TFII B. Genes Dev. 19, 2418–2423 (2005).
4. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev. Biol. 339, 225–229 (2010).
5. Kostrewa, D. et al. RNA polymerase II–TFIIB structure and mechanism of transcription initiation. Nature 462, 323–330 (2009).
6. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. Nature 495, 481–486 (2013).
7. Plassne, M. & Gann, A. Transcriptional activation by recruitment. Nature 386, 569–577 (1997).
8. Vannini, A. & Cramer, P. Conservation between the RNA polymerase II and III transcription initiation machineries. Mol. Cell 45, 439–446 (2012).
9. Kim, T. H. et al. A high-resolution map of active promoters in the human genome. Nature 436, 876–880 (2005).
10. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nature Genet. 36, 626–635 (2006).
11. Gilmore, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol. Cell. Biol. 6, 3984–3989 (1986).
12. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130, 77–88 (2007).
13. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339, 990–993 (2013).
14. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and RNA-antisense transcriptomes of human cells. Nucleic Acids Res. 37, D61–D65 (2009).
15. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic promoters. Annu. Rev. Biochem. 57, 873–914 (1988).
16. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic promoters. Annu. Rev. Biochem. 57, 873–914 (1988).
17. Ernst, J. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).
18. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D66 (2007).
19. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptome of human cells. Science 322, 1855–1857 (2008).
20. Seila, A. C. et al. Divergent transcription from active promoters. Science 322, 1849–1851 (2008).
21. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science 319, 1791–1792 (2008).
22. Fenouil, R. et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 22, 2399–2408 (2012).
23. Rozenberg, J. M. et al. All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. BMC Genomics 9, 67 (2008).
24. Sainsbury, S., Niessler, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II–TFIIB complex. Nature 493, 437–440 (2013).
25. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. Cell 116, 699–709 (2004).
26. Singer, V. L., Wobbe, C. R. & Struhl, K. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. Genes Dev. 4, 636–645 (1990).
27. Smale, S. T. & Baltimore, D. The “initiator” as a transcription control element. Cell 57, 103–113 (1989).
28. Kim, T. K. et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182–187 (2010).
29. Hamada, M., Huang, Y., Lowe, T. M. & Ma'ara, R. J. Widespread use of TATA elements in the core promoters for RNA polymerases III, II, and I in fission yeast. Mol. Cell. Biol. 21, 6870–6881 (2001).
30. Geiduschek, E. P. & Tocchini-Valentini, G. P. Transcription by RNA polymerase III. Annu. Rev. Biochem. 57, 873–914 (1988).
31. White, R. J. & Jackson, S. P. Mechanism of TATA-binding protein recruitment to a TATA-less class III promoter. Cell 71, 1041–1053 (1992).
32. Carrière, L. et al. Genomic binding of Pol III transcription machinery and relationship with TFIIIS transcription factor distribution in mouse embryonic stem cells. Nucleic Acids Res. 40, 270–283 (2012).
33. Berger, M. F. et al. Integrative analysis of the melanoma transcriptome. Genome Res. 20, 413–427 (2010).
34. Verrizer, C. P., Chen, J. L., Yokomori, K. & Tjian, R. Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. Cell 81, 1115–1125 (1995).
35. Kapranov, P. & St Laurent, G. Dark matter RNA: existence, function, and controversy. Front. Genet. 3, 60 (2012).
36. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Rev. Genet. 11, 446–450 (2010).
37. Berger, M. F. et al. Integrative analysis of the melanoma transcriptome. Genome Res. 20, 413–427 (2010).
38. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
39. Djebari, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Reja, S. Mahony, P. Albert and Y. Li for bioinformatic assistance, and M. Cousar and K.-Y. Chan-Salis for experimental support. This work was supported by National Institutes of Health grant GM059055.
Methods

Cell culture. Human chronic myelogenous leukaemia cells (K562, ATCC) were maintained between 1 × 10^3–1 × 10^5 cells ml^-1 in DMEM media supplemented with 10% bovine calf serum at 37 °C with 5% CO₂. Human adenocarcinoma cells from the cervix (HeLa S3, ATCC), liver (HepG2, ATCC) and breast (MCF7, ATCC) were grown in a similar manner as K562 cells except that they were maintained between 25% and 90% confluence. Cells were washed in PBS (1 × PBS, 8 mM Na₂HPO₄, 2 mM KH₂PO₄, 150 mM NaCl and 2.7 mM KCl) before incubation with formaldehyde in a final concentration of 1% for 10 min. Cells were lysed (10 mM Tris; pH 8, 10 mM NaCl; 0.5% NP40, and complete protease inhibitor cocktail (CPI; Roche)), and then the nuclei lysed (30 mM Tris, pH 8, 10 mM EDTA, 0.32% SDS, and CPI). Purified chromatin was resuspended in immunoprecipitation dilution buffer (40 mM Tris, pH 8.0, 7 mM EDTA, 56 mM NaCl, 0.4% Triton X-100, 0.2% SDS, and CPI) and sonicated with a Bioruptor (Diagenode) to obtain fragments with a size range between 100 and 500 bp.

ChIP-exo and antibodies. With the following modifications, ChIP-exo was carried out as previously described with chromatin extracted from 10 million cells, ProteinG MagSepharose resin (GE Healthcare), and 3 μg of TFIIB (Santa Cruz Biotech, sc-225), TBP (Santa Cruz Biotech, sc-204) or Pol II (Santa Cruz Biotech, sc-899), directed against the N terminus of the Pol II large subunit encoded by POL2RA).

Alignment to genome, peak calling and data access. Libraries were sequenced on an Illumina HiSeq sequencer. The entire length of the sequenced tags was aligned to the human hg18 reference genome using BWA with default parameters. Raw sequencing data are available at the NCBI Sequence Read Archive (accession SRA067908). The resulting sequence read distribution was used to identify peaks on the forward (W) and reverse (C) strand separately using the peak calling algorithm in GeneTrack (sigma = 20, exclusion zone = 40 bp). For strand-specific and strand-merged plots, sequencing tags were normalization to input. All 11,458 locations that were present in the ENCODE designated blacklist were removed from the analysis. Peaks were paired if they were 0–80 bp in the 3’ direction from each other and on opposite strands. Because patterns described here were evident among individual biological replicates, and replicates were well correlated, we merged all tags from biological replicate data sets to make final peak-pair calls. Between individual biological replicates, and replicates were well correlated, we merged all tags from biological replicate data sets to make final peak-pair calls. Within each group was considered. Distances between the two TFIIB peak-pair midpoint to the CP 3′ end) were then calculated. Their frequency distribution was then plotted as a 11-bp moving average.

Motif analysis. At each of these 6,511 promoters, using the MEME suite of tools, we searched for TATA elements within 80 bp of the midpoint of TFIIB-bound locations on the sense strand, first by searching for the consensus TATAAWWR (Supplementary Data 2), then sequentially for one to three mismatches to the consensus, if an element was not found. In rare cases in which multiple elements were found, we chose the one closest to the TFIIB peak. This rule had no qualitative effect on the data because such events were rare and choosing the furthest element gave the same result (not shown). Moreover, peak motif detection for BREu, TATA and INR was not centred over TFIIB, indicating that this distance criterion was not driving the observed motif enrichment at TFIIB locations. Using a similar strategy, we searched for candidate BREd elements (Supplementary Data 2) within 40 bp upstream of the 5,546 identified TATA elements, and searched for candidate BREd and INR elements (Supplementary Data 2) within 40 bp and 60 bp downstream of the 5,546 TATA elements, respectively. At Pol III promoters, candidate BREd elements were required to be within 20 bp of a TBP peak-pair midpoint, and in the same orientation as the TATA element.

Our searches infrequently picked up multiple motif instances within the search window. Where this did occur, we chose the motif with the best match to the published consensus (not the closest to TFIIB). In the situation where we obtained more than one motif with the same number of mismatches, we chose the one closest to TFIIB. Third, when we discard these multiple occurrences, the results qualitatively did not change. Fourth, the peak locations that we obtained for BREu, TATA and INR were not centred over TFIIB. Instead they peaked at the closest to TFIIB. This provided independent validation.

Using a core promoter (CP) PSPM matrix derived from individual CPE logos and spacing from Figs 2 and 3 (Supplementary Data 2), FIMO was used to find 39–47-bp CP sequences within 120 bp of a TFIIB peak pair, and had either a P value < 3 × 10^-4 or between 3 × 10^-4 and 10^-3. Only the strongest CP in each group was considered. Distances between the two (TFIIB peak-pair midpoint to the CP 3′ end) were then calculated. Their frequency distribution was then plotted as a 11-bp moving average.

41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

42. Albert, I., Wachi, S., Jiang, C. & Pugh, B. F. GeneTrack–a genomic data processing and visualization framework. Bioinformatics 24, 1305–1306 (2008).
Extended Data Figure 1 | Validation of ChIP-exo data and association with ENCODE annotated regions. a, Pie chart of all 159,117 TFIIB-bound locations in K562 cells parsed into ENCODE-annotated regions. b, Venn overlap among mRNA genes having TBP or TFIIB locations (<500 bp from its TSS) and genes with measured polyadenylated mRNA levels detected by RNA-seq. Data thresholding may contribute to non-overlapping sets. c, Moving average (100-gene) of mRNA levels versus TFIIB/TBP/Pol II occupancy levels on a median-centred log2 scale.
Extended Data Figure 2 | Distribution of TFIIB/TBP/Pol II in CpG islands that overlap mRNA TSSs. 

a, Peak-pair distribution for TFIIB, TBP and Pol II at the 5,095 CpG islands that overlap with the mRNA TSSs from Fig. 1b (78% overlap), and with the direction of transcription to the right. Rows are linked, and sorted by CpG island length. CpG island borders are indicated by blue and red bars, respectively. 
b, Shown is the averaged data from a. 
c, All 159,117 TFIIB locations were sorted by location, and inter-TFIIB distances calculated (red trace). Data were then sorted by distance, and the standard deviation of adjacent TFIIB occupancy ratios was calculated on a sliding window of 30 values. Peak calling parameters preclude detection of two separate TFIIB locations approximately <40 bp apart. Those that were 40–70 bp apart were correlated, whereas those >70 bp apart were less correlated or uncorrelated.
Extended Data Figure 3 | Properties of CPEs associated with RefSeq genes.

a. Average TFIIB and TBP occupancy parsed by the number of mismatches to the TATA consensus. b. Distribution of each candidate CPE relative to each other.
Extended Data Figure 4 | CPEs at non-coding loci bound by TFIIB. a, Bar graph showing the percentage of all 150,754 putative 'non-coding' TFIIB binding locations (>500 bp from an annotated RefSeq TSS) that have the indicated number of CPEs. b, Distribution of ChIP-exo peaks on each strand relative to the indicated CPE, for 150,754 putative non-coding TFIIB locations. Opposite strand traces (red) are inverted. c, Distribution of TBP (purple) and Pol II (black) peak-pair midpoints relative to the TATA motif midpoint derived from the 150,754 TFIIB putative non-coding locations. d, TFIIB occupancy versus percentage of locations that code for proteins. All 159,117 TFIIB locations were sorted by occupancy level, and the percentage of locations linked to an annotated RefSeq feature was plotted as a moving average.
RETRACTED
Extended Data Figure 5  |  Enrichment of different RNA fractions at 159,117 TFIIB locations throughout the human genome. Frequency distribution RNA 5′ ends for poly(A)⁺ (ref. 38) (top) and ENCODE project RNA fractions as indicated to the far left. Traces in the left panels are separated by sense (blue) and antisense (red, inverted) orientations relative to the corresponding mRNA TSS, which is directed to the right. Because the TSS orientation is not known for the poly(A)⁻ ncRNA loci, positive and negative strand tags were plotted relative to the TFIIB midpoint. The percentage of putative TFIIB locations that exist within 2 kb of an RNA tag are indicated in the top right corner of each plot.
Extended Data Figure 6 | TFIIB core promoter distances. Candidate CP at varying distances from all 159,117 TFIIB locations, for the indicated spacing variants (not all possible combinations were tested). Digits within spacing variant schematic reflect the base-pair spacing (N) between elements. CPE with high P values (less correlated to the PSPM matrix) have thin lines, whereas low/strong P values (≤3 × 10^{-4}) have thick lines.
Extended Data Figure 7 | Promoter complexes across cancer cell lines.

a, b, Occupancy levels for TFIIB linked to coding genes (a) and non-coding regions (b) in the indicated cell type were normalized by column. The colour scales represent the range of average-centred, log2-transformed values within each respective column. Detection in all four cell types defines group 1. Groups 2–4 were parsed by k-means clustering. Rows were sorted within groups based on TFIIB occupancy averaged across the four cell types (yellow, black, cyan and grey denote high, medium, low and zero occupancy, respectively). For clarity in b, TFIIB locations that were detected in only one cell line were excluded from clustering. Columns were hierarchically clustered. The MCF7 data set had 20–30% of the coverage of other cell lines (reported in Supplementary Data 3), which probably accounts for an excessive number of zero-occupancy loci (grey).
Extended Data Table 1 | Statistics of Illumina sequencing

| Factor | Antibody | Cell Line | Total Reads | Uniquely Mapped Reads | Unique Mapping Rate |
|--------|----------|-----------|-------------|-----------------------|--------------------|
| Input  | none     | K562      | 126,007,656 | 104,591,819           | 83%                |
| Input  | none     | K562      | 109,745,112 | 91,160,835            | 83%                |
|        | Totals:  |           | 235,752,768 | 195,752,654           |                    |
| TBP    | sc-204   | K562      | 97,896,951  | 60,581,579            | 62%                |
| TBP    | sc-204   | K562      | 181,420,753 | 132,655,896           | 73%                |
| TBP    | sc-204   | K562      | 200,167,837 | 115,213,419           | 58%                |
|        | Totals:  |           | 479,485,541 | 308,450,894           |                    |
| TFIIB  | sc-225   | K562      | 64,473,390  | 43,727,825            | 68%                |
| TFIIB  | sc-225   | K562      | 129,513,614 | 80,930,721            | 62%                |
|        | Totals:  |           | 193,987,004 | 124,658,546           |                    |
| Pol II | sc-899   | K562      | 40,833,504  | 31,260,456            | 77%                |
| Pol II | sc-899   | K562      | 119,799,682 | 88,431,598            | 74%                |
|        | Totals:  |           | 160,633,186 | 119,692,054           |                    |
| TFIIB  | sc-225   | HeLa-S3   | 62,249,055  | 41,815,431            | 67%                |
| TFIIB  | sc-225   | HeLa-S3   | 185,240,056 | 123,002,393           | 66%                |
|        | Totals:  |           | 247,489,111 | 164,817,824           |                    |
| TFIIB  | sc-225   | HepG2     | 78,313,847  | 50,505,201            | 64%                |
| TFIIB  | sc-225   | HepG2     | 264,530,278 | 172,112,282           | 65%                |
|        | Totals:  |           | 342,844,125 | 222,617,483           |                    |
| TFIIB  | sc-225   | MCF7      | 25,615,261  | 14,780,271            | 58%                |
| TFIIB  | sc-225   | MCF7      | 120,958,757 | 28,600,410            | 24%                |
|        | Totals:  |           | 146,574,018 | 43,380,681            |                    |

Summary of uniquely mapped sequencing reads for each biological replicate.
Retraction: Genomic organization of human transcription initiation complexes

Bryan J. Venters & B. Franklin Pugh

We reported the presence of degenerate versions of four well known core promoter elements (BRE\textsubscript{ω}, TATA, BRE\textsubscript{θ} and INR) at most measured TFIIB binding locations found across the human genome. However, it was brought to our attention by Matthias Siebert and Johannes Söding in the accompanying Brief Communication Arising (Nature 511, E11–E12, http://dx.doi.org/10.1038/nature13587; 2014) that the core-promoter-element analyses that led to this conclusion were not correctly designed. Consequently, the individual core promoter elements were not statistically validated, and therefore there is no evidence of specificity for most reported core-promoter-element locations. To the best of our knowledge, the raw and processed human TFIIB, TBP and Pol II ChIP-exo data are valid, but subject to standard false discovery considerations. We therefore retract the paper. We sincerely apologize for adverse consequences that may have arisen from the error in our analyses.