doi:10.1093/nar/gkq956

Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq

Yi Liu[1], Dali Han[1], Yixing Han[1], Zheng Yan[2], Bin Xie[3], Jing Li[1], Nan Qiao[1], Haiyang Hu[2], Philipp Khaitovich[2,4], Yuan Gao[3] and Jing-Dong J. Han[1,2,*]

[1]Chinese Academy of Sciences Key Laboratory of Molecular Developmental Biology, Center for Molecular Systems Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Lincui East Road, Beijing, 100101, [2]Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai, 200031, China, [3]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA and [4]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig, Germany

Received August 7, 2010; Revised September 29, 2010; Accepted September 30, 2010

ABSTRACT

Rhesus macaque is a widely used primate model organism. Its genome annotations are however still largely comparative computational predictions derived mainly from human genes, which precludes studies on the macaque-specific genes, gene isoforms or their regulations. Here we took advantage of histone H3 lysine 4 trimethylation (H3K4me3)'s ability to mark transcription start sites (TSSs) and the recently developed ChIP-Seq and RNA-Seq technology to survey the transcript structures. We generated 14 013 757 sequence tags by H3K4me3 ChIP-Seq and obtained 17 322 358 paired end reads for mRNA, and 10 698 419 short reads for sRNA from the macaque brain. By integrating these data with genomic sequence features and extending and improving a state-of-the-art TSS prediction algorithm, we *ab initio* predicted and verified 17 933 of previously electronically annotated TSSs at 500-bp resolution. We also predicted approximately 10 000 novel TSSs. These provide an important rich resource for close examination of the species-specific transcript structures and transcription regulations in the Rhesus macaque genome. Our approach exemplifies a relatively inexpensive way to generate a reasonably reliable TSS map for a large genome. It may serve as a guiding example for similar genome annotation efforts targeted at other model organisms.

INTRODUCTION

Rhesus macaque is a widely used model species for primate studies, including drug and virus infection tests, evolutionary sequence comparisons and so on. Since its full genome sequencing became available in 2007 (1), various computational gene predictions, mainly through comparing with human gene structures, have been included in the UCSC genome browser, including SGP gene prediction (2,3), Ensembl gene prediction (4) and the aligned non-Rhesus RefSeq genes (5). However, electronic gene annotations not only overlook many macaque specific genes but are also unreliable. Large fractions of the human genes are not alignable to the macaque genome (1). Meanwhile, experimentally validated transcript information is available for only no more than 1000 genes in the macaque genome (6).

An important application of the macaque genome sequence is to use it to compare the transcription regulations in different primate species, especially to those of our own. Transcription start sites (TSSs) are very important landmarks for locating transcription regulatory regions and elements for genes. Relying only on comparative genomic annotation, many novel or alternative TSSs specific to the Rhesus Macaque genome are missed or labeled incorrectly. Here, we aim to perform an

*To whom correspondence should be addressed. Tel: +86 10 64845843; Fax: +86 10 64845797; Email: jdhan@genetics.ac.cn; jackie.han@research.dfci.harvard.edu

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

unbiased survey and *ab initio* predictions of TSSs in the macaque genome.

H3K4me3 has been shown to specifically and sharply mark the TSS of genes (7,8). It has recently been used to identify long intergenic non-coding RNAs (lincRNA) in the mouse and human genomes (9,10). Although the level of promoter H3K4me3 is largely correlated with the expression level of the genes (7,8,11), this mark also marks repressed genes, especially for genes having high CpG in their promoters (8,12,13). We therefore carried out genome-wide H3K4me3 ChIP-Seq (chromatin immunoprecipitation followed by deep sequencing) to identify the TSSs in macaque brains. We further used RNA-Seq (massively parallel deep sequencing) of mRNA and small RNA (sRNA) to validate and further refine the TSS predictions derived from H3K4me3 ChIP-Seq data.

Finally, to facilitate the usage of our *de novo* TSS predictions, we deposited our predictions at http://hanlab.genetics.ac.cn/Rhesus-TSS for querying and visualizing these TSSs together with genome annotations and our ChIP-Seq and RNA-Seq data. We have also uploaded all the new deep sequencing data generated in this study (input DNA control, ChIP-Seq, and mRNA-Seq) to NCBI Gene Expression Omnibus under accession no. GSE24538.

MATERIALS AND METHODS

Tissue sample collection and preparation

The rhesus macaque samples were obtained from the Suzhou Experimental Animal Center (Suzhou, China). All macaque individuals used in this study suffered sudden deaths for reasons other than their participation in this study and without any relation to the tissue used. For all individuals the brain tissue was frozen in liquid nitrogen within 20 min from the time of death and then stored at −80°C. The cerebral cortex samples were dissected from postmortem frozen brain on dry ice. All samples had excellent tissue preservation and contained RNA of comparable and high quality.

ChIP-Seq

An amount of 0.25 g of a 9-year-old male macaque cortex were grinded in liquid nitrogen, suspended in cold 1× PBS, and chemically cross-linked by addition of formaldehyde to a final concentration of 1% for 15 min at room temperature, then the cross-linking reaction was quenched by adding glycine to a 125 mM final concentration. The chromatin preparation and immuneprecipitation procedure is as described in ref. (14). An amount of 6 µg antibodies (anti-H3K4me3, Abcam ab8580) were incubated with the sonicated chromatin fragments.

A sample of input genomic DNA without the ChIP procedure above was used to construct a control library, which was also sequenced at similar depth with 13 894 402 and 9 314 545 total and unique reads.

Sequencing library construction for the ChIP DNA, cluster generation and sequencing analysis using the Illumina 1G Genome Analyzer were performed following manufacture's protocols. Sequencing tags of 36-mer were obtained by the single-end pipline.

Gene annotations

Computationally predicted genes (non-Rhesus RefSeq Genes) were downloaded from UCSC genome browser on 13 October 2009 (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid = 151755641&c = chr7&g = xenoRefGene).

The experimentally validated transcription units were downloaded from http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid = 151755641&c = chr7&g = refGene.

ChIP-Seq tag mapping and H3K4me3 peak detection

We used SOAP2 (15) with parameters '−r 0' to align H3K4me3 reads to rheMac2 genome (downloaded from UCSC genome browser), discarding multiple alignment reads. Modification peaks were detected by using the SICER software with $P < 10e-3$ (16), where the sequence tags of input genomic DNA library were used for background subtraction in the SICER program.

Training TSS predictors

Similar to the CoreBoost_MH algorithm (12), the intensity profile of H3K4me3 (and RNA-Seq) signals for a centered genomic position is represented as a 49 dimension vector. Each bin in the vector corresponds to the sum of reads [−100 bp, +100 bp] surrounding 1 of the 25-bp uniformly spaced genomic positions within [−600 bp, +600 bp] of the profile center.

We used decision stumps trained on single H3K4me3/sequence features as the weak learners in the GentleBoost algorithm. Unlike general classification and regression trees, decision stumps are single node decision trees which are very conservative and do not tend to overfit in Boosting algorithms. Similarly, the GentleBoost algorithm is also very conservative in additively combining many weak learners into a strong classifier (17). Partly due of this reason, we find the performance of the final CpG/non-CpG TSS classifier is not sensitive to the number (>30) of weak learners. For computational purpose, we always use 200 weak learners for classifier training throughout this work. Besides, similar to the Real Adaboost and Logitboost algorithm, weak learners in the GentleBoost classifier generate real-valued predictions which are linearly combined by the Boosting algorithm to form the strong hypothesis. As a result, the absolute value of the final prediction score reflects the confidence that the instance is classified correctly (17).

Detecting candidate TSS positions from GentleBoost scores

After sliding the GentleBoost classifier through the H3K4me3 peaked regions in the genome, scores are assigned to the genomic positions at 10-bp step intervals to represent the tendency that each position is a true TSS. As expected, genomic coordinates in the vicinity of a true TSS will also receive high scores. To eliminate these false positive detections, we adopt the following approach to detect candidate TSS positions: (i) eliminate the genomic positions whose GentleBoost scores <1.5 *Z*-scores from being a true TSS. In other words, the threshold is set to be the average score plus 1.5 times of the standard variation; (ii) find the highest scoring genomic position from candidate TSS positions (Z-score > 1.5) and report this position as a true TSS; (iii) eliminate the possibility that another TSS appears in the [−200 bp, +200 bp] region of this TSS. If there is other candidate TSS appearing in this region, we continuously expand the region 200-bp upstream and 200-bp downstream until no candidate TSS is encountered during the expansion; and (iv) go to Step 2 and find another TSS.

RNA-Seq for mRNA

Total RNA was extracted from ~100 mg of the dissected frozen prefrontal cortex tissue using Trizol® reagent (Invitrogen, Carlsbad, CA, USA) from five male individuals of 8-, 9-, 10-, 11- and 14-year old. An amount of 4 µg of total RNA isolated from each individual were pooled together to perform twice Oligo(dT) selection using Oligotex® mRNA Midi Kit (Qiagen). After selection, 100 ng mRNA was first fragmented by addition of 5× fragmentation buffer (200 mM Tris-acetate, pH 8.2, 500 mM potassium acetate and 150 mM magnesium acetate) and heating at 94°C for 2 min 30 s in a thermocycler, then transferred to ice and run over a Sephadex-G50 column (USA Scientific) to remove the fragmentation ions (18). We used random hexamer primers (Invitrogen, Cat. No. 48190-011) for reverse transcription of fragmented mRNA to double-stranded cDNA. Sequencing libraries were prepared according to the paired-end sample preparation protocol (http://www.illumina.com) and sequenced as 75-mer X2 using the Illumina 1G Genome Analyzer paired-end pipeline.

RNA-Seq for small RNA

We have previously deposited this data to NCBI Gene Expression Omnibus with ID GSM450615 through an independent study (see NCBI GEO series GSE18013 and reference therein). Specifically, low molecular weight RNA isolated from 10 µg total RNA isolated from the 9-year-old male macaque was ligated to the 5′ and 3′ adapters separately. After reverse transcription and 15 cycles of amplification, the small RNA library was sequenced as 36-mer using the Illumina 1G Genome Analyzer single-end pipeline.

RNA-Seq tag mapping to the human and macaque genomes

MicroRNA sequence tags were mapped to rheMac2 genome by SOAP2 (15), discarding multiple alignment reads. Pair-end mRNA sequence tags were mapped to rheMac2 genome by TopHat1.0.12 (19) with the following parameters: '-r 100 -a 8 -m 0 -g 100 –solexa1.3-quals –coverage-search –microexon-search –segment-mismatches 2 –segment-length 25'.

TSS validation and refinement

We checked whether there exists an mRNA/sRNA-Seq tag within the [−500 bp, +500 bp] region of each TSS. If the answer is true, the TSS is believed to be 'validated' by the mRNA/sRNA-Seq data.

An upward edge is defined to be located in a position where there is no mRNA-Seq signal in [−20 bp, −1 bp] but at least one tag in [1 bp, 20 bp], and there is precisely an mRNA-Seq tag at this position. All validated predictions were refined to the nearest mRNA upward edge instead of sRNA edge, as the mRNA tags are shorter and therefore more precisely mapped to the genome.

For mRNA validated TSSs, the mRNA-seq data were used in computing the transcription direction; while for sRNA validated TSSs, the union of mRNA/sRNA-Seq signals were used in this computation.

RESULTS

H3K4me3 ChIP-Seq to mark TSSs

An outline of our TSS identification procedure is briefly summarized in Figure 1. We first performed ChIP with anti-H3K4me3 antibodies using macaque cerebral frontal cortex samples, and subjected the ChIP DNA to deep sequencing by Illumina Genome Analyzer II ('Materials and Methods' section). A total of 14 013 757 and 13 894 403 reads, and 10 023 993 and 9 162 762 unique reads were obtained for H3K4me3 ChIP and control input DNA, respectively. As expected, H3K4me3 tag counts show very sharp peaks at TSSs, with different shapes for CpG and non-CpG genes (Supplementary Figure S1). The CpG genes display two asymmetric peaks at TSSs, with the one after TSSs much broader and higher than the one before TSSs. The non-CpG genes contain only one peak at TSSs, which is slightly shifted toward the downstream of TSSs (Supplementary Figure S1).

Positive and negative training data sets

Within the 530 genes whose structures have been experimentally validated, 162 are CpG promoters, whereas the rest (368) are non-GpG promoters. These promoter regions are used as positive training cases for a TSS predictor. Similar to the CoreBoost_HM approach (12), which was designed for TSS re-annotation in the human genome, we combined sequence features as well as the intensity profile of H3K4me3 to predict TSSs for CpG and non-CpG promoters. Specifically, we used the cosine similarity and dot product of each H3K4me3 intensity (tag counts) profile from ±600-bp region surrounding the annotated TSSs to the average profiles of known CpG and non-CpG promoters to characterize the histone modification signal at the TSSs ('Materials and Methods' section). To characterize sequence property at core promoter regions, we used the CoreBoost package (20) to generate 29/31 features at each CpG/non-CpG TSS. These features include CpG-island, TATA box or Inr scores, k-mer frequencies, energy properties of nucleotides and so on (20). The union of these two types of features was used to train a Gentle Adaboost (GentleBoost) classifier for TSS prediction (17). Specifically, besides the 162/368 true positive examples of known CpG/non-CpG TSSs, two different ways were employed to generate negative training cases. The first way is similar to the one used in (12), that is, we generated six negative training cases (non-TSSs) randomly in the

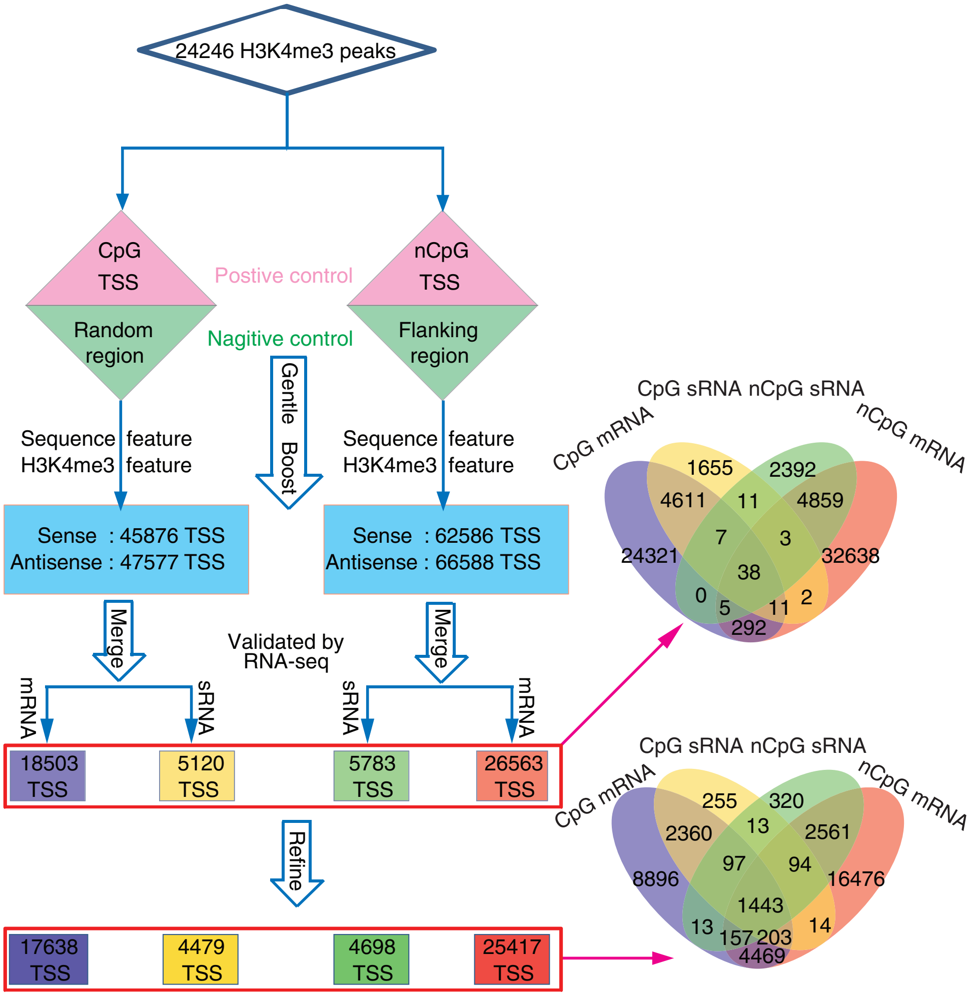

Figure 1. The flowchart of our *ab initio* TSS prediction approach for the Rhesus macaque genome. First, the SICER software (16) is employed to detect 24 246 H3K4me3 peaks throughout the Rhesus Macaque genome. Then, combined with raw sequence features, the dot product and cosine similarity features of the H3K4me3 profile are used to train two GentleBoost classifiers for predicting CpG and non-CpG TSS separately. Here, the 162 and 368 known CpG/non-CpG TSS were used as positive control in classifier training, while negative examples were randomly sampled from background genomic regions and from ([−1200 bp, −300 bp], [300 bp, 1200 bp]) flanking genomic regions for CpG and non-CpG TSS, respectively. Using the CpG and non-CpG classifiers to scan both positive and negative strands of the 24 246 H3K4me3 peak regions, we predict 45 876 (on positive strand) and 47 577 (on negative strand) CpG TSS, 62 586 (on positive strand) and 66 588 (on negative strand) non-CpG TSS. Then, positive and negative strand TSS predictions within a 400-bp moving window were merged to the genomic position with the largest GentleBoost score to reduce the redundancies. These predictions were further validated with the existence of mRNA-/sRNA-Seq signals. Finally, the precise genomic locations of the validated TSSs were refined to the nearest upward edge of mRNA-Seq signals in the predicted transcription direction. Upper right panel: the Venn diagram of the mRNA/sRNA-Seq validated CpG/non-CpG TSSs. Lower right panel: the Venn diagram of the final refined sets of CpG/non-CpG TSS predictions.

flanking region (three from [−1200 bp, −300 bp] and three from [+300 bp, +1200 bp]) of each known TSS. The second way was to generate (H3K4me3/sequence) features at approximately 10 000 genomic positions, which are randomly distributed on the genome.

Classification algorithm

Gentleboost algorithm has been suggested to be more numerically immune to mislabeled examples than Real Adaboost and Logitboost algorithm and generally yields better classification performance (17). In our approach, there is no warrantee that all negative training data are strictly non-TSSs. As a result, we used the GentleBoost algorithm rather than the Logitboost algorithm as in (12,20) to obtain more robust and accurate TSS prediction results ('Materials and Methods' section). This is indeed the case when the performances of the two algorithms are compared on the macaque data (See Supplementary Figure S2 and Section 'ROC to evaluate and compare the performances of predictors and improvements' below).

Genome-wide scan of H3K4me3 peaks for *ab initio* TSS prediction

The above algorithm was applied to the Rhesus macaque genome to predict CpG and non-CpG TSS on both positive and negative strands. Specifically, we first detected the peaks of H3K4me3 signal in the Rhesus macaque genome to pinpoint genomic regions that are highly enriched for TSS by using the SICER software (16) ('Materials and Methods' section). Then, the

GentleBoost classifier was applied to scan genomic regions enclosing these H3K4me3 peaks at 10-bp resolution. The classifier assigns for each position a real-valued score (17), which quantifies the likelihood that a real TSS is located at this position. Then the positions with local maximal GentleBoost scores are reported as candidate TSSs ('Materials and Methods' section) (Figure 1).

Cross-validation experiments demonstrate the effectiveness of the TSS detector

To investigate the accuracy and effectiveness of the aforementioned approach to TSS detection, we first examined the accuracy of predicting a random subset (~30%) of known CpG and non-CpG TSSs by using the remaining (~70%) CpG/non-CpG TSSs as positive training data. After the learning step, the two GentleBoost classifiers were used to scan the [−2 kb, +2 kb] region of the two held-out sets of TSS, either along or reversely with the direction of transcription. Then, we compared the distance between the true TSS sites with the positions with local maximal GentleBoost scores. The receiver operating characteristic (ROC) curves of the distance gap versus the percentage of within-gap true TSS predicted clearly demonstrate the effectiveness of our approach for detecting CpG and non-CpG TSS at relatively high accuracy, which reaches >80% within 500-bp prediction gap (Supplementary Figure S3).

Validating TSS prediction scores with the presence of electronically annotated TSSs

After obtaining these *ab initio* TSS predictions, we compared them with the recent Rhesus macaque TSS annotation derived from the human genome. Briefly, this comparison has two functions: one is to offer global statistics for quantifying the accuracy of our prediction and the other is to exemplify that our prediction indeed detects novel TSSs or refines the positioning of the existing electronically annotated TSSs.

We find a strong correlation [Pearson correlation coefficient (PCC) = 0.879 along 100 tiles of prediction scores] of the prediction scores to the percentages of predicted TSSs that have nearby electronically annotated TSSs, suggesting that the higher the prediction score, the more likely the TSS can be validated by homology-based electronic gene annotation transfer (Figure 2A and B).

RNA-Seq for mRNA and sRNA

Genome-wide transcripts so far have not been studied for Rhesus macaque, and it might be useful for increasing the accuracy of TSS prediction and providing further experimental support for newly predicted TSS. We therefore carried out RNA-Seq experiments also using mRNA or sRNA from macaque frontal cortex, to see whether there exist sharp RNA-Seq signals near the detected TSSs. Using Illumina Genome Analyzer, we obtained 17 322 358 paired end reads for mRNA, and 10 698 419 short reads for sRNA, among which 22 931 989 and 3 310 618 single-end tags can be uniquely mapped to the macaque genome, respectively.

Note that although RNA-Seq data consistently map to TSSs, using RNA-Seq data alone, there is no way to distinguish TSSs versus exon–intron boundaries, which are present very frequently in almost all primate genes. In our study, using H3K4me3 ChIP data greatly reduced such false positive TSS predictions at the splicing sites.

Validating TSS prediction scores with the presence of nearby mRNA or sRNA

We find a strong correlation (PCC = 0.898 along 100 tiles of prediction scores) of the prediction scores to the percentages of predicted TSSs that have nearby (within 500 bp) (s)RNA-Seq reads, suggesting that the higher the prediction score, the more likely the TSS can be validated by RNA-Seq signals (Figure 2A and B).

ROC to evaluate and compare the performances of predictors and improvements

The overall distance gaps of the *ab initio* predicted TSSs to annotated genes (a surrogate for ROC curves, and referred to as ROC below) can be used to evaluate the performance of different predictors.

Indicated by the ROCs, using the flanking negative training data is better for predicting non-CpG TSSs, while negative training data from random genomic backgrounds is good at predicting CpG TSSs (Figure 2C and D and Supplementary Figure S4). Since the G/C percentages of non-CpG TSSs and flanking regions of CpG TSSs are generally higher than that of the genomic background, the training data generation strategy above reflects the fact that CpG-like features are of uttermost importance for predicting CpG TSSs, while the G/C abundance at non-CpG TSS regions might be a confounding factor for non-CpG TSS prediction. It is not surprising that the above hybrid training strategy achieves good performance since it highlights the CpG features for predicting CpG TSS while reduces it for predicting non-CpG TSS.

We also find that the use of cosine similarity between two vectors yields much better prediction accuracy than PCC, [which was used in (12)] for predicting non-CpG TSSs (Figure 2E and F). TSS prediction accuracy using the Gentleboost algorithm is also enhanced compared with the Logitboost algorithm used in ref. (12) for both CpG/non-CpG TSS prediction (Supplementary Figure S2). The above three improvements on the classification algorithm over the algorithm described by Wang *et al.* (12) greatly enhanced its performance.

The ROC plot clearly indicates that TSS predictions supported by nearby RNA-Seq signals overlap better with the known TSSs (Figure 3A and B), hence filtering using mRNA-Seq and sRNA-Seq data can further increase the accuracy of TSS mapping. Interestingly, we find that TSS predictions overlapped with sRNA signals tend to have higher log-odds scores, suggesting that sRNA signal is a better TSS predictor than mRNA signal (Figure 3A and B). This might be because mRNA signals can be mapped to many different exons, whereas sRNA signals more often map to a single exon for a particular gene and that many sRNAs are associated with TSS (see below).

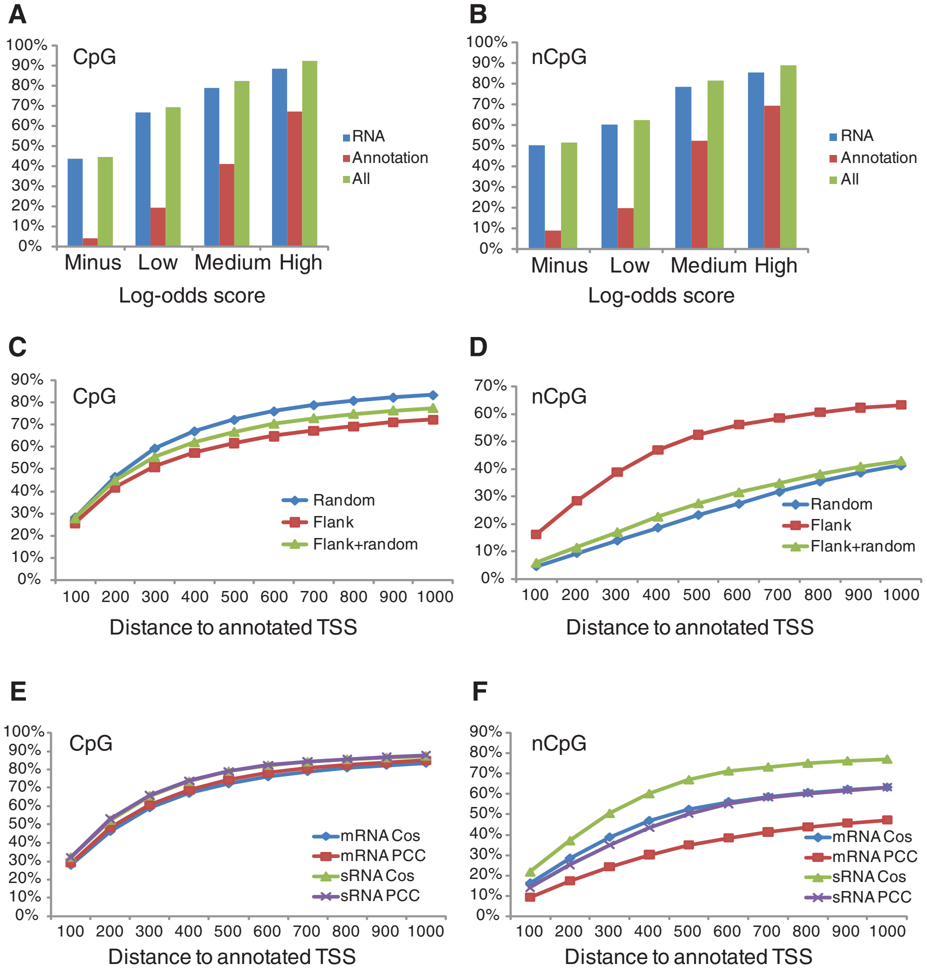

Figure 2. Evaluating the performance of the TSS classifier. (**A**) and (**B**) Correlation of the CpG (A) or non-CpG (B) TSS log-odds scores with the probability or percentage of the predicted TSSs containing electronically annotated TSSs, RNA-Seq signals, or either annotated TSSs or RNA-Seq signals within 500 bp. 'Low', 'medium', 'high' correspond to the sets of TSSs whose probability of being a true TSS (implied by log-odds scores) are >50, 95, 99%, while 'minus' denotes the predictions with negative log-odds scores. (**C**) ROCs of using different negative training samples for predicting CpG TSS. 'Random' or 'Flanking' means the negative training samples were randomly selected from the whole genome or the flanking regions of known TSS. 'Flanking + Random' means we combined the two sets above as negative training examples. Each point in an ROC shows the percentage of the TSS predictions that are supported by electronic TSS annotations within a certain distance. Given the different numbers of predictions made in different training strategies, only the top 10 000 predictions with the largest GentleBoost scores in each training scenarios are compared. (**D**). ROCs of using different negative training examples for predicting non-CpG TSS. The format of the graph is the same as in C. (**E**) and (**F**). ROCs of using cosine similarity compared with using PCC for predicting CpG and non-CpG TSSs, respectively. The format of the graphs is the same as in C. All TSS predictions in C–D are based on mRNA-seq validation (see section 'TSS validation and refinement', and those based on sRNA-seq are shown in Supplementary Figure S4). The TSS predictions in E and F are based on either mRNA-seq or sRNA-seq validation.

Directionality of TSS

The directionality of a TSS is important for understanding the transcriptional structure of the associated gene. The H3K4me3 profile at TSS shows a clear directionality for CpG genes: the peak downstream of a TSS is much higher than the one upstream. In addition, the RNA-Seq signal downstream of TSSs is generally much higher than the upstream counterpart for both CpG and non-CpG genes. Based on these directional features, we designed a principled approach to predict the direction of TSSs: for CpG genes, we computed the dot products of the average RNA-Seq profile with the RNA-Seq profile at a predicted TSS on both positive and negative strands (RNA-Seq profiles are computed in the same way as the H3K4me3 profile, see 'Materials and Methods' section). Then, the gene is judged to be located on the strand with a higher dot product. For those CpG genes without RNA-Seq signal flanking the TSSs, we compared the two dot products of H3K4me3 profile (one on each strand) with the average H3K4me3 shape and assign the gene to the strand with a higher dot product. Because the prediction accuracy of RNA-Seq signal is higher than the H3K4me3 signal, we preferentially used the RNA-Seq signal. For non-CpG genes, similarly, the RNA-Seq signal was used to decide on which strand the gene is transcribed. However, the second step prediction based on H3K4me3 signal was abandoned since its prediction accuracy for non-CpG genes is only slightly better than random,

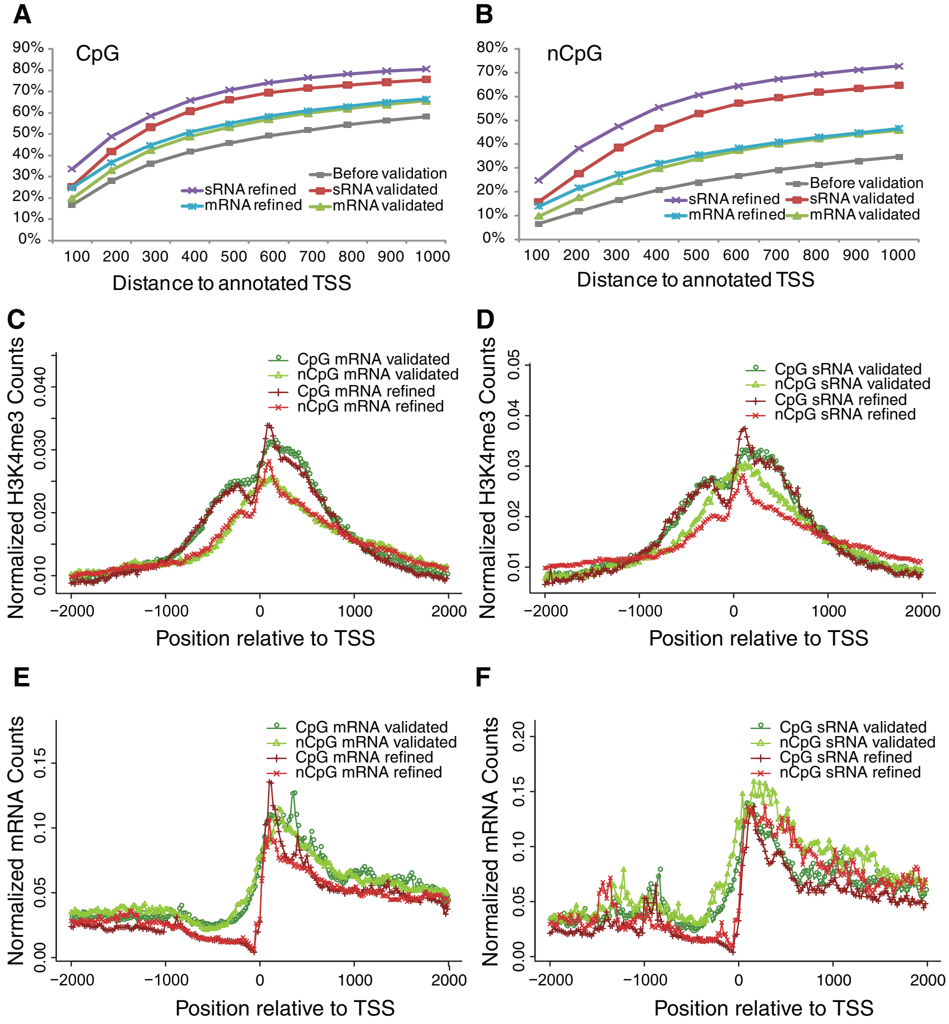

Figure 3. TSS prediction accuracy enhanced by RNA validation and refinement. (**A**) and (**B**) ROCs for CpG TSS and non-CpG TSS predictions before and after RNA validation and refinement. 'sRNA validated' and 'mRNA validated' refer to the TSS predictions validated by the sRNA-Seq and mRNA-Seq signals, respectively, while 'sRNA refined' and 'mRNA refined' refer to the final sets of TSS predictions which were further refined to the mRNA upward edges in the predicted direction of transcription. Each point in a ROC denotes the percentage of the set of TSS predictions that contain a homology-based electronic TSS annotation within the indicated distance. Only the TSS predictions with >0 GentleBoost scores are included. (**C–F**). The average profile of normalized H3K4me3 ChIP-Seq (C and D)/mRNA-Seq (E and F) tag counts for different sets of TSS predictions in the [−2 kb, +2 kb] region.

although the H3K4me3 peak is slightly biased towards downstream direction of a TSS. In this case, the directionalities of these TSSs are left unsolved.

Merging positive and negative strand TSS predictions

Since TSS prediction is performed by sliding the classifier on both positive and negative strands, it is important to merge nearly duplicate predictions. More precisely, if the distance between two predicted TSS is <400 bp, we only retain the TSS with a larger GentleBoost score to eliminate some redundant TSS predictions.

TSS refinement using the RNA-Seq data

As RNA-Seq could increase the prediction accuracy, we used it to further refine the position of the predicted TSSs. We first validated each predicted TSS by checking the existence of mRNA/sRNA-Seq tags in the [−500 bp, +500 bp] flanking region ('Materials and Methods' section). Then, the directions of each validated TSS were determined based on mRNA/sRNA-Seq signal (and the H3K4me3 signal for CpG genes), as described above. For these validated TSS, we further refined their positions by analyzing the detailed shape of the mRNA-Seq signal. The idea is fairly simple, we search for the nearest upward edge of the mRNA signal along the predicted direction of this TSS, also within the [−500 bp, +500 bp] region ('Materials and Methods' section). If an up-forward edge exists, we refine the predicted TSS to the position of the edge to signify the fact that transcripts typically start at true TSSs. Finally, we re-compute the direction of transcription at the refined TSS positions ('Materials and Methods' section).

We are able to verify 18 587 (35.42%) of homology-based TSS annotations within 500 bp after the validation procedure described above. When the refinement is included, the predictor can verify 17 933 (34.18%) of previously annotated TSSs at 500-bp resolution. The ROCs clearly demonstrate that after the sRNA/mRNA refinement, the precision of the predicted position of both the predicted CpG and non-CpG TSSs has much improved (Figure 3A and B). Meanwhile, the average profiles of normalized H3K4me3 tag counts surrounding the predicted TSSs display much sharper peaks after the refinement, and because of resolving the directionality of the TSSs, the average profile starts to display the characteristic asymmetry at TSS (Figure 3C and D). Similarly, the average profiles of normalized mRNA tag counts also display sharper peaks at TSS after the refinement (Figure 3E and F). We also note that due to RNA-Seq refinement, the overlap between different sets of TSS prediction increases (Figure 1), which suggests that the uncertainties in our raw predictions are greatly reduced. All of the above point to significantly enhanced prediction accuracy when directionality and RNA-Seq data are considered in the refinement step. Interestingly, the majority of the TSSs validated or refined by sRNAs are included by those validated or refined by mRNAs (Figure 1), suggesting that most sRNAs sequenced were derived from short transcripts at mRNA TSS, a notion in agreement with the recent findings by the ENCODE project (21).

Precise localization of macaque-specific TSSs

A large number of primate genes encode multiple transcripts via alternative TSSs or exon–intron splicing. Previous computational gene annotation approaches purely based on sequence alignment are unable to identify macaque-specific alternative TSSs, while our *ab initio* prediction approach is better for pinpointing such species-specific transcriptional events that are important for comparative genomic studies. An example is bromodomain PHD finger transcription factor (BPTF), the largest subunit of nucleosome-remodeling factor complex NURF (Figure 4A). BPTF is an 'H3K4me3 reader' and can recognize and tightly bind H3K4me3 (22). This gene was also found to be essential in early embryo and embryonic stem cells in mouse (23). By our approach, a TSS was predicted at the start of BPTF's third exon but not in the first exon as given by the electronic annotation. The large number of mRNA-Seq reads further supports our prediction. As the BPTF alternative splicing event in Drosophila has been reported to generate distinct NURF chromatin remodeling complexes (24), the distinctive transcriptional events in the macaque brain might imply existence of macaque-specific complex containing the BPTF protein. Other examples of alternative TSSs in the serine/threonine–protein kinase PLK2 and PAR domain protein 1PDP1 gene identified by our *ab initio* approach are shown in Supplementary Figure S5A and B.

Besides revealing macaque-specific transcriptional events, our approach also locates a large number of TSSs more precisely than previous homology-based gene annotations. For example, we found that the TSS of SLC24A4, an ergothioneine transporter, is located hundreds of base pairs upstream of the previous annotation (Figure 4B), which is firmly supported by the mRNA-Seq signal. Similarly improved TSS positioning can also be found for Sfrs2 and C3orf78 genes (Supplementary Figure S5C and D).

Finally, our *ab initio* predictions at previously annotated intergenic or intronic regions also suggest many novel transcripts (Figure 4C, Supplementary Figure S5E and F). Judging by the ROCs, as 80% of our best TSS predictions coincide with electronic gene annotations, the other 20% may contain novel TSSs at similar prediction accuracy (80%, Figure 2C and D). This extrapolates to approximately 1600 novel TSSs in the top 10 000 predictions, and approximately 10 000 novel TSSs if we look at all our 37 371 RNA-Seq refined predictions (with log odds score >0, at >50% accuracy) (Figure 3A and B). All the TSSs predicted by the H3K4me3 profile and sequence features, and further RNA-seq validation and refinement (Figure 1) can be found in Supplementary Tables S1 and S2.

Web interface for the Rhesus macaque TSS

To facilitate the usage of the TSSs identified in this study, we compiled the TSS coordinate information and the ChIP-Seq and RNA-Seq tag density to .bed files, which can be downloaded in batch to visualize in customized genome browsers. We also provided a web interface for users with no programming experience to query for user-defined genomic coordinates and genes of interest, then visualize the TSSs together with electronic homology based TSS annotations, as well as all the sequence tags of H3K4me3 ChIP-Seq and mRNA/sRNA-Seq. The web interface for these data is available at http://hanlab.genetics.ac.cn/Rhesus-TSS.

DISCUSSION

In this study, to discover macaque TSS, we generated H3K4me3 ChIP-Seq and mRNA and sRNA RNA-Seq data genome-wide for the Rhesus macaque frontal cortex, and based on these data we improved a computational method previously designed for TSS re-annotation (12) and extended it for *ab initio* TSS prediction.

Compared to the CoreBoost_MH algorthm (12), we made four major improvements/extensions, which significantly enhanced the performance of the classifier. (i) We use different ways to generate negative examples for training two TSS classifiers: for CpG promoters, negative training samples are cropped randomly from the genomic background; while for non-CpG promoters, we simply use random examples from the flanking region of known TSSs as negative training data. In this way, the most distinctive features for predicting the two classes of promoters are fully extracted by the two classifiers. (ii) We use the cosine similarity to quantify the similarity of the current H3K4me3 profile to the average H3K4me3 profile rather than the PCC used in the CoreBoost_MH algorithm (12). This turns out to be more robust to the fluctuations of ChIP-Seq reads. (iii) Leveraging on the asymmetry of the H3K4me3 and the RNA-Seq signal, we are able to

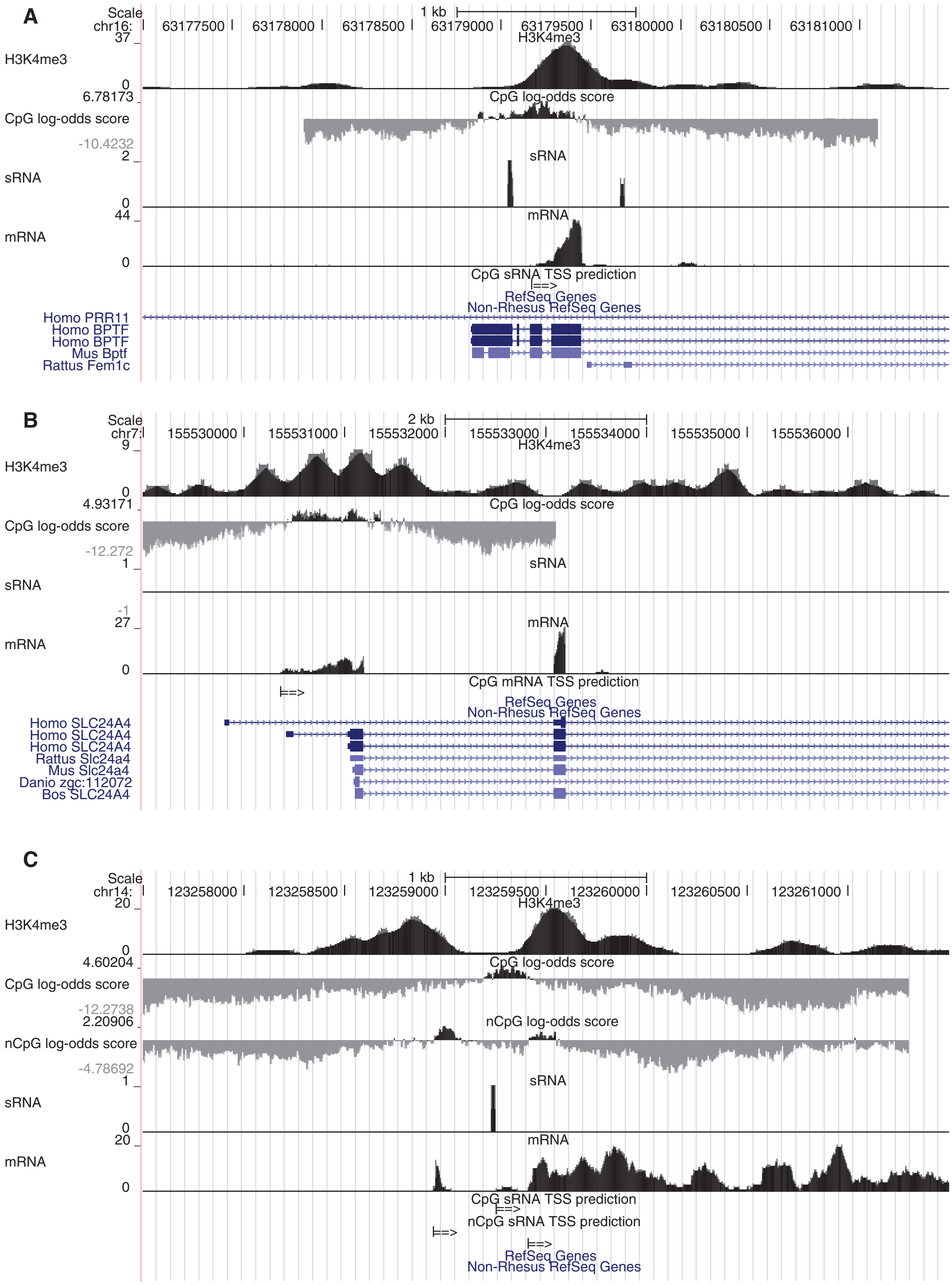

Figure 4. Exemplary TSSs identified by our approach. (**A**) A predicted macaque-specific alternative TSS. ‘H3K4me3’, ‘sRNA’ and ‘mRNA’ denote the H3K4me3 ChIP-Seq reads, sRNA-Seq and mRNA-Seq reads distribution around the predicted TSS. ‘CpG log-odds score’ visualizes the log-odds of the probabilities that each genomic position is a real TSS at 10-bp resolution. ‘CpG_sRNA TSS prediction’ denotes the predicted CpG TSS and its direction of transcription as indicated by the arrows. ‘Non-Rhesus RefSeq Genes’ denote the electronically annotated Rhesus macaque genes derived by homologies from other species. The mRNA signals indicate that our prediction is more accurate than previous annotations. (**B**) A predicted TSS that shows better positioning than homology-based electronic TSS annotations. (**C**) A novel TSS predicted at a previously annotated intergenic region.

further predict the directionality of transcription for each CpG/non-CpG TSS, which has not been addressed by previous TSS prediction approaches. (iv) Unlike the previous algorithm, which is based only on genomic and epigenomic features, we now also incorporated the transcriptomic features from the RNA-Seq data for TSS prediction. As demonstrated in the results, the RNA-Seq data can not only serve to validate our TSS classification scores, but more importantly be able to pinpoint the location of the TSS to single base pair resolution, and help determine the transcript directionality. Surprisingly, the majority of sRNAs is perhaps associated with TSSs and therefore can be used to identify both small RNA and mRNA TSSs at higher accuracy than mRNA sequence tags.

As a consequence of the technical improvements to the TSS prediction algorithm, the TSS identification accuracy is greatly increased. This enables us to *ab initio* predict, based on our H3K4me3 and RNA-Seq profiles, approximately 10 000 new TSSs and verify 17 933 (34.18%) of previously electronically annotated TSSs at 500-bp resolution, 52.96% of which are precisely located within 100 bp. If the number of TSSs in the Rhesus macaque genome is similar to that of the electronic gene annotation, among our predicted TSSs, the 17 933 validated TSSs together with the estimated approximately 10 000 new TSSs have covered approximately 27 933 (53.23%) of the whole set of TSSs.

Although we only detected H3K4me3 modification and RNA tags in macaque brain, the H3K4me3 modification has been shown to be able to identify both expressed and non-expressed CpG genes (8,12,13). Therefore the TSSs we identified for CpG promoters might also include genes that are expressed in tissues other than the brain.

The TSSs identified in this study provide a map and a rich resource for close examination of the species-specific transcript structures and transcription regulations in the Rhesus macaque genome, as well as a starting point for comparing them to other primate species, including us humans. Our approach constitutes a relatively inexpensive way to generate a reasonably reliable TSS map for a large genome and may serve as a guiding example for similar genome annotation efforts targeted at other model organisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Christopher Green for critical reading of the article.

FUNDING

China National Science Foundation (grant nos 30890033, 30588001 and 30620120433); Chinese Ministry of Science and Technology (grant no 2006CB910700); Chinese Academy of Sciences, Bai Ren Outstanding Achievement Award (to J.D.J.H.). Funding for open access charge: Chinese Academy of Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Gibbs,R.A., Rogers,J., Katze,M.G., Bumgarner,R., Weinstock,G.M., Mardis,E.R., Remington,K.A., Strausberg,R.L., Venter,J.C., Wilson,R.K. *et al.* (2007) Evolutionary, biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.
2. Guigo,R., Dermitzakis,E.T., Agarwal,P., Ponting,C.P., Parra,G., Reymond,A., Abril,J.F., Keibler,E., Lyle,R., Ucla,C. *et al.* (2003) Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl Acad. Sci. USA*, **100**, 1140–1145.
3. Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigo,R. (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.
4. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
5. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
6. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
7. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
8. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
9. 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.
10. Khalil,A.M., Guttman,M., Huarte,M., Garber,M., Raj,A., Rivea Morales,D., Thomas,K., Presser,A., Bernstein,B.E., van Oudenaarden,A. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 11667–11672.
11. Yu,H., Zhu,S., Zhou,B., Xue,H. and Han,J.D. (2008) Inferring causal relationships among different histone modifications and gene expression. *Genome Res.*, **18**, 1314–1324.
12. Wang,X., Xuan,Z., Zhao,X., Li,Y. and Zhang,M.Q. (2009) High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res.*, **19**, 266–275.
13. Karlic,R., Chung,H.R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.
14. Fei,T., Xia,K., Li,Z., Zhou,B., Zhu,S., Chen,H., Zhang,J., Chen,Z., Xiao,H., Han,J.D. *et al.* (2010) Genome-wide mapping of SMAD target genes reveals the role of BMP signaling in embryonic stem cell fate determination. *Genome Res.*, **20**, 36–44.
15. Li,R., Yu,C., Li,Y., Lam,T.W., Yiu,S.M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
16. Zang,C., Schones,D.E., Zeng,C., Cui,K., Zhao,K. and Peng,W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.
17. Friedman,J., Hastie,T. and Tibshirani,R. (2000) Additive logistic regression: a statistical view of boosting. *Annals Stat.*, **28**, 337–407.

18. 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.
19. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
20. Zhao,X., Xuan,Z. and Zhang,M.Q. (2007) Boosting with stumps for predicting transcription start sites. *Genome Biol.*, **8**, R17.
21. Taft,R.J., Kaplan,C.D., Simons,C. and Mattick,J.S. (2009) Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle*, **8**, 2332–2338.
22. Landry,J., Sharov,A.A., Piao,Y., Sharova,L.V., Xiao,H., Southon,E., Matta,J., Tessarollo,L., Zhang,Y.E., Ko,M.S. *et al.* (2008) Essential role of chromatin remodeling protein Bptf in early mouse embryos and embryonic stem cells. *PLoS Genet.*, **4**, e1000241.
23. Li,H., Ilin,S., Wang,W., Duncan,E.M., Wysocka,J., Allis,C.D. and Patel,D.J. (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, **442**, 91–95.
24. Kwon,S.Y., Xiao,H., Wu,C. and Badenhorst,P. (2009) Alternative splicing of NURF301 generates distinct NURF chromatin remodeling complexes with altered modified histone binding specificities. *PLoS Genet.*, **5**, e1000574.