- Supplementary Material

SUPPLEMENTARY METHODS

Source and Treatment of Annotations

We wanted to take a conservative approach when identifying transcribed regions to avoid erroneously labeling transcribed regions as ’novel’. Hence, we used the union of the RefSeq (71), UCSC (72) and Ensembl (73) annotations for both human and mouse. Our representation is such that each isoform is represented separately in the annotation and hence, many of the entries in our annotation will be mutually overlapping. To improve the coverage of the less well annotated ncRNAs, we decided to include two collections of long ncRNAs from mouse that have previously been reported in the literature: the macroRNAs (74) and the lincRNAs (75).

In addition to protein-coding genes and ncRNAs, we also considered distal enhancers that had previously been defined experimentally. For mouse neurons, we used the list from Kim *et al.*(76) and for HeLa cells, we used the list from Heintzman *et al.*(77). Since the latter list contained many loci near annotated TSSs and close to H3K4me3 peaks, we first filtered the list to remove any peaks that were near either of those two elements.

In the human and mouse genomes there are a few thousand very short ($<$100 bp) ncRNAs (sRNA, snRNA, rRNA, tRNA, snoRNAs) and some of them are expressed at extremely high levels. A complete list of these was obtained from the repeat masker (78). Many of these very short RNAs overlap much longer annotated genes and hence they will cause a spike in the read density profile which complicates the identification of transcribed regions. We address this issue by assigning reads to these regions prior to any further analyzes and then removing the reads overlapping these regions before running HaTriC. A summary of the statistics for this category can be found on the third line in Tab. 1 in the main text and Tab. S8.

There are still a handful of short regions with very high read densities not just for the RNA-Seq data, but also for all of the ChIP-Seq data sets, including the negative control ChIP-Seq input. Closer inspection of the DNA sequences at these loci reveals that they are identical to mitochondrial DNA or that they code for enzymes that are exported to the mitochondria. We assume that these regions are experimental artifacts and the reads are removed altogether from the analysis.

HaTriC algorithm for *de novo* transcript calling

We have developed an iterative algorithm that can detect transcribed regions, *i.e.* regions with high RNA-seq read densities, without relying on the annotation. For each iteration, the algorithm first uses a multi-scale wavelet-based approach for detecting break points; sites with sharp changes in RNA-seq read-densities. A wavelet (79) can be thought of as a brief oscillation, and they are frequently used in signal processing to extract features with certain characteristics. The Haar-wavelet corresponds to a square-wave and as such it is ideal for picking up abrupt changes in read densities.

The break points are used to partition the genome into segments of low or high read-densities. Empirically, we have found that the distribution of the segment read-density is bimodal, making it straightforward to separate the segments into two classes, the ones with high or low read density (Fig. S1). The segments with low read density are considered background or noise and hence they are ignored. The remaining high read density segments are retained, and if two of them are directly adjacent, they are merged. Following each iteration, the reads overlapping transcribed regions are removed, and in the next iteration, regions with lower density will be deemed significant. The algorithm terminates when no new transcribed regions are detected.

The algorithm takes as input a set of mapped RNA-Seq reads from one strand of one chromosome. This speeds up the computations significantly as each strand and chromosome can be analyzed independently. Additionally, HaTriC has four parameters that need to be specified by the user, a length scale over which the data is coarse grained to reduce the noise (L_{bin}), a minimum (L_{min}) and a maximum (L_{max}) length scale for the wavelets and a false detection rate (FDR) type cut-off for the number of breakpoints and what densities to include (P_{FDR}). The steps are outlined below:

1. Calculate a histogram for each bin i with $c_i = \log(1 + \sum_{j=iL_{\mathrm{bin}}}^{(i+1)L_{\mathrm{bin}}} r_j)$, where r_j is the number of reads overlapping bin j. In regions with high expression levels, such as exons, the variability can be very high. Empirically, we have found that the performance of the algorithm is improved significantly if the reads are averaged over non-overlapping bins of length $2^{L_{\mathrm{bin}}}$.

2. Calculate the Haar-wavelet coefficients (79), h_{ij}, for each bin i and length scale $j = L_{\mathrm{min}} \dots L_{\mathrm{max}}$

$$h_{ij} = \frac{1}{\sqrt{2^{j+1}}} \left(\sum_{k=i}^{k+2^j-1} \log(1+r_k) - \sum_{i-1}^{k=i-2^j} \log(1+r_k) \right)$$

3. Next, we collect a set of break points, $\mathbb{B}$, corresponding to locations with abrupt changes of the read density. For each scale, j, we first find all extrema of h_{ij} (i.e. local minima and local maxima) and starting from the largest absolute value, the P_{FDR} fraction of the extrema with the highest $|h_{ij}|$ are selected. Since the same break point can be picked up by different length scales, a break point will only be included if there are no other points on the list within 2^{j-1} bps.

4. At this point we may opt to add additional break points that have been obtained through other means, such as RNAPII peaks or transcription start sites from the annotation. We have found that the performance of the algorithm can be significantly improved by incorporating RNAPII ChIP-Seq data. For each RNAPII peak we add two break points, one halfway between the center of the peak and its 5´ edge and one halfway between the center of the peak and its 3´ edge (as defined by the forward strand). The additional break points are added regardless of the presence of nearby break points. The extra information provided by the RNAPII peaks is particularly useful in gene-rich regions.

5. The average log density is calculated for each region between two adjacent break points. Empirically, we have found that the density has a bimodal profile which can be accurately represented by a Gaussian mixture model (GMM) with three components. Without loss of generality, we assign the component with the lowest mean index 0, the one with the second lowest mean index 1 and the one with the highest mean index 2. An expectation-maximization (EM) algorithm (80) was used to fit the parameters of the GMM. The probability p_i that any region belongs to component i of the GMM can be found as

$$p_i = \frac{w_i \phi(\mu_i,\sigma_i)}{\sum_{j=1}^{3} w_j \phi(\mu_j,\sigma_j)}, \quad (1)$$

   where w_i are the weights and $\phi(\mu,\sigma)$ is the Gaussian probability density for a function with mean μ and standard deviation σ. Segments that have $p_0 < P_{\mathrm{FDR}}$, that is they do not belong to the first component of lowly expressed segments, are considered transcribed regions. Regions without any reads are excluded when fitting the GMM and they are automatically set as being not transcribed.

6. Segments that are adjacent are merged (including segments identified during previous iterations) and segments that are shorter than L_{bin} bps are removed. Short segments may be created due to the addition of break-points based on the RNAPII peaks in step 4.

7. If no new segments were found, the algorithm is terminated. Otherwise, reads that were included in transcribed regions are removed and the algorithm proceeds with the next iteration from 1.

Finally, we can improve the resolution of the start and end for each transcribed region. This is achieved by calculating the Haar-wavelet coefficients with scale $2^{L_{\mathrm{bin}}}$ for each basepair in the region $[s - 2^{L_{\min}}, s + 2^{L_{\min}}]$ bps, where s is the start of the transcribed region. For the start, we select the first location where the coefficient is greater than a given threshold which has been chosen in such a way to require $L_{\min}/35$ reads in a region of length $2^{L_{\mathrm{bin}}}$ bps. This allows us to pick the first site where the read density increases significantly. We have found this procedure to be more robust as the simple approach of picking the first local maxima for the coefficients in the window is overly sensitive to noise. For the ends of transcribed regions, we use a similar strategy, but we instead search for a location where the coefficient has a negative value exceeding the threshold and the search begins from the 3´-end.

Classifying transcribed regions identified by HaTriC To determine if a transcribed region t_j uniquely corresponds to an annotated gene, g_i, we first find all k genes on the same strand that overlap t_j. For each pair, (t_j,g_i), we calculate the degree of overlap based on the number of reads

$$r_{ij} = \frac{N_i}{N_j}, \quad i = 1 \ldots k \quad (2)$$

where N_i is the number of reads in t_j that also overlap g_i and N_j is the total number of reads in t_j. We also compute the degree of overlap for the number of base pairs

$$b_{ij} = \frac{B_i}{B_j}, \quad i = 1 \ldots k \quad (3)$$

where B_i is the number of bps covered by t_j that also overlap g_i and B_j is the total number of bps covered by t_j.

Each transcribed region, t_j, was assigned to a category as described below.

Annotated, correct If t_j only overlaps one gene ($k = 1$) and if the gene is only overlapped by one transcribed region and if $r_{ij} > .8$ and $b_{ij} > .8$, then the pair (t_j,g_i) is considered uniquely matching. In many cases, however, there will be multiple isoforms of a single gene and there may be a set, $\mathbb{I}_o$, of isoforms for which $r_i > .8$ and $b_i > .8$. If all of the members of $\mathbb{I}_o$ overlap one another, then we assume that they correspond to different isoforms and we consider t_j a unique match to the gene with the largest r_{ij}.

If a gene is covered by more than one region, and one of the transcribed regions satisfies $r_{ij} > .8$ and $b_{ij} > .8$ and the others have $r_{ij} < .1$ and $b_{ij} < .1$ then the t_j with the largest r_{ij} is considered uniquely corresponding to the gene. Similarly, if t_j covers two or three non-overlapping genes and for one of them $r_{ij} > .8$ and $b_{ij} > .8$ while for the other $r_{ij} < .1$ and $b_{ij} < .1$, then it is again considered uniquely matching.

Annotated, incorrect There are two ways in which t_j may end up in this category; if a gene has been incorrectly split into multiple regions or if t_j spans multiple non-overlapping genes.

- If a t_j covers more than one gene and it does not fulfill the above criteria for being uniquely matching, then it is categorized as **Multiple annotated genes**.
- If a gene is covered by more than one transcribed region and at least two of them have $r_i > .1$ and $b_i > .1$, then it is considered to be split into two or more transcripts, each of the transcripts are categorized as **Fragment of annotated gene**. Similarly, t_j will be assigned this category in the few cases when either of the coverages r_i or b_i drops below .8 while the other one is above the threshold.

Annotated ncRNA We use the same criteria as for **Annotated, correct** when classifying transcripts overlapping annotated non-coding RNAs. Transcribed regions that overlap ncRNAs in an ambiguous manner are categorized as **Annotated, incorrect** (as described above).

Unannotated, proximal Transcribed regions that start within 10 kb upstream of an annotated TSS on the anti-sense strand.

Unannotated, distal Transcribed regions that are found in extragenic regions or are anti-sense (AS) with respect to annotated genes further than 10 kb away from the nearest annotated TSS. These regions are further broken down into three sub-categories:

eRNAs Any t_j starting within 2 kb of any of an extragenic enhancer or on the anti-sense genic strand within 2 kb of an intragenic enhancer is categorized as an eRNA.

Other AS Remaining t_j that overlap an annotated gene on the anti-sense strand.

Novel Remaining t_j that do not fall into any other category.

A central part of our analyzes involves categorizing reads, transcribed regions, RFBSs and conserved islands. In all of these cases we want to avoid double-counting and hence each entity (read, RFBS, etc) is assigned to only one category.

Table S1 shows the number of transcribed regions found in each category. For Tab. 1 in the Main Text, we started from this result, but we used the annotation to correct the miscategorized transcripts and genes. We also used the more sensitive method described below to find additional transcribed regions.

Characterizing the transcriptome by combining the annotation, HaTriC and regions found near Regulatory Factor Binding Sites (RFBSs)

To characterize the transcriptome, we assigned each read uniquely to a transcribed region. To define the transcribed regions in the most accurate way possible, as reported in Tab. S7, Fig. S4 and Tab 1 in the main text, we combined the annotation, the HaTriC transcript-caller, and a targeted search close to enhancers and RFBSs. Below, we describe how each category of transcribed regions was defined, as well as the criteria for assigning reads to each category. We considered the categories sequentially, and at each step we identified (and removed from further analysis) all reads that overlapped regions in the current category.

1. Protein-coding gene Annotated protein-coding genes were first separated into non-overlapping clusters. From each cluster the longest region, g_i, was extracted as a representative of that cluster (this was done to avoid double counting, and the majority of clusters contain only one gene). If the average read density of g_i fell below a threshold, the region was ignored and the reads were retained and made available for inclusion in another category.

Next, we applied HaTriC and merged the identified regions that had been categorized as corresponding to a part of a gene (**Fragment of annotated gene**) or uniquely to a gene (**Annotated, correct**) with their overlapping genes. When two regions g_i and g_i' are merged, they are removed and a new region $\widetilde{g}_i$ is created. The new region contains the union of the reads from g_i and g_i', and it extends from the 5´-most end of g_i and g_i' to the 3´-most end of g_i and g_i'. The transcribed regions $\widetilde{g}_i$ were frequently longer than the annotation would have predicted.

2. Annotated non-coding gene Having removed all reads corresponding to annotated coding genes, we carried out the same procedure for annotated non-coding genes.

When counting the number of reads in the two categories relating to annotated genes (as reported in Tab. 1, but not for the transcript read density reported in Fig. S4, S7 and Fig. 2), we also assigned all sense reads found within 10 kb upstream or downstream of $\widetilde{g}_i$ to the (protein-coding or ncRNA) genic category. As reported by van Bakel *et al.*(81), these regions often have a read density that is above the background levels found in more distal regions.

3. Promoter AS Next, we searched for promoter AS transcribed regions, *i.e.*, divergent transcribed regions (82, 83). We started by searching all windows located 2 kb upstream of all annotated TSSs. If a window contained more than r_0 reads, it was considered significant. Most transcribed regions that were not detected by HaTriC are shorter than 2 kb (see Fig. S4), but to account for longer regions we extended the search to the next 2 kb window upstream of the TSS if the TSS proximal window contained more than r_0 reads. Additional windows were investigated until a window containing fewer than r_0 reads was found. For a set of adjacent 2 kb windows, the length of the transcribed region is defined as the maximum distance between all pairs of reads found in these windows. We refer to the procedure where subsequent 2 kb windows are scanned as a *window-based search*. The threshold was set to $r_0 = 9$ reads in a 2 kb window for the mouse neurons and $r_0 = 5$ reads for the HeLa cells, corresponding to an FDR of .001. The regions detected using the window-based search were merged with all regions identified by the transcript caller with the label **Unannotated, proximal**.

4. Novel (HaTriC-defined) transcript This category corresponds to long unannotated transcripts and hence we simply assign all regions categorized by HaTriC as **Unannotated, distal** to this class. Since there are occasionally low numbers of reads close to the starts and ends of the unannotated transcribed regions (similar to how promoter AS reads are found near annotated TSSs), we carried out a window based search upstream and downstream of the transcribed regions. Any reads found from the window-based search was included in the total read count reported in Tab. 1 in the Main Text and Tab S1.

5. Other (HaTriC-defined) AS transcript We first applied the window-based search to the AS strand downstream of all RFBSs overlapping annotated genes. The regions obtained using the window-based method are merged with the ones found by HaTriC and categorized as **Other AS**.

6. Extragenic enhancer RNA, Intragenic enhancer RNA For extragenic enhancers, we applied the window-based search in the downstream directions on both strands and for intragenic enhancers, only the anti-sense downstream window was considered. The regions obtained using the window-based method are merged with all **eRNA** regions identified by HaTriC.

7. Associated with other H3K4me3 peaks Since the H3K4me3 mark is strongly associated with active promoters, we wanted to make sure that we did not miss any significant transcription initiated from these loci. For all extragenic RFBSs that were within 2 kb of a H3K4me3 peak, we used the window-based method on both strands to extract a set of transcribed regions.

8. Other RFBSs-associated RNA For the remaining extragenic RFBSs that did not have a H3K4me3 peak nearby, we again used the window-based method on both strands to extract a set of transcribed regions.

9. Insulator-associated RNA Finally, for HeLa cells where we also have access to CTCF data, we applied the window-based method on both strands at CTCF peaks.

Characterizing the remaining reads Having removed all reads that were associated with any of the nine categories above, we calculated the number of reads in non-overlapping 2 kb bins covering the remainder of the genome. We fitted the distribution across the 2 kb bins to either a Poisson or a negative binomial distribution and as shown in Fig. S2, the latter provides a good fit for both mouse neurons and HeLa cells.

Identifying and classifying RFBSs

As part of our goal of understanding the relationship between transcription, binding and sequence conservation, we needed to categorize the regulatory factor binding sites (RFBSs). For mouse neurons, we used the same list of peaks as in our previous work where a description of the peak calling algorithm can be found (76). For the HeLa cells, we obtained the peaks by downloading .narrowPeak-files for DNaseI, H3K4me3 and the TFs listed in Tab. S2 from the ENCODE website[1]. Peaks for different regulatory factors were identified independently but when creating the final list of RFBSs for a particular cell-type, overlapping peaks were merged to avoid double-counting. Each peak from either collection was assigned to a category according to the scheme outlined below, and the results are reported in Tab. S3. We start with the full set of peaks, and once a peak has been assigned to a category, it cannot be assigned to any further categories.

1. If the peak was within 1 kb of an annotated TSS, it is considered *“Promoter of annotated protein-coding gene”* or “*Promoter of annotated ncRNA*” for annotated coding genes and ncRNAs, respectively.

2. If the peak was within 1 kb of an enhancer it is classified as either “*Intragenic enhancer”* or *“Extragenic enhancer”*.

3. If the peak was within 1 kb of the start of a **Novel** transcribed region it is assigned to the “*Promoter of novel ncRNA”* category.

4. If the peak overlaps an annotated protein-coding gene but is further than 1 kb from the start, it is assigned to the *“Overlaps exon of annotated protein-coding gene”* or *“Overlaps intron of annotated protein-coding gene”* depending on its overlap with exons. Similarly, peaks overlapping annotated ncRNAs are considered either *“Overlaps exon of annotated ncRNA* or *“Overlaps intron of annotated ncRNA”*.

5. If the peak does not fit into any of the previous categories it is classified as *”Unannotated extragenic”*.

Identifying and classifying conservation islands

The third part of our analysis concerns the highly conserved elements of the genome. Since our resolution for identifying both transcribed regions and RFBSs is limited, it is not meaningful consider conservation at the level of individual bases. Instead, we take a coarse-graining approach to identify regions that are highly conserved, starting from the PhastCons scores (as compared to 30 other vertebrates) (84). For both human and mouse, the global distribution of PhastCons scores is bimodal with few bases showing intermediate levels of conservation. To obtain a more coarse-grained binary representation, we processed the PhastCons scores as follows:

1. First, the PhastCons scores are coarse-grained by calculating the average for non-overlapping 10 bp bins across the genome.

2. Any bin with an average conservation greater than .9 is considered highly conserved. Highly conserved bins are labeled conserved islands.

3. To avoid fragmenting the genome too finely, bins that are within 100 bp are merged into conservation islands. This reduces the total number of islands by ∼50%, but the number of genomic bases covered increase by ∼20%.

Estimating the number of transcripts per cell

Following Mortazavi *et al* (85), we estimated the number of transcripts based on the read density and length for each transcribed region i. The normalized read density (reads per kilobase per million, RPKM) is defined as

$$r_i = \frac{10^9 c_i}{NL}, \qquad (4)$$

where r_i is RPKM, c_i is the number of reads that were found in the region, N is the total number of reads in the experiment and L is the total genome length. To calculate the number of transcripts that are present from a given region, we note that the probability of obtaining c_i out of N reads for a region of length l_i, assuming that the reads are uniformly distributed to the mappable regions of the genome is equal to the proportion of the total length of the genome multiplied by the number of transcripts

[1] http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/

$$P(c_i|N) = \frac{x_i l_i}{L}, \quad (5)$$

where x_i is the copy-number for the transcript and l_i is the length of the transcript. Assuming a uniform distribution of reads across transcripts, the fraction of reads should equal c_i/N. Combining with Eq. (**5**) and solving for x_i, we obtain

$$x_i = \frac{c_i L}{N l_i}. \quad (6)$$

The scale of x_i here is arbitrary and the relative amounts are reported in Tab. 1 in the main text. To fix the scale, we need to find the total number of transcripts in the cell. To obtain the numbers used in Fig. S4, we assumed that there are 240,000 polyadenylated mRNAs in the cell (86).

Estimating overlap of novel ncRNAs and RFBSs with conserved islands

Our goal was to estimate how much of the extragenic conservation could be explained by exons of unannotated ncRNAs. Calculating the number of conserved bases under unannotated exons is complicated by the fact that HaTriC does not provide us with the intron-exon structure of transcribed regions. We assumed that the novel transcribed regions are spliced (75), and thus it is unlikely that all of the conserved islands within an unannotated transcribed region are located at exons.

We took a statistical approach to estimate the number of conserved islands explained by exons of novel transcribed regions. We assume that the novel ncRNAs have the same conservation characteristics as the long ncRNAs (longer than 1 kb) found in the annotation. From the annotation, we estimate the average number of exons per transcribed kb, e, the average number of conserved islands per exon, c, and the average number of conserved islands per promoter, p, defined as the 1 kb region upstream of the TSS (Tab. S7). Given a novel ncRNA of length l_i kb, the expected number of conserved islands covered is $p + ecl_i$. From Tab. S7, we find that the number of conserved islands/kb of novel ncRNAs is 0.43 and 0.51 for Hs and Mm, respectively.

As a the worst case scenario, we consider all conserved islands that overlap with the annotated ncRNAs longer than 1 kb. The number of conserved islands is proportional to the length of the ncRNA and the slope is $\sim$0.7 conserved islands/kb for both Hs and Mm.

In contrast, estimating the overlap of RFBSs and conserved islands is straightforward. Since a typical RFBS and a typical conserved island are of comparable size, we label any RFBS overlapping a conserved island as highly conserved.

Identifying MARs and estimating overlap with conserved islands

Identifying MARs genome-wide using computational approaches is challenging and the predictive power of all methods available to date remains poor as reported in a comparative study by Evans *et al* (87). To detect MARs, we implemented the so-called H-rule, which designates any stretch of 20 or more bases with only A, C or T as a MAR. We used the H-rule to identify MARs in both the mouse and human genomes. Figure S10 shows the overlap between MARs and conserved islands for different stringencies and it is clear that no more than 4% of the unannotated conserved islands can be explained by MARs. To determine the significance of the overlap, we carried out a permutation study whereby all the MARs were distributed randomly throughout the genome. As shown in Fig. S10, this reveals that the enrichment is between 30% and 50% greater than expected by chance.

Estimating the overlap of RFBSs from multiple cell types

For the DHSs, H3K4me3 and CTCF peaks, we downloaded data for 11 cell lines, Hepg2, GM12878, Jurkat, K562, Mcf7, Nb4V2, Nhek, Panc1, Saec, SkmcV2, and Sknshra, from ENCODE (88), and we categorized the RFBSs as described in Section . To assess the number of new RFBSs that are discovered when a new cell-type is investigated, we first selected one cell-type at random. Next, additional cell types were selected at random and the number of new RFBSs detected at each iteration was computed. As this procedure depends on the order in which the cell types are selected, the procedure was repeated 1,000 times and the average number of discovered elements is reported in Fig. 3 in the Main text.

P-value for the overlap between conservation islands and RFBSs

We used a hypergeometric test to assess the significance of the overlap between conserved islands and RFBSs. We assumed that there were L/100 loci in the genome that can overlap a RFBSs and/or a conserved island, where L is the total number of bps in the genome. Given the total number of conserved islands and RFBSs, it is possible to calculate the probability of observing a given overlap under the assumption that both conserved islands and RFBSs are randomly distributed throughout the genome. The number of RFBSs found at conserved islands is much higher than one would expect based on a hypergeometric distribution. For example, in a 3 Gb genome with 1 million conserved islands and 40,000 RFBSs one would only expect 1,300 of the RFBSs to be conserved whereas we observe that 4,000 of the extragenic HeLa ucRFBSs are conserved (Tab. S2a). In this case, a hypergeometric test with an overlap of 4,000 betweem subsets of sizes 40,000 and 1,000,000 out of a total of 30,000,000 elements yields a highly significant p-value that is smaller than 10^{-16}. Similar calculations for the estimated number of ncRNA exons and promoters suggests a highly significant p-value ($< 10^{-16}$).

Table S1. Categorization of transcribed regions detected using HaTriC for mouse neurons and HeLa cells. The table shows the number of HaTriC-defined transcribed regions that correspond to coding and non-coding genes as well as to unannotated regions. For protein-coding genes, it is indicated whether the transcribed regions span multiple genes or correspond to an incomplete fragment (portion) of a gene. Here the fraction of reads is calculated with respect to the number of reads left after having removed the ones found at highly expressed short ncRNAs, (S) is sense and (AS) is anti-sense.

Mouse neurons

Category	**# regions**	**Fraction of transcripts**	**Fraction of reads**	**Fraction of genome**
Matching a single protein-coding gene (S)	7492	0.612	0.544	0.094
Matching multiple annotated genes (S)	374	0.031	0.009	0.003
Matching portion of annotated gene (S)	2213	0.181	0.363	0.073
Matching a single annotated non-coding gene (S)	240	0.020	0.001	0.001
Starting within 10 kb of an annotated TSS (S or AS)	1508	0.123	0.002	0.004
Starting within annotated coding or non-coding gene (AS)	104	0.008	0.000	0.000
Starting within 2kb of an enhancer	52	0.004	0.000	0.000
Novel (HaTriC-defined) ncRNA	255	0.021	0.001	0.001

HeLa

Category	**# regions**	**Fraction of transcripts**	**Fraction of reads**	**Fraction of genome**
Matching a single protein-coding gene (S)	5987	0.692	0.587	0.060
Matching multiple annotated genes (S)	377	0.044	0.006	0.001
Matching portion of annotated gene (S)	1464	0.169	0.304	0.035
Matching a single annotated non-coding gene (S)	62	0.007	0.002	0.000
Starting within 10 kb of an annotated TSS (S or AS)	628	0.073	0.004	0.001
Starting within annotated coding or non-coding gene (AS)	26	0.003	0.000	0.000
Starting within 2kb of an enhancer	1	0.000	0.000	0.000
Novel (HaTriC-defined) ncRNA	103	0.012	0.002	0.000

Table S2. Categorization of DNaseI hypersensitive sites (DHSs) in HeLa cells and overlap between RFBSs and DNaseI hypersensitive sites. (*a*) Each DHS was assigned to one of the non-overlapping categories represented at each row (Methods). The two columns show the total number of DHSs in each category, including the fraction of the DHSs that overlap a conserved island. (*b*) Around 90% regulatory factor binding sites overlap with DHSs. The only significant exception is the insulator binding protein CTCF that appears to bind in many locations that are not DHSs.

a

Annotation/HaTric-based category	**HeLa**	
	Peaks	Conserved
Promoter of annotated protein-coding gene	29001	0.19
Promoter of annotated ncRNA	3030	0.16
Promoter of novel (HaTriC-defined) ncRNA	286	0.07
Extragenic enhancer	17607	0.16
Intragenic enhancer	12767	0.17
Overlaps exon of annotated protein-coding gene	1577	0.44
Overlaps exon of annotated ncRNA	213	0.14
Overlaps intron of annotated protein-coding gene	23983	0.12
Overlaps intron of annotated ncRNA	2510	0.11
Overlaps MAR	1445	0.15
Unannotated extragenic	32413	0.10

b

Factor	**HeLa Binding sites**	**Overlap of factor with HeLa DHSs**
Ap2alpha	13701	0.873
Ap2gamma	18244	0.828
Cfos	20701	0.938
Cmyc	19039	0.946
Max	24789	0.921
E2f4	3455	0.930
E2f6	5766	0.881
Ctcf	44809	0.565

Table S3. Categorization of (*a*) RFBSs, (*b*) CTCF peaks and (*c*) H3K4me3 peaks in mouse neurons and HeLa cells. See Section and caption for Tab. S2a for description of categories. The HeLa RFBSs include AP2α, AP2γ, MAX, cFOS, cMYC, E2F4, and E2F6 and the mouse RFBSs include CBP, CREB, NPAS4 and SRF.

a

Annotation/HaTric-based category	**Mouse neurons**		**HeLa**	
	Peaks	Conserved	Peaks	Conserved
Promoter of annotated protein-coding gene	15944	0.25	12025	0.24
Promoter of annotated ncRNA	2833	0.27	1134	0.17
Promoter of novel (HaTriC-defined) ncRNA	223	0.31	87	0.05
Extragenic enhancer	6960	0.35	7443	0.15
Intragenic enhancer	7758	0.34	5498	0.15
Overlaps exon of annotated protein-coding gene	1298	0.41	545	0.41
Overlaps exon of annotated ncRNA	252	0.27	63	0.11
Overlaps intron of annotated protein-coding gene	14848	0.23	7035	0.09
Overlaps intron of annotated ncRNA	1496	0.29	747	0.09
Overlaps MAR	1218	0.31	268	0.12
Unannotated extragenic	19398	0.23	9740	0.08

b

Annotation/HaTric-based category	**HeLa**	
	Peaks	Conserved
Promoter of annotated protein-coding gene	8137	0.19
Promoter of annotated ncRNA	953	0.14
Promoter of novel (HaTriC-defined) ncRNA	56	0.02
Extragenic enhancer	1771	0.14
Intragenic enhancer	1257	0.13
Overlaps exon of annotated protein-coding gene	1365	0.48
Overlaps exon of annotated ncRNA	112	0.16
Overlaps intron of annotated protein-coding gene	11958	0.10
Overlaps intron of annotated ncRNA	1223	0.11
Overlaps MAR	649	0.14
Unannotated extragenic	15131	0.11

c

Annotation/HaTric-based category	**Mouse neurons**		**HeLa**	
	Peaks	Conserved	Peaks	Conserved
Promoter of annotated protein-coding gene	11213	0.18	32974	0.16
Promoter of annotated ncRNA	1016	0.21	2825	0.12
Promoter of novel (HaTriC-defined) ncRNA	104	0.23	290	0.04
Extragenic enhancer	147	0.33	0	NaN
Intragenic enhancer	192	0.34	0	NaN
Overlaps exon of annotated protein-coding gene	232	0.39	488	0.53
Overlaps exon of annotated ncRNA	63	0.11	47	0.13
Overlaps intron of annotated protein-coding gene	1379	0.14	3948	0.08
Overlaps intron of annotated ncRNA	195	0.16	314	0.08
Overlaps MAR	69	0.14	192	0.07
Unannotated extragenic	1436	0.15	4827	0.07

Table S4. Categorization of conserved islands based on (*a*) mouse neurons and (*b*) HeLa cells. The first column shows the number of islands in each category, and the second shows the fraction of the genome covered by those islands.

a

Category	#Conserved islands	Percentage of genome
Promoter of annotated protein-coding gene	113645	0.54
Promoter of annotated non-coding gene	42796	0.32
Exon of annotated protein-coding gene	190296	1.02
Exon of annoated ncRNA	12153	0.06
Enhancer (Kim et al., 2010)	7051	0.05
Other (unannotated) RFBS	14450	0.10
MARs	44910	0.06
Intronic conserved island	297578	1.14
Extragenic conserved island	383289	1.87
Total	1106168	5.17

b

Category	#Conserved islands	Percentage of genome
Promoter of annotated protein-coding gene	143422	0.63
Promoter of annotated non-coding gene	20915	0.09
Exon of annotated protein-coding gene	179032	0.82
Exon of annoated ncRNA	8074	0.03
Enhancer (Heintzmann et al., 2009)	7481	0.04
Insulator (defined by presence of CTCF)	9621	0.02
Other (unannotated) RFBS	9324	0.04
MARs	49547	0.05
Intronic conserved island	318337	1.03
Extragenic conserved island	400340	1.45
Total	1136472	4.22

Table S5. Read counts and the number of novel transcripts detected using total RNA RNA-Seq data from human tissues. To detect novel ncRNAs, we applied HaTriC to RNA-Seq data from ten different human tissues. The RNA-Seq was performed as described in (76). We obtained the RNA from the Ambion FirstChoice Human total RNA Survey Panel. For the Survey panel, since we lacked H3K4me3 occupancy for these tissues, we used RefSeq, UCSC, and Ensembl gene annotations in lieu of H3K4me3 for HaTriC optimization. We used the data from chr 21 with the data from the brain sample to establish the a parameter set that was used for all other samples.

Tissue	**#Reads (M)**	**#novel ncRNAs**
Brain	12.17	110
Heart	11.82	111
Kidney	17.72	135
Liver	16.09	81
Lung	11.78	85
Ovary	6.92	159
Placenta	10.04	245
Spleen	27.64	91
Testis	15.43	211
Thymus	16.91	75

Table S6. Data sets used. A summary of the different data sets used in this study and where they were obtained.

Data set	**Type**	**Origin**
Mouse total RNA	RNA-Seq	Kim *et al* (76)
Mouse polyA$^+$ RNA	RNA-Seq	Kim *et al* (76)
HeLa total RNA	RNA-Seq	This study
Total RNA from 10 tissues	RNA-Seq	This study
Mouse RFBSs	ChIP-Seq	Kim *et al* (76)
Mouse histone modifications	ChIP-Seq	Kim *et al* (76)
DNaseI hypersensitive regions	ChIP-Seq	ENCODE (88)
Human histone modifications	ChIP-Seq	ENCODE (88)
Human RFBSs	ChIP-Seq	ENCODE (88)
Mouse enhancers	list	Kim *et al* (76)
HeLa enhancers	list	Heintzman *et al* (77)

Table S7. Conservation properties of ncRNAs. Empirically, it was found that the number of exons per kb has a fat-tailed distribution. To reduce the impact of outliers, the table reports the geometric mean rather than the arithmetic mean for the number of exons/kb. The last three lines correspond to the statistics from a collection of lincRNAs (89) which only contains data for human.

	Mm	**Hs**
Exons/kb, coding	.73	1.05
Conserved islands/exon, coding	1.53	1.41
Conserved islands/promoter, coding	1.30	1.04
Exons/kb, ncRNA	.62	.59
Conserved islands/exon, ncRNA	.82	.73
Conserved islands/promoter, ncRNA	1.07	1.08
Exons/kb, lincRNA	N/A	.34
Conserved islands/exon, lincRNA	N/A	.41
Conserved islands/promoter, lincRNA	N/A	.41

Table S8. Comprehensive accounting of RNA-Seq reads by genomic locus for HeLa cells. See Tab. 1 in main text for legend. Transcribed loci were required to have 5 RNA-Seq reads and a read density of at least 1 per kb.

Transcript category	Percentage of RNA-Seq reads	#Loci	Percentage of genome
Protein-coding gene	59.591	7423	9.77
Annotated non-coding gene	1.614	629	0.32
snRNAs, tRNAs, scRNAs, srpRNAs, rRNAs	37.649	3058	0.01
Promoter AS transcript	0.469	1783	0.26
Other (HaTric-defined) AS transcript	0.115	189	0.08
Novel (HaTric-defined) transcript	0.117	91	0.03
Extragenic enhancer RNA	0.161	306	0.03
Intragenic enhancer RNA	0.018	65	0.01
Other RFBSs-associated RNA	0.040	182	0.01
Insulators associated RNA	0.031	368	0.02
Associated with other H3K4me3 peaks	0.021	289	0.01
Total	99.8253	14383	10.5507

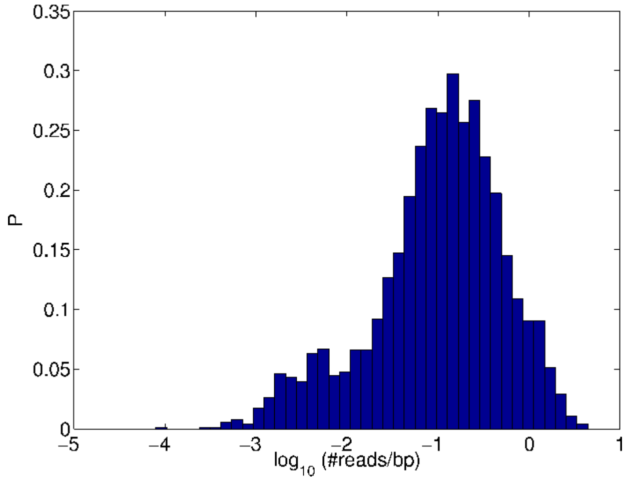

Figure S1. Distribution of average RNA-Seq read densities in candidate transcribed regions defined in the first iteration of HaTriC. In the first iteration of HaTriC, read densities across candidate transcribed regions from mouse chromosome 19 show a bimodal distribution: HaTriC calls the candidate regions in the high density mode (right) as transcribed. Note that the high-density mode is larger for two reasons. First, regions with zero reads are not shown, since the scale on the x-axis is logarithmic. Second, the lengths of the regions vary, with candidate regions in the low-density mode (left) typically being much longer than those in the high-density mode. In later iterations of HaTriC, the distribution shifts and becomes uni-modal (not shown), prompting the algorithm to terminate as no new transcribed regions are detected.

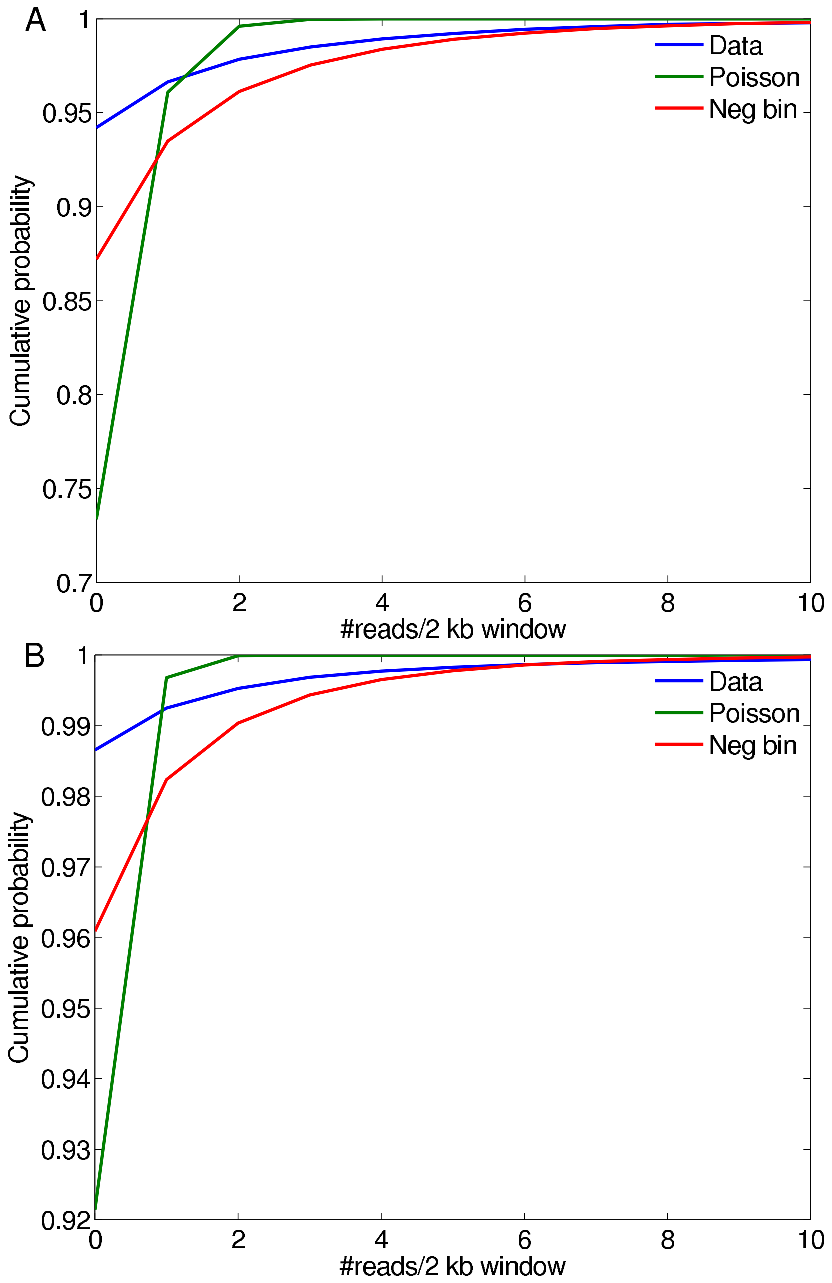

Figure S2. Distribution of the reads that were not accounted for by HaTriC, the annotation, enhancers or RFBSs. (*A*) Mouse neurons. (*B*) HeLa cells. The .2% of RNA-Seq reads that remained after constructing Tab. 1 in the main text (or Tab. S8 for HeLa) were placed into non-overlapping 2 kb bins. We then fitted the data to Poisson and negative binomial distributions and it is clear that the latter provides a better fit. Moreover, the vast majority of the bins with higher than expected read counts have neighboring bins that are empty (not shown), arguing against them being part of longer lowly expressed transcripts.

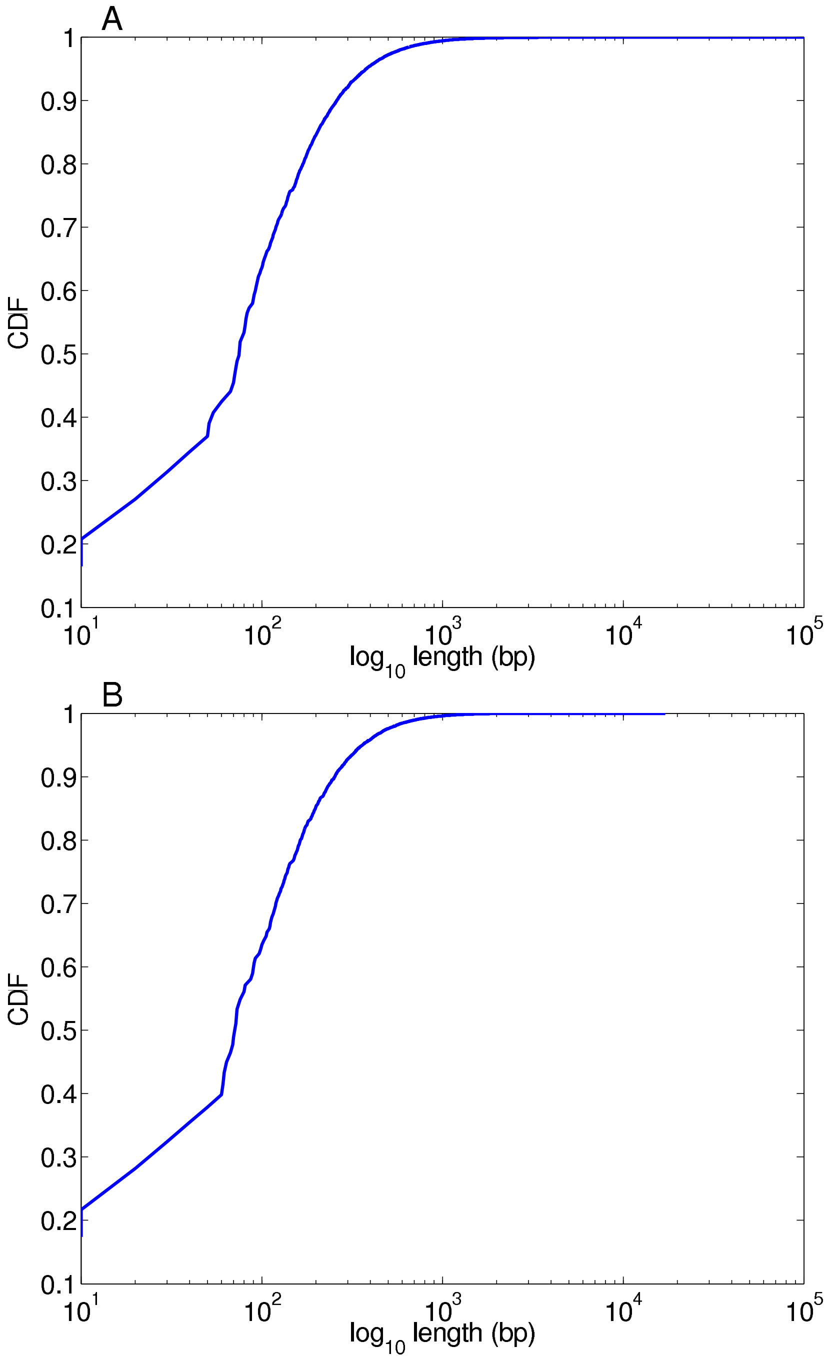

Figure S3. Cumulative distribution of the length of conserved islands found in mouse (a) and human (b). The total number of conserved islands can be found in Tab. S4.

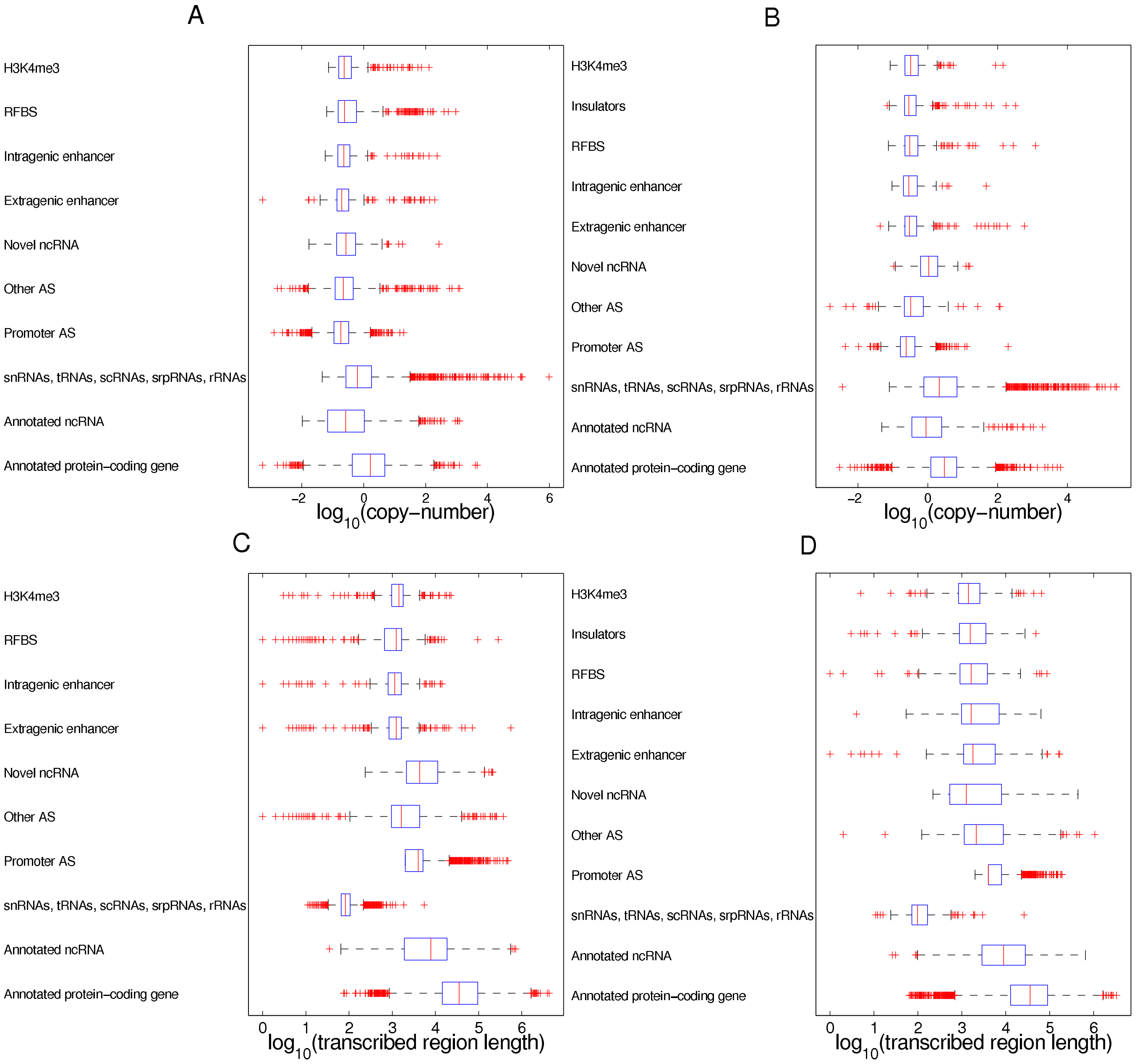

Figure S4. Estimated transcript copy number per cell and lengths of transcribed regions. Box-plot showing the distribution of the estimated copy number per cell for (*A*) mouse neurons and (*B*) HeLa cells. The number of transcripts for each category are log-normally distributed with a median close to or below one for most regions. The transcribed regions and categories are the same as in Tab. 1. The red line shows the median and the box spans the 25th to 75th percentiles of the data, with the whiskers covering 99.3% of the data and the '+' representing outliers. The total number of loci in each category can be found in Tab. 1 and S8, and it is clear that most non mRNA transcripts are present at levels of fewer than one transcript per cell. Box-plot showing the distribution of the lengths of transcribed regions for (*C*) mouse neurons and (*D*) HeLa cells. The lengths for each category are log-normally distributed. The transcribed regions and categories are the same as in Tab. 1.

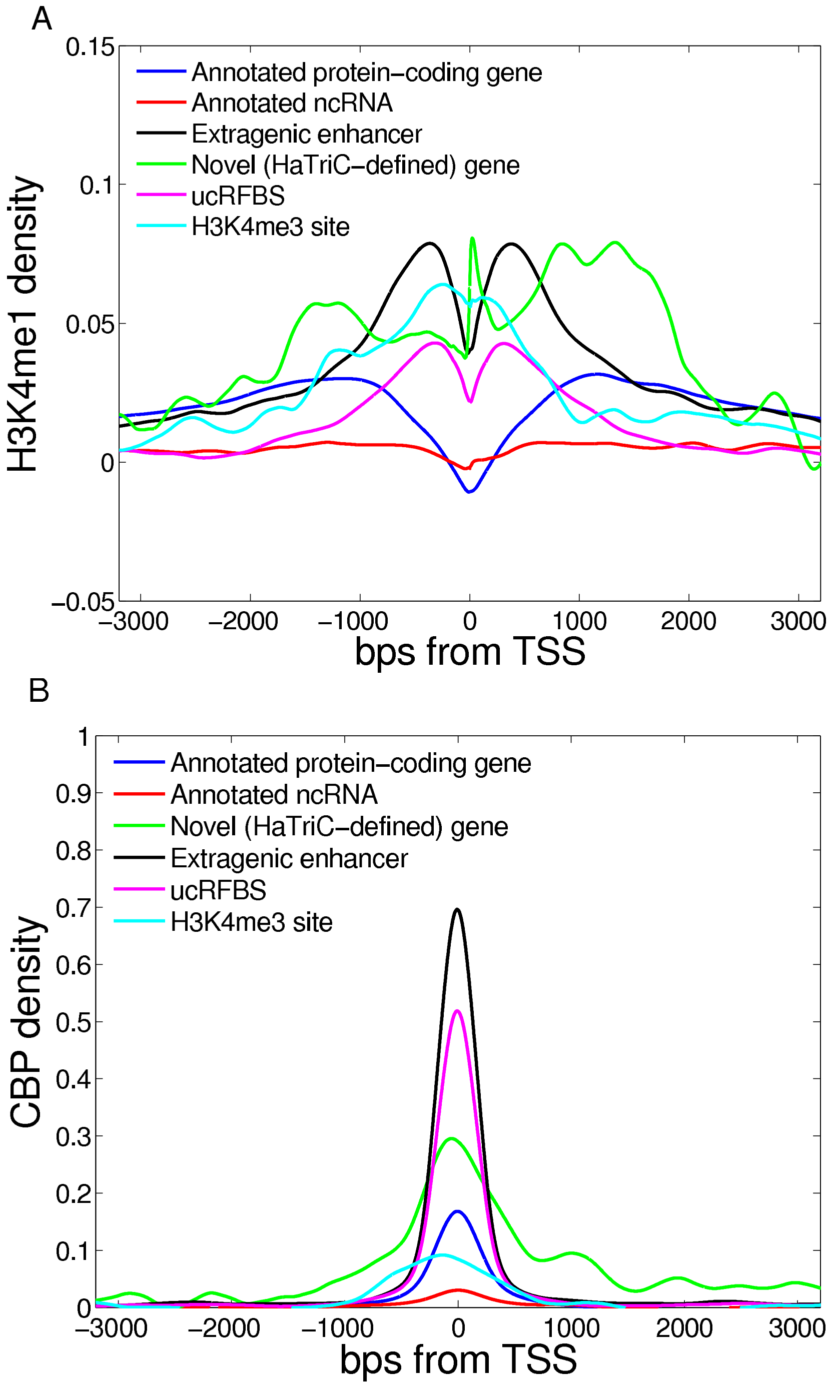

Figure S5. Profiles at conserved islands (*A*) H3K4me1 (*B*) CBP. These plots are similar to those in Fig. 2 in the main text and they show the density of H3K4me1 and CBP in the vicinity of conserved islands.

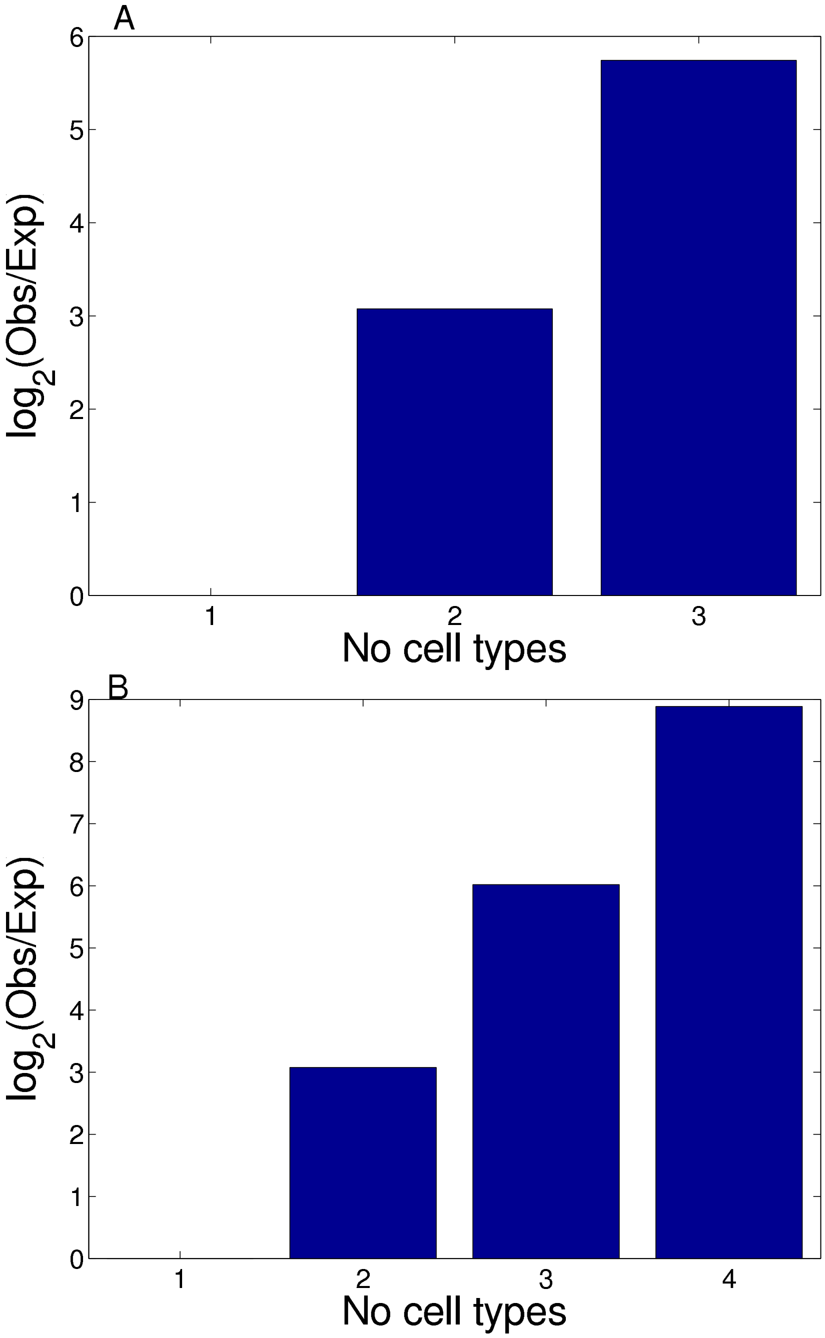

Figure S6. TF binding sites from different cell types are likely to overlap. (*A*) cFOS (*B*) MAX. The y-axis shows the fold-enrichment of the number of binding sites found in two or more cell types compared to what one would expect if the choice of binding sites in different cell-types was independent. The binding probability for each cell type was estimated as the total number of binding sites divided by the total number of DHSs.

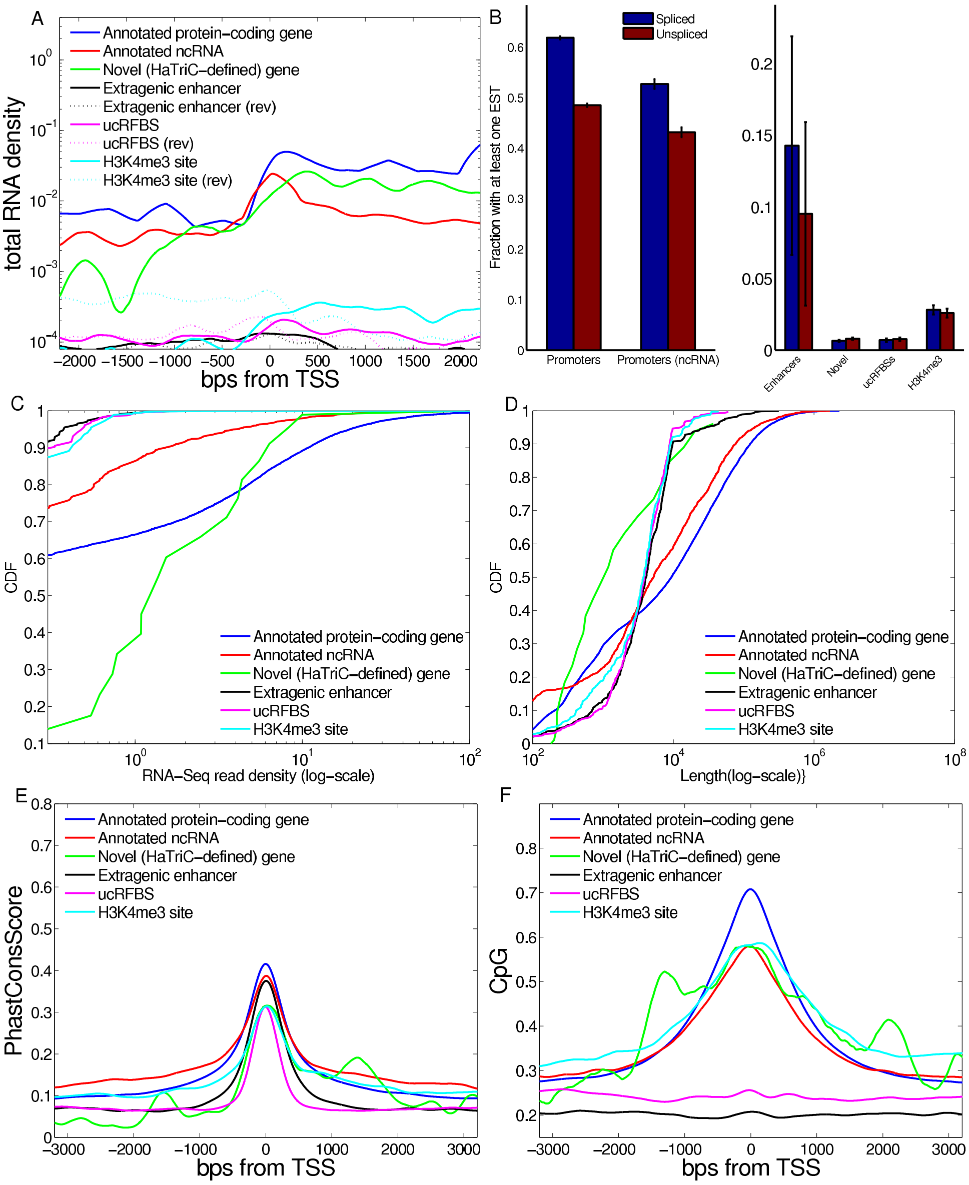

Figure S7. Properties of transcribed conserved islands in HeLa cells. This figure is similar to Fig.2 in the main text except that the H3K4Me3 profile and the polyadenylation ratio are missing. Thresholds for defining expressed loci were 5 RNA-Seq reads and a read density of at least 1 per kb.

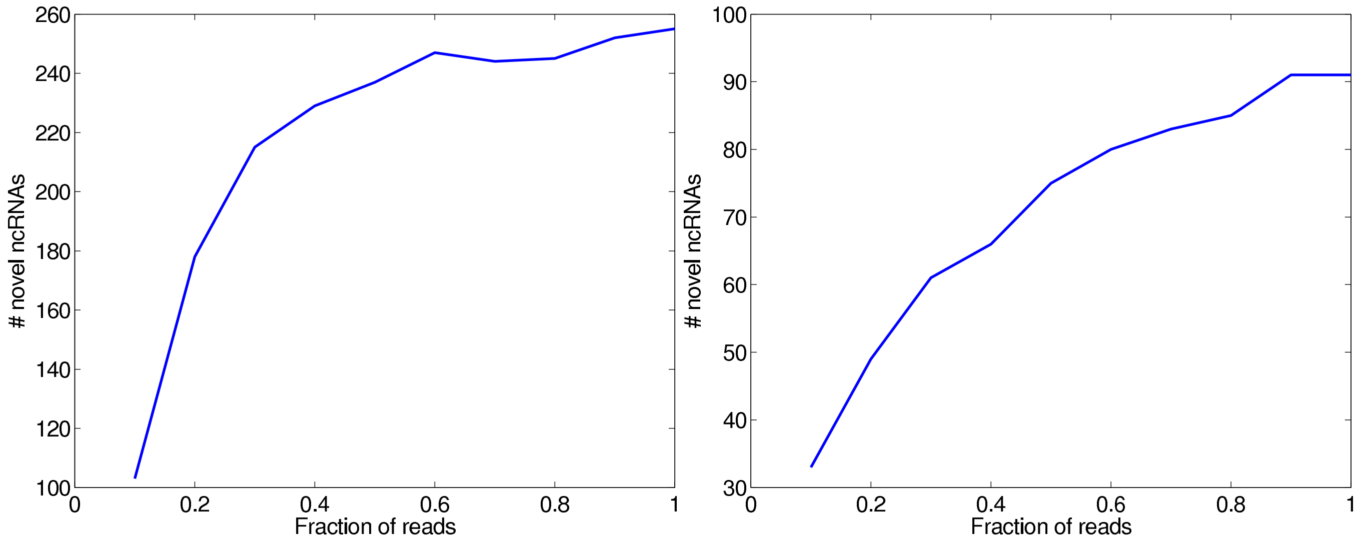

Figure S8. Few additional transcripts are discovered as a result of deeper sequencing We randomly down-sampled the RNA-Seq reads to between 10% and 90% of the original number (140 million for mouse neurons and 50 million for HeLa) and re-ran HaTriC using the same parameters as for the full set of reads for mouse neurons (left) and HeLa (right). The y-axis shows the number of novel ncRNAs that were discovered for each sub-sample fraction. Each sub-sample was repeated ten times and the results shown are the averages. For both human and mouse the slopes are relatively flat near the current sequencing depth, suggesting that the number of additional novel ncRNAs that will be found from additional sequencing is relatively low.

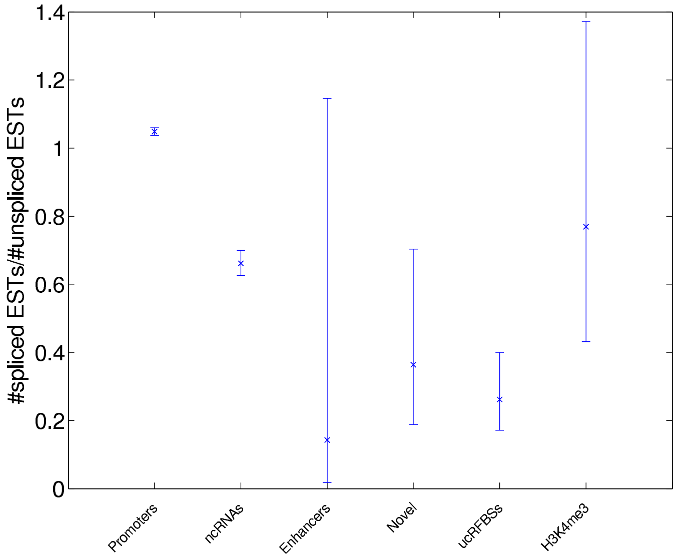

Figure S9. Ratio of spliced and unspliced ESTs. The plot shows the ratio between spliced and unspliced ESTs (the blue and the red bars in Fig. 2B in the main text). Only the mRNA category is significantly above 1. The error bars represent a 95% confidence interval and they were calculated using the binomial ratio test (90). In some cases the error bars are very large since we are taking the ratio between relatively small values.

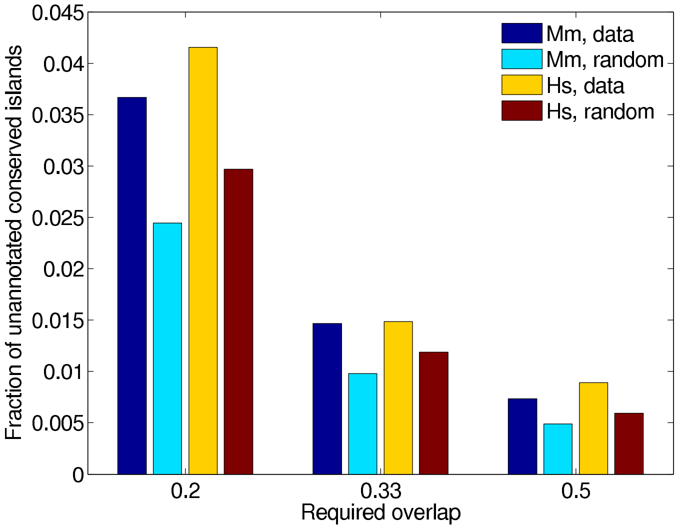

Figure S10. Fraction of conserved islands overlapping MARs. We used the H-rule (87) to identify MARs in the mouse and human genomes. The fraction of conserved islands that are explained by MARs depends on how large fraction of the conserved island that we require to overlap with the MAR and we present the result for three different stringencies (x-axis). The bars labeled data represent the overlap between conserved islands and MARs as observed in the mouse and human genomes. The bars labeled random are shuffle controls where the total number and the length of the MARs was the same as in the data, but their locations were randomized.

REFERENCES

71. Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue): D61–5, Jan 2007.
72. Brooke Rhead, Donna Karolchik, Robert M Kuhn, Angie S Hinrichs, Ann S Zweig, Pauline A Fujita, Mark Diekhans, Kayla E Smith, Kate R Rosenbloom, Brian J Raney, Andy Pohl, Michael Pheasant, Laurence R Meyer, Katrina Learned, Fan Hsu, Jennifer Hillman-Jackson, Rachel A Harte, Belinda Giardine, Timothy R Dreszer, Hiram Clawson, Galt P Barber, David Haussler, and W James Kent. The ucsc genome browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–9, Jan 2010.
73. Paul Flicek, Bronwen L Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, James Smith, and Stephen M J Searle. Ensembl’s 10th year. *Nucleic Acids Res*, 38(Database issue): D557–62, Jan 2010.
74. P Carninci, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, T Ravasi, B Lenhard, C Wells, R Kodzius, K Shimokawa, V B Bajic, S E Brenner, S Batalov, A R R Forrest, M Zavolan, M J Davis, L G Wilming, V Aidinis, J E Allen, A Ambesi-Impiombato, R Apweiler, R N Aturaliya, T L Bailey, M Bansal, L Baxter, K W Beisel, T Bersano, H Bono, A M Chalk, K P Chiu, V Choudhary, A Christoffels, D R Clutterbuck, M L Crowe, E Dalla, B P Dalrymple, B de Bono, G Della Gatta, D di Bernardo, T Down, P Engstrom, M Fagiolini, G Faulkner, C F Fletcher, T Fukushima, M Furuno, S Futaki, M Gariboldi, P Georgii-Hemming, T R Gingeras, T Gojobori, R E Green, S Gustincich, M Harbers, Y Hayashi, T K Hensch, N Hirokawa, D Hill, L Huminiecki, M Iacono, K Ikeo, A Iwama, T Ishikawa, M Jakt, A Kanapin, M Katoh, Y Kawasawa, J Kelso, H Kitamura, H Kitano, G Kollias, S P T Krishnan, A Kruger, S K Kummerfeld, I V Kurochkin, L F Lareau, D Lazarevic, L Lipovich, J Liu, S Liuni, S McWilliam, M Madan Babu, M Madera, L Marchionni, H Matsuda, S Matsuzawa, H Miki, F Mignone, S Miyake, K Morris, S Mottagui-Tabar, N Mulder, N Nakano, H Nakauchi, P Ng, R Nilsson, S Nishiguchi, S Nishikawa, F Nori, O Ohara, Y Okazaki, V Orlando, K C Pang, W J Pavan, G Pavesi, G Pesole, N Petrovsky, S Piazza, J Reed, J F Reid, B Z Ring, M Ringwald, B Rost, Y Ruan, S L Salzberg, A Sandelin, C Schneider, C Schönbach, K Sekiguchi, C A M Semple, S Seno, L Sessa, Y Sheng, Y Shibata, H Shimada, K Shimada, D Silva, B Sinclair, S Sperling, E Stupka, K Sugiura, R Sultana, Y Takenaka, K Taki, K Tammoja, S L Tan, S Tang, M S Taylor, J Tegner, S A Teichmann, H R Ueda, E van Nimwegen, R Verardo, C L Wei, K Yagi, H Yamanishi, E Zabarovsky, S Zhu, A Zimmer, W Hide, C Bult, S M Grimmond, R D Teasdale, E T Liu, V Brusic, J Quackenbush, C Wahlestedt, J S Mattick, D A Hume, C Kai, D Sasaki, Y Tomaru, S Fukuda, M Kanamori-Katayama, M Suzuki, J Aoki, T Arakawa, J Iida, K Imamura, M Itoh, T Kato, H Kawaji, N Kawagashira, T Kawashima, M Kojima, S Kondo, H Konno, K Nakano, N Ninomiya, T Nishio, M Okada, C Plessy, K Shibata, T Shiraki, S Suzuki, M Tagami, K Waki, A Watahiki, Y Okamura-Oho, H Suzuki, J Kawai, Y Hayashizaki, FANTOM Consortium, RIKEN Genome Exploration Research Group, and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science*, 309(5740): 1559–63, Sep 2005.
75. M Guttman, I Amit, M Garber, C French, M Lin, D Feldser, M Huarte, O Zuk, B Carey, J Cassady, M Cabili, R Jaenisch, T Mikkelsen, T Jacks, N Hacohen, B Bernstein, M Kellis, A Regev, J Rinn, and E Lander. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, Feb 2009.
76. Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F Worley, Gabriel Kreiman, and Michael E Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, Apr 2010.
77. Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, Keith A Ching, Jessica E Antosiewicz-Bourget, Hui Liu, Xinmin Zhang, Roland D Green, Victor V Lobanenkov, Ron Stewart, James A Thomson, Gregory E Crawford, Manolis Kellis, and Bing Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243): 108–12, May 2009.
78. AFA Smit, R Hubley, and P Green. Repeatmasker open-3.0. http://www.repeatmasker.org.
79. Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.
80. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximim likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodlogical)*, 39(1):1–38, 1977.
81. Harm van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Most ”dark matter” transcripts are associated with known genes. *PLoS Biol*, 8(5):e1000371, Jan 2010.
82. Amy C Seila, J Mauro Calabrese, Stuart S Levine, Gene W Yeo, Peter B Rahl, Ryan A Flynn, Richard A Young, and Phillip A Sharp. Divergent transcription from active promoters. *Science*, 322(5909):1849–51, Dec 2008.
83. Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–8, Dec 2008.
84. Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, Aug 2005.
85. A Mortazavi, B Williams, K McCue, L Schaeffer, and B Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, May 2008.
86. Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21:1160–1167, 2011.
87. Kenneth Evans, Sascha Ott, Annika Hansen, Georgy Koentges, and Lorenz Wernisch. A comparative study of S/MAR prediction tools. *BMC Bioinformatics*, 8(71), 2007.
88. ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K G Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan

Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ulaş Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W James Kent, Eric A Stone, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, Alexander E Urban, Peter Kraus, Sara Van Calcar, Nate Heintzman, Tae Hoon Kim, Kun Wang, Chunxu Qu, Gary Hon, Rosa Luna, Christopher K Glass, M Geoff Rosenfeld, Shelley Force Aldred, Sara J Cooper, Anason Halees, Jane M Lin, Hennady P Shulha, Xiaoling Zhang, Mousheng Xu, Jaafar N S Haidar, Yong Yu, Yijun Ruan, Vishwanath R Iyer, Roland D Green, Claes Wadelius, Peggy J Farnham, Bing Ren, Rachel A Harte, Angie S Hinrichs, Heather Trumbower, Hiram Clawson, Jennifer Hillman-Jackson, Ann S Zweig, Kayla Smith, Archana Thakkapallayil, Galt Barber, Robert M Kuhn, Donna Karolchik, Lluis Armengol, Christine P Bird, Paul I W de Bakker, Andrew D Kern, Nuria Lopez-Bigas, Joel D Martin, Barbara E Stranger, Abigail Woodroffe, Eugene Davydov, Antigone Dimas, Eduardo Eyras, Ingileif B Hallgrímsdóttir, Julian Huppert, Michael C Zody, Gonçalo R Abecasis, Xavier Estivill, Gerard G Bouffard, Xiaobin Guan, Nancy F Hansen, Jacquelyn R Idol, Valerie V B Maduro, Baishali Maskeri, Jennifer C McDowell, Morgan Park, Pamela J Thomas, Alice C Young, Robert W Blakesley, Donna M Muzny, Erica Sodergren, David A Wheeler, Kim C Worley, Huaiyang Jiang, George M Weinstock, Richard A Gibbs, Tin. Identification and analysis of functional elements in % of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007.

89. Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development*, 25, 2011.
90. P.A.R. Koopman. Confidence intervals for the ratio of two binomial proportions. *Biometrics*, 40:513–517, 1984.