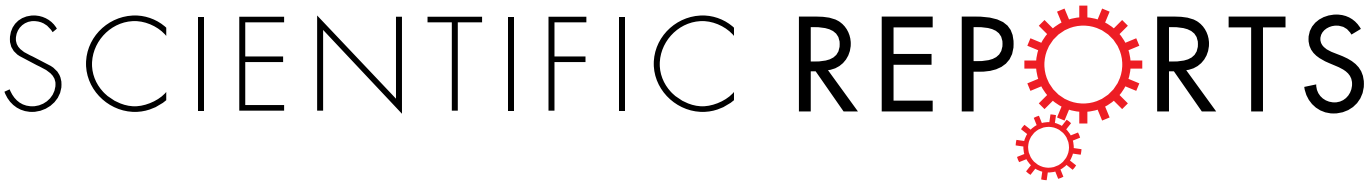


OPEN

The PRC2-binding long non-coding RNAs in human and mouse genomes are associated with predictive sequence features

Received: 11 October 2015
Accepted: 28 December 2016
Published: 31 January 2017

Shiqi Tu[1,2], Guo-Cheng Yuan[3,4,5] & Zhen Shao[1]

Recently, long non-coding RNAs (lncRNAs) have emerged as an important class of molecules involved in many cellular processes. One of their primary functions is to shape epigenetic landscape through interactions with chromatin modifying proteins. However, mechanisms contributing to the specificity of such interactions remain poorly understood. Here we took the human and mouse lncRNAs that were experimentally determined to have physical interactions with Polycomb repressive complex 2 (PRC2), and systematically investigated the sequence features of these lncRNAs by developing a new computational pipeline for sequences composition analysis, in which each sequence is considered as a series of transitions between adjacent nucleotides. Through that, PRC2-binding lncRNAs were found to be associated with a set of distinctive and evolutionarily conserved sequence features, which can be utilized to distinguish them from the others with considerable accuracy. We further identified fragments of PRC2-binding lncRNAs that are enriched with these sequence features, and found they show strong PRC2-binding signals and are more highly conserved across species than the other parts, implying their functional importance.

Polycomb group (PcG) proteins are important epigenetic regulators in development and disease[1,2]. In mammalian cells, although quite a few transcription factors has been found to be linked with the chromatin binding and function of PcG proteins[1,3–6], yet the underlying mechanisms controlling their site-specific chromatin recruitment remain incompletely understood. Since the identification of XIST and HOTAIR[7,8], non-coding RNA-mediated recruitment of Polycomb repressive complex 2 (PRC2) has become a plausible, potentially sequence-dependent mechanism for Polycomb protein and H3K27me3 target regulation[1]. Recently, a set of RNA coimmunoprecipitation and chip hybridization (RIP-chip) experiments were published, which examined the expression and function of hundreds of lncRNAs in three different human cell types, and found more than 200 of them can physically interact with the core subunits of PRC2[9]. This result provided the first population-scale evidence of the interaction between lncRNA and PRC2.

Although a number of models have been proposed to elucidate how lncRNAs interact with their protein partners, especially chromatin remodeling factors, and participate in epigenetic regulations[10–12], only a few large-scale RIP experiments have been published[9,13], which makes it extremely difficult to study the role of interactions between lncRNAs and chromatin remodeling factors across different cell types. In particular, the precise mechanism through which lncRNAs may be targeted by chromatin remodeling factors, such as Polycomb proteins, is unclear. For example, it remains under debate whether PRC2 binds to RNA in a sequence dependent manner[14–17], and it has been proposed that promiscuous and specific RNA binding may both exist for PRC2[15]. Moreover, quite a number of PRC2-binding lncRNAs have been discovered in human and mouse genomes[7–9,13], but it is still not clear whether the mechanisms mediating *in vivo* PRC2-lncRNA interactions are evolutionarily conserved[15].

[1]Chinese Academy of Sciences Key Laboratory of Computational Biology, Collaborative Innovation Center for Genetics and Developmental Biology, Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [5]Harvard Stem Cell Institute, Cambridge, MA 02138, USA. Correspondence and requests for materials should be addressed to Z.S. (email: shaozhen@picb.ac.cn)

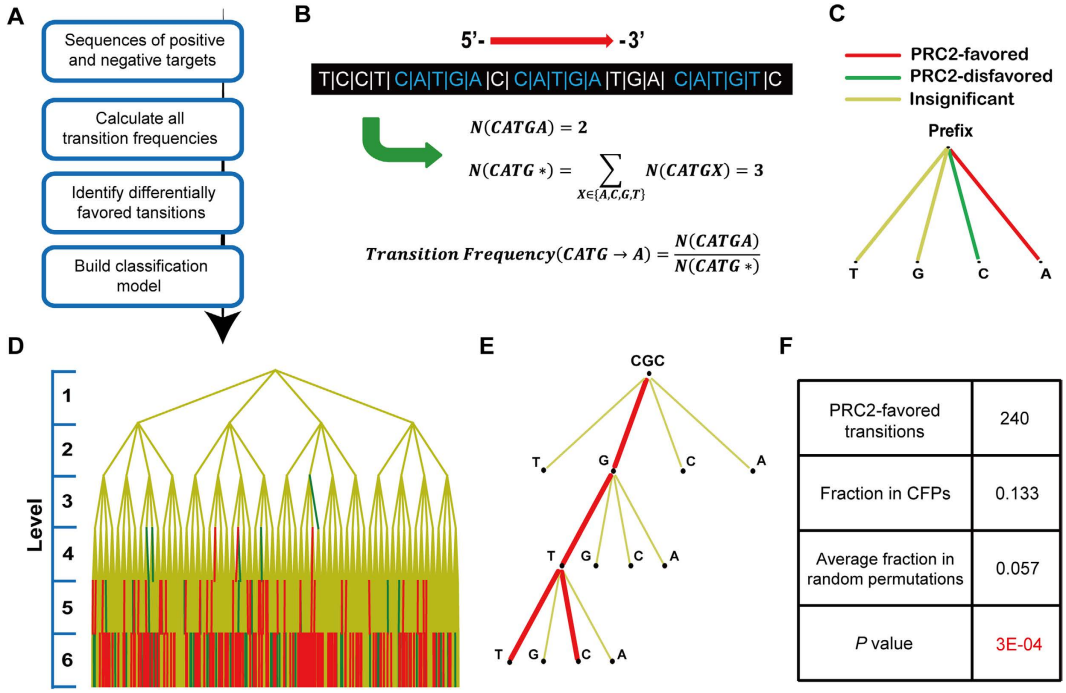


Figure 1. Analysis of the sequence features of human PRC2-binding lncRNAs. (**A**) Workflow of the sequence composition analysis pipeline. (**B**) Calculation of transition frequency, which is defined as the frequency of observing a transition in the given sequence (here order-4 transition CATG→A is used as an example). (**C**) A building block of quad-tree comprised of 4 transitions with the same prefix. Each line represents a transition and the color indicates whether the transition is significantly favored or disfavored by human PRC2-positive lncRNAs. (**D**) The complete quad-tree of height 6 constituted by all possible transitions of order 0–5 (placed on level 1-6 accordingly). Particularly, the root is an empty string as the prefix of 4 order-0 transitions. (**E**) A branch cut from the quad-tree shown in (**D**), which starts from level 3 and contains two consecutively favored paths (CFPs) CGC→G→T→T and CGC→G→T→C. (**F**) Summary statistics of the CFPs observed in (**D**), which suggest the human PRC2-favored transitions significantly prefer to connect with each other and form CFPs.

In order to address these important questions, we carry out a systematic analysis of the DNA sequence patterns associated with PRC2-binding lncRNAs in both human and mouse genomes. In particular, we have developed a new computational pipeline for analyzing the composition of long DNA and RNA sequences of variable length using a Markov-chain based approach[18]. It considers each sequence as a series of transitions between adjacent nucleotides and uses the frequency of observing each possible transition to characterize the composition of this sequence. Through application of this pipeline to the PRC2-binding and non-binding lncRNAs identified from publicly available RIP data in human and mouse, we discovered a number of transitions that are differentially favored by these two classes of lncRNAs as the sequence features associated with PRC2-lncRNA interactions. By mapping all possible transitions to a complete quad-tree, we found a considerable fraction of transitions favored by PRC2-binding lncRNAs are located in consecutive paths, and these transitions are more likely to be simultaneously favored by human and mouse PRC2-binding lncRNAs than the others. We further built prediction models using the sequence features of PRC2-binding lncRNAs as predictors, which could distinguish these lncRNAs from the others with considerable accuracy. Remarkably, the fragments of PRC2-binding lncRNAs that are highly enriched with these sequence features show significant conservation across species, indicating the importance of these fragments.

Results

PRC2-lncRNA interactions in human are associated with significant sequence specificity.

Figure 1A shows an overview of our computational pipeline for sequence composition analysis. It takes two distinct groups of sequences as input, e.g. the DNA sequences of genes that are associated and not associated with a specific biological function. In this pipeline, a systematic analysis is applied to study the compositional patterns of input sequences by modeling each sequence as a Markov chain[18–20], which can be dissected into a series of transitions between adjacent nucleotides (Fig. 1B). To avoid arbitrarily selecting the exact order of Markov chain model, all possible transitions of order 0 through m are utilized (here we chose m = 5, which led to 5460 possible transitions in total). Next, transitions differentially favored by two sequence groups are selected as their sequence features (Fig. 1C; see Methods). Finally, a classification model is constructed by applying Bayesian additive regression trees (BART)[21] analysis to test whether these sequence features can be used to predict the group label of each sequence.

Level	Count	[Q1, Q3]	*P* value
1	0	[0, 0]	1
2	0	[0, 0]	1
3	0	[0, 1]	1
4	3	[2, 7]	0.63
5	34	[16, 30]	0.19
6	203	[87, 115]	0.007
All	240	[108, 151]	0.048

Table 1. Distribution of human PRC2-favored transitions on each level of the quad-tree. Here the count refers to the number of PRC2-favored transitions on each level. Q1 and Q3 corresponds to the first and third quartile of this count obtained from 1000 randomized lncRNA sets, respectively. The *P* value associated with each level was calculated as the fraction of randomized lncRNA sets having an equal or larger number of PRC2-favored transitions on this level than that observed from real data.

To investigate the role of lncRNAs' sequence composition in mediating their interactions with PRC2, we first collected 261 human lncRNAs that can physically interact with the core subunits of PRC2 in three human cell types from Khalil *et al.*[9], together with 227 lncRNAs that were expressed in these cell types but failed to show detectable interaction with PRC2. These two groups of lncRNAs were labeled as human PRC2-positive and PRC2-negative lncRNAs, respectively. Next, we applied our pipeline to compare their sequence composition, with the purpose to uncover the underlying sequence features associated with PRC2-lncRNA interactions. As a result, we identified 240 transitions that are significantly favored by human PRC2-positive lncRNAs compared to PRC2-negative ones, together with 87 transitions significantly disfavored by them (using P-value < 0.05 as cutoff), and named them as human PRC2-favored and disfavored transitions, respectively. To make a global visualization of these transitions, we constructed a complete quad-tree of height 6 comprised of all 5460 possible transitions of order 0 through 5, which were placed on level 1 through 6 of the tree accordingly. Here the PRC2-favored and disfavored transitions were specially colored as red and green, respectively (Fig. 1C,D).

Besides serving as a platform for visualization, the quad-tree can also be utilized to test whether the selected transitions form a nontrivial subset of all 5460 possible transitions, by inspecting their distribution on the tree. We first examined the number of PRC2-favored and disfavored transitions on each level, and found the vast majority of these transitions are on level 6 (Table 1 and Supplementary Table S1). To estimate the significance of this observation, we generated 1000 sets of randomized PRC2-positive and PRC2-negative lncRNAs, each of which was derived by randomly shuffling the original group labels of 488 human PRC2-positive and PRC2-negative lncRNAs (see Methods), and repeated the same feature selection process on each randomized lncRNA set. By this means, only the number of PRC2-favored transitions on level 6 was found to be significantly higher than that observed from randomized lncRNA sets ($P = 0.007$, Table 1). Furthermore, although the selected transitions showed a rather sparse distribution on quad-tree, we still observed that these transitions, especially the PRC2-favored ones, tend to connect with each other across adjacent levels and form consecutive paths. To validate this finding, we define consecutively favored/disfavored paths (CFPs/CDPs) as the consecutive paths on quad-tree that are completely constituted by PRC2-favored/disfavored transitions (Fig. 1E and Supplementary Fig. S1A), respectively. Interestingly, a considerable fraction of PRC2-favored transitions are located in CFPs (13.3%, Fig. 1F), and this value is significantly higher than that observed in random simulations, in which all the PRC2-favored transitions were randomly re-distributed on each level ($P = 3E\text{-}4$ by permutation test, see Methods). On the other hand, PRC2-disfavored transitions only exhibited a weak enrichment in CDPs compared to that expected by chance (9.2% and $P = 0.023$, Supplementary Fig. S1B).

To understand why PRC2-favored transitions prefer to form consecutive paths, we additionally applied our pipeline to analyze the sequence features associated with transcription factor CTCF's DNA binding in human cells (see Supplementary text for details), as the sequence specificity of this interaction is largely known[22]. Strikingly, the vast majority of CTCF-favored transitions identified by our pipeline are located in CFPs and, particularly, 52 CFPs are of full length. We compared a representative full-length CFP with CTCF's binding motif obtained from JASPAR database[23], and found the 6-mer formed by this full-length CFP can be well matched with the motif (Supplementary Fig. S1C). Inspired by this observation, we calculated the motif score of each full-length path, which is defined to measure the similarity between the 6-mer formed by all the 6 transitions on this path and CTCF's binding motif (see Supplementary text for details). Interestingly, strong correlation was observed between the length of the longest CFP lying on each full-length path and the motif score of this path (Supplementary Fig. S1D), suggesting the preference of those favored transitions to form consecutive paths may be intrinsically connected with the sequence specificity mediating the binding of corresponding proteins and the CFPs observed in this study can be of biological importance.

The sequence features of human PRC2-binding lncRNAs are predictive of PRC2-lncRNA interactions. In order to evaluate whether the sequence features identified by our pipeline can be used to predict PRC2-binding lncRNAs, we took the frequencies of all human PRC2-favored and disfavored transitions as predictors, and employed BART analysis[21] to fit a prediction model. Based on a standard 10-fold cross-validation (CV) process, we found this model is able to distinguish human PRC2-positive lncRNAs from the negative ones with good accuracy, and the area under the receiver operating characteristic (ROC) curve (AUC) is 0.82 (Fig. 2A), which is close to that achieved by the prediction model published previously[24]. In addition, we also adopted a

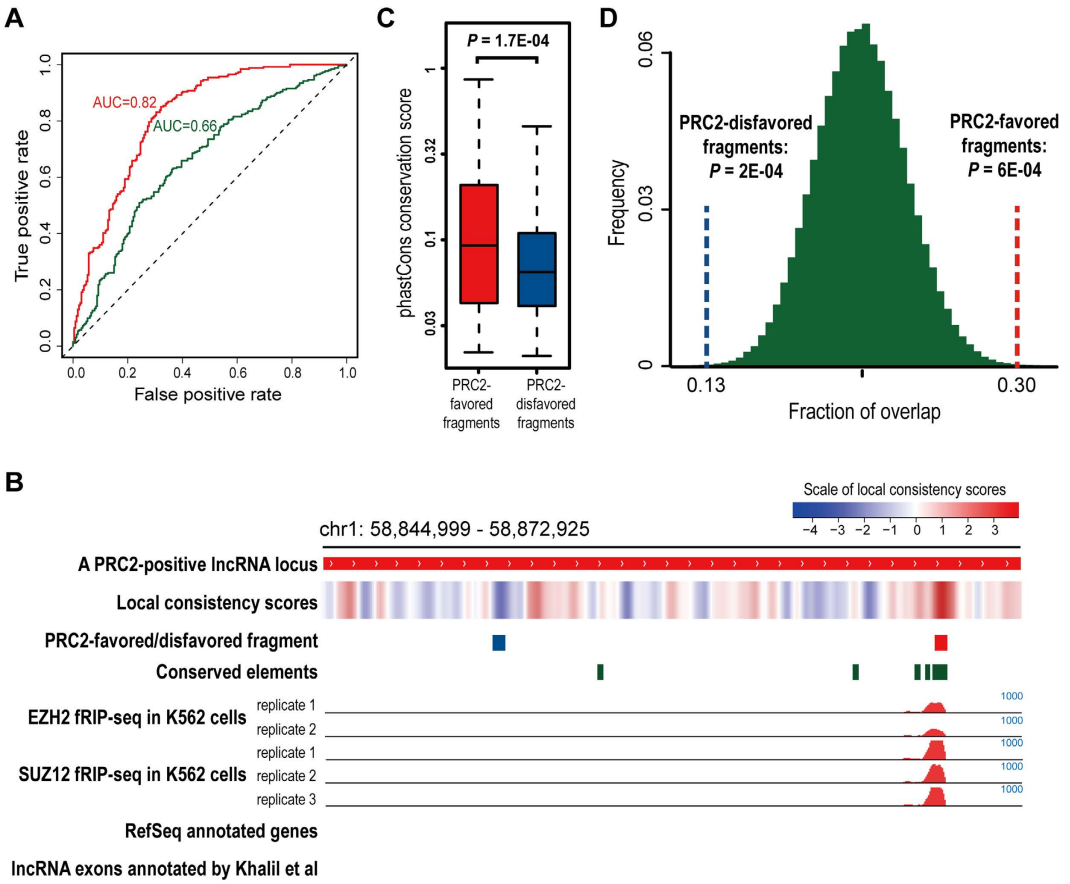


Figure 2. Prediction of the PRC2-lncRNA interactions in human genome based on transition frequencies. (**A**) ROC curves and corresponding AUC values of the prediction models built by the non-blind CV (red line) and the fully blind method (green line) in predicting human PRC2-binding lncRNAs. (**B**) A representative PRC2-positive lncRNA locus. Here its PRC2-favored and disfavored fragment are indicated by the red and blue bar, respectively, and the red tracks in the middle show the fRIP-seq read counts of EZH2 and SUZ12 in human K562 cell line. (**C**) Boxplot of the average PhastCons conservation scores of the PRC2-favored and disfavored fragments identified from human PRC2-binding lncRNAs. (**D**) Distribution of the fraction of the 500 bp fragments randomly selected from human PRC2-binding lncRNAs that overlap with the conserved elements. Here the distribution was draw from 10^5 times of random sampling and dash lines represent the fraction of PRC2-favored/disfavored fragments that overlap with the conserved elements.

more stringent approach of model building, which is called as fully blind method here. Similar to the original 10-fold CV process, all human PRC2-positive and PRC2-negative lncRNAs are divided into 10 subgroups and at each cross-validation step, only one subgroup is selected as the testing set, leaving the other 9 subgroups as training set. The key difference of the fully blind method is that predictor selection is repeatedly performed at each cross-validation step and only lncRNAs in the training set can be used to identify PRC2-favored and disfavored transitions as predictors (see Methods), which means the predictors used at each step may not be exactly the same as the 240 PRC2-favored and 87 PRC2-disfavored transitions that were used as predictors in the original 10-fold CV process (it will be called as "non-blind CV" from now on in this study). In this way, human PRC2-binding lncRNAs were predicted with moderate accuracy (AUC = 0.66). More specifically, 76% of the top 261 lncRNAs predicted by the non-blind CV approach are true PRC2-positive ones, and this fraction decreased to 66% for the top 261 lncRNAs predicted by the fully blind approach. To explain why prediction models built by these two methods exhibited distinct performance, we drew the ROC curve for each of the 10 lncRNA subgroups separately and calculated the corresponding AUC value. Remarkably, compared to the non-blind CV method, the AUC values of 10 lncRNA subgroups got from the prediction model built by fully blind method exhibited much higher variation, with a range from 0.53 to 0.8 (Supplementary Fig. S2A). On the other hand, we devised two empirical classification models using the non-blind CV and also the fully blind method, respectively, but without involving BART to perform sophisticated model training (see Supplementary text for details). By comparing the performance of these empirical models on the same set of human lncRNAs, it could be clearly viewed that whether or not to exclude lncRNAs in the testing set from predictor selection can strongly affect prediction accuracy (Supplementary Fig. S2B,C). Since the fully blind method is more stringent, we think it's better to use prediction models built by this method to infer how accurately PRC2-binding lncRNAs can be predicted by their sequence composition.

Next, we further investigated the distribution of the sequence features of human PRC2-binding lncRNAs along their gene bodies. Following this direction, each PRC2-positive lncRNA was scanned by a sliding window of size 500 bp and a local consistency score was assigned to the sequence fragment in the window, which is defined as the sum of the frequencies of all PRC2-favored transitions in this sequence fragment minus those of all PRC2-disfavored ones. In this way, sequences with high consistency scores should be highly enriched for PRC2-favored transitions and also depleted of PRC2-disfavored ones. Interestingly, these lncRNAs exhibit highly non-uniform local consistency scores along their gene bodies, and some fragments of them have clearly higher scores than the others (Fig. 2B). Inspired by this finding, we defined the fragment with the highest/lowest consistency score in each human PRC2-positive lncRNA as its PRC2-favored/disfavored fragment (Fig. 2B), respectively. To know whether they can be potentially important for PRC2-lncRNA interactions, we examined the RNA binding of PRC2 on these fragments as well as their conservation level across vertebrate genomes (Fig. 2B). For the first analysis, we incorporated a recently published RIP-seq dataset of PRC2 core subunit EZH2 and SUZ12 in K562 human leukemia cell line[14], and calculated the RIP-seq read density at each PRC2-favored and disfavored fragment. Interestingly, binding of EZH2 and SUZ12 at PRC2-favored fragments was found to be stronger than that at PRC2-disfavored ones (Supplementary Fig. S2D). Meanwhile, we also observed PRC2-favored fragments have significantly higher average conservation scores than PRC2-disfavored ones ($P = 1.7E-04$ by paired t-test; Fig. 2C). More explicitly, 30% of PRC2-favored fragments overlap with conserved elements[25,26], which is significantly higher than that of the 500 bp fragments randomly selected from the same lncRNAs ($P = 6E-04$), and this fraction for PRC2-disfavored fragments is only 13% ($P = 2E-04$; Fig. 2D).

To check whether the high conservation level of PRC2-favored fragments are directly linked with the aggregation of sequence features associated with PRC2-lncRNA interactions, we again took the 1000 sets of randomized PRC2-positive and PRC2-negative lncRNAs, and reselected a group of pseudo PRC2-favored and disfavored fragments for each set of randomized PRC2-positive lncRNAs using the pseudo PRC2-favored and disfavored transitions associated with this randomized lncRNA set, which were identified by comparing the sequence composition of these lncRNAs with the corresponding randomized PRC2-negative lncRNAs. Then, the same analyses as shown in Fig. 2C,D were applied to each set of pseudo PRC2-favored and disfavored fragments identified from the randomized PRC2-positive lncRNAs. Remarkably, when comparing the average conservation scores of pseudo PRC2-favored and disfavored fragments, only 1.7% of the 1000 randomized lncRNA sets achieved *P*-values lower than that shown in Fig. 2C (Supplementary Fig. S2E,F), which is taken as an empirical estimate of the false positive rate (FPR) of the test conducted in Fig. 2C (Supplementary Fig. S2F). Similarly, we calculated the *P*-value for the overlap between each set of pseudo PRC2-favored fragments and conserved elements by comparing with the fragments randomly selected from the same lncRNAs, and found only 0.8% of the 1000 sets of pseudo PRC2-favored fragments got *P*-values lower than that shown in Fig. 2D (Supplementary Fig. S2F). Taken together, these findings indicate the PRC2-favored fragments, which are highly enriched with sequence features associated with PRC2-lncRNA interactions, are generally more conserved than the other parts of the lncRNAs they belong to, and, thus, are more likely to be of functional importance.

Comparison of the sequence features of human and mouse PRC2-binding lncRNAs. The core subunits of PRC2 as well as their roles in transcriptional repression are highly conserved from Drosophila to mammals[1]. Besides, interactions between PRC2 and lncRNAs are detected in both human and mouse cells, and some of them are shared between two species[7,8]. Thus, it would be interesting to know whether the PRC2-binding lncRNAs in human and mouse genomes tend to share common sequence features, despite that the sequences of lncRNAs are known to be generally much less conserved than protein-coding genes[27]. To answer this question, we first studied a RIP-seq dataset of EZH2 generated from mouse embryonic stem cells (mESCs)[13]. Based on this dataset and the mouse lncRNAs that were discovered in parallel with the human ones used in this study[28], we obtained 153 mouse lncRNAs having physical interactions with EZH2 in mESCs, together with 387 lncRNAs that are expressed in mESCs but failed to show detectable interaction with EZH2, which were labeled as mouse PRC2-positive and PRC2-negative lncRNAs, respectively (see Supplementary text for details). Subsequently, the same sequence composition analysis was applied to these lncRNAs, and 175 mouse PRC2-favored transitions as well as 116 PRC2-disfavored ones were identified as the sequence features of mouse PRC2-binding lncRNAs. Then, we used the frequencies of these transitions as predictors and employed BART analysis to fit a prediction model of mouse PRC2-binding lncRNAs. Similar to what we observed from human lncRNAs, prediction model built by the fully blind method exhibited moderate accuracy (AUC = 0.64), clearly lower than the model built by the non-blind CV method (AUC = 0.88, Fig. 3A). Meanwhile, we fitted a prediction model using all human PRC2-positive and PRC2-negative lncRNAs as training set and all the human PRC2-favored and disfavored transitions as predictors, and applied this model on mouse lncRNAs to perform cross-species prediction. Interestingly, by this means mouse PRC2-positive lncRNAs can be distinguished from the PRC2-negative ones with considerable accuracy (AUC = 0.60, Fig. 3A). Since in the cross-species prediction only human lncRNAs were involved in predictor selection, its performance should be compared with the prediction model trained with mouse lncRNAs using the fully blind method. Thus, we speculate some sequence features are shared between human and mouse PRC2-binding lncRNAs.

Inspired by this hypothesis, we analyzed the overlap between the sequence features of human and mouse PRC2-binding lncRNAs. In general, only 10% of the PRC2-favored transitions are shared between human and mouse (18 of 240/175, Supplementary Fig. S3A), though still significantly higher than expected by chance (Fig. 3A). As we have shown, human PRC2-favored transitions prefer to be located on level 6 of the quad-tree and connect with each other to form CFPs. It's reasonable to pay special attentions to these transitions. Interestingly, being on level 6 itself doesn't significantly increase the likelihood of human PRC2-favored transitions being also favored by mouse PRC2-binding lncRNAs. However, 19% of the 32 human PRC2-favored transitions located in CFPs remained to be mouse PRC2-favored ones (which means 1/3 of the PRC2-favored transitions shared

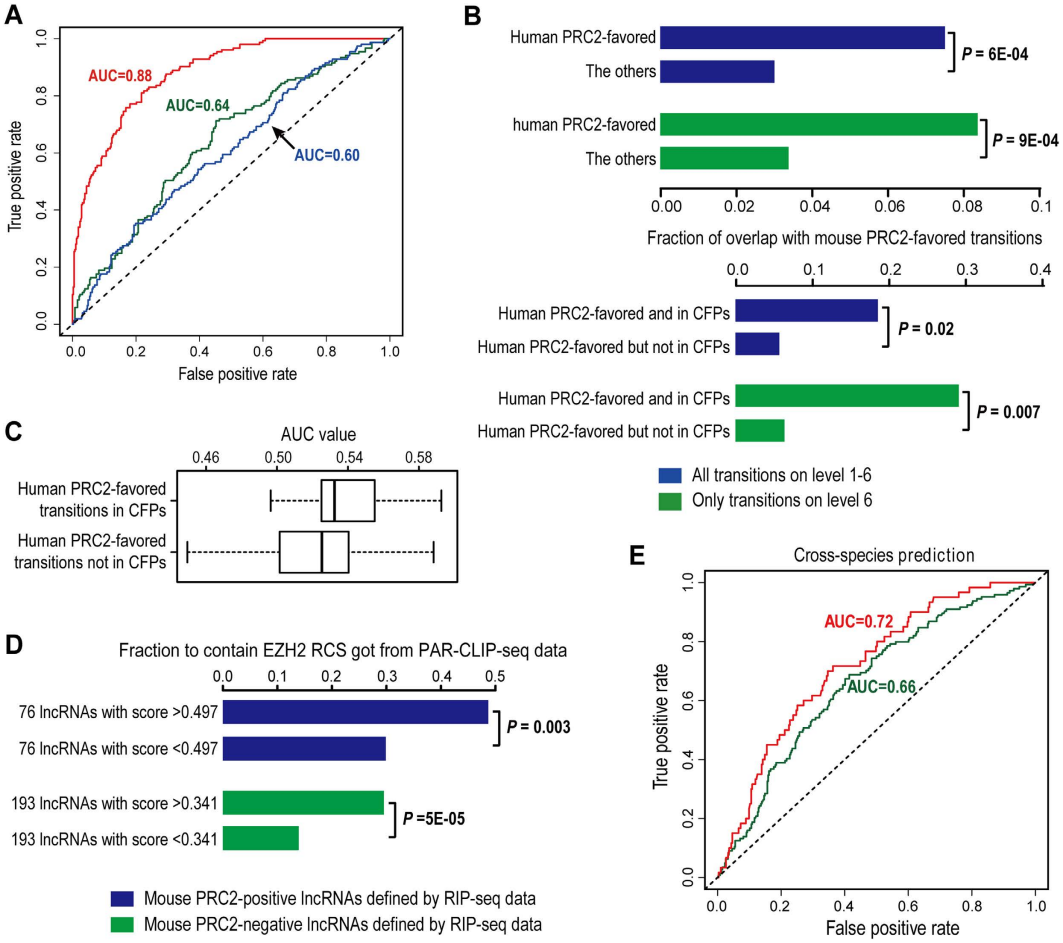


Figure 3. Human PRC2-favored transitions in CFPs are more likely to be also favored by mouse PRC2-binding lncRNAs than the others. (**A**) ROC curves and corresponding AUC values of different prediction models in predicting mouse PRC2-binding lncRNAs. The red and green curve correspond to mouse prediction models built by the non-blind CV and the fully blind method, respectively, in which mouse PRC2-positive and PRC2-negative lncRNAs were used for predictor selection and model training. The blue curve corresponds to the human prediction model using human PRC2-positive and PRC2-negative lncRNAs for predictor selection and model training. (**B**) Fractions of different groups of transitions that are identified as mouse PRC2-favored transitions. Here, the *P*-values were computed by right-tailed Fisher's exact test based on hypergeometric distribution. (**C**) Boxplot of the AUC values of human PRC2-favored transitions in predicting mouse PRC2-binding lncRNAs. Here the human PRC2-favored transitions are divided into 2 groups based on whether or not they are located in CFPs, and the AUC value of a transition is calculated by directly using its frequency in each sequence as the prediction score of this sequence. (**D**) Fraction of mouse PRC2-positive and PRC2-negative lncRNAs that contain EZH2 RCS identified from PAR-CLIP-seq data. Here each group of lncRNAs were split into two subgroups of equal size by the median of their cross-species prediction scores derived from the prediction model trained with human lncRNAs, and the *P*-values were calculated by right-tailed Fisher's exact test to measure whether the subgroup of lncRNAs with high prediction scores are significantly more likely to contain EZH2 RCS compared to the subgroup with low prediction scores. (**E**) ROC curve and corresponding AUC value of the human prediction model in predicting mouse RCS-containing lncRNAs from the RCS-null ones (green), and also that in predicting high-confidence mouse PRC2-positive lncRNAs from high-confidence mouse PRC2-negative ones (blue).

between human and mouse are located in human CFPs), and this fraction increases to 29% for the human PRC2-favored transitions located on level 6 and also in CFPs (Fig. 3B). Moreover, we tried building cross-species prediction models using all human PRC2-favored transitions and only those located in CFPs as predictors, respectively, and tested the performance of these models on mouse lncRNAs. Of note, they achieved AUC values (0.61 and 0.59) very close to the cross-species prediction model using all human PRC2-favored and disfavored transitions as predictors (Supplementary Fig. 3C). Then, we calculated the AUC value of each human PRC2-favored transition in predicting mouse PRC2-binding lncRNAs, which is computed by directly assigning the frequency of observing this transition in the sequence of each mouse lncRNA as the score of this lncRNA. Remarkably, the majority of human PRC2-favored transitions have AUC values greater than 0.5, which suggests these transitions tend to also be positively favored by mouse PRC2-binding lncRNAs, and, especially, the ones

falling in CFPs achieved clearly higher AUC values than those out of CFPs (Fig. 3C). Thus, we conclude that the association between transitions falling in CFPs and PRC2-binding lncRNAs are more highly conserved between human and mouse than those out of CFPs. Again, these findings strongly support the biological significance of CFPs.

In recent years, *in vivo* UV light cross-linking and immunoprecipitation followed by high-throughput sequencing (CLIP-seq) experiments have also been widely used to study genome-wide protein-RNA interactions[29–32]. To make a more comprehensive assessment of our cross-species comparison of PRC2-binding lncRNAs, we additionally incorporated a recently published EZH2 PAR-CLIP-seq dataset in mESCs[31]. We obtained 13,764 putative RNA-contact sites (RCSs) of EZH2 from this dataset and mapped them to the mouse PRC2-positive and PRC2-negative lncRNAs we defined from the EZH2 RIP-seq data in mESCs. 39.2% of mouse PRC2-positive lncRNAs were found to contain at least one EZH2 RCS, and this fraction for mouse PRC2-negative ones is only 21.7% (Supplementary Fig. 3D, see Supplementary text for details), indicating a moderate consistency between these two datasets. Next, we divided both mouse PRC2-positive and PRC2-negative lncRNAs into two subgroups of equal size, based on their cross-species prediction scores derived from the prediction model trained with human lncRNAs. Interestingly, almost half of the mouse PRC2-positive lncRNAs with high cross-species prediction scores have EZH2 RCS identified from the PAR-CLIP-seq data (Fig. 3D), which is significantly higher than that of the PRC2-positive lncRNAs with low prediction scores (29.0%, $P = 0.003$ by Fisher's exact test), indicating they are more likely to be true PRC2-binding lncRNAs. On the other hand, still a considerable faction of the mouse PRC2-negative lncRNAs with high cross-species prediction scores were found to contain EZH2 RCS (29.5%, Fig. 3D), which is also significantly greater than that of the PRC2-negative lncRNAs with low prediction scores (13.9%, $P = 5E\text{-}5$), implying many of them may actually have the potential to physically interact with PRC2 as predicted by their sequence similarity with the human PRC2-binding lncRNAs. Inspired by these findings, we defined high-confidence mouse PRC2-positive lncRNAs as the mouse PRC2-positive lncRNAs that also contain RCS of EZH2, and high-confidence mouse PRC2-negative lncRNAs as the mouse PRC2-negative lncRNAs with no EZH2 RCS. By taking only these high-confidence lncRNAs into account, we found the accuracy of cross-species prediction is even higher (AUC = 0.72, Fig. 3E), which strongly supports that a considerable proportion of the sequence patterns associated with PRC2-lncRNA interactions are shared between human and mouse.

Compare the performance of prediction models based on transition and K-mer frequencies. In previous studies, the composition of a sequence was usually analyzed by counting the occurrence of different K-mers in it[18,33], and typically the count of each K-mer would be further normalized by sequence length to represent the frequency of observing this K-mer in the sequence. Here, sequence composition analysis based on K-mer frequencies was also applied to study human PRC2-binding lncRNAs. Technically, we calculated the frequencies of all possible K-mers of length from 1 through 6 in the DNA sequence of each lncRNA, and searched for K-mers that occur in human PRC2-positive lncRNAs with significantly higher or lower frequencies than in the PRC2-negative ones (see Supplementary text for details). By using P-value < 0.05 as cutoff, 129/83 K-mers that are significantly over/under-represented in human PRC2-binding lncRNAs were identified, respectively. The prediction model using the frequencies of these K-mers as predictors showed slightly lower accuracy (AUC = 0.63 using the fully blind method) than the model based on transition frequencies (Fig. 4A). To assess the impact of lncRNA length, we separately divided the human PRC2-positive and PRC2-negative lncRNAs into two subgroups of equal size according to their length, which were named as moderately long and extremely long subgroup, respectively (the cutoff of lncRNA length to separate these two lncRNA subgroups is 18.6 kb for human PRC2-positive lncRNAs and 13.1 kb for PRC2-negative ones). Then, the performance of prediction was evaluated on the two subgroups separately. Interestingly, two prediction models achieved similar AUC values on the moderately long subgroup (Fig. 4A), while the model based on transition frequencies exhibited a better accuracy on the extremely long subgroup of lncRNAs (AUC = 0.70) than the one based on K-mer frequencies (AUC = 0.63; Fig. 4A). In addition, we also performed the same analysis on mouse lncRNAs and got a similar result (Fig. 4B), suggesting this finding is not specific to the human lncRNAs chosen by us and sequence composition analysis based on transition frequencies can have plausible performances on extremely long sequences.

It should be noted that besides finding the sequence features associated with PRC2-lncRNA interactions in mammalian cells, another main purpose of this study is to develop a new computational pipeline for sequence composition analysis based on splitting each sequence into transitions between adjacent nucleotides. We have demonstrated sequence composition analysis based on transition frequencies can have different downstream analyses from that based on K-mers, e.g. to examine the distribution of selected transitions on a quad-tree, which led to the identification of CFPs (Supplementary Fig. S1C and S4D). However, which of these two types of methods can have a superior performance highly depends on the context (a theoretical example is given in Supplementary text) as well as the implementation of these methods, since the analysis based on K-mers has been widely used for years and a large number of computation models have been developed to improve its performance and to extend its applications[18].

Discussion

In this study, we conducted a systematic sequence composition analysis on known PRC2-binding lncRNAs in both human and mouse genomes. To be noted, identifying characteristic sequence features from the lncRNAs associated with a specific biological function is important and also computationally challenging. One of the main reasons is that the length of lncRNA genes can be quite long and highly variable (here we chose to use the whole gene body of lncRNAs for sequence analysis, and a detailed explanation can be found in Supplementary text). In our case, the human lncRNAs used here are of size 33.4 ± 41.3 kb (average gene length $\pm$ standard deviation, the median length is 15.8 kb), and a large fraction of them may not be well annotated (two examples can be seen in

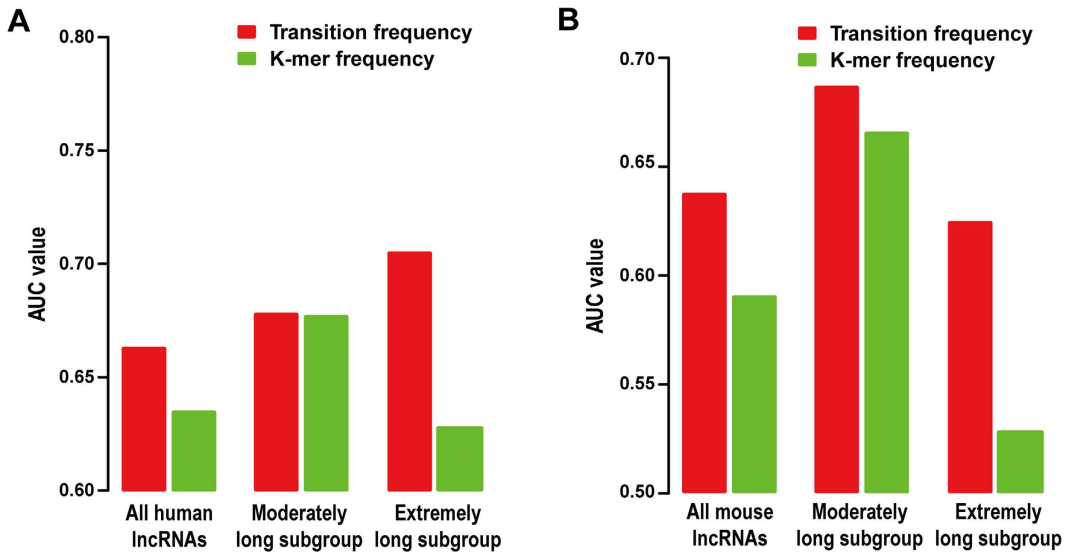


Figure 4. Compare the performance of prediction models based on K-mer and transition frequencies. (**A**,**B**) AUC values of the prediction models based on transition (red bars) or K-mer (blue bars) frequencies, which were trained and tested by the human (**A**) and mouse (**B**) lncRNAs, respectively. Here the prediction models were built by the fully blind method, and all human/mouse PRC2-positive and PRC2-negative lncRNA were separately divided into two subgroups of equal size according to their length, termed as the moderately long and the extremely long subgroup, to access the performance of these models on lncRNAs of different length.

Fig. 2B and Supplementary Fig. S2H), which makes the sequence analysis even more complicated. Here, we presented a new computational pipeline for analyzing the compositional patterns of long sequences, which considers each sequence as a series of transitions between adjacent nucleotides and can systematically search for transitions that occur in the sequences of interest with significantly different frequencies compared to the control sequences. Besides, the pipeline is incorporated with a set of computational analyses to visualize all candidate transitions using a complete quad-tree and then to dissect the distribution of selected transitions on the tree. Applying it to compare the sequences of PRC2-binding and non-binding lncRNAs in human and mouse genomes, we identified a large pool of transitions as features of PRC2-binding lncRNAs in each species, and found those transitions favored by PRC2-binding lncRNAs exhibit a significant preference to connect with each other and form CFPs on the quad tree, which seems not to be sufficiently appreciated by other similar studies. Interestingly, although the sequence features of PRC2-binding lncRNAs show a low overlap between human and mouse, the majority of human PRC2-favored transitions have AUC values higher than 0.5 in predicting mouse PRC2-binding lncRNAs, especially for those falling in CFPs (Fig. 3C). Although lncRNAs generally are thought to be poorly conserved[27,34–36], our findings suggest PRC2-lncRNA interactions in mammalian cells are clearly associated with specific sequence patterns and these patterns tend to be conserved across species, which can be further supported by the good performance of cross-species predictions (Fig. 3D).

Another interesting aspect of our findings is that the sequence composition of lncRNAs can be highly complex along their gene bodies, which supports the hypothesis that such a great complexity might be necessary for its functions[37]. For example, as shown in Fig. 2D and Supplementary Fig. S2G,H, the sequence features of PRC2-binding lncRNAs showed a highly non-uniform distribution along their gene bodies and, particularly, some regions are significantly more enriched with these features than the other parts. Taking a step further, we recognized a set of fragments that are highly enriched with these features from human PRC2-binding lncRNAs, and found these fragments are significantly more highly conserved than the other parts of these lncRNAs, implying they may be potentially important for the function of these lncRNAs. This observation can provide a different viewpoint to understand the low conservation levels of lncRNAs in mammals, and implies evolutionary analyses can still serve as a useful tool for identifying functional elements of lncRNAs[38]. Taken together, our analysis indicates that, although the sequences of lncRNAs are of tremendous complexity, they still share quite a number of recurring patterns. Using these patterns as clues, our predictions based on global and local sequence compositions can serve as a useful guide for experimental biologists to investigate the potential connections between Polycomb group proteins and lncRNAs in a tissue-specific manner, and also to further dissect how these connections are established. For future studies, even more sophisticated models, e.g. nonhomogeneous Markov model[39], may be employed to further understand the heterogeneous sequence composition patterns of lncRNAs.

Material and Methods

Selection of human PRC2-positive and PRC2-negative lncRNAs. Khalil *et al.* used RIP-chip experiments to examine the interaction between ~900 human lncRNAs and SUZ12 or EZH2, two well-known core subunits of PRC2, in three human cell types: HeLa, lung fibroblasts and foot fibroblasts[9]. In these lncRNAs, 261 were found to have physical interactions with PRC2 in at least one cell type, which are defined as human PRC2-positive

lncRNAs here. Besides, 227 lncRNAs that are expressed in these cells but failed to show detectable interaction with PRC2 are defined as human PRC2-negative lncRNAs.

To infer the transcription start site (TSS) and then the coding strand of each lncRNA, we collected all available ChIP-Seq data of histone mark H3K4me3, which is known to be mainly associated with active gene promoters, from ENCODE project[40,41]. All ChIP-Seq reads were mapped to both ends of each annotated lncRNA locus and the one with higher overall H3K4me3 signal intensities was considered as putative TSS, leaving the other end as putative transcription end site (TES). To assess the validity of this approach, we applied it to RefSeq annotated protein-coding genes[42] and found for the 15655 genes longer than 5 kb, the accuracy is around 86%. Additionally, since many lncRNAs used in this study lack reliable exon annotations, we use the whole gene body of lncRNAs to perform sequence composition analysis (a detailed explanation can be found in Supplementary text).

Decomposition of long DNA sequences into transitions between adjacent nucleotides and selection of differentially favored transitions as sequence features. To start sequence decomposition, each sequence is considered as Markov chain of transitions between neighboring nucleotides. Then, the composition of this sequence can be described by the frequencies of observing all possible transitions in it, which are produced by enumerating the order of Markov chain from 0 to m (here m = 5 was used, resulting in a total number of 5460 different transitions). Taking order-4 transition CATG→A as an example (Fig. 1B), its transition frequency in a given sequence is calculated as

$$\textit{Transition frequency}\,(\boldsymbol{CATG \rightarrow A}) = \frac{\boldsymbol{N}(\boldsymbol{CATGA})}{\boldsymbol{N}(\boldsymbol{CATG^{*}})} = \frac{\boldsymbol{N}(\boldsymbol{CATGA})}{\sum_{\boldsymbol{X}\in\{\boldsymbol{A},\boldsymbol{C},\boldsymbol{G},\boldsymbol{T}\}} \boldsymbol{N}(\boldsymbol{CATGX})}, \quad (1)$$

Here, *N(CATGA)* is the times of observing 5-mer CATGA in this sequence.

To find transitions differentially favored by the sequences of PRC2-positive and PRC2-negative lncRNAs as sequence features of PRC2-binding lncRNAs, Welch's two-sample t-test is applied to compare the frequencies of each transition between these two sequence groups. If the frequencies of a transition in PRC2-positive lncRNAs are significantly higher (lower) than those in PRC2-negative ones with P-value < 0.05, it will be classified as a PRC2-favored (disfavored) transition.

Examine the distribution of PRC2-favored and disfavored transitions on the quad-tree. The distribution of human PRC2-favored and disfavored transitions on the quad-tree was examined from two perspectives. First, the number of favored and disfavored transitions on each level was counted and compared with that got from 1000 sets of randomized PRC2-positive and PRC2-negative lncRNAs. In each randomized lncRNA set, the original group label of 488 human PRC2-positive and PRC2-negative lncRNAs were randomly shuffled, and transitions differentially favored by these randomized lncRNAs were re-selected using the same criteria. Finally, an empirical *P*-value was calculated for each level as the fraction of randomized lncRNA sets that resulted in an equal or higher number of PRC2-favored/disfavored transitions on this level.

To check whether the PRC2-favored/disfavored transitions prefer to connect with each other and form consecutively favored/disfavored paths (CFPs/CDPs), the fraction of them located in CFPs/CDPs was calculated and compared to that got from 1E + 06 times of random permutations. In each random permutation, all PRC2-favored/disfavored transitions were randomly re-selected from the tree, with keeping the number of selected transitions on each level unchanged. Then, an empirical *P*-value was calculated as the fraction of random permutations that led to an equal or higher proportion of PRC2-favored/disfavored transitions falling in CFPs/CDPs.

Prediction of PRC2-binding lncRNAs based on the sequence composition of lncRNAs. To build the prediction model of PRC2-positive lncRNAs, the frequencies of all PRC2-favored and disfavored transitions/K-mers were used as predictors, and Bayesian additive regression trees (BART)[21] analysis was applied to perform model fitting. Here BART was called by using its R package implementation with the default parameter settings, except the number of regression trees was set to be 500. The overall performance of each model was quantified by the area under the receiver operating characteristic (ROC) curve (AUC), which equals 1 if the model made a perfect prediction and 0.5 if the prediction was random. Besides, a more stringent approach of model building, termed as fully blind method, was also used here. The only difference of this approach is that predictor selection was repeatedly performed at each cross-validation step and only lncRNAs in the training set can be used to identify transitions differentially favored by PRC2-positive and PRC2-negative lncRNAs as predictors. In the cross-species prediction, no cross-validation was performed and all human PRC2-positive and PRC2-negative lncRNAs were used to train the prediction model, which was then applied to mouse PRC2-positive and PRC2-negative lncRNAs.

Definition of PRC2-favored/disfavored fragments and conservation analyses. To find lncRNA fragments that are highly enriched for PRC2-favored transitions and depleted of PRC2-disfavored ones, each human PRC2-positive lncRNA was scanned by a sliding window of size 500 bp, and a local consistency score was assigned to the DNA sequence in the window, which is defined as the sum of the frequencies of all PRC2-favored transition in this sequence fragment minus the sum of the frequencies of all PRC2-disfavored ones. Then, the 500-bp fragment with the highest/lowest score of each PRC2-positive lncRNA was defined as its PRC2-favored/disfavored fragment, respectively.

To measure the conservation levels of these fragments, PhastCons conservation scores of human genome were downloaded from UCSC genome browser (the 44-way version was used here)[43]. Additionally, 1354034 conserved elements annotated by GERP (Genomic Evolutionary Rate Profiling) software were obtained from its website[25,26], which cover about 7% of the human genome. To assess whether the selected lncRNA fragments contain more

conserved elements than expected by chance, 1E + 05 times of random simulations were performed. At each time, a 500-bp fragment was randomly chosen from each PRC2-positive lncRNA, and the fraction of these random fragments that overlap with the conserved elements was calculated. Finally, an empirical *P* value was calculated as the proportion of simulations showing an equal or higher fraction of overlapping.

References

1. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–349 (2011).
2. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419,** 624–629 (2002).
3. Li, G. *et al.* Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev* **24,** 368–380 (2010).
4. Liu, Y., Shao, Z. & Yuan, G. C. Prediction of Polycomb target genes in mouse embryonic stem cells. *Genomics* **96,** 17–26 (2010).
5. Arnold, P. *et al.* Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res* **23,** 60–73 (2013).
6. Pinello, L., Xu, J., Orkin, S. H. & Yuan, G. C. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *Proc Natl Acad Sci USA* **111,** E344–353 (2014).
7. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
8. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322,** 750–756 (2008).
9. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* **106,** 11667–11672 (2009).
10. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482,** 339–346 (2012).
11. Lee, J. T. Epigenetic regulation by long noncoding RNAs. *Science* **338,** 1435–1439 (2012).
12. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43,** 904–914 (2011).
13. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40,** 939–953 (2010).
14. Hendrickson, D. G., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biol* **17,** 28 (2016).
15. Davidovich, C. *et al.* Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol Cell* **57,** 552–558 (2015).
16. Davidovich, C., Zheng, L., Goodrich, K. J. & Cech, T. R. Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol* **20,** 1250–1257 (2013).
17. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res* **26,** 896–907 (2016).
18. Shan, G. & Zheng, W. M. Counting of oligomers in sequences generated by markov chains for DNA motif discovery. *J Bioinform Comput Biol* **7,** 39–54 (2009).
19. Kleffe, J. & Langbecker, U. Exact computation of pattern probabilities in random sequences generated by Markov chains. *Comput Appl Biosci* **6,** 347–353 (1990).
20. Kleffe, J. & Borodovsky, M. First and second moment of counts of words in random texts generated by Markov chains. *Comput Appl Biosci* **8,** 433–441 (1992).
21. Chipman, H. A., George, E. I. & McCulloch, R. E. Bart: Bayesian Additive Regression Trees. *Annals of Applied Statistics* **4,** 266–298 (2010).
22. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128,** 1231–1245 (2007).
23. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32,** D91–94 (2004).
24. Glazko, G. V., Zybailov, B. L. & Rogozin, I. B. Computational prediction of polycomb-associated long non-coding RNAs. *PLoS One* **7,** e44878 (2012).
25. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15,** 901–913 (2005).
26. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP plus. *Plos Computational Biology* **6** (2010).
27. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136,** 629–641 (2009).
28. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
29. Guil, S. *et al.* Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat Struct Mol Biol* **19,** 664–670 (2012).
30. Kaneko, S. *et al.* Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Mol Cell* **53,** 290–300 (2014).
31. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20,** 1258–1264 (2013).
32. Mondal, T. *et al.* MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* **6,** 7743 (2015).
33. Robin, S. & Daudin, J. J. Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability* **36,** 179–193 (1999).
34. Bentwich, I. *et al.* Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* **37,** 766–770 (2005).
35. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8,** e1002841 (2012).
36. Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22,** 1–5 (2006).
37. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20,** 300–307 (2013).
38. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
39. Rajagopalan, B., Lall, U. & Tarboton, D. G. Nonhomogeneous Markov Model for Daily Precipitation. *Journal of Hydrologic Engineering* **1,** 33–40 (1996).
40. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
41. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
42. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35,** D61–65 (2007).
43. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15,** 1034–1050 (2005).

Acknowledgements

We gratefully thank Dr. John L. Rinn for assistance with the RIP-chip data, as well as Dr. Stuart H. Orkin, Dr. Weimou Zheng and Dr. Xiaohua Shen for helpful discussion and instruction. We also thank the anonymous reviewers for their valuable suggestions. This work was supported by the Shanghai Pujiang Talent Program (14PJ1410000 to Z.S.) and the "100-Talent Program" (Y516C11851 to Z.S.) of the Chinese Academy of Science. This work was also supported by NIH R01HL119099 (to G.-C.Y.).

Author Contributions

G.-C.Y. and Z.S. conceived the study; S.T. implemented the algorithms and performed the analysis; Z.S. supervised the analysis; S.T., G.-C.Y. and Z.S. wrote the manuscript; all authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Tu, S. *et al.* The PRC2-binding long non-coding RNAs in human and mouse genomes are associated with predictive sequence features. *Sci. Rep.* **7**, 41669; doi: 10.1038/srep41669 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The PRC2-binding long non-coding RNAs in human and mouse genomes are associated with predictive sequence features

Shiqi Tu[1, 2], Guo-Cheng Yuan[3, 4, 5], Zhen Shao[1,*]

[1]Chinese Academy of Sciences Key Laboratory of Computational Biology, Collaborative Innovation Center for Genetics and Developmental Biology, Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

[3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA.

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.

[5]Harvard Stem Cell Institute, Cambridge, MA 02138, USA

Supplementary text

Analysis of the sequence features associated with transcription factor CTCF's DNA binding

To dissect whether the sequence features identified by our sequence composition analysis pipeline, as well as the consecutively favored paths (CFPs) formed by them with significant preference, have any connection with the sequence specificity of protein binding, we applied our pipeline to analyze the DNA sequences bound by transcription factor (TF) CTCF in human H1 embryonic stem cells. We applied MACS to the ChIP-Seq data of CTCF in H1 cell line generated by ENCODE project[1] , and picked 1000 top CTCF non-promoter binding peaks from MACS output. Then, we extracted the 1-kb long DNA sequences around the summit of these peaks, and defined them as CTCF-positive sequences. Here, we removed CTCF promoter peaks that fall within 1.5kb from the transcription start site of any RefSeq annotated gene, as usually a large fraction of CTCF binding sites in mammalian cells are located in gene promoters[2] and the sequence composition of gene promoters is quite different from the other parts in genome. On the other hand, we randomly selected the same number of 1-kb sequences from the intergenic regions of human genome that don't overlap with any CTCF peak, and defined them as CTCF-negative sequences. Next, we applied the sequence composition analysis pipeline to these two groups of sequences, and identified 558/389 transitions that are significantly favored/disfavored by CTCF-positive sequences compared to the negative ones with P-

value lower than 1E-15 (by 2-tailed Welch t-test). Surprisingly, a very large fraction of CTCF-favored transitions are located in CFPs and, especially, 52 CFPs formed by these transitions are of full length, which means they start from the root of the quad-tree and end at the 6th level (Supplementary Fig. S1C).

A direct by-eye comparison between the DNA binding motif of CTCF obtained from JASPAR database[3] and a representative full-length CFP suggests that the 6-mer formed by all the 6 consecutive transitions on this CFP can be well aligned with a highly informative part of CTCF's binding motif (Supplementary Fig. S1C). To systematically investigate this associations, we defined a motif score for each 6-mer to measure its similarity with the given motif as

$$\textbf{Motif score} = \max_{t\in\{0,1,\ldots,L-6\}}\{\sum_{i=1}^{6}\log_{10}(P(t+i,S_i))\} \quad (1)$$

Here L is the length of motif, *P(i,j)* is the element of its position frequency matrix at row *i* and column *j*, and S_i is the numerical expression of the i^{th} nucleotide of this 6-mer (S_i=1, 2, 3, 4 corresponds to nucleotide A, C, G, T, respectively). Additionally, if a 6-mer's motif score is lower than that of its reverse complementary, the higher score will be assigned to both of them. Then, we divided all the 4^6 possible full-length paths on the quad-tree into 4 groups, based on the length of the longest CFP on each path (here isolated CTCF-favored transitions were taken as CFPs of length 1), and calculated the motif score of the 6-mer corresponding to each full-length path. Interestingly, the CTCF motif scores of these K-mers obviously correlate with the maximum CFP lengths of these paths, and the 52 full-length CFPs have the highest motif sores among the 4 groups (Supplementary Fig. S1D), suggesting the preference of those favored transitions to be connected with each other across adjacent tree levels is not a trivial observation and is intrinsically connected with the sequence specificity mediating protein bindings.

Building empirical classification models of PRC2-binding lncRNAs

To investigate whether PRC2-binding lncRNAs can be predicted by their sequence features in a simple way, as well as to infer the improvement of prediction accuracy achieved by the model fitting process using BART, we additionally devised an empirical classification model to predict human PRC2-binding lncRNAs without involving BART to perform sophisticated model fitting, which is named as "reward-and-punish" model here. We first build this model using the fully blind method as described in main text. Again, all human PRC2-positive and PRC2-negative lncRNAs are divided into 10 subgroups. At each step, one subgroup is selected as the testing set, leaving the other 9 subgroups to serve as the training set, and only lncRNAs in the training set can be used to identify PRC2-favored and disfavored transitions as predictors (here the same cutoff P<0.05 is used). Then, for each lncRNA in the testing set, the frequency of observing each PRC2-favored/disfavored transition in its sequence,

saying $F_{testing}$, is compared with the frequencies of this transition in the sequences of training lncRNAs, and the reward and punishment to this lncRNA are determined based on following rules:

1) For each PRC2-favored transition, check whether $F_{testing}>\mu_{negative}+b*\sigma_{negative}$ is true ($\mu_{negative}$ and $\sigma_{negative}$ is the mean and standard deviation of the frequencies of this transition of PRC2-negative lncRNAs in the training set, respectively, and b is a constant), and give the testing lncRNA a reward 1/M if the answer is yes (M is the number of predictors identified from the training lncRNAs at this step); meanwhile, check whether $F_{testing}<\mu_{positive}-b*\sigma_{positive}$ is true ($\mu_{positive}$ and $\sigma_{positive}$ is the mean and standard deviation of the frequencies of this transition of PRC2-positive lncRNAs in the training set, respectively), and give the lncRNA a punishment 1/M if the answer is yes.

2) For each PRC2-disfavored transition, check whether $F_{testing}<\mu_{negative}-b*\sigma_{negative}$ is true, and give the testing lncRNA a reward 1/M if the answer is yes; meanwhile, check whether $F_{testing}>\mu_{positive}+b*\sigma_{positive}$ is true, and give the lncRNA a punishment 1/M if the answer is yes.

Finally, the prediction score of each testing lncRNA is calculated as the sum of all the reward got by it minus the sum of all the punishment got by it. In this way, we found human PRC2-positive lncRNAs can be better distinguished from PRC2-negative ones using the empirical model with b lying between 0 and 0.5 (Supplementary Fig. S2B), and the highest AUC value was close to 0.60, which is still lower than that got by the prediction model fitted by BART using the same fully blind method. Interestingly, the empirical models with b>1 (which means the testing lncRNAs have low likelihood to get reward and punishment) got AUC values close to or lower than 0.5 (Supplementary Fig. S2B), implying prediction of PRC2-binding lncRNAs based on their sequence composition can be better achieved by combining a number of weak classifiers.

On the other hand, we also tried building an empirical classification model using the non-blind CV process as described in main text. Here, we directly chose the 240 and 87 human PRC2-favored and disfavored transitions identified in main text as predictors, and divided all PRC2-positive and PRC2-nagative lncRNAs into 10 subgroups to perform a 10-fold cross-validation. Next, we adopted the same rules to calculate the reward and punishment for each testing lncRNA. By this means, the empirical model exhibited a similar performance (AUC~0.8-0.85 by models with b lying between 0 and 2, Supplementary Fig. S2C) to that achieved by the prediction model fitted by BART using the non-blind CV method.

Evaluate the *P*-value cutoff used for predictor selection

To see whether the prediction can be made with a smaller number of predictors compared with the original prediction model shown in main text, we tried building a prediction model using only the human PRC2-favored and disfavored transitions with *P*-value<0.01 as predictors. By this means, 66 transitions in total were selected, and the prediction model exhibited a similar (AUC=0.81 using the

non-blind CV method) or reduced accuracy (AUC=0.61 using the fully blind method). Considering the second method is more stringent, we speculate that *P*-value<0.05 can serve as a more reasonable cutoff for predictor selection than *P*-value<0.01. Besides, this finding also implies that the relatively low accuracy of the prediction model built by the fully blind method may not be simply explained by over-fitting, which is usually indicated by an observation that the model using more predictors tends to show a lower accuracy, as evaluated by the testing set[4].

Explanation of using the whole gene body of lncRNAs for sequence composition analysis

In this study, we extracted the sequence of the entire gene body of each lncRNA to perform sequence composition analysis. This is mainly due to two reasons. First, several recent studies published by different labs suggested that EZH2/PRC2 may directly interact with the nascent transcripts of many genes [5-7], and a very large fraction of the interaction sites identified from corresponding CLIP-seq experiments were found to be located in regions annotated as introns[6,8,9]. Second, the lncRNAs used here were initially profiled using custom-design tiling arrays, which are known to have a lower resolution and higher noise levels than sequencing based platforms. For example, we collected the annotation of 4859 lncRNA exons from Khalil *et al.*[10] (in total they cover about 16.24% of the gene body of these lncRNAs), and found quite a number of the 488 human lncRNAs used in this study have no exon annotations (Fig. 2B). Moreover, these exon annotations show clear differences from other gene annotations, such as those provided by RefSeq (Supplementary Fig. S2G-H). On this account, we finally decided to use the entire gene body of lncRNAs to perform sequence composition analysis, in order to reduce the risk of missing regions potentially important for PRC2-lncRNA interactions.

On the other hand, we also found that the difficulty posed by using the entire gene body for sequence analysis can be largely overcome by focusing the study on sequence features associated with PRC2-lncRNA interactions. Here we use the PRC2-favored and disfavored fragments identified by us as examples. 16.9% of the PRC2-disfavored fragments overlap with the exons obtained from Khalil *et al*., which is close to the fraction of the lncRNA gene bodies covered by these exons, and this fraction for PRC2-favored ones is as high as 26.2%. Of note, we have shown ~30% of the PRC2-favored fragments overlap with annotated conserved elements, indicating a considerable part of the important building blocks of these lncRNAs may not be covered by these 4859 exons. Beside the lncRNA shown in Fig. 2B, we additionally selected two representative human PRC2-binding lncRNAs to illustrate the power of sequence composition analysis. For the well-studied PRC2-binding lncRNA XIST, its PRC2-favored fragment is at the second exon (based on RefSeq gene annotations, Supplementary Fig. S2G), which is not in the exon list obtained from Khalil *et al.* On the other hand, we found the PRC2-favored fragment of lncRNA chr13:90799274-90818300 is not covered by either RefSeq annotated gene exons (the closest RefSeq annotated lncRNA is MIR17HG) or the exons got from Khalil *et al.* (Supplementary Fig. S2H). However, by incorporating a recently published RIP-seq

dataset of EZH2 and SUZ12 in K562 cells, we found the transcript generated from this fragment can strongly interact with these two PRC2 core subunits (Supplementary Fig. S2H), suggesting our analysis can largely recover the missing building blocks important for the function of PRC2-binding lncRNAs.

Definition of mouse PRC2-binding lncRNAs

We collected 1666 putative mouse lncRNAs from Guttman *et al.*[11], which were discovered by using a similar method to the human lncRNAs studied in Khalil *et al.*[10], and lifted-over them to mm9 mouse genome assembly. Next, we obtained 8670 PRC2-associated RNA tags from Zhao *et al.*[12], which were generated by RIP-seq experiments against EZH2 in mouse embryonic stem cells (mESCs), and mapped them to these mouse lncRNAs. In total, we got 283 mouse lncRNAs that overlap with PRC2-associated RNA tags. Meanwhile, we also mapped Pol2 ChIP-Seq peaks of mESCs to these mouse lncRNAs, and found 540 mouse lncRNAs contain at least one Pol2 peak, which were taken as lncRNAs expressed in mESCs[13]. Finally, we defined mouse PRC2-positive lncRNAs as the 153 mouse lncRNAs that overlap with both PRC2-associated RNA tags and Pol2 ChIP-Seq peaks, and defined mouse PRC2-negative lncRNAs as the other 540-153=387 mouse lncRNAs that overlap with Pol2 ChIP-Seq peaks but contain no PRC2-associated RNA tag.

To further validate our cross-species prediction, we collected 13,764 putative RNA-contact sites (RCSs) of EZH2 in mESCs from Kaneko *et al* [6] (by personal communication with the authors. The list of RCSs obtained from the GEO webpage of Kaneko *et al* contains only 6784 RCSs, which was said to be generated using slight different and more stringent parameters), which were identified from the corresponding PAR-CLIP-seq data, and mapped them to the 540 putative mESC-expressed lncRNAs. As a result, 144 lncRNAs were found to contain at least one RCS, and 60 of them (41.7%) were previously classified as mouse PRC2-positive lncRNAs based on the RIP-seq data of EZH2 published in Zhao *et al* [12]. More explicitly, 39.2% of mouse PRC2-positive lncRNAs contain at least one RCS of EZH2, while this fraction for mouse PRC2-negative lncRNAs is only 21.7%. Furthermore, mouse PRC2-positive lncRNAs are also more likely to contain more than one RCSs than PRC2-negative ones (Supplementary Fig. 3D). Interestingly, we found for both mouse PRC2-positive and PRC2-negative lncRNAs, lncRNAs with high prediction scores derived from the model trained with human lncRNAs (Fig. 3D and Supplementary Fig. 3F) and also from the model trained with mouse lncRNAs (Supplementary Fig. 3E) are more likely to contain EZH2 RCS than those with low prediction scores, indicating the definition of PRC2-positive ad PRC2-negative lncRNAs only based on single RIP/CLIP-seq dataset may not be quite reliable and inherent sequence patterns can potentially be utilized to correct the false positives and false negatives in them. Then, we defined the 144 lncRNAs that contain at least one RCS as RCS-containing lncRNAs and the other 540-144=396 lncRNAs as RCS-null ones. By using the prediction model trained with human lncRNAs, we found mouse RCS-containing

lncRNAs can be distinguished from RCS-null ones with considerable accuracy (AUC=0.66, Fig 3E). Finally, to combine these two classes of lncRNA labels, we defined the 60 mouse PRC2-positive lncRNAs that also contain EZH2 RCS as high-confidence mouse PRC2-positive lncRNAs, and the 303 mouse PRC2-negative lncRNAs without any EZH2 RCS as high-confidence mouse PRC2-negative lncRNAs. Again, by using the human prediction model to perform cross-species prediction, we found high-confidence mouse PRC2-positive lncRNAs can be distinguished from the high-confidence PRC2-negative ones with clearly better accuracy (AUC=0.72, Fig 3E).

Compare RNA contact sites of EZH2 with PRC2-favored fragments

Here, we tried using the EZH2 RCSs to assess the method we proposed to recognize PRC2-favored fragments in human lncRNAs. Again, a 500bp sliding window was used to scan each mouse RCS-containing lncRNA, and the local consistency score of the sequence fragment in the sliding window was calculated as the sum of the frequencies of all human PRC2-favored transitions in this sequence fragment minus those of all human PRC2-disfavored ones. Then, we selected the fragment with the highest and the lowest consistency score of each mouse RCS-containing lncRNA as its PRC2-favored and disfavored fragment, respectively. Of note, here we used the sequence features of human PRC2-binding lncRNAs to detect PRC2-favored/disfavored fragments for mouse lncRNAs, so that these mouse fragments were identified by using the same sequence model as the human ones. Interestingly, similar to what we observed in human, 37.5% of mouse PRC2-favored fragments (54 of 144) overlap with mouse conserved elements (also annotated by GERP program), and this fraction is significantly higher than that expected by chance (right-tailed *P*-value<1E-06 by the same random permutation test as that shown in Fig. 2D). Meanwhile, only 10.4% of mouse PRC2-disfavored fragments overlap with conserved elements (15 of 144), which is slightly lower than that expected by chance (left-tailed *P*-value=0.03). Again, these findings can support our hypothesis that a considerable proportion of the sequence patterns associated with in vivo PRC2-lncRNA interactions are shared between human and mouse.

On the other side, we only observed a small number of mouse PRC2-favored fragments (9.7%, 14 of 144) directly cover EZH RCS, though it's still significantly higher than expected by chance (right-tailed *P*-value=0.0097 by random permutation test) and also higher than that of mouse PRC2-disfavored fragments (4.2%, 6 of 144, left-tailed *P*-value =0.3931). (If we switched to use those 6784 more stringent RCSs, we got 92 RCS-containing lncRNAs. Among these lncRNAs, 11 of the PRC2-favored fragments directly cover EZH2 RCS, with empirical *P*=0.0013, and only 1 PRC2-disfavored fragment overlaps with RCS.) It should be of note that the main purpose of defining PRC2-favored fragments is to investigate whether the aggregation of the sequence features associated with PRC2-lncRNA interactions at certain genomic regions is potentially linked with the functional importance of these regions, and for simplicity, we only took one 500bp fragment with the highest score of each

PRC2-binding lncRNA as the representative region for downstream analysis (here PRC2-disfavored fragment with the lowest score of each lncRNA is used as control). As we have found PRC2-lncRNA interactions in human and mouse exhibit clear sequence specificities, it's reasonable to speculate the direct contact sites of PRC2 on lncRNAs are very likely to be enriched with sequence features associated with PRC2-lncRNA interactions, but we do not intend to say PRC2-lncRNA interactions should always happen at the 500bp fragment with the highest score of each lncRNA. Following this direction, we calculated the local consistency score around the 400 EZH2 RCSs falling in the 540 mouse lncRNAs used in this study, which was defined to the 500bp sequence fragment centered at the middle of each RCS. Besides, we generated 100 sets of random control regions for these RCSs and at each time a 500bp control region was randomly selected for each RCS from the same lncRNA without overlapping with any identified EZH2 RCS. Interestingly, the 400 EZH2 RCSs falling in lncRNAs in general exhibit obviously higher local consistency scores than the random control regions (Supplementary Fig. 3G), suggesting the local sequences around these RCSs are more enriched with the sequence features associated with PRC2-lncRNA interactions compared to the other parts of the lncRNAs they belong to.

Sequence feature analysis of the RNA contact sites of EZH2 falling in lncRNAs

To directly find the sequence features associated with the RNA contact sites of PRC2 on lncRNAs, we specially analyzed the composition of sequences surrounding the EZH2 RNA contact sites (RCSs) identified from PAR-CLIP-seq data in mESCs. First, EZH2 RCSs located within 500bp from each other were merged together, and we defined EZH2 RCS fragment as the 500bp sequence fragment surrounding the center of each merged RCS. Since this study mainly focuses on interactions between PRC2 and lncRNAs, we only took the 310 RCS fragments falling in the 540 putative ESC-expressed lncRNAs for sequence analysis. As negative controls, we randomly chose the same number of 500bp sequence fragments from the high-confidence mouse PRC2-negative lncRNAs defined by us. By using our sequence composition analysis pipeline to compare the 310 EZH2 RCS fragments with the negative control sequences, we identified 190/110 transitions that are significantly favored/disfavored by EZH2 RCS fragments (using *P*-value<0.05 as cutoff). However, this time we found thymine (T) is significantly over-represented in these fragments and cytosine (C) is significantly under-represented (Supplementary Table 7). Specifically, for those transition of order 0-2, all the transitions that are significantly favored by RCS fragments end up with T, and all the transitions are significantly disfavored end up with C, though not all the transitions that end up with T/C are significantly favored/disfavored. As these RCSs were originally detected by taking advantage of the T-to-C transitions produced by PAR-CLIP procedure[6,14], we speculate the enrichment of T and depletion of C in EZH2 RCSs may not be real sequence features of PRC2's RNA contact, and moved to rescale the frequencies of all the transitions of order>0 by the frequency to observe their last nucleotide in each sequence (the 4 order-0 transitions, i.e. the nucleotide frequencies of A,T,G and C, were

excluded from the following analysis). For example, rescaled transition frequency CATG→A of a sequence is calculated as the original transition frequency CATG→A divided by the frequency of observing nucleotide A in this sequence. After rescaling, we repeated the feature selection procedure, and got 181/99 transitions that are significantly favored/disfavored by EZH2 RCS fragments, which were name as RCS-favored and disfavored transitions (Supplementary Table 7).

Again, we used a complete quad-tree of height 6 to visualize the distribution of these selected transitions (Supplementary Fig. 4A). The RCS-favored transitions were also found to strongly prefer to form consecutively favored paths (CFPs, Supplementary Fig. 4B), while the RCS-disfavored ones showed a relatively weak preference to form consecutively disfavored paths (CDPs, Supplementary Fig. 4C). Finally, we applied the same fully blind approach and built a prediction model to distinguish the 310 EZH2 RCS fragments falling in lncRNAs from the corresponding negative control sequences. In this way, the model achieved a clearly lower accuracy (AUC=0.59, Supplementary Fig. 4C) compared to the prediction of RCS-containing lncRNAs.

In addition, we also performed de novo motif discovery around the RNA contact sites of EZH2. Again, we took the 310 500bp EZH2 RCS fragments falling in lncRNAs as well as the corresponding negative control sequences selected from high-confidence mouse PRC2-negative lncRNAs. Then, we randomly split both the RCS fragments and the negative control sequences into two subgroups of equal size, and only used the first subgroup as input for the MEME suite (both MEME and DREME in the suite were used here) to perform de novo motif finding[15]. MEME was ran with the default parameter setting, and the analysis with DREME was carried for two times. In the first time, we used the first subgroup of negative control sequences as input control sequences, and in the second time, we asked DREME to create a set of control sequences by randomly shuffling the RCS sequences while preserving their dimer frequencies. Finally, all the 78 motifs detected by MEME and DREME were collected and applied to the second subgroup of RCS fragments and the corresponding negative control sequences to perform motif enrichment analysis[16]. For each candidate motif, we compared the fraction of RCS fragments that contain this motif to that of the corresponding negative control sequences, and used Fisher's exact test to check whether this motif is significantly over-represented in the RCS fragments compared to the negative control sequences (the first column of *P*-values shown in Supplementary Table 8). Besides, we generated another set of random sequences by shuffling the sequences of RCS fragments, as they have been found to contain significantly more thymine (T) and less cytosine (C) than the negative control sequences, and used them as controls to perform a second motif enrichment test with the RCS fragments (the second column of *P*-values shown in Supplementary Table 8). Unfortunately, none of the 78 motifs obtained from de novo motif finding showed significant enrichment in both of the two tests. This finding indicates it could be quite difficult to find the RNA sequence motifs that can be directly recognized by EZH2, which is consistent

with the hypothesis suggested by several recent studies that the mechanisms governing PRC2's RNA binding may be very complicated[17,18].

Decomposition of long sequences into K-mers and prediction of PRC2-binding lncRNAs based on K-mer frequencies

To compare with transition based sequence decomposition, traditional K-mer based method was also applied to analyze the sequence composition of human lncRNAs. In this approach, the composition of a sequence is described by the frequencies of observing all the possible K-mers of length from 1 to m+1 (here m=5, resulting in a total number of 5460 different K-mers, which can also be visualized by a complete quad-tree of height 6). Taking the 5-mer CATGA as an example, the corresponding K-mer frequency was defined as the frequency of observing it in the given sequence

$$\mathbf{K-mer\ frequency\ (CATGA)} = \frac{N(CATGA)}{N-K+1} \qquad (2)$$

Here, *N(CATGA)* is the times of observing 5-mer CATGA in this sequence, and N is the length of this sequence. Finally, K-mers significantly under- or over-represented in the sequences of PRC2-positive lincRNAs compared to PRC2-negative ones were identified by using the same statistical test and *P*-value cutoff as that used for transition frequencies.

In main text, we have shown the prediction models based on transition frequencies using the fully blind method showed superior accuracies in predicting both human and mouse PRC2-binding lncRNAs (Fig. 4A-B), especially for the extremely long lncRNAs. Additionally, we also checked the performance of the prediction models built by the non-blind CV method on the moderately long and extremely long subgroup of lncRNAs separately, and again got similar results (Supplementary Fig. S5A-B), indicating this finding is not specific to the method we chose to build the prediction model. Moreover, to control the impact of model complexity, we chose to build prediction models using a fixed number of top K-mer or transition frequencies (ranked by the P-value of Welch's t-test) as predictors at each time. Still, a consistent improvement of prediction accuracy on the extremely long subgroup of human lncRNAs was observed for the models based on transition frequencies, as compared to the K-mer based models using the same number of predictors (Supplementary Fig. S5C, here the prediction models were built by the non-blind CV method).

In the following part, we use a highly idealized theoretical model to illustrate the difference between transition and K-mer based feature selections. Suppose a 6-mer, saying CAGTCT, is the sequence feature mediating the interactions between a protein and DNA/RNA sequences. In positive sequences, the frequency of observing this K-mer can be expected as

$$Frequency_{positive}^{K-mer} = \frac{M + N\prod_{i=1}^{6} P(S_i)}{N} = \prod_{i=1}^{6} P(S_i) * \left(1 + \frac{M}{N\prod_{i=1}^{6} P(S_i)}\right) \qquad (3)$$

. Here we ignored the difference between *N* and *N-K+1* as typically *K*<<*N*, and $P(S_i)$ is the frequency of observing the *i*-th nucleotide of this 6-mer, and *M* stands for the extra number of occurrences of the 6-mer in each positive sequence as needed for the interaction, which is set to be 0 for negative sequences. Thus, its frequency in negative sequences can be expected as

$$\boldsymbol{Frequency_{negative}^{K-mer} = \frac{N\prod_{i=1}^{6} P(S_i)}{N} = \prod_{i=1}^{6} P(S_i)} \quad (4)$$

. For the transition-based method, we use the last transition $CGCGC \rightarrow A$ of this 6-mer as an example. Its frequency in positive sequences can be expected as

$$\boldsymbol{Frequency_{positive}^{Transition} = \frac{M + N\prod_{i=1}^{6} P(S_i)}{M + N\prod_{i=1}^{5} P(S_i)} = P(S_6) * \left(1 + \frac{M(1 - P(S_6))}{M * P(S_6) + N\prod_{i=1}^{6} P(S_i)}\right)} \quad (5)$$

, and its frequency in negative sequences can be expected as

$$\boldsymbol{Frequency_{negative}^{Transition} = \frac{N\prod_{i=1}^{6} P(S_i)}{N\prod_{i=1}^{5} P(S_i)} = P(S_6)} \quad (6)$$

. For both of the K-mer and the transition based methods, it's easy to find the frequencies in positive and negative sequences can be expressed as *C**(1+Δ) and *C**1, respectively. Here C is a constant between two sequence groups. Moreover, we simplify the nucleotide frequencies of A, C, G and T as 1/4. Then, $\Delta_{k\text{-mer}}$ and $\Delta_{transition}$ can be expressed as

$$\boldsymbol{\Delta_{K-mer} = \frac{M * 4^6}{N}} \quad (7)$$

and

$$\boldsymbol{\Delta_{Transition} = \frac{M * 4^6}{(N + M * 4^5)} * \frac{3}{4}} \quad (8)$$

This idealized model suggests, when *N* is a constant for all sequences, the difference in K-mer frequencies between positive and negative sequences should be greater than that in transition frequencies.

For our analysis, however, the problem is that *N* is quite large and varies dramatically across different lncRNAs (10^0-10^2 Kb). Specifically, we put the length of 261 human PRC2-binding lncRNAs into above formulas and used one-sample Students' t-test to separately compare $\Delta_{k\text{-mer}}$ and $\Delta_{transition}$ with zero (here *M* is set to be 1 to mimic the most extreme situation). Interestingly, the t-statistics for $\Delta_{k\text{-mer}}$ and $\Delta_{transition}$ equals to 16.9 and 23.7, respectively, which is opposite from that expected for the case with constant N. Besides, protein-DNA interactions usually happen on small sequence fragments of length 10-30 base pairs, and they are often associated with clear sequence specificities[3].

However, the interaction between a protein complex and a long RNA molecular may take place at multiple loci on this RNA transcript[6], as they usually have highly complicated high-order structures, and the sequence features associated with such interactions might be quite diffusive on lncRNAs, which suggests the sequence composition analysis based on transition frequencies can also be a plausible way for lncRNAs.

Prediction of human PRC2-binding lncRNAs based on the composition of sequences near the transcription start site of lncRNA genes

It has been revealed that PRC2 may preferentially interact with lncRNAs at regions close to their 5' ends[19,20]. Inspired by these findings, we specifically investigated the sequences close to the transcription start site (TSS) of human PRC2-binding lncRNAs. First, we define the TSS region of each lncRNA as the region spanning from the TSS to 1kb downstream of the TSS (Supplementary Fig. S6A). Then, we applied our sequence composition analysis pipeline to compare the sequences extracted from the TSS region of human PRC2-positive and PRC2-negative lncRNAs, and identified transitions differentially favored by their TSS regions using the same criteria as the analysis for the whole gene bodies, which were then used to build a prediction model of PRC2-binding lncRNAs using the fully blind method. In this way, the prediction model using transitions with *P*-value<0.05 as predictors showed a clearly reduced accuracy (AUC=0.57, typically ~250 predictors are used at each cross-validation step), compared with the model using only transitions with *P*-value<0.01 as predictors (AUC=0.61, typically ~60 predictors are used), indicating the former one may suffer from overfitting[4]. As a comparison, we also defined the transcription end site (TES) region of each lncRNA as the region spanning from 1kb upstream of the TES to the TES (Supplementary Fig. S6A), and applied the same analysis to the sequences extracted from the TES region of PRC2-positive and PRC2-negative lncRNAs. Interestingly, the prediction model using 56 transitions differentially favored by their TES regions (with P-value<0.01) as predictors achieved a slightly lower accuracy (AUC=0.59) than the TSS region based model, implying the sequence composition close to the TSS of each lncRNA may be more predictive of whether this lncRNA can potentially interact with PRC2 than the other regions. To test this hypothesis, we randomly selected 1000 relative positions from [0.1, 0.9] (here the relative position 0 and 1 represent the TSS and TES of each lncRNA, respectively, and the relative positions out of [0.1, 0.9] were excluded to make the selected positions far enough from TES and TSS). For each of these selected positions, we extracted the 1kb sequence around the corresponding position of each human lncRNA, and built a prediction model based on the composition of these sequences to predict whether the sequence is from a PRC2-positive or negative lncRNA. Interestingly, none of the prediction models based on these 1000 randomly selected positions achieved a better accuracy than that got by the TSS region based prediction model, while 29 of them achieved higher AUC values than that got by the TES region based prediction model (Supplementary Fig. S6B). On the other hand, quite a number of these prediction models selected more sequence features as predictors than the

TSS region based model (Supplementary Fig. S6C). This finding suggests the sequence close to the TSS of lncRNAs may be better used to predict whether this lncRNA can potentially interact with PRC2 than the other parts. Recent studies have suggested that some lncRNAs can interact with multiple chromatin modifying complexes at different regions, implying a potential role of these lncRNAs to serve as molecular scaffold to link different functional modules together[20,21]. Then, it's reasonable to speculate that different parts of these lncRNAs may be associated with different functions and, thus, have different sequence patterns. However, due to the resolution of the lncRNA annotations used in this study, it's still difficult to reliably address this question.

Supplementary Figure Legends

Supplementary Figure S1. Consecutively disfavored paths (CDPs) formed by human PRC2-disfavored transitions and the sequence features associated with transcription factor CTCF's DNA binding. (A) A branch of the complete quad-tree that starts from level 4 and contains a CDP of length 2: CCAC➔C➔A. (B) Summary statistics of the CDPs observed in Fig. 1D, which indicate the human PRC2-disfavored transitions have relatively weak preference to connect with each other and form CDPs. (C) A representative full-length CFP formed by 6 consecutive transitions that are significantly favored by the CTCF-binding DNA sequences identified from human H1 cells compared to the sequences not bound by it. These 6 consecutive transitions in together form a 6-mer CCCCCT that can be well matched with the binding motif of CTCF obtained from JASPAR database. (D) Box plot of the motif scores of all the possible full-length paths extracted from the quad-tree, which start from the root and end at each level-6 node. Here the motif score of a path was defined to measure the similarity between CTCF's binding motif and the K-mer formed by all the 6 consecutive transitions on this path, and the full-length paths are grouped by the length of the longest CFP found on them.

Supplementary Figure S2. Analysis of the PRC2-favored and disfavored fragments identified from human PRC2-binding lncRNAs. (A) AUC value of each of the 10 lncRNA subgroups as the testing set. The prediction scores were derived from the prediction model build by the non-blind CV (red stars) or the fully blind (blue stars) method. (B-C) AUC value achieved by the empirical classification model build by the fully blind (B) or the non-blind CV (C) method with different b. (D) Boxplot of EZH2 and SUZ12 fRIP-seq signal intensities at PRC2-favored and disfavored fragments. Here the signal intensity of each fragment was measured as the number of reads mapped to it per million total reads (RPM) and then averaged over 2 (for EZH2) or 3 (for SUZ12) biological replicates. (E) Distribution of the P-values got by using pairwise Students' t-test to compare the average PhastCons conservation scores between the PRC2-favored and disfavored fragments identified from each of the 1000 sets of randomized PRC2-positive lncRNAs. Here 1.7% of the randomized lncRNA

sets achieved *P*-values lower than 1.7E-04, which is got by using the same test to compare the average conservation scores between the PRC2-favored and disfavored fragments identified from the original human PRC2-positive lncRNAs as shown in Fig. 2C, and this fraction is used as an empirical estimate of the false positive rate (FPR) for the test shown in Fig. 2C. (F) List of three tests performed in Fig. 2C-D, together with their empirical P-values and FPR values, which were estimated by using the same method as shown in (E). (G-H) Two representative PRC2-positive lncRNA loci, including the well-known PRC2-binding lncRNA XIST (G). Their PRC2-favored and disfavored fragments are indicated by the red and blue bars, respectively. (I-J) Distribution of PRC2-favored (I) and disfavored (J) fragments in human PRC2-positive lncRNAs. Here relative position 0 and 1 correspond to the 5' and 3' end of each lncRNA, respectively.

Supplementary Figure S3. Overlap between human and mouse PRC2-favored and disfavored transitions. (A) Venn diagram to show the overlap between human and mouse PRC2-favored and disfavored transitions. (B) Scatter plot of the prediction score of each mouse lncRNA obtained from the mouse prediction model trained with mouse lncRNAs using the fully blind method and the prediction score got from cross-species prediction using the human prediction model trained with human lncRNAs. (C) AUC values of three human prediction models in predicting mouse PRC2-binding lncRNAs. Here the models were trained with human lncRNAs using different groups of transitions as predictors, and then applied to mouse lncRNAs. (D) Fraction of mouse PRC2-positive and PRC2-negative lncRNAs that contain exactly one and more than one RCSs of EZH2 identified from the PAR-CLIP-seq data. (E) Fraction of mouse PRC2-positive and PRC2-negative lncRNAs that contain EZH2 RCSs. Here each group of lncRNAs are split into two subgroups of equal size by the median of their prediction scores derived from the prediction model trained with mouse PRC2-positive and PRC2-negative lncRNAs using the fully blind method, and the P-values were calculated by right-tailed Fisher's exact test to measure whether the subgroup of lncRNAs with high prediction scores are significantly more likely to contain EZH2 RCS compared to the subgroup with low prediction scores. (F) The same analysis as that shown in Fig. 3D. Here we switched to use the 6784 RCSs of EZH2 obtained from the GEO webpage of Kaneko *et al*, which were suggested to be generated with more stringent parameters. (G) Cumulative distribution of the local consistency scores of 400 EZH2 RCSs falling in the 540 mouse lncRNAs used in this study (red curve). Here the local consistency score of each RCS was defined as the sum of frequencies of all human PRC2-favored transitions in the 500bp sequence fragment centered at the middle of this RCS minus those of all human PRC2-disfavored ones in this sequence fragment, and each blue dash line represents the cumulative distribution of the local consistency scores of the corresponding 500bp random control regions selected from the same lncRNAs. In total 100 sets of random control regions were generated.

Supplementary Figure S4. Sequence features of EZH2 RNA contact sites (RCSs) falling in lncRNAs. (A) The complete quad-tree of height 6 constituted by all possible transitions of order 0-5 (placed on level 1-6 accordingly). Left panel shows a building block of the quad-tree, which comprises of 4 transitions with the same prefix. Each line represents a transition and the color indicates whether the transition is significantly favored or disfavored by the sequences surrounding the EZH2 RCSs falling in lncRNAs. (B) Summary statistics of the CFPs observed from the quad-tree shown in (A). The empirical *P*-value is estimated by permutation test and suggests the transitions that are significantly favored by EZH2 RCSs strongly prefer to connect with each other and form CFPs. (C) Summary statistics of the CDPs observed from the quad-tree shown in (A). The *P*-value suggests the transitions that are significantly disfavored by EZH2 RCSs have a weak preference form CDPs. (D) ROC curve and corresponding AUC value of the prediction model built by the fully blind method in predicting EZH2 RCS fragments falling in lncRNAs.

Supplementary Figure S5. Compare the performance of prediction models based on K-mer and transition frequencies. (A-B) AUC value of the prediction models based on transition (red bars) or K-mer (blue bars) frequencies, which were trained and tested by the human (A) and mouse (B) lncRNAs, respectively. Here the prediction models were built by the non-blind CV method, and all the human/mouse PRC2-positive and PRC2-negative lncRNA were further divided into two subgroups of equal size by their length, termed as the moderately long and the extremely long subgroup, to access the accuracy of these models on lncRNAs of different length. (C) AUC value of the human prediction models on the moderately long and the extremely long subgroup of human lncRNAs, respectively. Here the models were built by the non-blind CV method, using a fixed number of top transition or K-mer frequencies (ranked by the P-value of Welch's t-test) as predictors. (D) A branch of K-mers cut from the quad-tree, which is constituted by all the possible K-mers of length 1-6, from the same position as the branch shown in Supplementary Figure 1C. Here the edge color indicates whether the K-mer is significantly over-represented (red) or under-represented (green) in CTCF-binding DNA sequences compared to the non-binding sequences.

Supplementary Figure S6. Prediction of human PRC2-binding lncRNAs based on the sequence in the TSS and TES region of each lncRNA. (A) Definition of the TSS and TES region of each lncRNA. (B) AUC values of the prediction models based on the sequences at the TSS/TES region of human lncRNAs (red/blue stars, respectively), as well as the prediction models based on sequences extracted from 1000 randomly chosen positions of the lncRNAs (blue dots shown in the middle), which are used to estimate the empirical *P*-values for the performance of the TSS and TES region based prediction model. (C) Number of sequence features identified from the sequences of TSS and TES

regions (red/blue stars, respectively), as well as the sequences extracted from 1000 randomly chosen positions of the lncRNAs (blue dots shown in the middle). All the sequence features are selected by using the same method and *P*-value cutoff (0.01).

Supplementary Table S1. Distribution of human PRC2-disfavored transitions on each level of the quad-tree.

Level	Count	[Q1, Q3]	*P* value
1	0	[0, 0]	1
2	0	[0, 0]	1
3	1	[0, 1]	0.40
4	4	[2, 7]	0.49
5	14	[16, 29]	0.86
6	68	[84, 111]	0.97
All	87	[103, 146]	0.94

Supplementary Table S2. Summary statistics of using two-sample Welch t-test to compare the frequencies of each transition between human PRC2-positive and PRC2-negative lncRNAs.

Supplementary Table S3. List of the human lncRNAs used in this study and their prediction scores obtained from the prediction models built by the non-blind CV and the fully blind method (based on hg18 genome assembly).

Supplementary Table S4. List of the PRC2-favored and disfavored fragment of each human PRC2-binding lncRNA (based on hg18 genome assembly).

Supplementary Table S5. Summary statistics of using two-sample Welch t-test to compare the frequencies of each transition between mouse PRC2-positive and PRC2-negative lncRNAs.

Supplementary Table S6. List of the mouse lncRNAs used in this study and their prediction scores obtained from the prediction models built by the non-blind CV and the fully blind method (based on mm9 genome assembly).

Supplementary Table S7. Summary statistics of using two-sample Welch t-test to compare the frequencies of each transition between the EZH2 RCS fragments falling in lncRNAs and the corresponding negative control sequences.

Supplementary Table S8. List of motifs detected by de novo motif discovery from the EZH2 RCS fragments as well as their relative enrichment in EZH2 RCS fragments compared to two sets of control sequences generated by different methods.

1 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, (2008).

2 Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245, (2007).

3 Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-94, (2004).

4 Yuan, Y., Guo, L., Shen, L. & Liu, J. S. Predicting gene expression from sequence: a reexamination. *PLoS Comput Biol* **3**, e243, (2007).

5 Kaneko, S., Son, J., Bonasio, R., Shen, S. S. & Reinberg, D. Nascent RNA interaction keeps PRC2 activity poised and in check. *Genes Dev* **28**, 1983-1988, (2014).

6 Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20**, 1258-1264, (2013).

7 Kanhere, A. *et al.* Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell* **38**, 675-688, (2010).

8 Guil, S. *et al.* Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat Struct Mol Biol* **19**, 664-670, (2012).

9 Kaneko, S. *et al.* Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Mol Cell* **53**, 290-300, (2014).

10 Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672, (2009).

11 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227, (2009).

12 Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**, 939-953, (2010).

13 Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-1851, (2008).

14 Corcoran, D. L. *et al.* PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* **12**, R79, (2011).

15 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, (2009).

16 Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* **23**, 796-811, (2012).

17 Betancur, J. G. & Tomari, Y. Cryptic RNA-binding by PRC2 components EZH2 and SUZ12. *RNA Biol* **12**, 959-965, (2015).

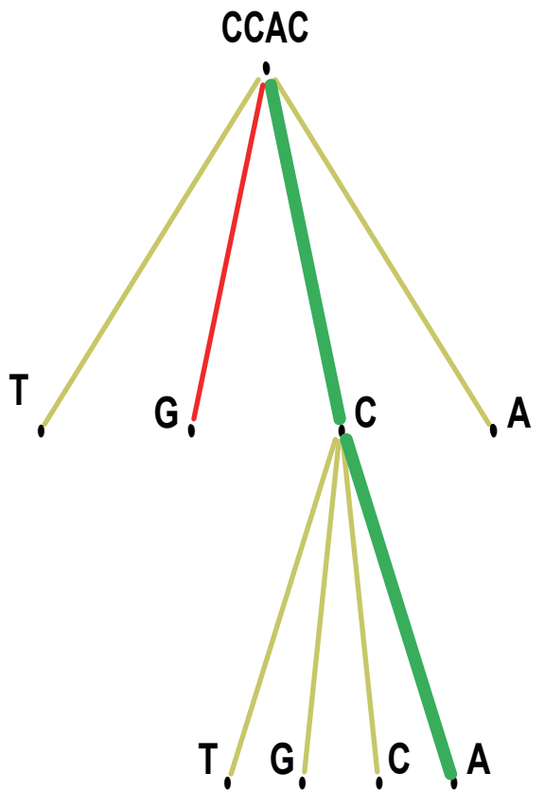
18 Cifuentes-Rojas, C., Hernandez, A. J., Sarma, K. & Lee, J. T. Regulatory interactions between RNA and polycomb repressive complex 2. *Mol Cell* **55**, 171-185, (2014).

19 Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756, (2008).

20 Tsai, M. C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693, (2010).

21 Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20**, 300-307, (2013).

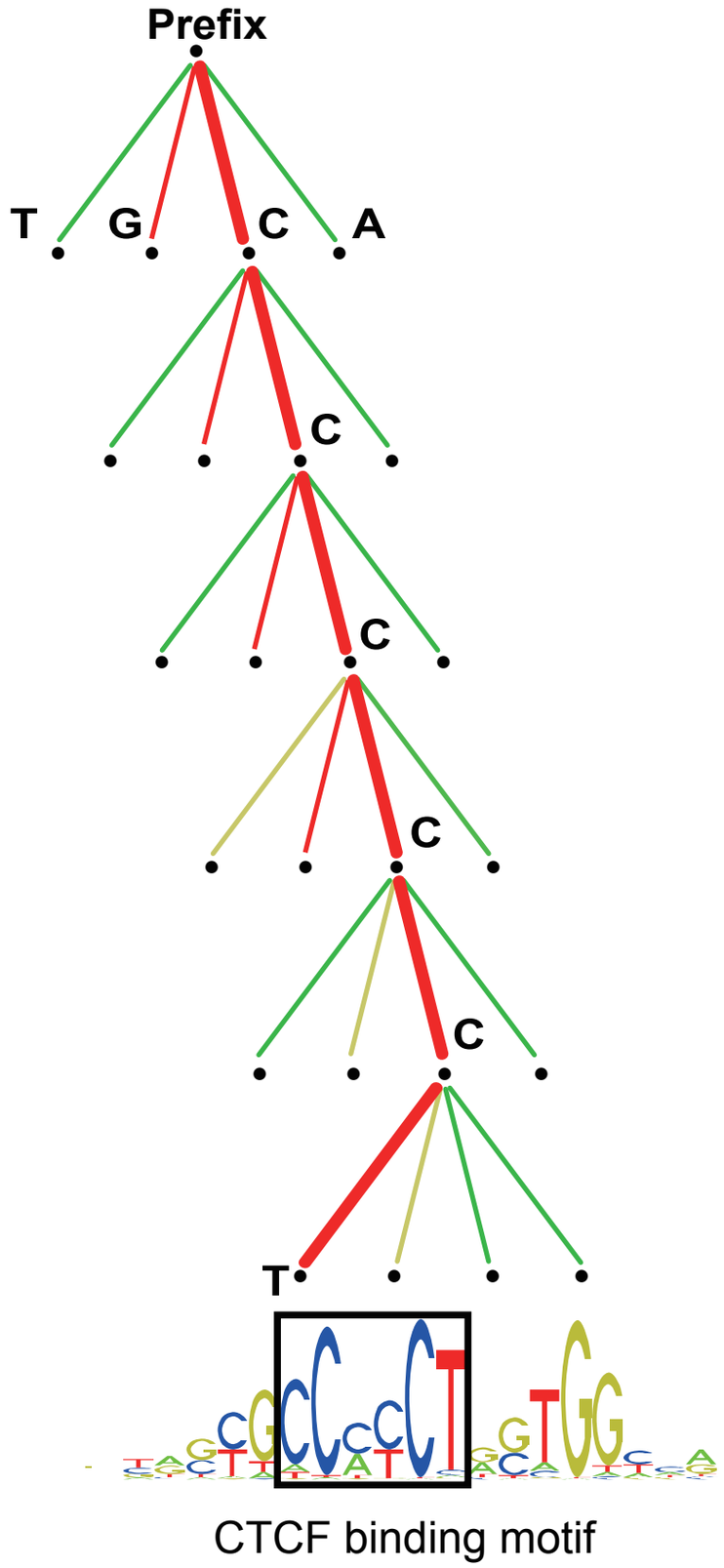
A



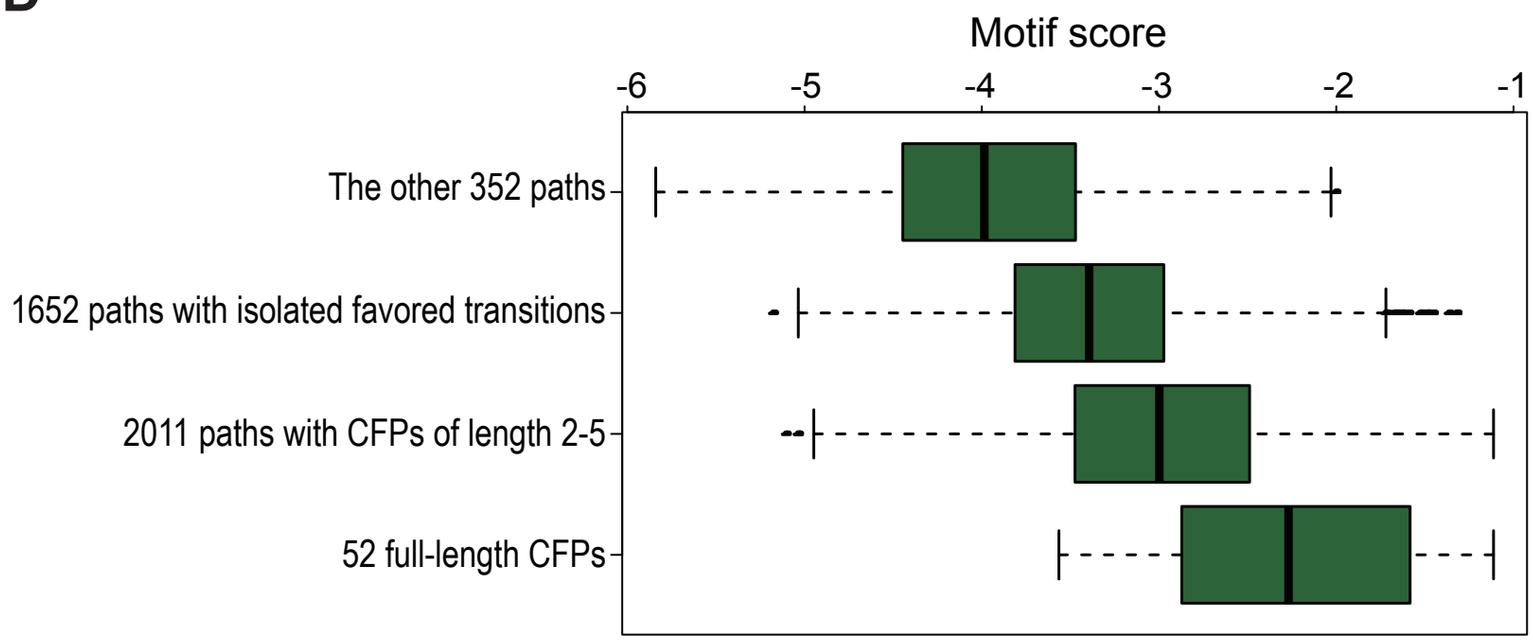
B

PRC2-disfavored transitions	87
Fraction in CDPs	0.092
Average fraction in random permutations	0.028
P value	0.023

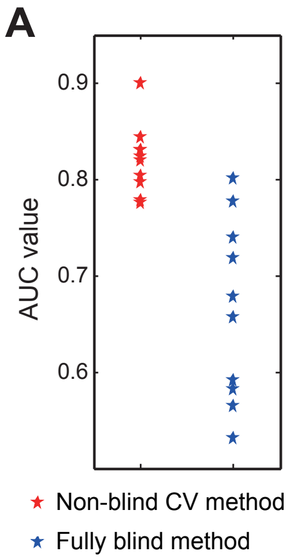
C

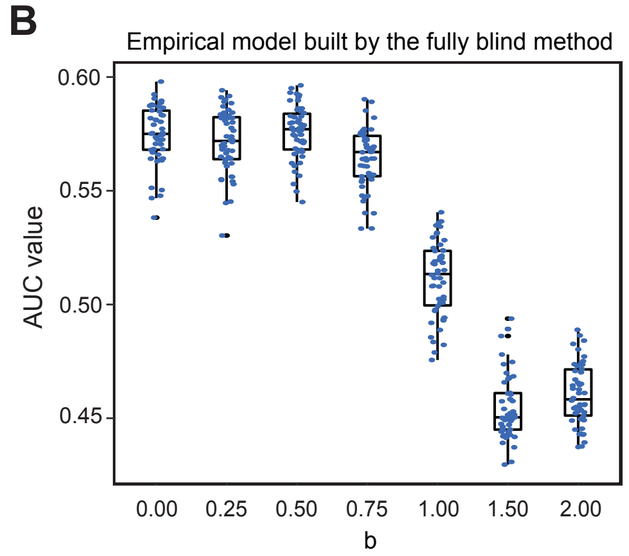


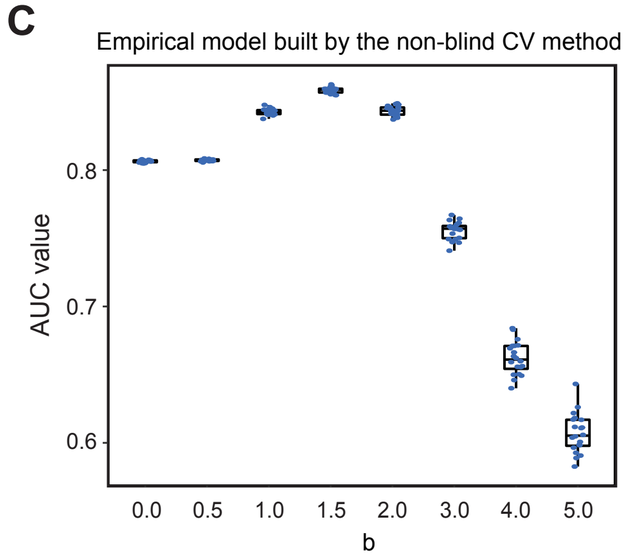
D

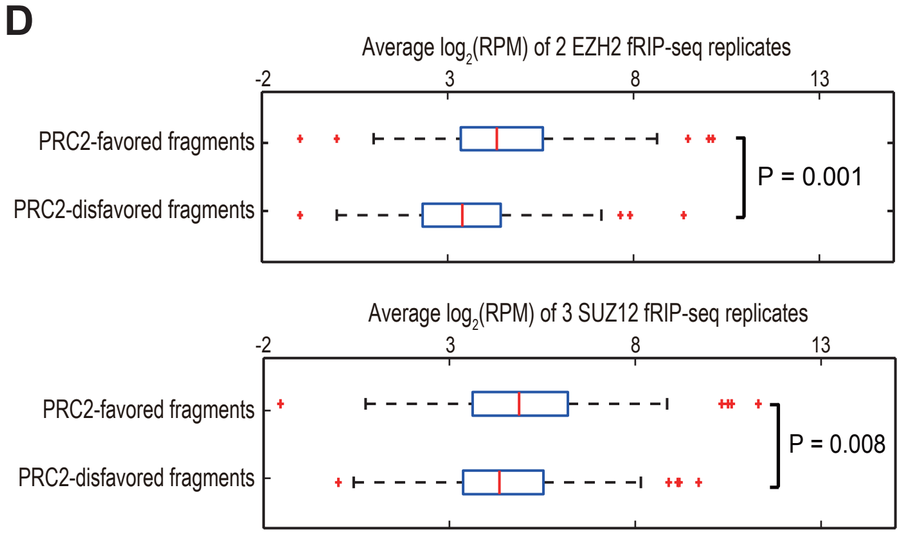


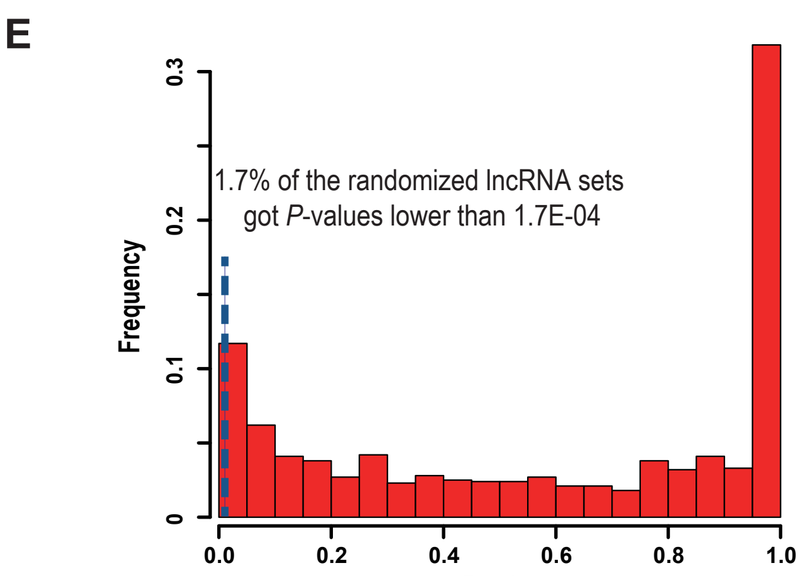
Supplementary Figure S1.





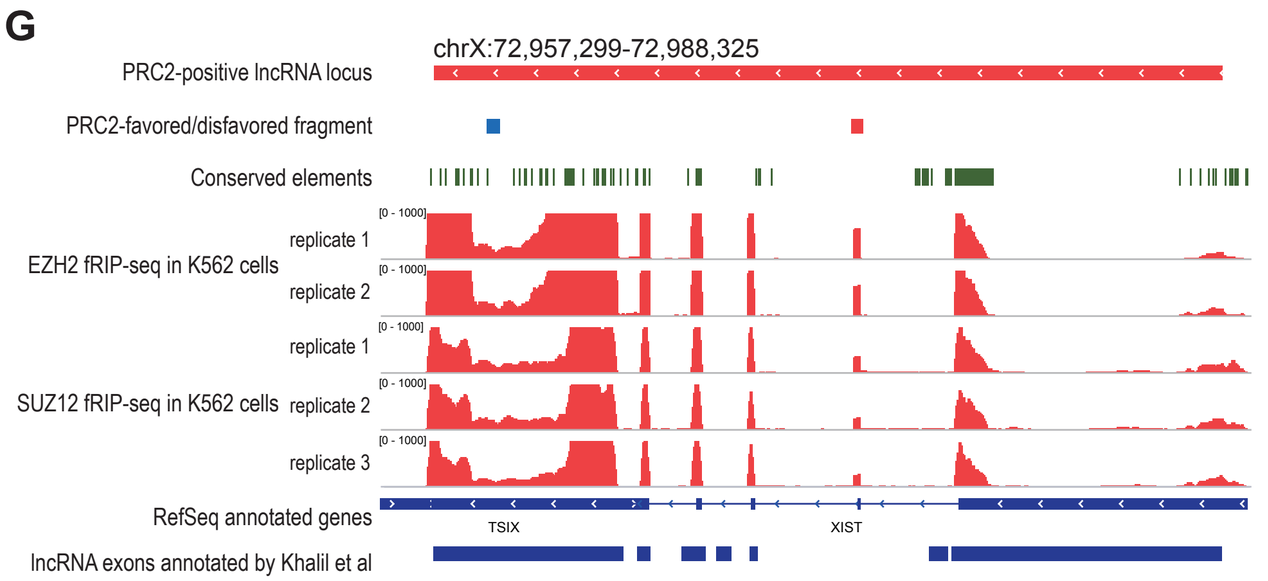


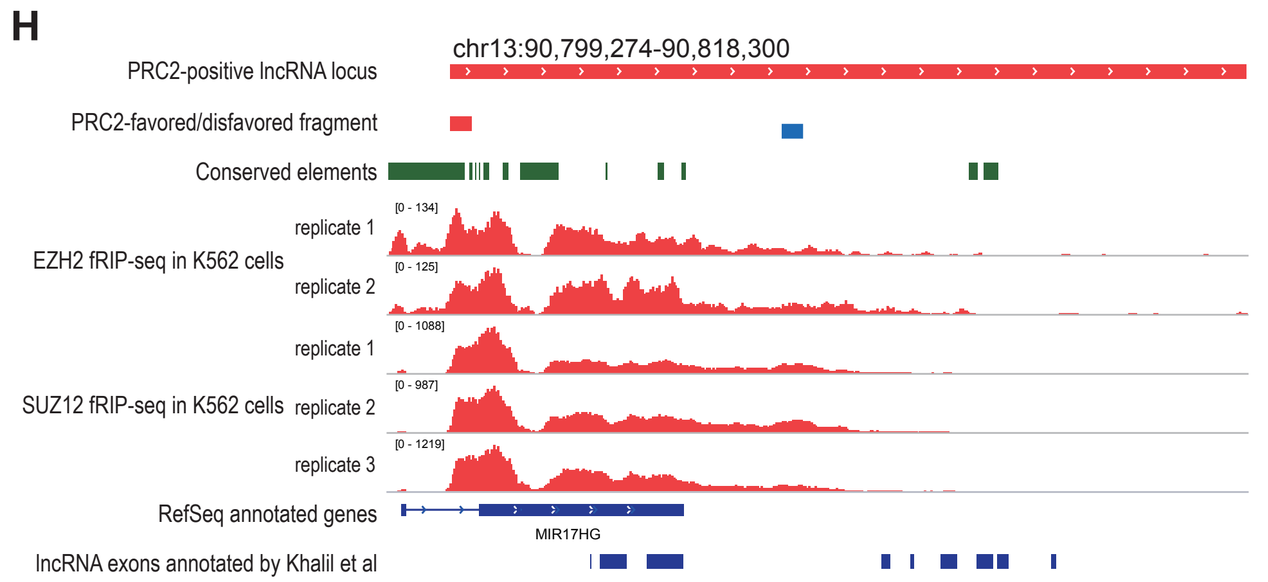


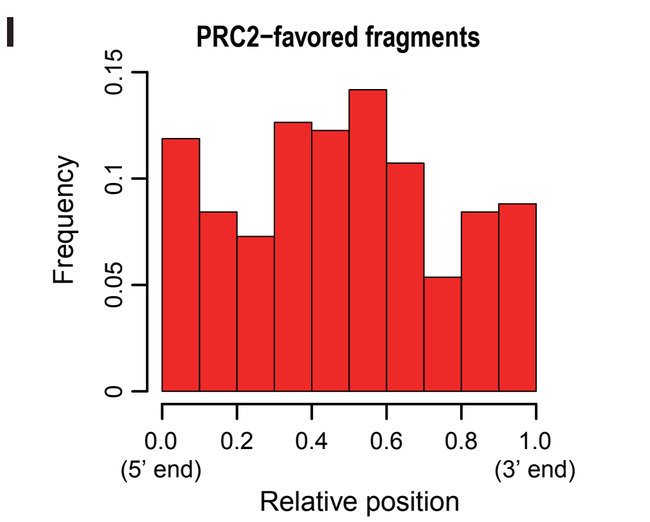


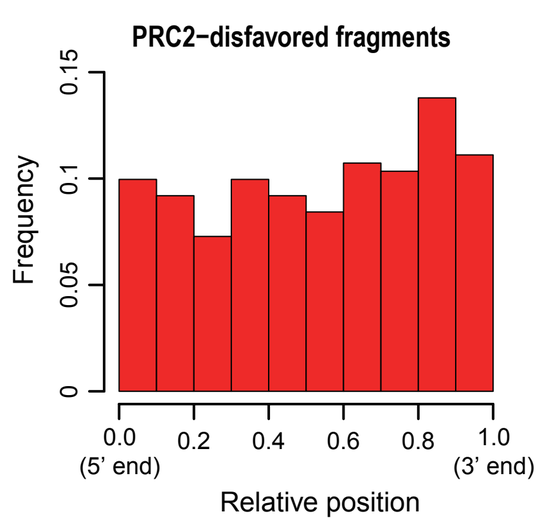
F

Test	Empirical *P*-value	FPR estimated from 1000 sets of randomized lncRNA
PRC2-favored fragments have higher average conservation score than PRC2-disfavored fragments (Figure 2C)	**1.7E-04**	**1.7%**
PRC2-favored fragments are more likely to contain conserved elements than the other parts of lncRNAs (Figure 2D)	**6E-04**	**0.8%**
PRC2-disfavored fragments are less likely to contain conserved elements than the other parts of lncRNAs (Figure 2D)	**2E-04**	**12.2%**

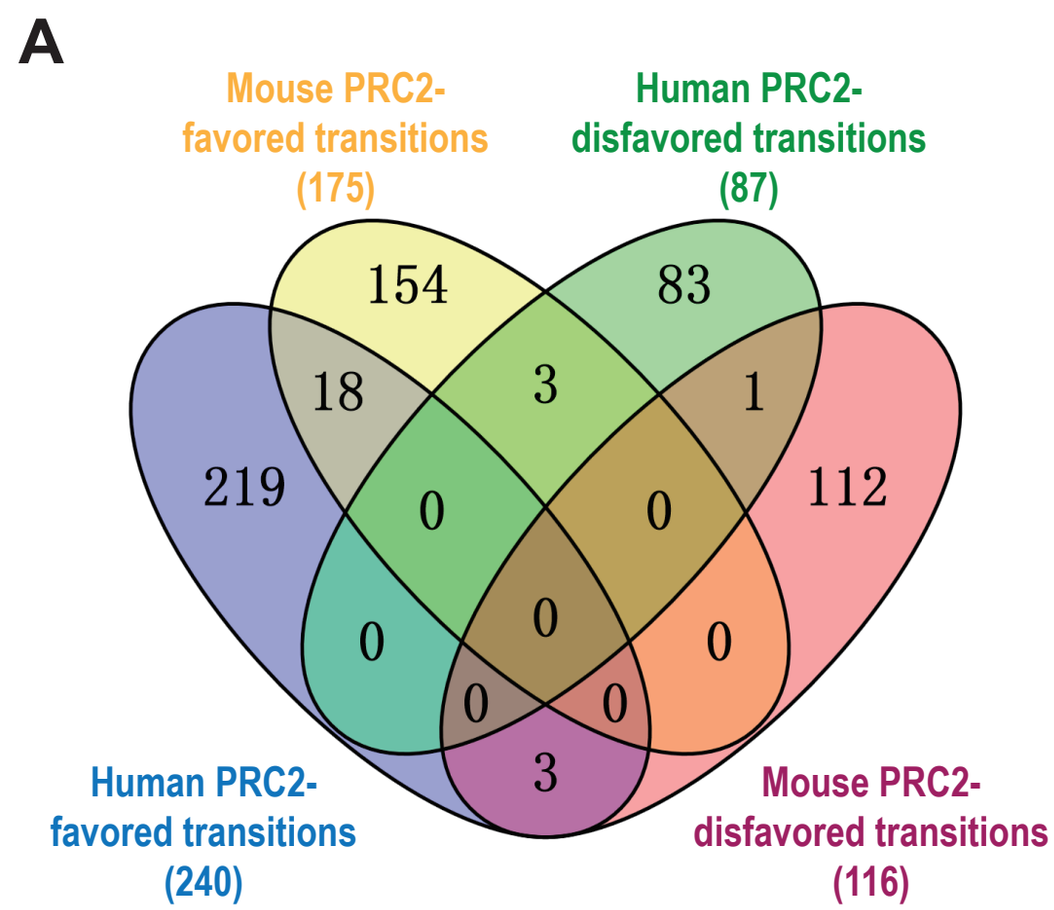


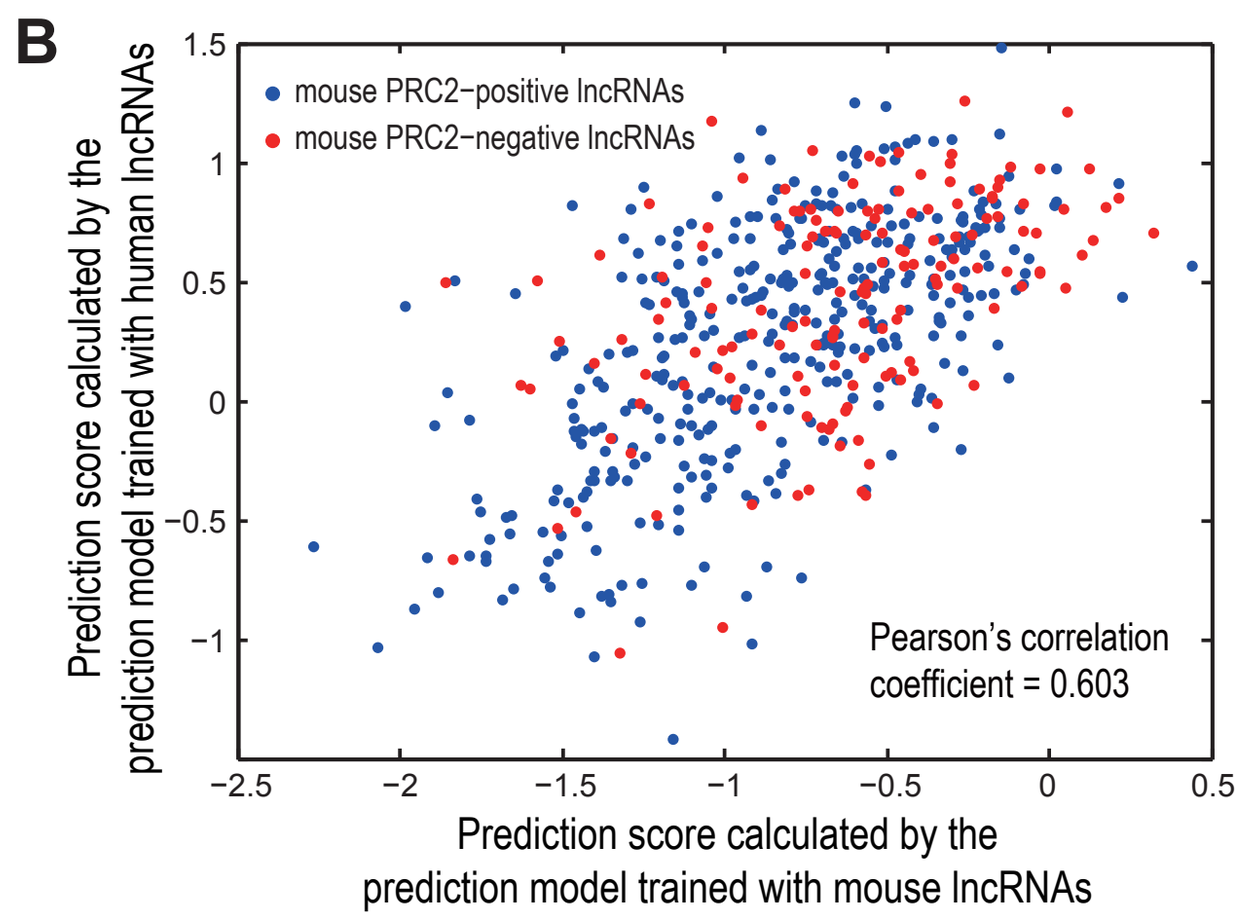


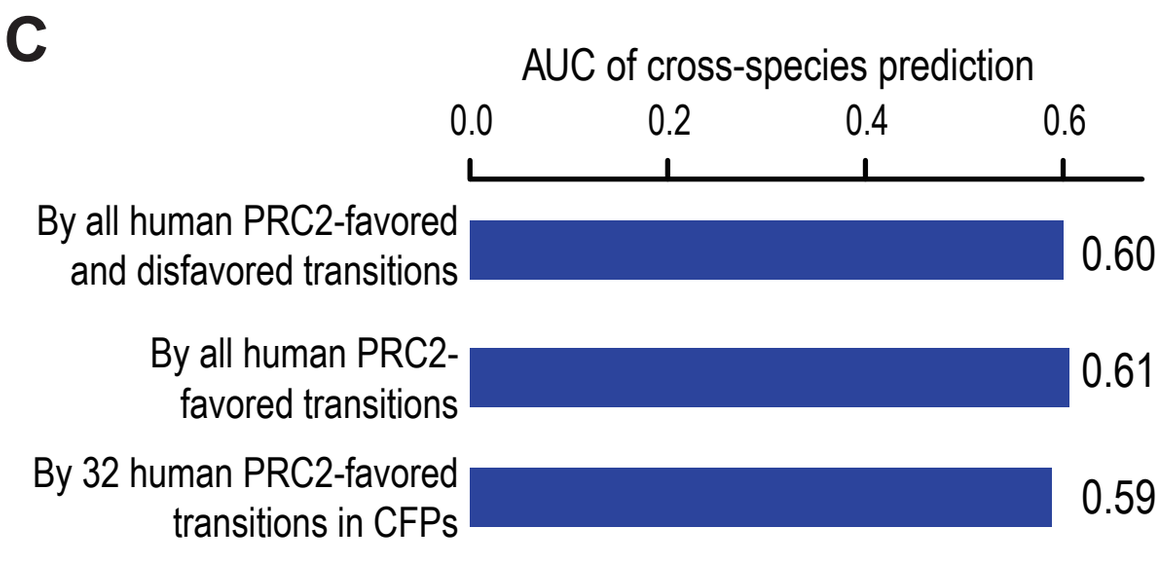


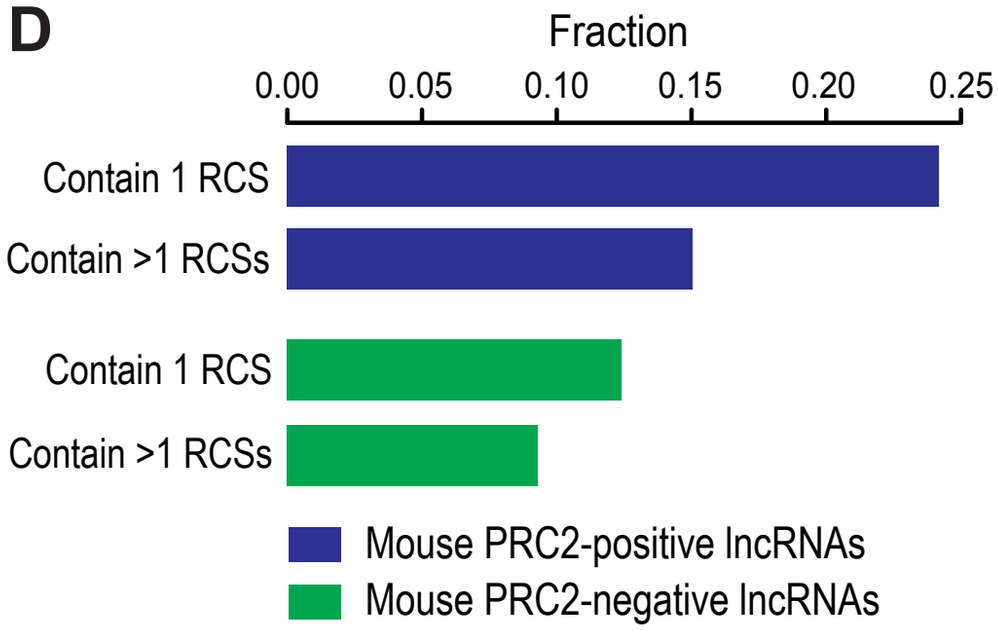


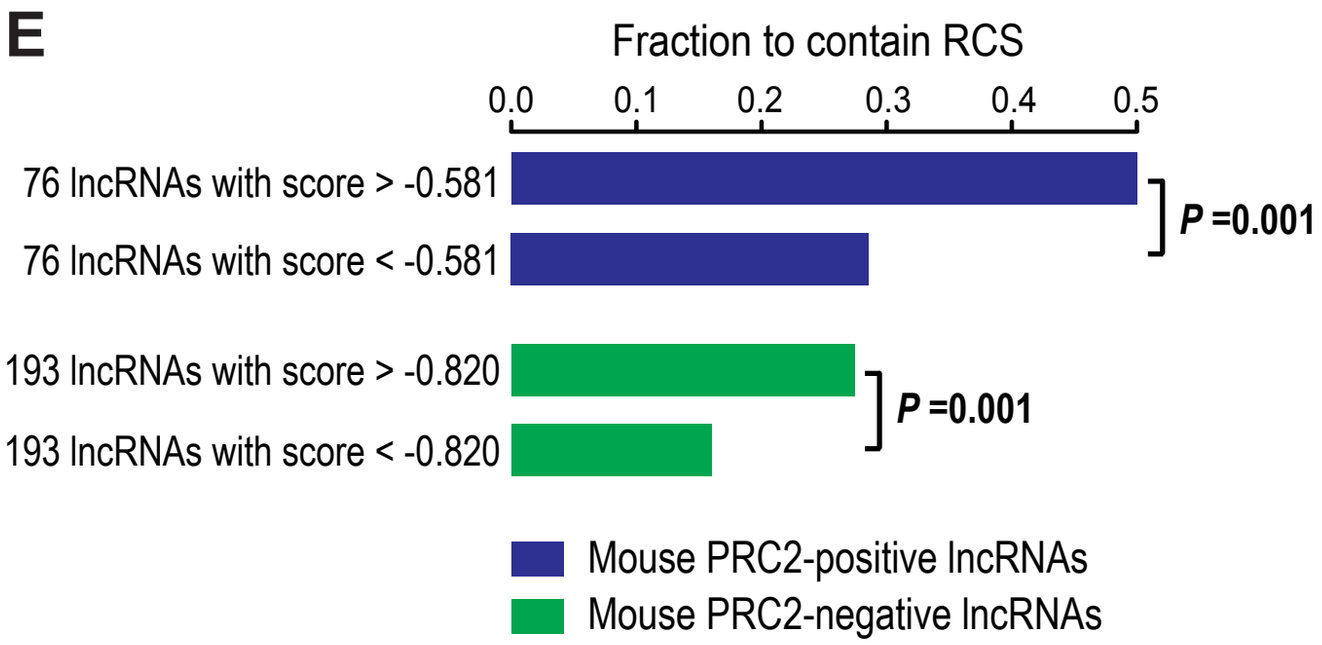
Supplementary Figure S2.

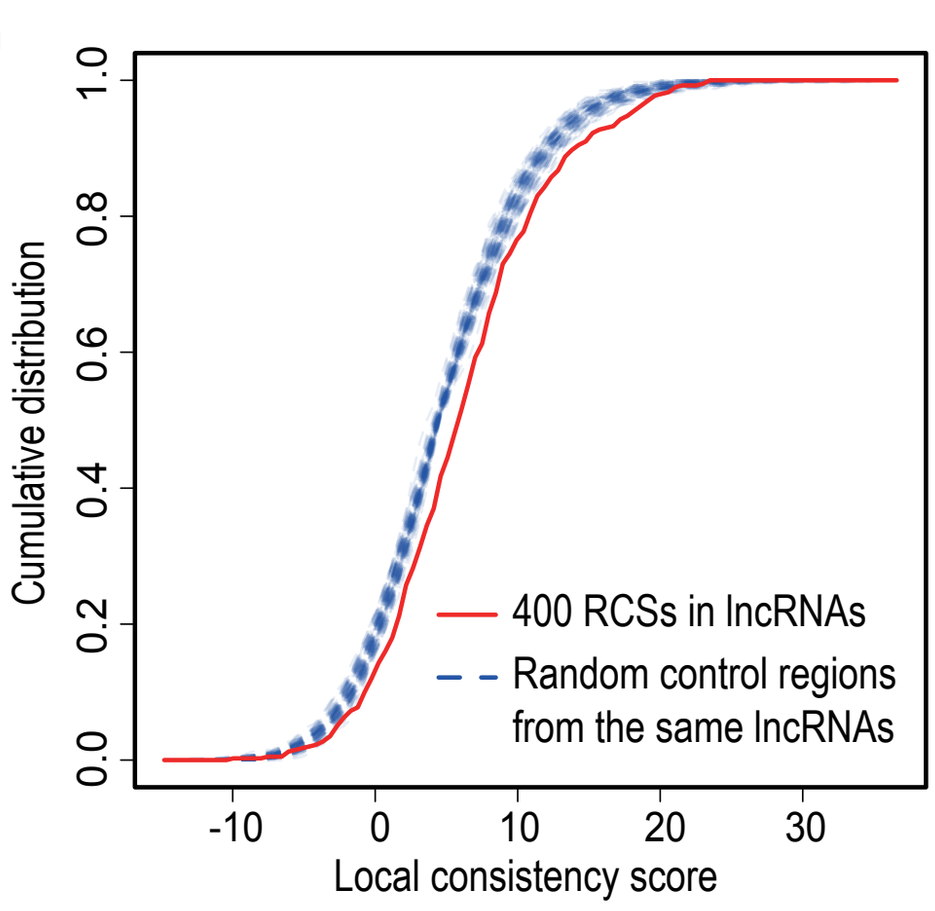


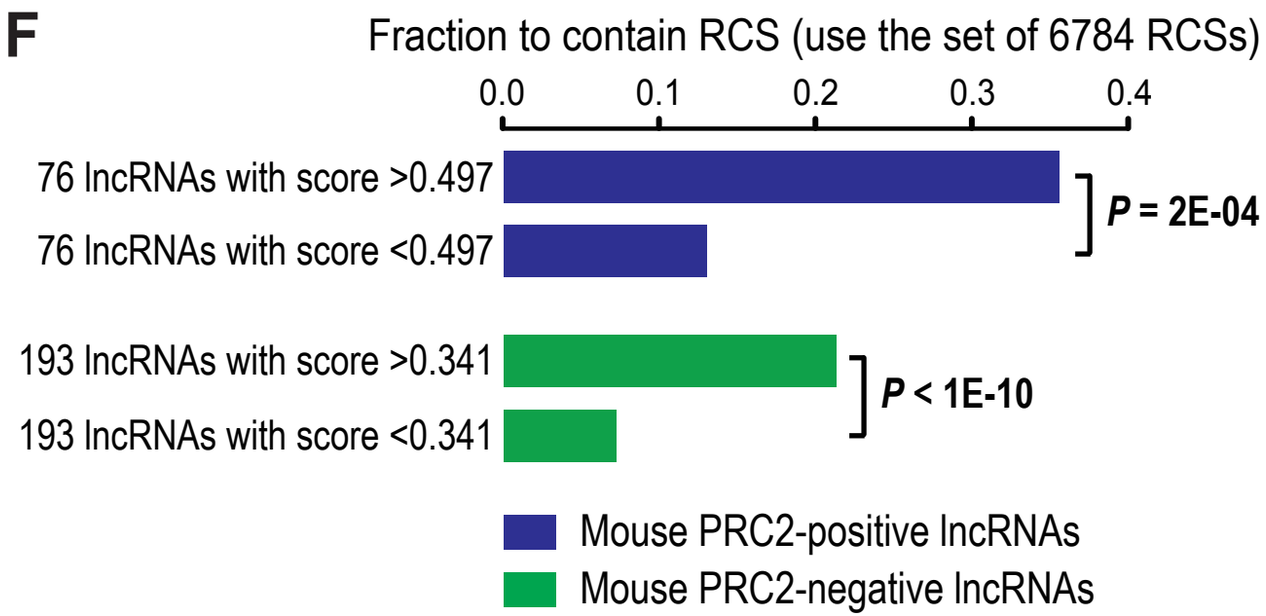






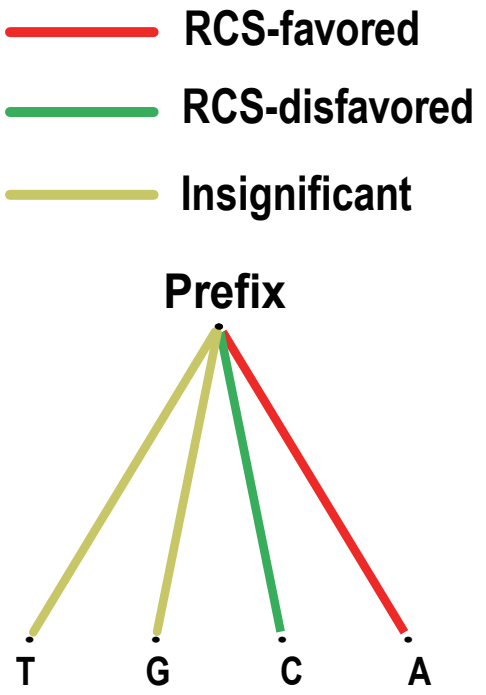


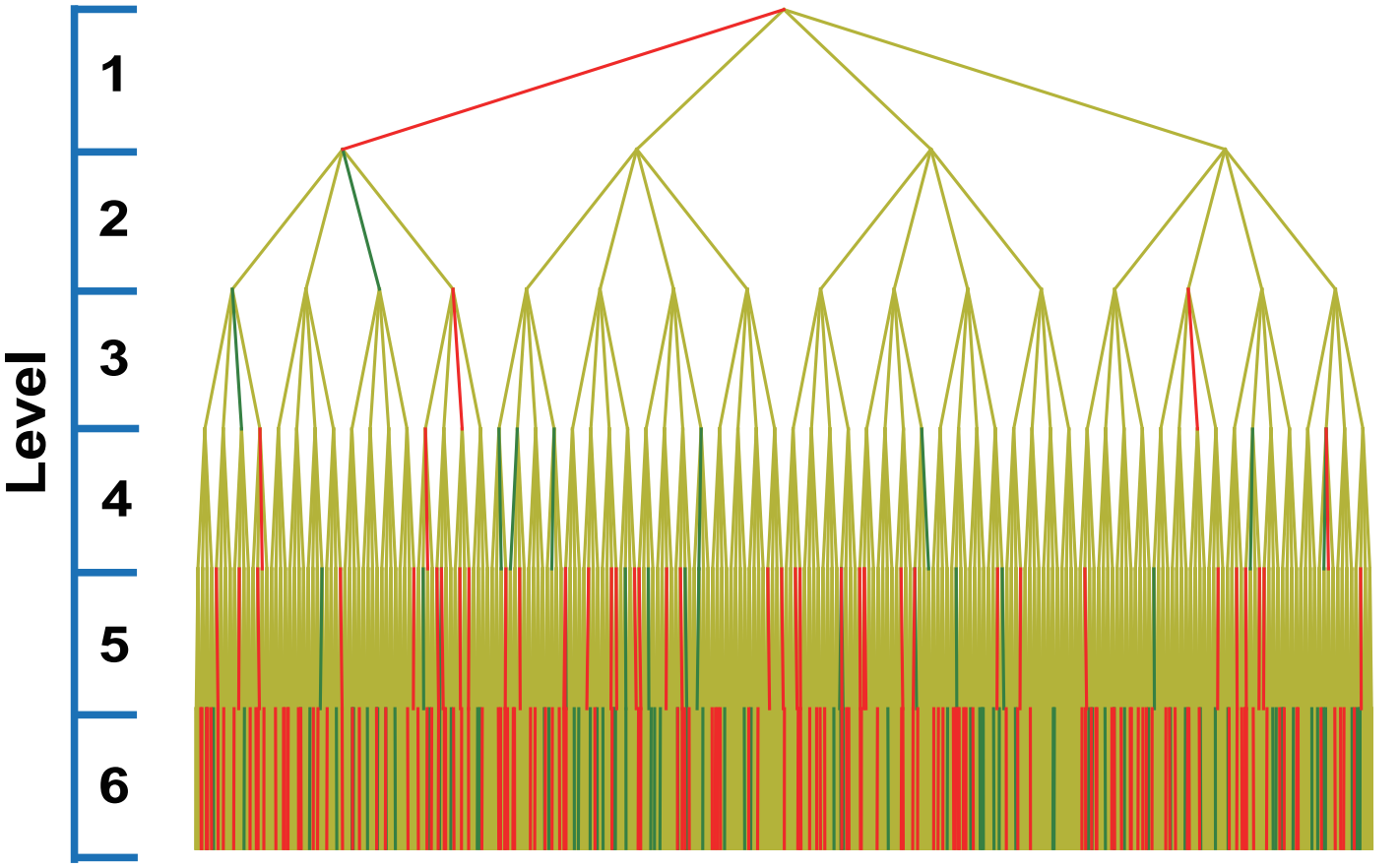




Supplementary Figure S3.

A





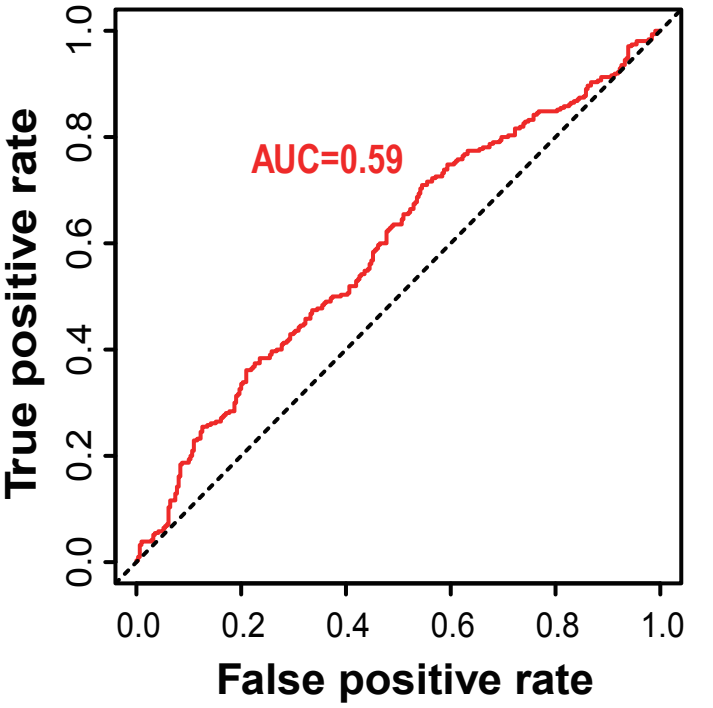
B

RCS-favored transitions	181
Fraction in CFPs	0.188
Average fraction in random permutations	0.062
P value	6E-06

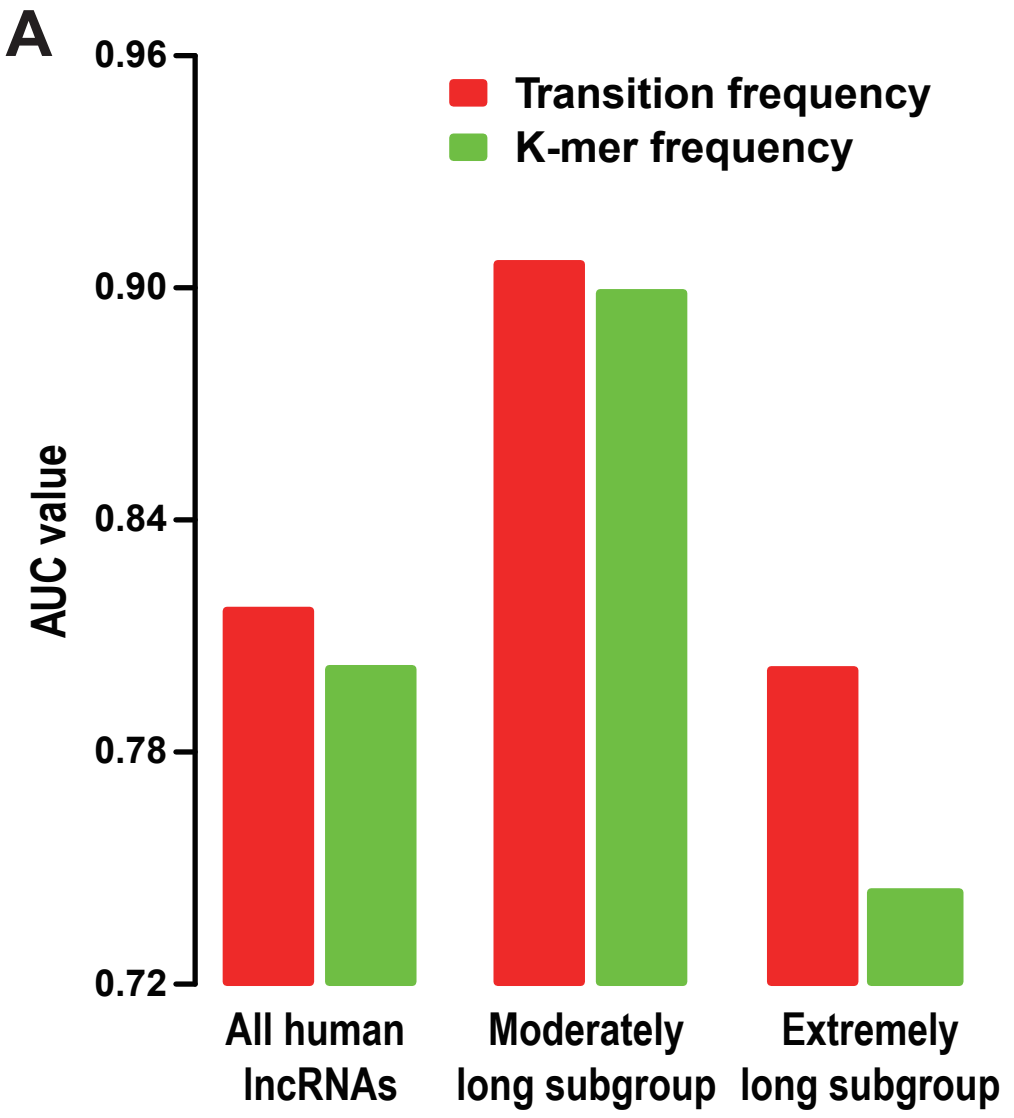
C

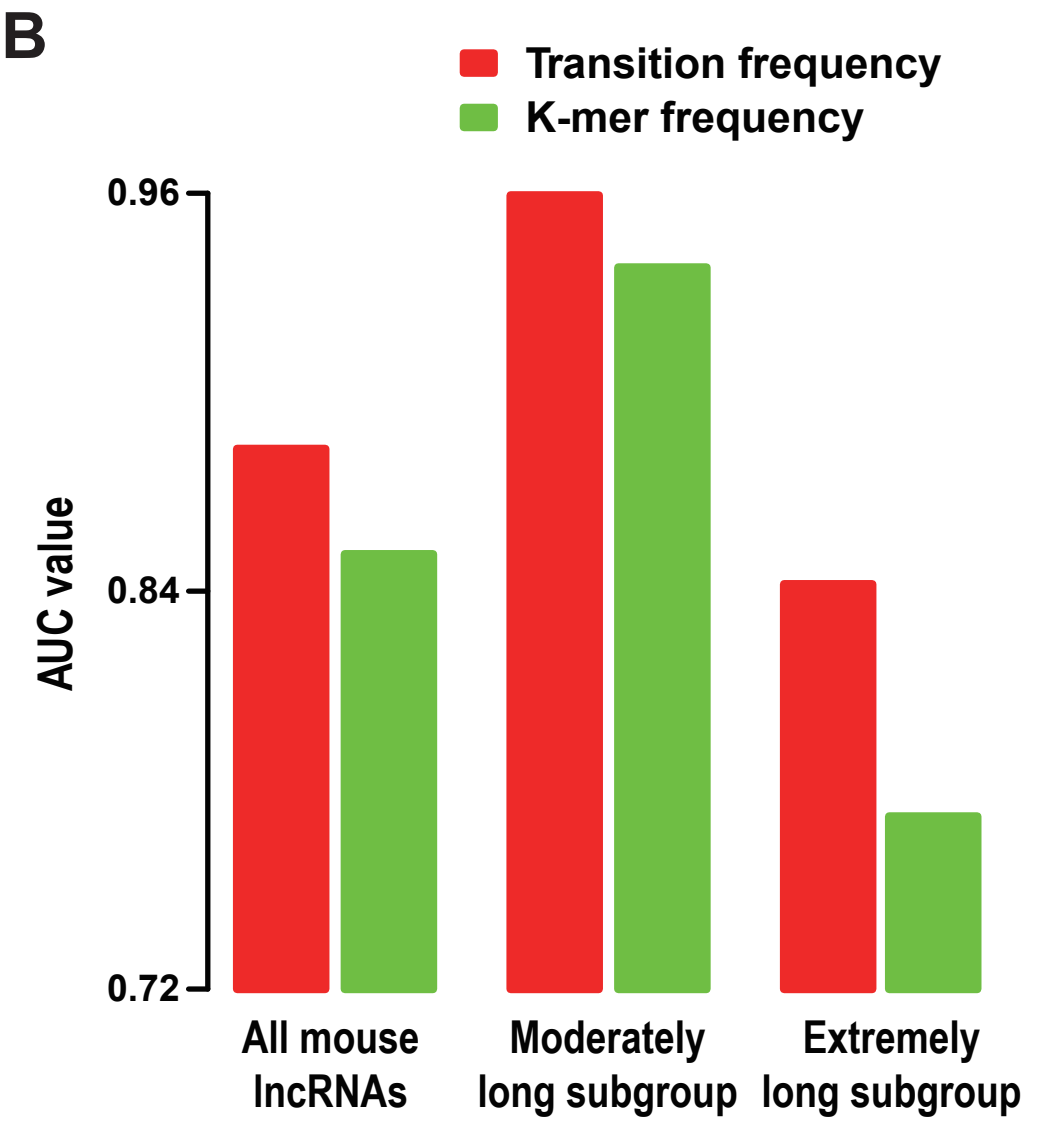
RCS-disfavored transitions	99
Fraction in CDPs	0.101
Average fraction in random permutations	0.031
P value	0.013

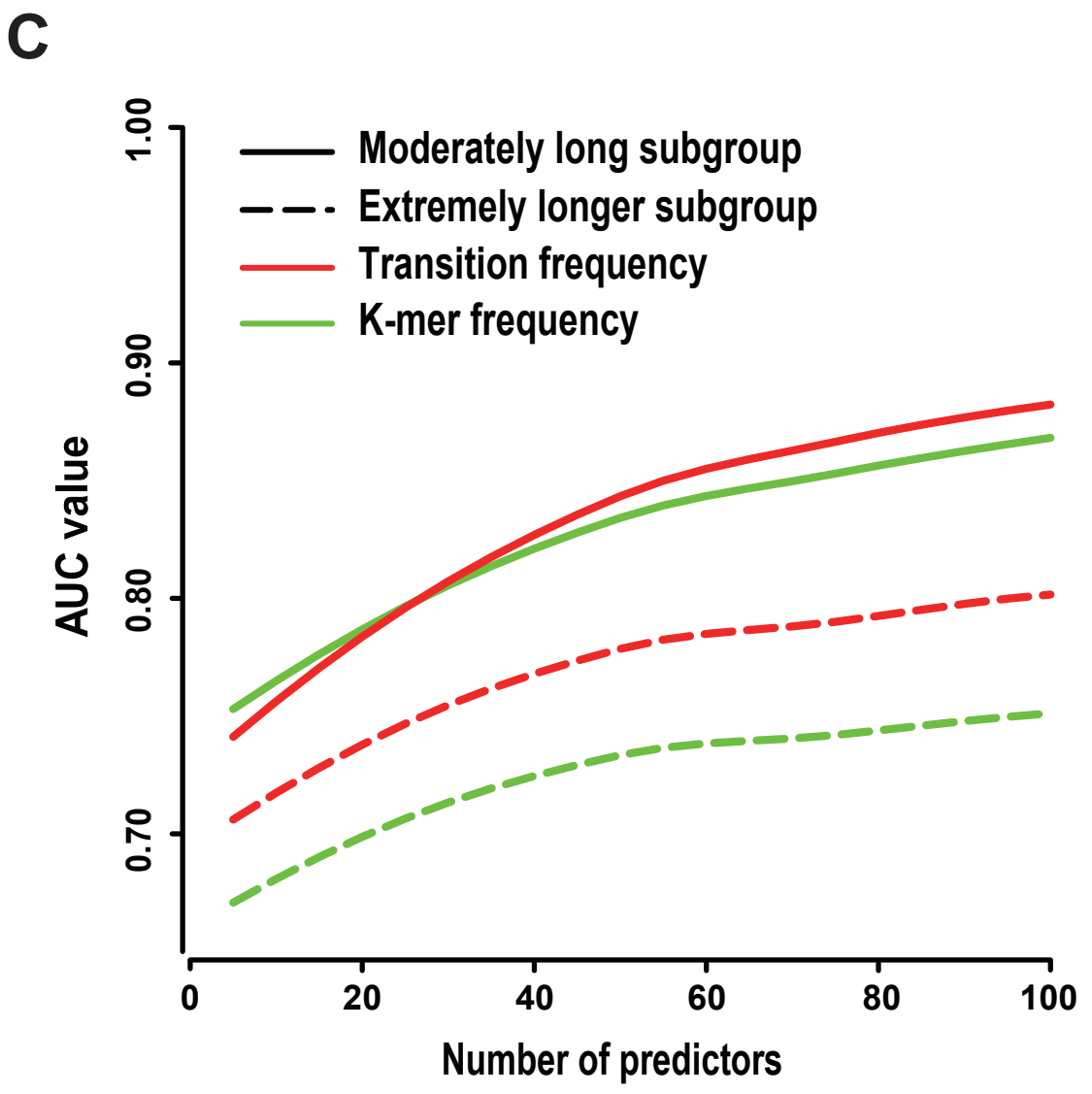
D

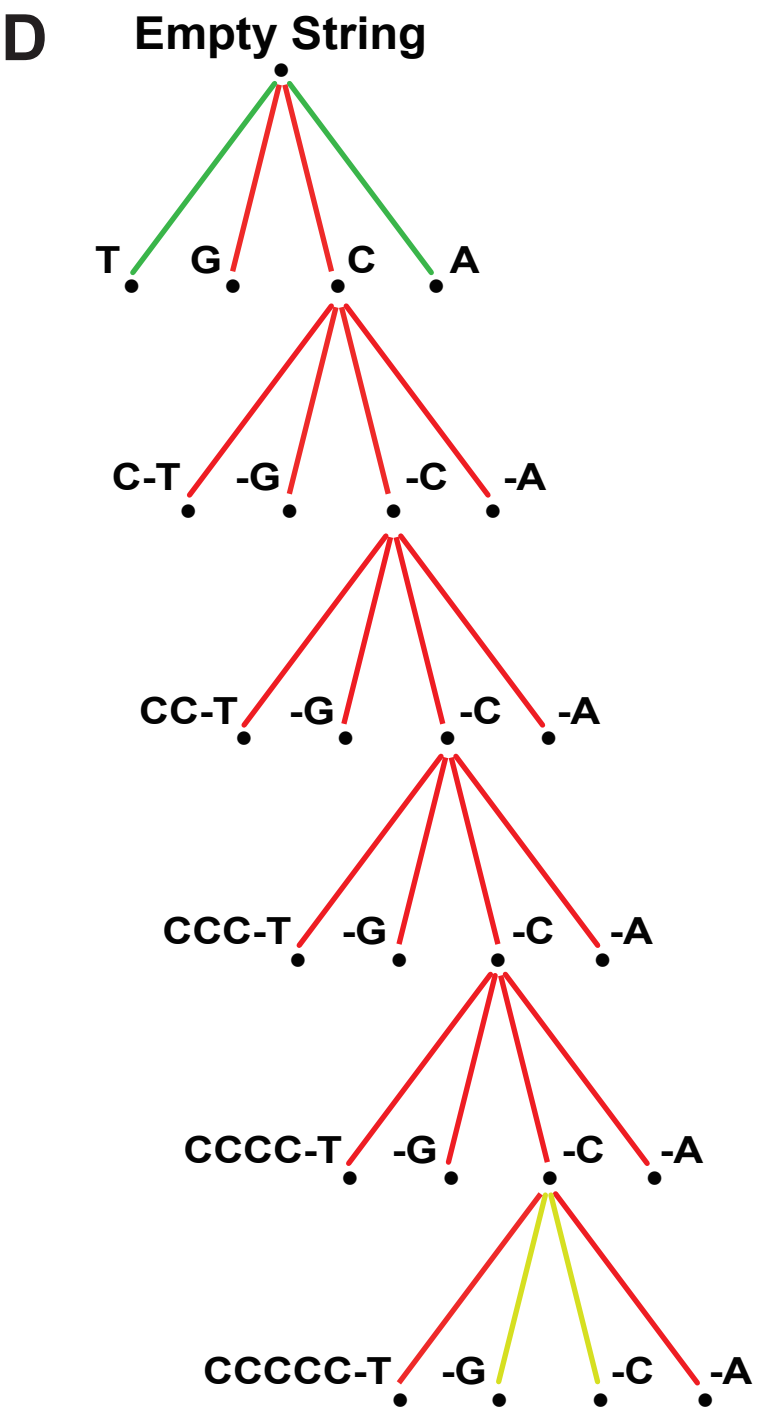


Supplementary Figure S4.

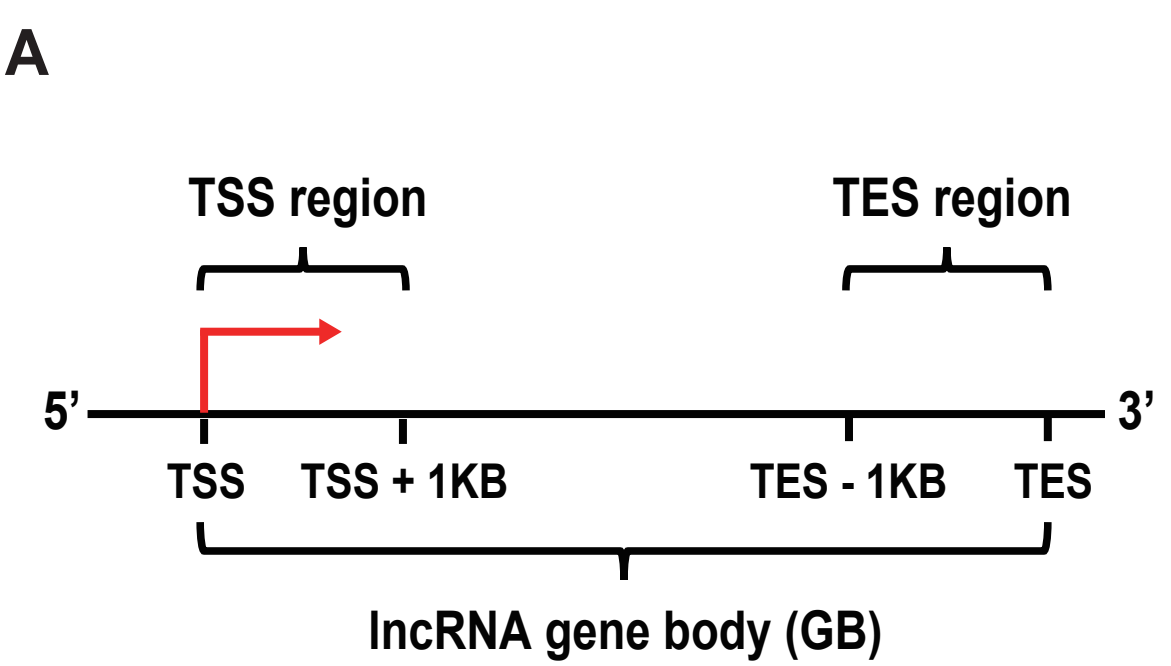


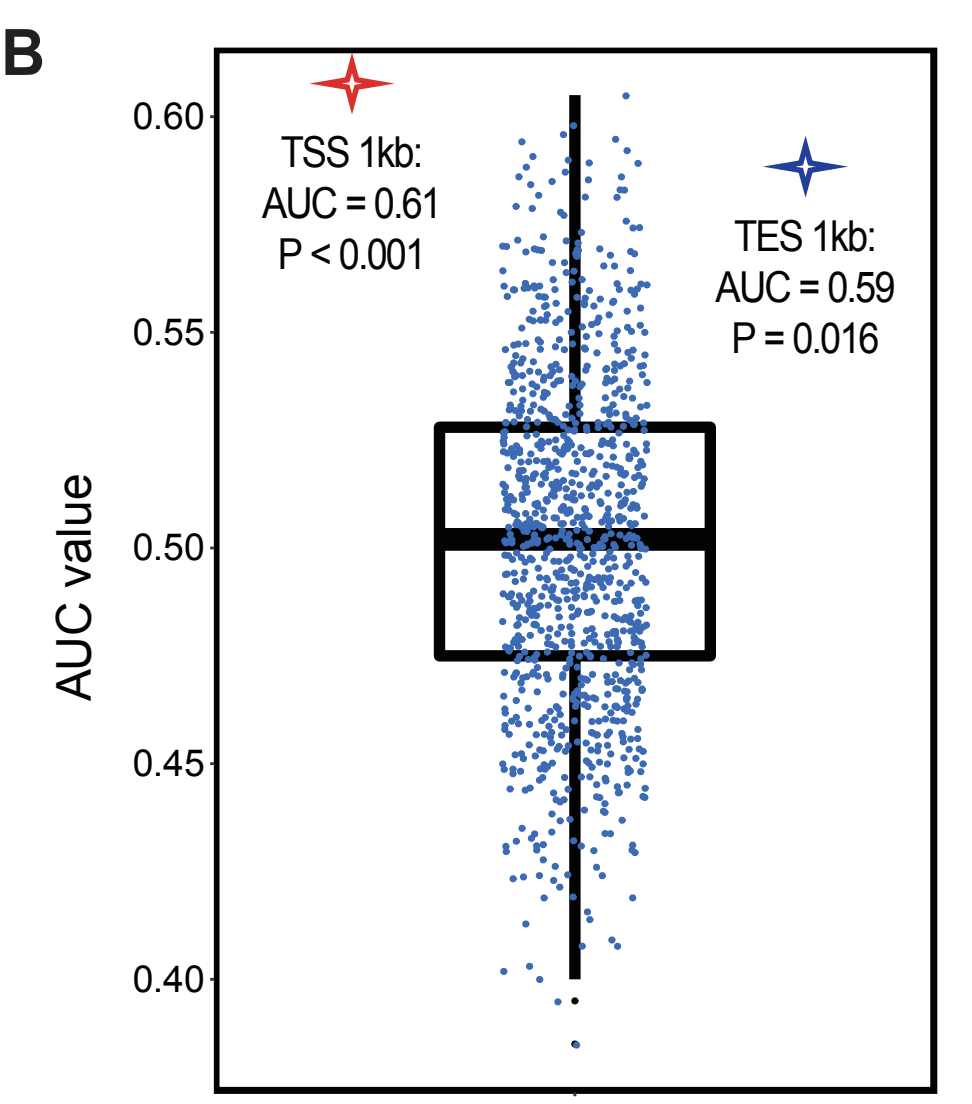


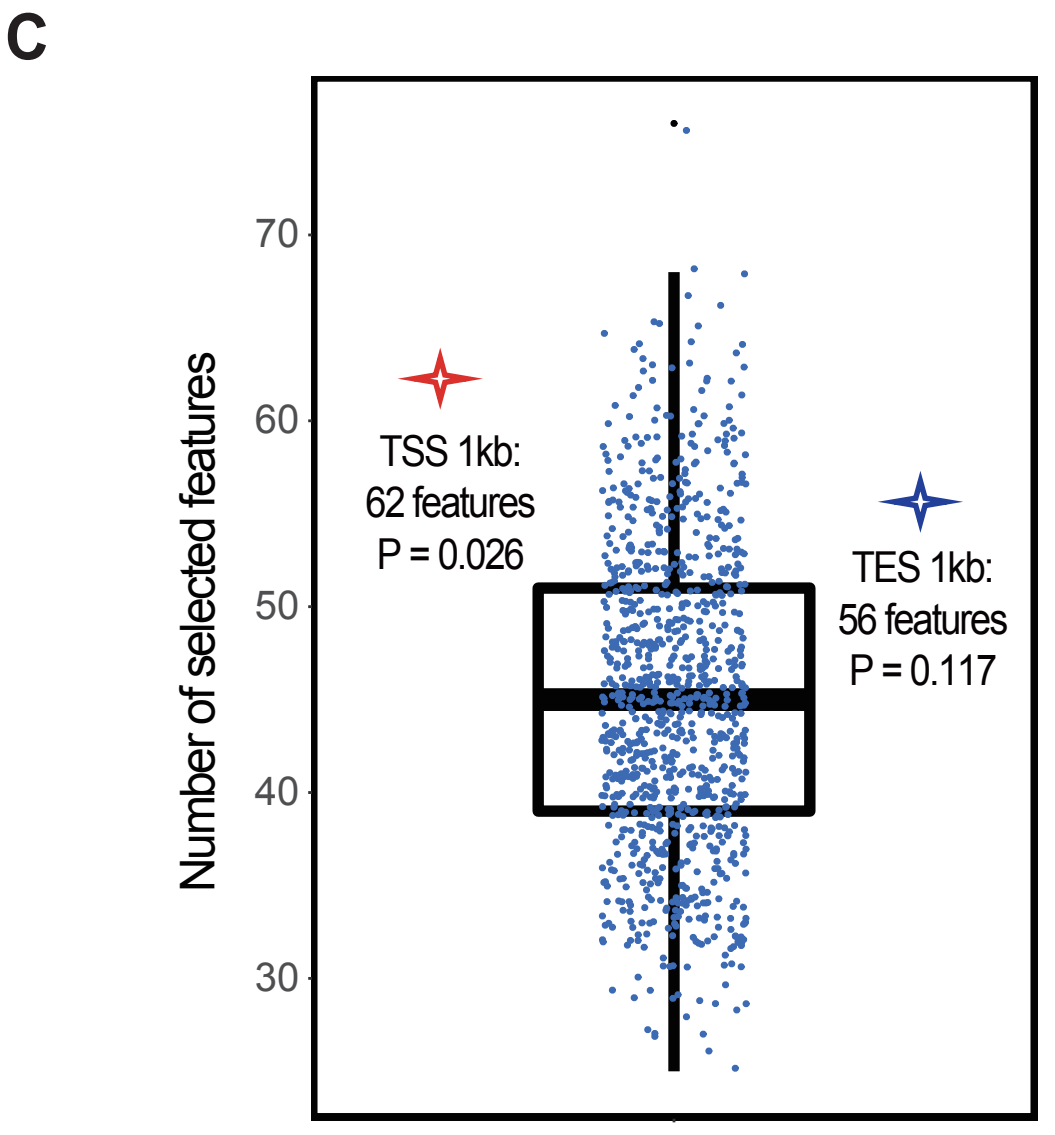




Supplementary Figure S5.







Supplementary Figure S6.