Published online 30 June 2011

Nucleic Acids Research, 2011, Vol. 39, No. 17 ***e116***
doi:10.1093/nar/gkr516

Decoding the genome with an integrative analysis tool: Combinatorial CRM Decoder

Keunsoo Kang[1,*], Joomyeong Kim[2], Jae Hoon Chung[1] and Daeyoup Lee[1,*]

[1]Department of Biological Sciences, Korea Advanced Institute of Science and Technology, 335 Gwahak-ro, Yuseong-gu, Daejeon 305-701, South Korea and [2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA

Received February 14, 2011; Revised May 31, 2011; Accepted June 9, 2011

ABSTRACT

The identification of genome-wide *cis*-regulatory modules (CRMs) and characterization of their associated epigenetic features are fundamental steps toward the understanding of gene regulatory networks. Although integrative analysis of available genome-wide information can provide new biological insights, the lack of novel methodologies has become a major bottleneck. Here, we present a comprehensive analysis tool called combinatorial CRM decoder (CCD), which utilizes the publicly available information to identify and characterize genome-wide CRMs in a species of interest. CCD first defines a set of the epigenetic features which is significantly associated with a set of known CRMs as a code called 'trace code', and subsequently uses the trace code to pinpoint putative CRMs throughout the genome. Using 61 genome-wide data sets obtained from 17 independent mouse studies, CCD successfully catalogued ~12 600 CRMs (five distinct classes) including polycomb repressive complex 2 target sites as well as imprinting control regions. Interestingly, we discovered that ~4% of the identified CRMs belong to at least two different classes named 'multi-functional CRM', suggesting their functional importance for regulating spatiotemporal gene expression. From these examples, we show that CCD can be applied to any potential genome-wide datasets and therefore will shed light on unveiling genome-wide CRMs in various species.

INTRODUCTION

A *cis*-regulatory module (CRM) is a short DNA fragment which governs spatial and temporal expression of nearby genes by interacting with transcription factors (TFs) (1,2). As the basic unit of the gene regulatory network (3,4), the CRM contains multiple transcription factor binding sites (TFBSs) to which a set of TFs binds as an input signal (5–8). Deciphering the relationship between CRMs and associated input signals is a fundamental step toward understanding the precise mechanisms of these gene regulatory networks.

Owing to the popularity of the ChIP-seq method, which generates a snapshot of genome-wide DNA–protein interactions in high resolution, genome-wide occupancy profiles of various TFs and histone modifications have accumulated in public data repositories such as the gene expression omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and UCSC genome browser (http://genome.ucsc.edu/). These ChIP-seq datasets may be an ideal source for identifying CRMs since the multiple TFBSs are thought to represent locations of particular CRMs (9). In this regard, recent studies have accurately predicted tissue-specific CRMs (enhancers) using a few ChIP-seq data sets (10,11), or endeavored to achieve some improvement in CRM prediction with machine learning algorithms (12,13). However, these studies used only a handful of data sets, and their methodologies are not well established to be applicable to other available genome-wide data sets in an unbiased manner.

To illustrate the great potential inherent in the integrative analysis of genome-wide data sets, we have developed a comprehensive analysis tool for identifying genome-wide CRMs in a species of interest, called combinatorial CRM decoder (CCD). The feasibility of CCD is assessed in this study by using nine types of known CRMs (training sets) and 61 feature sets (genome-wide occupancy profiles of 39 TFs and 19 histone modifications in various cell types as well as three computational annotations) that are obtained from 17 independent mouse studies. We validated the CCD algorithm in various aspects and demonstrated key features of CCD.

*To whom correspondence should be addressed. Tel: +82 42 350 2623; Fax: +82 42 350 2610; Email: daeyoup@kaist.ac.kr
Correspondence may also be addressed to Keunsoo Kang. Tel: +82 42 350 2663; Fax: +82 42 350 2610; Email: chaperon@kaist.ac.kr

MATERIALS AND METHODS

CCD tutorials and additional information can be found in the website (http://decode.kaist.ac.kr/).

CCD

CCD is a stand-alone program running on Windows, Linux and Mac OS. The executable versions and its source code can be downloaded freely at the website (http://decode.kaist.ac.kr/).

Definition of the context

CCD requires a single Refseq file of a species of interest for defining genomic context. The genome is divided into five sections ('upstream', 'promoter', 'genebody', 'downstream' and 'intergenic') based on transcription start sites of genes. The knowledge of the context is further used to generate pseudo sets. More detailed information about the context is well-described in Supplementary Figure S1A.

Input data sets

CCD requires training sets and feature sets as inputs. The training set ('trainingset' directory) is a collection of known CRMs, function of which has known *a priori*. The feature set comprises two types of genome-wide datasets: experimental-driven sets and annotation sets which are stored in the 'chipseq' and 'annotation' directories, respectively. We did not manipulate the raw data sets and only used the processed data sets which have already been confirmed in the original papers. The annotation set contains genome coordinates of the same type of elements predicted via computational approaches such as CpG islands and conservation. CCD only needs position information of elements, thereby adopting the BED format as the standard format (http://decode.kaist.ac.kr/). All inputs and the Refseq file should be in the same version of genome assembly.

In the present study, the feature sets were selected if they met one of the following criteria.

(1) The feature has been reported to be associated with any of the training sets.
(2) The enriched regions of the feature have been provided as an additional file.

Total 58 genome-wide ChIP-seq and three annotation data sets were collected from the literature and used as feature sets. Full list of the feature sets and references can be found in Supplementary Table S1.

CCD score

Each feature set contains the genome coordinates of elements. In case of the ChIP-seq data sets used in the present study, the average length ranged from 10 to 7894 bp, and the number of the elements varied from 446 to 48 670. Most features (90%, 53 out of 58) showed <0.04-fold genome coverage. These observations suggest that the occurrences of the features would be very rare. Therefore, CCD assumes that distribution of the features follows the negative binomial distribution. Each significant feature is scored with the CCD score S_{ccd}, which is based on cumulative probabilities from the negative binomial distribution P_{nb} as well as normalized occurrence rate O_{norm}. To calculate the cumulative probabilities P_{nb}, the 'pnbinom' function in the Perl module (Math-CDF-0.1, http://search.cpan.org/~callahan/Math-CDF-0.1/) has been incorporated into CCD. Detailed procedure for the calculation is described in Supplementary Figure S1B. The normalized occurrence rate O_{norm} is estimated as the difference between the occurrences of the feature and the pseudo-feature in the training set, and further divided by the total number of the training set. Finally, the CCD score S_{ccd} is calculated as follows:

$$S_{ccd} = -\log(P_{nb}) \times O_{norm}$$

Ensembl regulatory build database

Ensembl regulatory build database (Mouse Regulatory Build version 4) was downloaded from the website (www.ensembl.org). The database comprises of a total of 140 603 unique clusters which are bound (or enriched) by at least one of the following 27 epigenetic features—CTCF, c-MYC, E2F1, ESRRB, KLF4, NANOG, n-MYC, OCT4, STAT3, SMAD1, SOX2, SUZ12, TCFCP2L1, ZFX, p300 and DNase1 in ES cells; H3K4me3, H3K9me3 and H3K36me3 in ES hybrid cells; H3K4me3, H3K9me3, H3K27me3 and H3K36me3 in NPC (or MEF). To get likely regulatory clusters, we only used the clusters which contain more than three of the above epigenetic features. For comparison, we converted their genome coordinates into mm8 genome assembly by using the liftOver tool (http://genome.ucsc.edu/). A total of 17 562 regulatory clusters were defined and used for comparison analysis.

VISTA enhancer database

We downloaded 745 experimentally validated enhancers ('positive' status) from the VISTA enhancer browser (http://enhancer.lbl.gov). These enhancers were used as a confirmed set for the performance comparison.

RESULTS

Algorithm for the combinatorial CRM decoder

For an analysis, CCD requires the following two types of information for inputs, sets of known CRMs in the same functional category (training sets) and genome-wide data sets of epigenetic features (feature sets). An epigenetic feature refers to one of the followings; a transcription factor, a histone modification or a computational annotation such as CpG islands. To integrate a wide assortment of the feature sets, CCD digitizes binding profiles of the features as one of two digits, either 0 (absence) or 1 (presence). The digitized information is the basic data used in CCD (Figure 1A, blue box). The CRMs function by interacting with designated TFs and are also associated with histone modifications (2,14). Therefore, each epigenetic feature that is significantly shared by the CRMs in a training set can be interpreted as a 'trace'. By defining the set of the traces as 'trace code' representing

the characteristics of the training set, CCD identifies genome-wide CRMs due to these 'genome-wide' properties of the feature sets. In summary, CCD first defines the trace code of each training set, and subsequently uses the trace codes to detect putative CRMs in the genome (Figure 1A).

The elements of inputs (training sets and feature sets) may substantially vary in lengths and numbers in general. To define the traces, an approach filtering out randomly occurred features is essential. Thus, CCD uses the Matthews correlation coefficient (MCC) which takes into account true and false positives and negatives of peaks (*a*, *b*, *c* and *d* in Step 2, Figure 1A). MCC is a balanced measure which can be used regardless of different sizes (15). To estimate MCCs of features, CCD generates pseudo sets by random sampling. The conventional random sampling method randomly selects regions from the entire genome. However, this approach may not be appropriate based on the observations that about half of the genome corresponds to the intergenic regions and the training sets show varying degrees of genomic context (Figure 1B). Therefore, we adopted a novel sampling method called context-dependent random sampling (CDRS). The CDRS method empirically constructs pseudo sets (0.1% genome coverage for the training set and the same number of instances with the feature set), the context of which is similar to the training set and feature set. Then, the pseudo sets are used to generate a confusion table (Figure 1A, Steps 1 and 2). To test whether this approach can be reliable in various circumstances, we generated more than 120 000 random training sets in different numbers (10–15 000; the typical number of elements for the training set), lengths and contexts. The random training sets were evaluated with 61 different feature sets. The result showed that almost all of the features' MCC values are <0.15 (Figure 1C), and therefore the

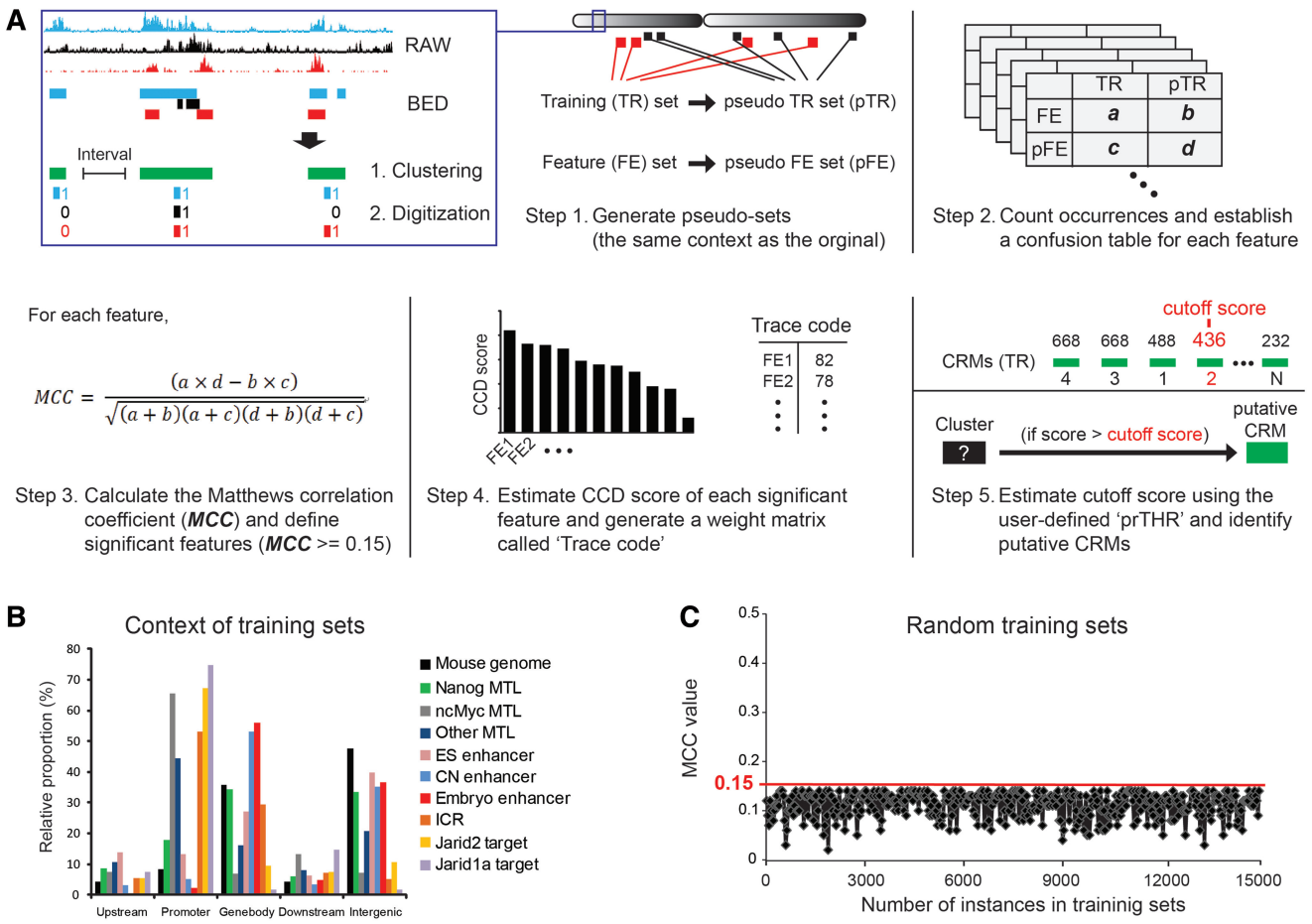

Figure 1. Overview of the CCD algorithm. CCD requires training sets and feature sets for analysis. (**A**) For the feature sets, CCD uses processed genome-wide data sets (BED format) obtained from the literature (see 'Materials and Methods' section). Clusters are defined as regions where features are located within a user-defined interval. Occurrence of each feature in a cluster is encoded as bits ('0' for absence and '1' for presence) (blue box). With this scheme, CCD generates pseudo sets to calculate the Matthews correlation coefficient (MCC) of each feature and further defines significant features (traces) (Step 1–3) (see 'Materials and Methods' section). Once the trace code has been established (Step 4), putative CRMs can be identified by searching the entire genome for the clusters where the calculated scores are above the cutoff score (Step 5). (**B**) The context of the training sets used in the present study differs in varying degrees. Therefore, CCD generates pseudo sets according to the context of the given set. (**C**) A total of 120 000 control sets were randomly generated and used as training sets. Each spot denotes the maximum MCC value. Due to uneven distribution of peaks, MCC values of some random sets (~0.2%) occasionally are >0.15 but their CCD scores are <1 which hardly affect final outcome. These non-significant features are not shown.

features greater than or equal to 0.15 will be regarded as significant features (traces).

Once the traces are defined, a measure which estimates the relative level of the traces' significance is required. Thus, the CCD score has been created. The CCD score is calculated using the cumulative probabilities from the negative binomial distribution, and the normalized occurrence rate which is empirically estimated for each trace (Figure 1A and Supplementary Figure S1B; see 'Materials and Methods' section). These traces and associated CCD scores are called 'trace code' and further used to identify genome-wide putative CRMs (Figure 1A, Step 4).

With the trace code, CCD scans the entire genome to identify clusters showing similar patterns of the traces as the training set. As a supervised approach, CCD first sums all CCD scores of the traces in each CRM and arranges the CRMs by the calculated scores. Then, a cutoff score is set by using the user-defined 'prTHR' parameter (Figure 1A, Step 5). Finally, CCD searches the entire genome for clusters where the sum of CCD scores of the traces is above the cutoff score and consequently defines them as putative CRMs (Figure 1A, Step 5; Supplementary Figure S1C). With this strategy, users can take advantage of the 'prTHR' parameter to adjust the expected level of validation and putative CRMs prior to run. For example, when prTHR is set to 20, the result always includes 80% (100–20) of CRMs in a training set (which automatically validates the model) as well as putative CRMs which contain traces similar to those found in the given training set. In this way, users can obtain putative CRMs with a certain confidence compared with the training sets.

In sum, CCD requires sets of known CRMs (training sets) and genome-wide data sets (feature sets) for analysis. CCD first defines trace codes of the training sets and subsequently searches the entire genome for the clusters showing similar trace codes which is controlled by the user-defined 'prTHR'. Theses clusters are designated as putative CRMs. To gain biological insights of the identified CRMs, the CCD outputs are specifically designed for the available related tools such as R, GREAT and UCSC genome browser (Figure 2).

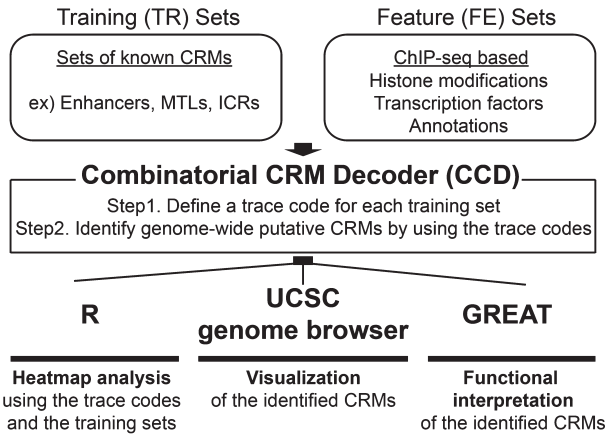

Figure 2. Outline of CCD framework. CCD requires two inputs, training sets and feature sets. With the inputs, CCD first defines trace codes for the training sets. Then, it scans the entire genome to identify putative CRMs by using the trace codes. CCD outputs can be analyzed with the following available tools; R (heatmap, http://www.r-project.org/), GREAT (functional annotation, http://great.stanford.edu/) and UCSC genome browser (visualization, http://genome.ucsc.edu/). Additional information can be found in the webpage (http://decode.kaist.ac.kr/).

Trace codes represent the properties of *cis*-regulatory modules

The prominent advantage of CCD is that any kind of epigenetic features can be evaluated as to whether or not they are significantly associated with particular types of CRMs. To demonstrate the key features of CCD, various types of known CRMs were obtained from five independent studies and used as training sets (Table 1). For feature sets, the following genome-wide data sets were selected to assess the reliability of the CCD algorithm in various aspects. First, 25 epigenetic features were collected from the same studies of the training sets to validate the trace code system (10,11,16–18). Second, 10 epigenetic features (at least two independent sets of Ezh2, Jarid2, Suz12 and H3K27me3 in ES cells) were chosen to confirm the unbiased performance of CCD (16–20). Third, five genome-wide binding profiles of enhancer-associated protein p300, each from ES cells, embryo tissues (forebrain, midbrain and limb) and adult liver, were included to assess whether target CRMs of the p300 are altered in different cell types (10,17,21). Furthermore, an additional 21 epigenetic features including transcription factors, histone modifications and computational annotations were also evaluated (16–27). Total nine training sets and 61 feature sets were analyzed in the current study (Supplementary Table S1).

Even with a wide assortment of the feature sets, CCD successfully identified 23 (Nanog MTL), 26 (ncMyc MTL), 17 (Other MTL), 12 (ES enhancer), 13 (CN enhancer), 10 (Embryo enhancer), 4 (ICR), 23 (Jarid2 target) and 26 (Jarid1a target) traces in the training sets (Supplementary Table S2 and Spreadsheet 1 in Supplementary Data). To validate the defined trace codes, the traces were compared with the known features described in the original studies of the training sets. As expected, CCD successfully detected all previously known features (100%, 29/29) as parts of the trace codes illustrating that the reliability of the trace code system is satisfactory (Figure 3A and Supplementary Table S2).

The quality and number of peaks in feature sets may vary depending on which programs (algorithms) are used. The addition of weak peaks may influence the outcome of the CCD algorithm. To evaluate the effect of additional weak peaks, we generated three or four different sets of Jarid1a (negative control), Jarid2, Ezh2 and Suz12 from the same original raw data (GSE18776) by using MACS with different *P*-value thresholds (1E-03, 1E-05, 1E-07 and 1E-09), and analyzed them with the Jarid2 target (TR8) set. Since CCD filters out non-significant features efficiently by using the MCC value, all of the Jarid1a sets were not included in the trace code (Figure 3B). In case of the Jarid2, Ezh2 and Suz12 sets which are the traces for the Jarid2 target set, the sets consisting of <40 000 peaks were

Table 1. List of the training sets

ID	Name	Description	Property	Count	Avg. bp	Ref.
TR1	Nanog MTL	MTL in mouse embryonic stem cells	(Nanog-Oct4-Sox2) clusters	1554	218	17
TR2	ncMyc MTL		Myc-specific (n-Myc or c-Myc) clusters	1178	223	
TR3	Other MTL		Other clusters	255	229	
TR4	ES enhancer	Enhancers	ES enhancer	25	357	
TR5	CN enhancer		Neuronal activity-regulated enhancer	12 631	1000	11
TR6	Embryo enhancer		Mixture of embryonic forebrain, midbrain and limb tissues specific enhancers	75	1163	10
TR7	ICR	ICRs	Putative or verified imprinting control regions	20	6619	16
TR8	Jarid2 target	Jarid2 binding sites	Jarid2 binding sites near promoters	1393	3601	18
TR9	Jarid1a target	Jarid1a binding sites	Jarid1a binding sites near promoters	2443	934	

ES, embryonic stem cell; CN, cortical neuron; Other, E2f1, Esrrb, Klf4, Smad1, Stat3, Tcfcp2l1, Zfx.

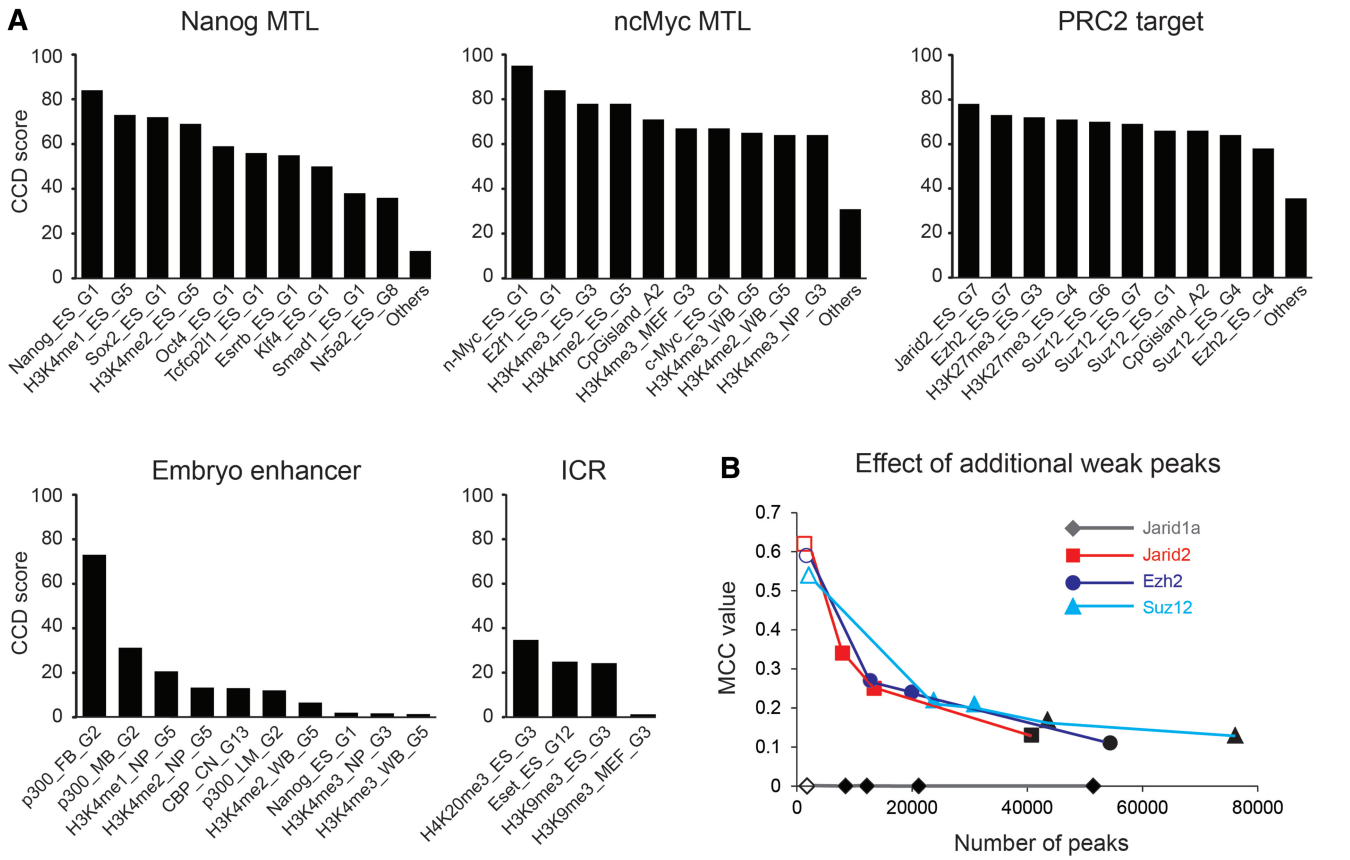

Figure 3. CCD identified all previously reported features (Supplementary Table S2) as parts of the trace codes for the training sets. (**A**) The significant features were weighted with CCD scores. (**B**) Several sets of Jarid1a, Jarid2, Ezh2 and Suz12, which contain different numbers of peaks, were obtained from the same original raw set (GSE18776) by using MACS. The MCC values of the sets were plotted. Unfilled marks and filled marks denote original sets and newly generated sets, respectively. Black marks indicate non-significant feature sets.

still regarded as traces. Although the number of predicted CRMs was increased as the additional weak peaks were included, the quality of the predicted CRMs can be controlled by using the 'prTHR'. For example, totals of 4512 and 7890 putative CRMs were predicted by using the '1E-09' and '1E-07' sets with default settings, respectively. By adjusting the 'prTHR', 98.8% (4457) of the putative CRMs in the former result was identified with the latter set. These results showed that the CCD algorithm is tolerable to the variations of peak numbers within a typical range (<40 000), and the addition of weak peaks results in an increase in the number of putative CRMs. In general, we recommend users to use <40 000 peaks for a single ChIP-seq data set. More number of peaks can be used only if users have confidence that weak peaks are also genuine binding regions of protein.

The trace code, a combination of significant features, can reflect the relationship between a particular type of

CRMs and features. For instance, recent studies have demonstrated that Jarid2 is a novel subunit of the polycomb repressive complex 2 (PRC2) (18,20,28,29). By using the Jarid2 target set (comprising 1393 binding sites of JARID2 protein) and 61 feature sets as inputs, CCD invariably pinpoints the associated epigenetic features of PRC2—the components of PRC2 (Jarid2, Suz12 and Ezh2), a component of polycomb repressive complex 1 (Ring1B), histone modifications (H3K27me3, H3K4me2 and H3K4me3), CpG islands and Eset as the trace code, whereas none of the other features are included (Figure 3A and Supplementary Table S2). Since most of the traces except Eset are known to be components or associated features of PRC2 (28–32), the result demonstrates the unbiased algorithm of CCD and further strengthens the connection between Jarid2 and PRC2. Thus, we designated the Jarid2 target set as PRC2 target set. In case of the enhancer sets (TR4, TR5 and TR6), enhancer-associated histone modifications (H3K4me1 or H3K4me2) are defined as parts of the trace codes consistent with the previous reports (11,14). Intriguingly, each type of enhancer is only related to p300 in the same cell type of the enhancer set, suggesting that target enhancers of p300 may vary with different cell types (Supplementary Table S2).

In addition, the trace code is capable of explaining some biological questions as well. For instance, TBX3 and the orphan nuclear receptor NR5A2 (also known as LRH-1) have been reported to share binding targets with OCT4 (also known as POU5F1), SOX2, NANOG, SMAD1 and ESRRB (26,27,33). Our integrative analysis via CCD simply reveals that the Nanog MTL set (TR1) is significantly occupied by all the above factors as well as TCFCP2L1, KLF4, E2F1, STAT3, p300, CTCF and active histone modifications (H3K4me1, H3K4me2 and H3K4me3) in ES cells (Supplementary Table S2). Therefore, the result implies that the known property of NR5A2 and TBX3, which enhance reprogramming efficiency, can be explained by the extensive binding of the NR5A2 and TBX3 to the Nanog MTLs (35.7 and 10.9% of the CRMs) (Spreadsheet 1 in Supplementary Data).

The great advantage of the trace code system is that it can discover unnoticed connections between CRMs and epigenetic features. For instance, imprinting control regions (ICRs) controlling monoallelic expression of genes in the imprinted domains can be regarded as CRMs (34–36). There is growing evidence that repressive (H3K9me3 and H4K20me3) histone modifications are enriched at the ICRs in an allele-specific manner (37,38). However, this is based on observations from a small fraction of the ICRs. To validate the above facts and identify unnoticed epigenetic features possibly associated with the ICRs, a set (TR7) of twenty ICRs was analyzed with 61 features by using CCD. Surprisingly, the analysis reveals that the following epigenetic features are extensively enriched at most ICRs in ES cells—H4K20me3 (18/20), H3K9me3 (17/20) and Eset (18/20) (Spreadsheet 1 in Supplementary Data). In addition, manual investigation of the ICRs confirmed that the *Sgce-Peg10* and *Rasgrf1* domains also contain high enrichment of the H3K9me3 and H4K20me3 in the ES cells (Supplementary Figures S2O and P), suggesting that the processed ChIP-seq data missed these marks due to the algorithm of peak identification (16). Notably, the histone H3 Lys 9 methyltransferase ESET, which was reported to bind to 15 ICRs in ES cells (25), turned out to be enriched at 18 ICRs with the above histone modifications (Supplementary Figures S2 and S3). Based on these observations, we propose that H3K9me3, H4K20me3 and Eset are the key epigenetic features associated with the ICRs in the early-stage embryo (ES cells).

Overall, these results strongly suggest that the trace code can represent the unique characteristics of certain types of CRMs.

Identification and characterization of genome-wide *cis*-regulatory modules

To specify the training sets with the defined trace codes, heatmap analysis was performed by using R with the CCD output. We also analyzed three randomly subsampled sets from the original training sets. The result indicates that some training sets may belong to the same functional classes according to the similar patterns of the trace codes (Figure 4). All of the subsampled sets show almost similar trace codes with the originals implying that the variation of number of instances is marginal. Based on the dendrogram in the heatmap, the following training sets are regarded as distinct classes; Nanog MTL (multi transcription factor-binding loci) (class I), embryo enhancer (class II), ICR (class III), ncMyc MTL (class IV) and PRC2 target (class V) sets.

With the defined trace codes, CCD is able to identify genome-wide putative CRMs. To catalogue genome-wide CRMs with high confidence, we empirically determined a cutoff score (prTHR) for each class to contain similar occurrence of significant features as compare to that on the given known CRMs ($R^2 > 0.75$) (Figure 5A). For instance, we set a cutoff score (prTHR = 10, 777.29 CCD score) for the PRC2 target set (class V) since the number of identified CRMs using a high cutoff score (prTHR = AVG; average of sum of CCD scores in the training set, 970.52 CCD score) were less than that of the training set (Supplementary Figure S4A). Manual investigation of the Hoxd cluster reveals that the defined cutoff score is enough to identify the previously known PRC2 target sites (Supplementary Figure S4B) (39), and hence we applied this strategy to the rest four classes.

Using the defined cutoff scores, CCD successfully pinpointed genome-wide CRMs in the mouse genome including 2797 (class I), 2455 (class II), 176 (class III), 5557 (class IV) and 2160 (class V) CRMs (Spreadsheet 2 in Supplementary Data). Due to the CCD algorithm, a subset of CRMs in the given training set is always guaranteed to be identified along with putative CRMs. For example, 40% of the known ICRs (8 out of 20, prTHR = 60) near the *Impact*, *Peg3*, *Airn*, *Peg13*, *Nnat*, *Snurf*, *H19* and *Meg3* imprinted genes were obtained along with 168 newly predicted CRMs in the ICR result. The newly predicted CRMs are not located around the computationally predicted imprinted genes (www.geneimprint.com). Nevertheless, it will be interesting to

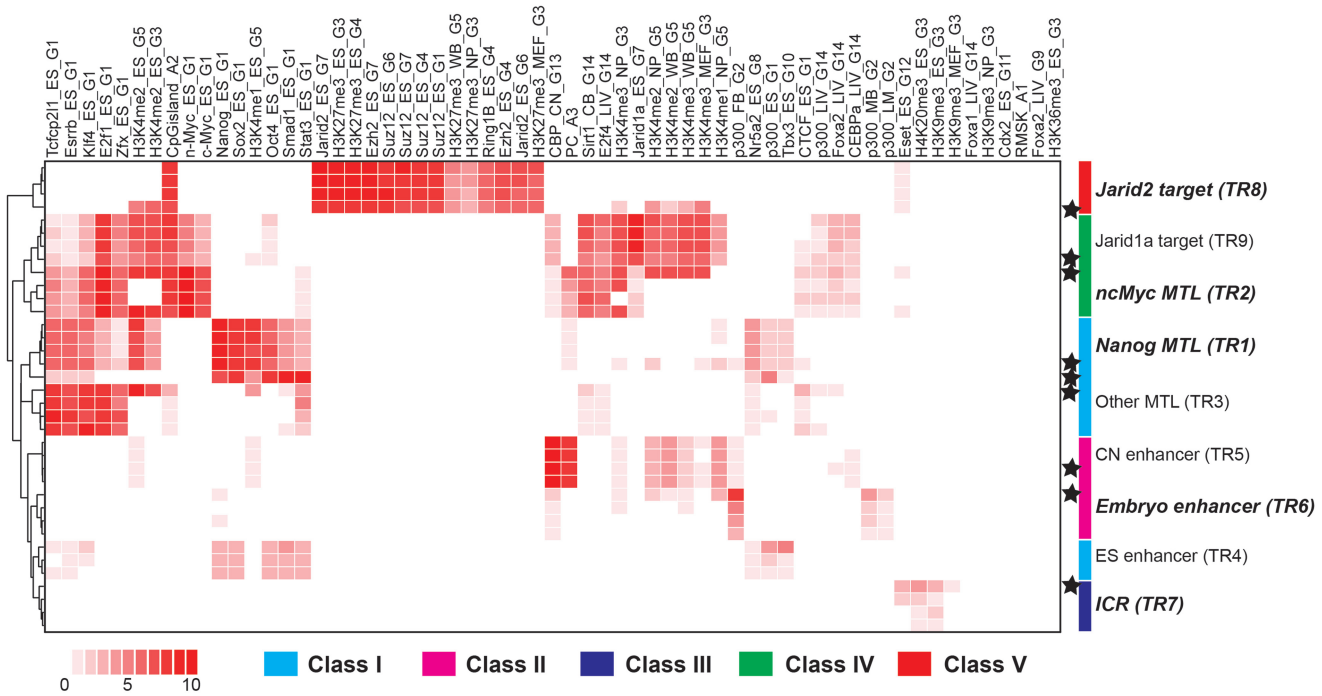

Figure 4. Unique trace codes are present in all training sets. Total 61 features extracted from the literature were used as feature sets (Supplementary Table S1). Both of the original (star marked row) and three randomly subsampled training sets show the similar trace codes. For each training set, the CCD scores of traces were rescaled into 10 levels and these rescaled scores were used to draw the heatmap. According to the patterns of the trace codes, the training sets were categorized into five classes (I–V, bottom).

examine the genes around these CRMs since they are enriched (or bound) by unique features (H3K9me3, H4K20me3 and Eset) similar to the known ICRs. In addition, the identified CRMs contain similar occurrence of significant features compared to the training sets. For instance, 74.9, 45.5 and 59.1% of the predicted Nanog MTL CRMs (class I) are occupied by NANOG, OCT4 and SOX2 of which the proportions are similar with the training set (NANOG—83.9%, OCT4—58.7% and SOX2—72.3%). The genome-wide distributions of the identified CRMs depend on the classes (Figure 5B). In case of the class IV (ncMyc MTL) CRMs, the genomic locations are biased toward the promoter regions, whereas the class I (Nanog MTL) CRMs seem to be distributed randomly with respect to genomic context.

To validate the putative CRMs in terms of biological relevance, functional annotation analysis was conducted by using GREAT (see 'Materials and methods' section) (http://great.stanford.edu/), which unpacks genomic regions based on the annotation of the nearby genes. The top 300 newly identified CRMs for each class were analyzed with the default parameters (FDR = 0.05). The analysis reveals that the annotated functions of the putative CRMs are well correlated with previously known facts, thereby confirming the CCD framework (Supplementary Table S3). For instance, the Nanog MTL (class I) candidates are involved in stem cell maintenance (binomial $P = 3.5E-05$) and differentiation (binomial $P = 8.1E-05$) in the GO Biological Process category. In case of the embryo enhancer (class II) candidates, the CRMs are located near the genes affecting 'abnormal morphology' (eight terms, binomial $P < 4.1E-04$) in the Mouse Phenotype category and 'compartment specification' (binomial $P = 7.4E-07$) in the GO Biological Process category. The most striking example was obtained from the analysis of the PRC2 target (class V) candidates. The majority of significantly associated genes near the CRMs are related to 'negative regulation' (or 'positive regulation') (21 terms, binomial $P < 7.2E-05$) in the GO Biological Process and 'abnormal morphology' (27 terms, binomial $P < 1.7E-04$) in the Mouse Phenotype categories, consistent with the known properties of PRC2 (40,41).

Comparison of CCD with the Ensembl regulatory build method

There are limited numbers of experimentally validated CRMs. The VISTA enhancer browser is a central resource for the experimentally validated CRMs showing enhancer activity in a single embryonic timepoint (http://enhancer.lbl.gov) (42). To evaluate the performance of CCD compared to the Ensembl regulatory build method, we used 745 experimentally validated enhancers from the VISTA enhancer browser as a confirmed data set (see 'Materials and Methods' section). The Ensembl regulatory database is comprised of best-guessed regulatory elements predicted by an overlapping approach using a variety of genome-wide epigenomic data sets. Despite a large number of predicted CRMs in the Ensembl regulatory database (17 562) as compared to CCD (12 636), the database

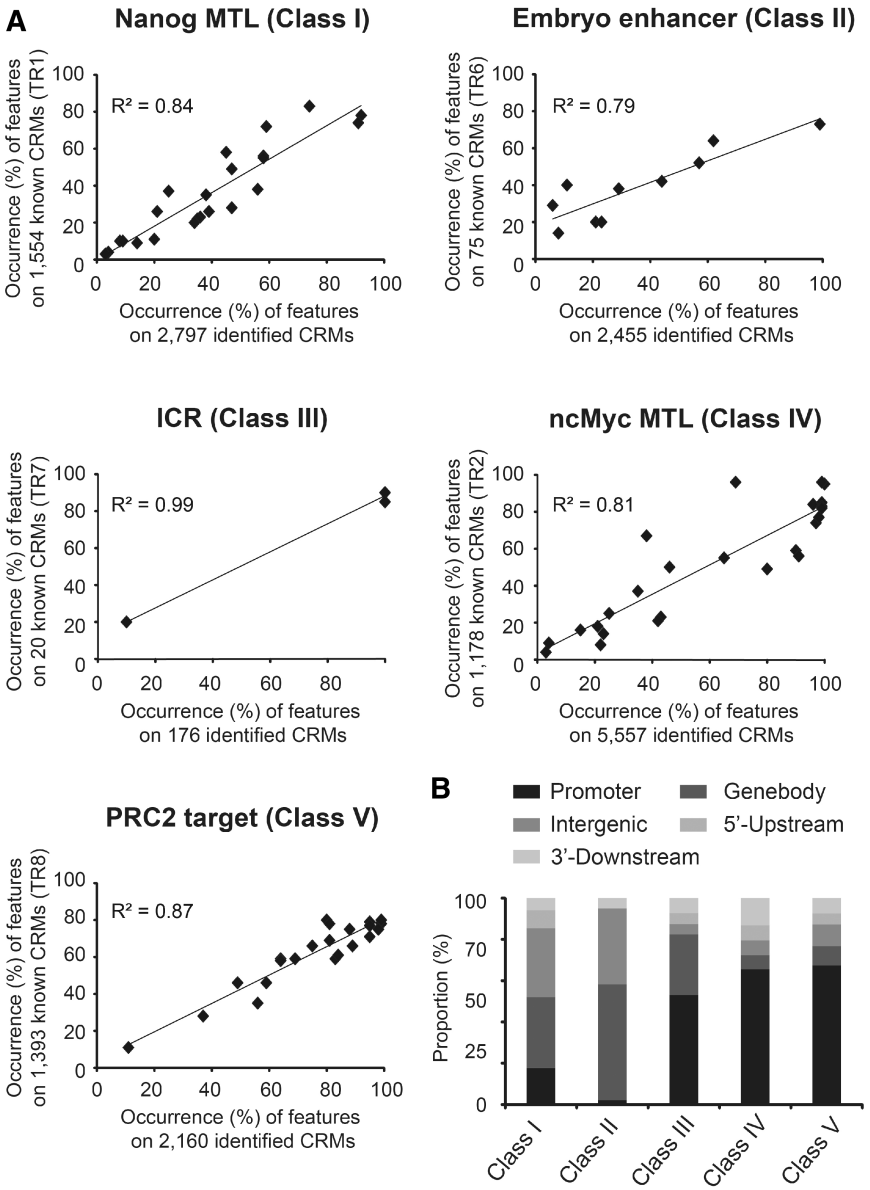

Figure 5. Identification of genome-wide CRMs. (**A**) CCD identified a total of 12 636 CRMs which belong to the five classes (I–V) in the mouse genome. The patterns of occurrences of features in the identified CRMs were compared with those of the known CRMs. The following prTHRs were used: 20 (class I), 30 (class II), 60 (class III), 30 (class IV) and 10 (class V). (**B**) Genome-wide distributions of the identified CRMs were examined in five sections (Supplementary Figure S1A).

only includes 2.1% of experimentally confirmed enhancers (16 out of 745), whereas the identified CRMs by CCD contain 27.0% of the enhancers (201 out of 745). The accurate identification of the CRMs is based on not only the CCD algorithm but also a wealth of epigenetic information used in the present study. The average number of associated features with the identified CRMs is significantly higher than the Ensembl regulatory database (Figure 6A). Therefore, more genome-wide CRMs can be discovered and classified by integrating more training sets and feature sets which will be available in near future with CCD.

In contrast to the Ensembl regulatory method, CCD has several unique properties. First, CCD can categorize identified CRMs according to their trace codes. For instance, a large domain (~240 kb) contains three genes (*Plk1s1*, *Xrn2* and *Nkx2-4*) and an experimentally validated enhancer in the 12th intron of the *Plk1s1* gene (Figure 6B). The Ensembl regulatory build method predicted three CRMs in this region, whereas CCD identified six CRMs. Notably, CCD exactly identified the validated enhancer (VISTA enhancer) which the Ensembl regulatory build method failed to detect, and precisely classified it as embryo enhancer. Second, CCD can also measure the relative contribution of the features to the CRMs using the CCD score. For example, p300 (CCD score = 72) is the top feature that contributes most significantly to the embryo enhancer CRM

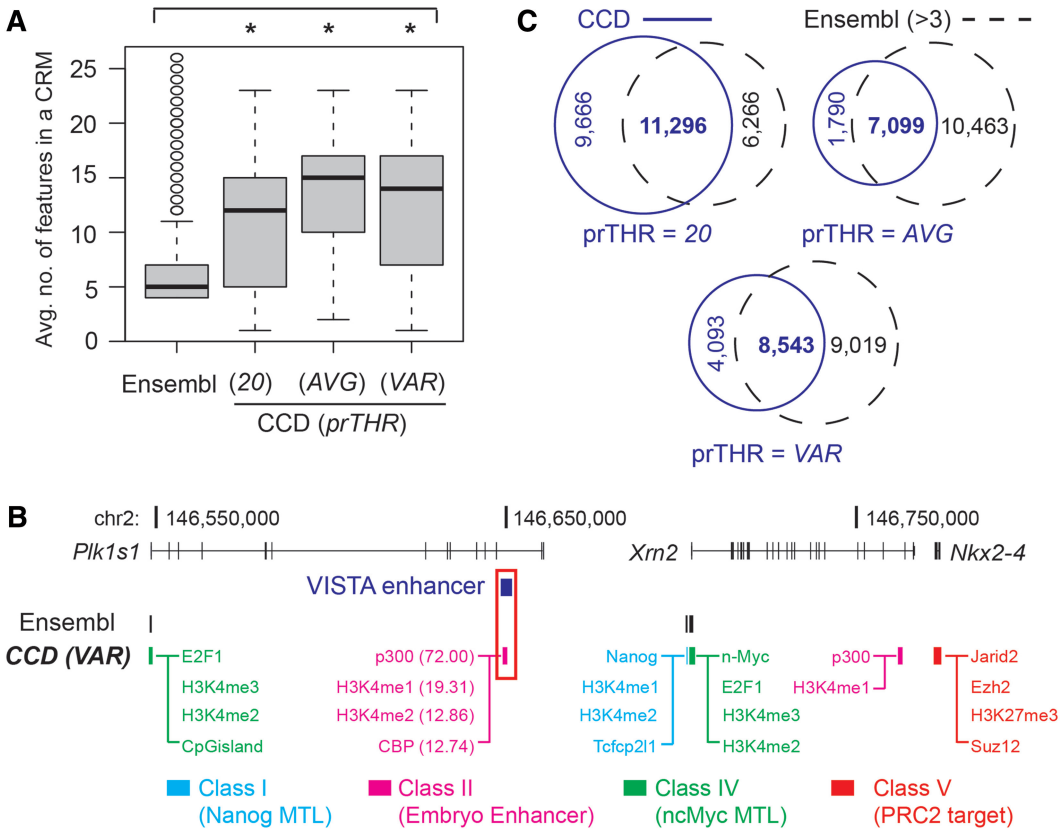

Figure 6. Comparison to the Ensembl regulatory build database. (**A**) The boxplot represents the average feature number of the CRMs which were identified by different methods (or parameters). The following prTHRs were used—20: 80% of the CRMs in a training set will be identified with putative CRMs, *AVG* (default): the cutoff score is set to average of sum of CCD scores in a training set and *VAR*: the same prTHRs used in Figure 5; $*P < 0.001$, two tailed *t*-test. (**B**) A locus (chr2:146 539 923–146 781 533, mm8) containing three genes (*Plk1s1*, *Xrn2* and *Nkx2-4*) and different types of CRMs. A CRM (blue box) showing enhancer activity experimentally is located in the 12th intron of the *Plk1s1* gene. Ensembl detected three CRMs (ENSMUSR00000182442, ENSMUSR00000182448 and ENSMUSR00000131065). CCD identified six CRMs and categorized them into four classes (cyan—class I, magenta—class II, green—class III and red—class V). The top four features which contain high CCD scores in each identified CRM are indicated by name (CCD score). Detailed information of the features' CCD score is provided as Spreadsheet 1 in Supplementary Data. (**C**) The identified CRMs by CCD and Ensembl were compared using Venn diagrams.

overlapped with the VISTA enhancer (Figure 6B). Third, CCD enables users to adjust the level of expected result by using the prTHR parameter. The default parameter (prTHR = AVG) is very stringent, and thus only 40.4% of the Ensembl regulatory elements were overlapped (Figure 6C). These numbers can be increased by lowering the prTHR parameter. With prTHR = 20, CCD identified 64.3% of the Ensembl regulatory elements. Overall, these results demonstrate that CCD is a flexible application and performs better than Ensembl at identifying the experimentally validated CRMs in VISTA.

Integrative analysis of the *cis*-regulatory modules in genome-wide level

Visualization is a powerful alternative approach to examine data in detail. Therefore, CCD provides an output ('UCSC_') to visualize all identified CRMs and features by means of the UCSC genome browser (http://genome.ucsc.edu/). With the advantage of the ChIP-seq technology, mapping of the CRMs produces a high-resolution map of genome-wide CRMs which can be a valuable source for researchers (Supplementary Figure S5). For instance, manual investigations of the regions near the transcription start site (TSS) of *Trit1*, *Coro1c* and *Klf7* genes reveal that a single CRM in each locus is occupied by p300s in embryo forebrain, midbrain and limb tissues (Supplementary Figure S6). Based on the trace code, these CRMs are highly likely to be enhancers which may affect expression of the genes in the embryo forebrain, midbrain and limb tissues.

The *H19-Igf2* and *Meg3-Dlk1* imprinted domains are known to share several intriguing features (43); (i) similar distances between the genes in each domain, (ii) non-coding RNAs (products of *H19* and *Meg3*), (iii) monoallelic expression pattern of the genes. By examining two domains on the high-resolution map, we found that both promoter regions of the protein-coding genes, *Igf2* and *Dlk1*, are bound by PRC2 in ES cells (Supplementary Figure S3). This result supports a previous report that *Igf2* expression depends on PRC2 (44) and also provides explicit evidence that these two domains are likely controlled by similar mechanisms, since both domains contain the same set of the CRMs

(ICR and PRC2 target) (Supplementary Figure S3). Interestingly, additional examination of the imprinted domains around the other 18 ICRs reveals that the promoter regions of *Rasgrf1*, *Grb10* and *Gnas* are also bound by PRC2 in ES cells (Supplementary Figures S2P–R). Further investigations are required to confirm whether PRC2 is involved in the general mechanism of genomic imprinting. Complete results for all 20 ICRs can be found in the Supplementary Figures S2 and S3.

Investigations of the regions encoding four transcription factors (OCT4, KLF4, NANOG and SOX2) expressed highly in ES cells show that CCD successfully pinpoints several CRMs regardless of genomic context (Supplementary Figure S7). For example, four Nanog MTL CRMs are located at ~1-, 2-, 3- and 15-kb upstream regions from the TSS of the *Oct4* gene. Three Nanog MTL CRMs are positioned at ~52-, 56- and 67-kb downstream from the TSS of the *Klf4* gene and three Nanog MTL CRMs are resided in ~0.1-, 4- and 42-kb away from the *Nanog* gene. Intriguingly, two CRMs (CRM3_102 and CRM3_105) positioned around the *Sox2* gene harbors two or three different trace codes (Embryo enhancer, Nanog MTL or ncMyc MTL) (Supplementary Figure S7D). This result hints that some CRMs may contain different trace codes together, and we call these 'multi-functional CRMs'.

To further elucidate the multi-functional CRMs, the five distinct classes (I–V) are intersected. Surprisingly, only 481 out of 12 636 identified CRMs overlap with at least two different trace codes demonstrating that a small number of the multi-functional CRMs do indeed exist (Spreadsheet 3 in Supplementary Data and Supplementary Figure S8). Due to the rarity, we postulate that the multi-functional CRMs may be located near genes that are critical for gene regulation. In consistent with the assumption, the functions of genes near the multi-functional CRMs are significantly related to establishment or maintenance of chromatin architecture (binomial $P = 2.4E-06$) such as *Jmjd1a, Jmjd3, Mbd3, Arid1a, Smarcc1, Smarcd1, Smarcd2* and *Chd3* (Supplementary Table S4). Among the genes, *Jarid2* should be specifically expressed during development according to its critical role involved in global gene silencing (40,41). Interestingly, five CRMs (four Nanog MTLs and one ncMyc MTL) and one multi-functional CRM (Nanog MTL/Embryo enhancer) are located within regions spanning from −53 to +22 kb around the gene's TSS, indicating that the gene appears to be regulated by a complex *cis*-regulatory network (Figure 7). The multi-functional CRM is highly conserved and bound by TCFCP2L1, NANOG, OCT4, SMAD1, SOX2 in early-stage embryo (ES cells) and CBP, p300 in later-stage embryo (cortical neurons, forebrain and midbrain). Based on the trace codes, this CRM appears to regulate the spatial and temporal expression of *Jarid2* by interacting with the above combinations of TFs in two different development stages, although additional investigations are required. The complete map of the identified genome-wide CRMs can be viewed on the website (http://decode.kaist.ac.kr/).

DISCUSSION

The progressive increase of the genome-wide data sets, especially from the ChIP-seq method, gives rise to a need for novel applications which fully exploit the data sets for particular purposes. Although integrative analysis of the genome-wide data sets holds great potential (45–48), there are no generalized methodologies to integrate a variety of genome-wide data sets in an unbiased manner. To resolve the above issue and apply it for identifying genome-wide CRMs, combinatorial CRM decoder (CCD) has been developed. As a generalized platform, CCD has several remarkable advantages. First, any kind of 'genome-wide' data sets can be used as the feature sets, since it independently models background distribution of each feature based on the negative binomial distribution coupled with the CDRS method (Figure 1). Second, previously unnoticed relationships between epigenetic features and CRMs can be identified by analyzing various data sets altogether. Owing to the rapid growth of genome-wide ChIP-seq data sets, this property will increasingly accelerate the identification of new associations between the epigenetic features and CRMs without prior knowledge. For example, based on the extensive binding of the NR5A2 and TBX3 to the Nanog MTLs (35.7 and 10.9% of the CRMs), we postulate that TCFCP2L1 may improve the reprogramming efficiency further due to its significant association with the Nanog MTLs (56.4% of the CRMs) compared to the pseudo set (0.3% of the pseudo-CRMs) (Spreadsheet 2 in Supplementary Data). Third, the performance can be superior to the other CRM prediction tools due to the basic data sources, experimentally derived (ChIP-seq) datasets rather than computational predictions. Fourth, the biological relevance of identified CRMs can easily be assessed with the available tools including R, GREAT and UCSC genome browser (Figure 2). Furthermore, it can also be used to decode the genomes of other species by utilizing appropriate input data sets. The CCD program and tutorials can be found on our website (http://decode.kaist.ac.kr/).

In the present study, our extensive evaluations demonstrated that the algorithm of CCD is robust and reliable. By using the MCC value, CCD will automatically discriminate features for the training set. If there is a single significant feature among input features, then the identified CRMs will be the binding regions of the single feature, which might be biased due to the lack of information. In this regard, we believe that a variety of features results in better outcomes as shown in this study (Figure 6A). Subsequent analysis of a large number of various data sets further verified the reliability of the algorithm by identifying all previously known features (100%, 29/29) as parts of the trace codes (Figure 3A and Supplementary Table S2). These remarkable performances are based on the 'trace code system'. With 9 training sets (Table 1), we showed that the trace code is sufficient to represent the characteristics of CRMs (Figure 3 and Supplementary Table S2). Accordingly, it enabled us to identify genome-wide CRMs including the PRC2 target sites (Spreadsheet 3 in Supplementary) in an unbiased manner.

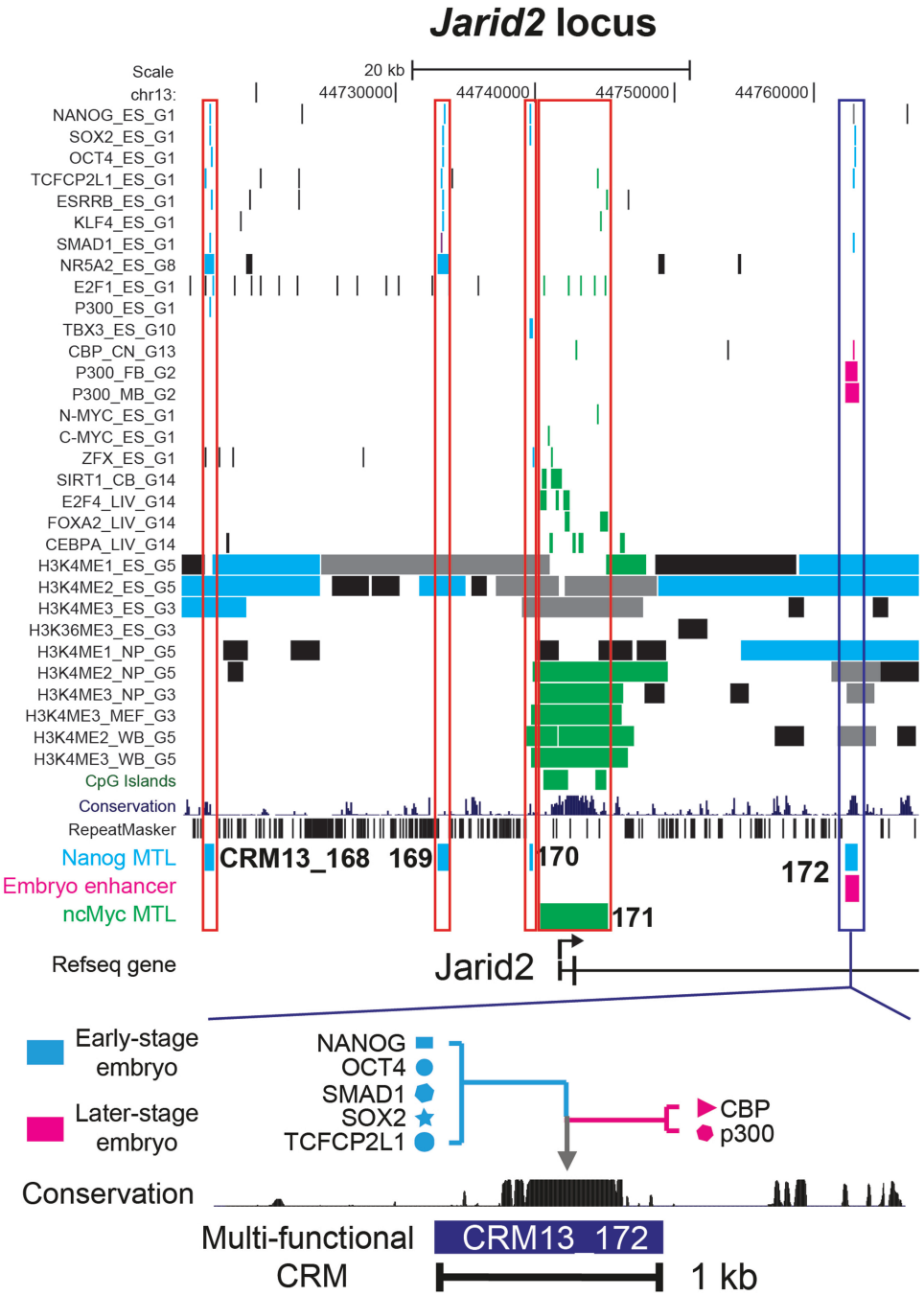

Figure 7. Identification of multi-functional CRMs. Visualization of the identified CRMs unveils a *cis*-regulatory network around the transcription start site of the *Jarid2* gene. CRM13_168, CRM13_169, CRM13_170 and CRM13_171 contain a single trace code (class I or IV), whereas CRM13_172 (multi-functional CRM, blue vertical box) includes two distinct trace codes, class I and II. The CRM13_172 is bound by various TFs in different developmental stages; early-stage embryo (ES cells) and later-stage embryo (cortical neurons, forebrain and midbrain). Bars represent genomic positions of features or CRMs. Colors correspond to the following classes and their associated trace codes (Figure 4)—green (class IV), cyan (class I), magenta (class II) and grey (shared by two classes).

The virtue of integrative analysis using CCD leads to unexpected findings including the ICR signature and multi-functional CRMs. The imprinting control regions have been known to be associated with the active (H3K4me3) and repressive (H3K9me3 and H4K20me3) histone modifications in allele-specific manner (37,38). Our results are well correlated with the previous reports and further suggest that ESET is a potential key factor involved in the mechanism of genomic imprinting (Supplementary Figures S2 and S3). However, Eset, H3K9me3 and H4K20me3 are also the signature of the endogenous retroviruses (ERVs) (49). Given this similarity, it should be interesting to test whether the mechanism of silencing the ERVs and maintaining (or establishing)

the ICRs are mediated by the same regulatory complexes. Another intriguing finding is the multi-functional CRMs (Figure 7 and Supplementary Figure S8). Although the analysis of data sets from various sources may lead to the identification of false positive CRMs, we showed that our approach is very effective and eventually discovers the multi-functional CRMs. Therefore, we argue that the data sets from different cell types still provide characteristic patterns of CRMs in the given time-point and can be used at least for identifying CRMs. Based on the distinct trace codes, the multi-functional CRMs belong to at least two different classes (Spreadsheet 3 in Supplementary Data). We propose that they are likely the key CRMs which determine the temporal and spatial expression of nearby genes by interacting with more than two combinations of TFs (input signals). Further investigations are needed to elucidate whether the multi-functional CRMs represent the general property of CRMs or a special type of CRMs, since the CRMs tend to harbor multiple TFBSs.

With the great capability of the integrative analysis, CCD will shed light on unveiling the gene regulatory networks by assisting the growth of genome-wide association studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Jennifer M. Huang for her excellent editing.

FUNDING

Epigenomic Research Program for Human Stem Cells (2007-2004134); Research Program for New Drug Target Discovery (2007-0052983); Ministry of Education, Science & Technology, South Korea. Funding for open access charge: Brain Korea 21.

Conflict of interest statement. None declared.

REFERENCES

1. Levine,M. and Davidson,E.H. (2005) Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA*, **102**, 4936–4942.
2. Segal,E., Raveh-Sadka,T., Schroeder,M., Unnerstall,U. and Gaul,U. (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, **451**, 535–540.
3. Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.*, **12**, 130–136.
4. Stathopoulos,A. and Levine,M. (2005) Genomic regulatory networks and animal development. *Dev. Cell*, **9**, 449–462.
5. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
6. Ochoa-Espinosa,A. and Small,S. (2006) Developmental mechanisms and *cis*-regulatory codes. *Curr. Opin. Genet. Dev.*, **16**, 165–170.
7. Janga,S.C., Collado-Vides,J. and Babu,M.M. (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc. Natl Acad. Sci. USA*, **105**, 15761–15766.
8. Kang,K., Chung,J.H. and Kim,J. (2009) Evolutionary conserved motif finder (ECMFinder) for genome-wide identification of clustered YY1- and CTCF-binding sites. *Nucleic Acids Res.*, **37**, 2003–2013.
9. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E. (2009) Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature*, **462**, 65–70.
10. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
11. Kim,T.K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
12. Kantorovitz,M.R., Kazemian,M., Kinston,S., Miranda-Saavedra,D., Zhu,Q., Robinson,G.E., Göttgens,B., Halfon,M.S. and Sinha,S. (2009) Motif-blind, genome-wide discovery of *cis*-regulatory modules in *Drosophila* and mouse. *Dev. Cell*, **17**, 568–579.
13. Won,K.J., Agarwal,S., Shen,L., Shoemaker,R., Ren,B. and Wang,W. (2009) An integrated approach to identifying *cis*-regulatory modules in the human genome. *PLoS ONE*, **4**, e5501.
14. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
15. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
16. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
17. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
18. Peng,J.C., Valouev,A., Swigut,T., Zhang,J., Zhao,Y., Sidow,A. and Wysocka,J. (2009) Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell*, **139**, 1290–1302.
19. Ku,M., Koche,R.P., Rheinbay,E., Mendenhall,E.M., Endoh,M., Mikkelsen,T.S., Presser,A., Nusbaum,C., Xie,X., Chi,A.S. *et al.* (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.*, **4**, e10000242.
20. Pasini,D., Cloos,P.A., Walfridsson,J., Olsson,L., Bukowski,J.P., Johansen,J.V., Bak,M., Tommerup,N., Rappsilber,J. and Helin,K. (2010) JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature*, **464**, 306–310.
21. MacIsaac,K.D., Lo,K.A., Gordon,W., Motola,S., Mazor,T. and Fraenkel,E. (2010) A quantitative model of transcriptional regulation reveals the influence of binding location on expression. *PLoS Comput. Biol.*, **6**, e10000773.
22. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
23. Wederell,E.D., Bilenky,M., Cullum,R., Thiessen,N., Dagpinar,M., Delaney,A., Varhol,R., Zhao,Y., Zeng,T., Bernier,B. *et al.* (2008) Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.
24. Nishiyama,A., Xin,L., Sharov,A.A., Thomas,M., Mowrer,G., Meyers,E., Piao,Y., Mehta,S., Yee,S., Nakatake,Y. *et al.* (2009) Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell*, **5**, 420–433.

25. Yuan,P., Han,J., Guo,G., Orlov,Y.L., Huss,M., Loh,Y.H., Yaw,L.P., Robson,P., Lim,B. and Ng,H.H. (2009) Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev.*, **23**, 2507–2520.
26. Han,J., Yuan,P., Yang,H., Zhang,J., Soh,B.S., Li,P., Lim,S.L., Cao,S., Tay,J., Orlov,Y.L. *et al.* (2010) Tbx3 improves the germ-line competency of induced pluripotent stem cells. *Nature*, **463**, 1096–1100.
27. Heng,J.C., Feng,B., Han,J., Jiang,J., Kraus,P., Ng,J.H., Orlov,Y.L., Huss,M., Yang,L., Lufkin,T. *et al.* (2010) The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell*, **6**, 167–174.
28. Landeira,D., Sauer,S., Poot,R., Dvorkina,M., Mazzarella,L., Jørgensen,H.F., Pereira,C.F., Leleu,M., Piccolo,F.M., Spivakov,M. *et al.* (2010) Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. *Nat. Cell Biol.*, **12**, 618–624.
29. Li,G., Margueron,R., Ku,M., Chambon,P., Bernstein,B.E. and Reinberg,D. (2010) Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev.*, **24**, 368–380.
30. Azuara,V., Perry,P., Sauer,S., Spivakov,M., Jørgensen,H.F., John,R.M., Gouti,M., Casanova,M., Warnes,G., Merkenschlager,M. *et al.* (2006) Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.*, **8**, 532–538.
31. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
32. Eskeland,R., Leeb,M., Grimes,G.R., Kress,C., Boyle,S., Sproul,D., Gilbert,N., Fan,Y., Skoultchi,A.I., Wutz,A. *et al.* (2010) Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell*, **38**, 452–464.
33. Ivanova,N., Dobrin,R., Lu,R., Kotenko,I., Levorse,J., DeCoste,C., Schafer,X., Lun,Y. and Lemischka,I.R. (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538.
34. Reik,W. and Walter,J. (2001) Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, **2**, 21–32.
35. Edwards,C.A. and Ferguson-Smith,A.C. (2007) Mechanisms regulating imprinted genes in clusters. *Curr. Opin. Cell Biol.*, **19**, 281–289.
36. Bartolomei,M.S. (2009) Genomic imprinting: employing and avoiding epigenetic processes. *Genes Dev.*, **23**, 2124–2133.
37. Delaval,K., Govin,J., Cerqueira,F., Rousseaux,S., Khochbin,S. and Feil,R. (2007) Differential histone modifications mark mouse imprinting control regions during spermatogenesis. *EMBO J.*, **26**, 720–729.
38. Regha,K., Sloane,M.A., Huang,R., Pauler,F.M., Warczok,K.E., Melikant,B., Radolf,M., Martens,J.H., Schotta,G., Jenuwein,T. *et al.* (2007) Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. *Mol. Cell*, **27**, 353–366.
39. Soshnikova,N. and Duboule,D. (2009) Epigenetic temporal control of mouse *Hox* genes in vivo. *Science*, **324**, 1320–1323.
40. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
41. Simon,J.A. and Kingston,R.E. (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.*, **10**, 697–708.
42. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
43. Paulsen,M., Takada,S., Youngson,N.A., Benchaib,M., Charlier,C., Segers,K., Georges,M. and Ferguson-Smith,A.C. (2001) Comparative sequence analysis of the imprinted *Dlk1-Gtl2* locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the *Igf2-H19* region. *Genome Res.*, **11**, 2085–2094.
44. Li,T., Hu,J.F., Qiu,X., Ling,J., Chen,H., Wang,S., Hou,A., Vu,T.H. and Hoffman,A.R. (2008) CTCF regulates allelic expression of *Igf2* by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loop. *Mol. Cell Biol.*, **28**, 6473–6482.
45. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
46. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
47. Alexander,R.P., Fang,G., Rozowsky,J., Snyder,M. and Gerstein,M.B. (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, **11**, 559–571.
48. Hawkins,R.D., Hon,G.C. and Ren,B. (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, **11**, 476–486.
49. Matsui,T., Leung,D., Miyashita,H., Maksakova,I.A., Miyachi,H., Kimura,H., Tachibana,M., Lorincz,M.C. and Shinkai,Y. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature*, **464**, 927–931.