Published online 11 December 2012

Nucleic Acids Research, 2013, Vol. 41, No. 3 ***e48***
doi:10.1093/nar/gks1214

EMdeCODE: a novel algorithm capable of reading words of epigenetic code to predict enhancers and retroviral integration sites and to identify H3R2me1 as a distinctive mark of coding versus non-coding genes

Federico Andrea Santoni*

Department of Genetic Medicine and Development, University of Geneva, 1211 Geneva 4, Switzerland

Received May 15, 2012; Revised and Accepted October 30, 2012

ABSTRACT

Existence of some extra-genetic (epigenetic) codes has been postulated since the discovery of the primary genetic code. Evident effects of histone post-translational modifications or DNA methylation over the efficiency and the regulation of DNA processes are supporting this postulation. EMdeCODE is an original algorithm that approximate the genomic distribution of given DNA features (e.g. promoter, enhancer, viral integration) by identifying relevant ChIPSeq profiles of post-translational histone marks or DNA binding proteins and combining them in a supermark. EMdeCODE kernel is essentially a two-step procedure: (i) an expectation-maximization process calculates the mixture of epigenetic factors that maximize the Sensitivity (recall) of the association with the feature under study; (ii) the approximated density is then recursively trimmed with respect to a control dataset to increase the precision by reducing the number of false positives. EMdeCODE densities improve significantly the prediction of enhancer loci and retroviral integration sites with respect to previous methods. Importantly, it can also be used to extract distinctive factors between two arbitrary conditions. Indeed EMdeCODE identifies unexpected epigenetic profiles specific for coding versus non-coding RNA, pointing towards a new role for H3R2me1 in coding regions.

INTRODUCTION

The role of epigenetic mechanisms as modulators of transcriptional activation and repression, DNA methylation, cell memory maintenance, stem cell differentiation, cancerogenesis and other major genomic processes have been widely demonstrated [reviewed in (1–3)]. Moreover, it has been already observed that epigenetic features such as histone post-translational modifications and DNA methylation can act in concert to promote or silence specific cellular functions (4). Accordingly, I have recently shown that retroviral insertional site selection into the host DNA could be strongly influenced by a set of specific histone modifications acting as a sort of beacon for incoming retroviruses (5). This information could be taken as the prelude to postulate the existence of a not-yet-well-defined epigenetic code (6).

Here I propose a statistical approach to identify patterns of epigenetic signatures associated with specific genomic functions. Relatively new technologies, notably ChIPSeq, are able to track the position of DNA binding proteins as transcription factors or histones carrying specific modifications and yield a genome-wide density profile. When searching for association to a given genomic process, this information might be probabilistically interpreted and combined to generate a new virtual mark of greater statistical power with respect to the 'real' mark alone. Several approaches have been developed. Some (7–10) are based on unsupervised genome segmentation where the aim is to find clusters of epigenetic factors repeated over the genome (transcription factor binding sites, histone marks, DNase profile etc.). These patterns are then associated 'a posteriori' with known annotations as promoter or enhancer regions. As another example, (11) uses Bayesian networks to explores causal relationships among clusters of DNA binding proteins in Drosophila and infers direct or indirect interactions among them. In general, unsupervised methods are useful to label large portions of the genome but they lack a false-positive correction resulting in poor performance especially in terms of positive predicted value (PPV or Precision <0.5) (10). On the other hand, supervised methods use training subsets of specific functional regions (enhancers

*To whom correspondence should be addressed. Tel: +41 223795719; Fax: +41 223795706; Email: federico.santoni@unige.ch and promoters in most cases) to predict the genomic position of other unknown functional sites (12,13), and they use a control (random) dataset to increase the Precision. Additionally these methods yield better Sensitivity (Recall) than unsupervised methods (12). Previously I implemented a supervised heuristic methodology based on F score statistics to create a 'supermarker' out of ChIPSeq profiles (5). This methodology requires the pre-calculation of individual F scores for each mark, the selection of a subset of significant marks and the eventual addition or deletion of some marks to the subset. EMdeCODE is a general extension of this idea, based on mixture modeling Expectation and Maximization (EM) and statistical selection. It is specifically designed to deal with post-translational histone modifications and it presents some peculiar features. The input data is a set of peaks calculated by an independent peak caller from row ChIPSeq data [similarly to (9)]. The rationale behind this design is to decouple the peak calling task that can be accomplished by several continuously improved algorithms [see (14) for a review] from the statistical combination of histone marks. The advantage is 2-fold: (i) EMdeCODE is independent from current and likely soon obsolete treatments of ChIPSeq data; (ii) EMdeCODE can work with peaks called from different algorithms and from different experiments. The peak caller can be chosen accordingly with the expected ChIPSeq profile and the experimental conditions, for example HOMER (15) for narrow peaked marks as H3K4me3, and RSEG (16) for wider distribution as H3K27me3. Differently from all other approaches, for each input peak set, EMdeCODE recreates a new histone mark distribution by interpreting each peak as a single or stretch of nucleosomes (accordingly to the genomic size of the peak) and modeling the nucleosome occupancy by Gaussian functions of equal amplitude and variance (= 200 bp). The idea here is to drastically reduce the experimental noise and facilitate the interaction of data obtained by different experiments. These new distributions are then combined during the training phase to generate a genome-wide probability mass distribution that approximate the unknown functional distribution. Here, another peculiarity of EMdeCODE is the maximization of the *F* Score, a weighted combination of precision (positive predictive value) and recall (Sensitivity), to reduce the influence of false negatives similarly to what presented in (5). The main difference is the addition of the EM procedure that combines a set of factors and selects the relevant subset without supervision. As shown in the Results section, a consequence of this approach is a gain of precision with respect to other methods. EMdeCODE can therefore be used to discover new associations between marks and genomic functions. Applications to enhancer identification in comparison with previous supervised methods, retroviral integration site prediction and to discriminate between coding versus non-coding genes are discussed.

MATERIALS AND METHODS

Input data

High-Throughput Sequencing–based applications usually require large amount of computational power and software tools. The minimal computational ChIPSeq pipeline consists of a sequence aligner that maps millions of reads produced by the sequencer to the proper reference genome and a peak finder to identify enriched regions with respect to a control background according to some user pre-defined parameters [see (17) for a review about ChIPSeq and other similar technologies]. EMdeCODE takes as input the final peak set and assumes that these steps are correctly performed with proper controls and sufficient profile quality. In particular, all ChIPSeq histone profiles used in this study have been pre-processed with Fseq (18) for it produces good-quality peak set from both narrow and broad read distributions (14).

The algorithm

Here *M* DNA-binding factors are considered. Formally, each ChIPSeq profile for factor X_j could be interpreted as the probability to find X_j bound to a specific genomic region of size w centered on $i = (chr, pos)$: $p_{X_j} \equiv p(X_j = i)$. This mass density is modeled by EMdeCODE with a sum of Gaussian functions, each one centered on one nucleosome. Nucleosome positions are extracted from ChIPSeq peaks by considering a genomic occupancy of 200 bp. This reduces the potential bias that can arise by combining ChIPSeq densities obtained over different experimental conditions. Briefly, each marker probability density function is written as

$$p(X = i) = \frac{1}{C}\sum_{p\in\Gamma} e^{\frac{-(i-p)^2}{2\sigma^2}}, \qquad (\mathbf{1})$$

where Γ is the peak set of factor X, and C is a normalization factor.

Similarly, the biological event could be modeled as a random variable B taking values on all chromosomal loci. In other words, the event modeled by B has an unknown (mass) probability to occur in the region centered on i: $p_B \equiv p(B = i)$.

The aim is to find a new mass probability functional that can approximate p_B by means of the p_{X_j}'s: $p_B \cong \Im[p_{X_1}, p_{X_2}, ..., p_M]$.

Assuming that B has been observed to occur on N loci $\{k_1,...,k_N\}$, p_B can be written as

$$\begin{aligned} p_B = p(B = k) &= \sum_i p(B = k|X_j = i)p(X_j = i) \\ &= p(B=k|X_j=k)p(X_j=k) + \sum_{i\neq k} p(B=k|X_j=i)p(X_j=i) \end{aligned} \qquad (\mathbf{2})$$

where i spans across all possible loci.

The first right term in (2) can be interpreted as how well p_{X_j} fits p_B. The second term is the related error expressing the spreading of p_{X_j} over loci unrelated to B.

Summing (2) over j:

$$\begin{aligned} p(B = k) &= \sum_j^M \frac{p(B = k|X_j = k)}{M} p(X_j = k) \\ &+ \sum_j \sum_{i\neq k} \frac{p(B= k|X_j = i)}{M} p(X_j = i) = \tilde{p}_{B|\Theta} + \varepsilon, \end{aligned} \qquad (\mathbf{3})$$

where $\tilde{p}_{B|\Theta} = \sum_j^M \alpha_j p(X_j)$, $\Theta = (\alpha_1,...,\alpha_M)$ and ε is the aggregate of all M approximation errors.

Θ can be estimated by a classical Maximum Likelihood approach. Indeed, maximizing the (log)likelihood of $\tilde{p}_{B|\Theta}$:

$$\log(L(\Theta|B)) = \log(\tilde{p}_{B|\Theta}) = \log\left(\sum_j^M \alpha_j p(X_j)\right).$$

is equivalent to a mixture-density parameter estimation problem and can be efficiently treated with the Expectation Maximization algorithm (EM) (19).

This algorithm works in a way that, at the t-esim iteration, the expectation function $Q(\Theta,\Theta^{(t-1)}) = E_Y\left[\log L(\Theta|B,Y)|B,\Theta^{(t-1)}\right]$ is maximized (Y is an auxiliary random variable), that is, $\Theta^{(t)} = \arg\max_\Theta Q(\Theta,\Theta^{(t-1)})$. As previously derived in (19), by introducing the probability that value k arises from the j-esim distribution,

$$P(j|k,\Theta) = \frac{\alpha_j p(X_j = k)}{\sum_l^M \alpha_l p(X_l = k)}, \tag{4}$$

$Q(\Theta,\Theta^{(t-1)})$ can be written as

$$Q(\Theta,\Theta^{(t-1)}) = \sum_j \sum_k \log(\alpha_j) p(j|k,\Theta^{(t-1)}) \tag{5}$$

The mixture coefficient vector Θ is evaluated by maximizing (5) by the Lagrange multiplier λ with constraints $\sum_j \alpha_j = 1$ and (4):

$$\begin{aligned}\alpha_j^{(t)} &= \frac{1}{N}\sum_k P(j|k,\Theta^{(t-1)}) = \frac{1}{N}\sum_k \frac{\alpha_j^{(t-1)} p(X_j = k)}{\tilde{p}(B = k|\Theta^{(t-1)})} \\ &= \frac{1}{N}\sum_k \frac{\alpha_j^{(t-1)} p(X_j = k)}{\sum_l^M \alpha^{(t-1)} p(X_l = k)}\end{aligned} \tag{6}$$

where N is the number of discrete genomic loci all densities have support.

From (3) the following identity holds:

$$\alpha_j^{(t-1)} p(X_j = k) = \frac{1}{M} p(B = k, X_j = k|\Theta^{(t-1)})$$

Introducing the concept of Sensitivity (Recall), defined as $R = \frac{P(B,X)}{P(B)}$, equation (6) can be simply written as:

$$\alpha_j^{(t)} = \frac{1}{M} P(B|X_j,\Theta^{(t-1)}) = \frac{1}{M} R_j(\Theta^{(t-1)}) \tag{7}$$

In other words, each factor is weighted proportionally to the respective Sensitivity.

The spreading error ε can be interpreted as the affinity of factor M_j for a control dataset C, defined as the set of random loci where $p(B = c) = 0, \forall c \in C$.

Indeed, considering a control dataset with N_C control loci, the spreading error could be estimated as:

$$\begin{aligned}\varepsilon &= \sum_j \sum_{i \neq k} \frac{p(B = k|X_j = i)}{M} p(X_j = i) \\ &\cong \sum_j \sum_i^{N_C} \frac{p(C = i|X_j = i)}{M} p(X_j = i) = \frac{1}{MN_C} \sum_j fp_j = \langle fp \rangle\end{aligned}$$

where $fp_j = N_C P(C,X_j)$ is the number of false positives, that is, the number of control loci localizing in the region where factor X_j is bound. Therefore, minimizing ε implies the minimization of false positives.

The strategy adopted here to approximate p_B is a two-step procedure where $\tilde{p}_{B|\Theta}$ is calculated by (6) and low-probability peaks of $\tilde{p}_{B|\Theta}$ are trimmed out to reduce the number of false positives (the whole procedure is sketched in Figure 1). Obviously this could also reduce the Sensitivity, that is, the number of peaks associated to B. To measure the quality of the approximation, the F_β score is evaluated after trimming to find the optimal tradeoff between Sensitivity and false positives reduction (Precision).

Formally, the F_β score is defined as the β-weighted harmonic mean of Precision(P) and Sensitivity(R): $F_\beta \equiv \left(1+\beta^2\right)\frac{PR}{\beta^2 P+R}$.

Here, β = 0.5 to give more weight to Precision than to Sensitivity. This balances type I and type II errors by adjusting for the high rate of False Positives (*fp*) inherent to the examination of large datasets for genome-wide binding sites according to statistical significance. Moreover, to make the F score insensitive to data size, the number of false positives *fp* has been normalized with respect to N [F score–based statistics and comparison with other measures have been extensively discussed in (5)].

Eventually, the algorithm is lined out as follows:

(1) $\tilde{p}_{B|\Theta}$ is evaluated by (6). This corresponds to weighting each factor distribution with the respective Sensitivity.
(2) $\tilde{p}_{B|\Theta}$ is trimmed to reduce the number of false positives by removing small peaks until $F_{0.5}$ score is maximized. Accordingly, the support of mass probability functions $p(X_j = i)$ is reduced.
(3) If $F_{0.5}$ score has increased, step 1) is repeated or else it ends.

EMdeCODE has been implemented in Python and can be downloaded from http://seaseq.unige.ch/~fsantoni/EMdeCODE.

RESULTS

Comparison with previous methods

Cross validation

To compare the performance of EMdeCODE with previously published algorithms, a 5-fold cross validation has been implemented over a set of 74 well-defined enhancers according to the experimental settings proposed in (20), but replacing ChIP-ChIP signals for H3K4me3 and H3K4me1 with the corresponding ChIPSeq profiles (21). As background, 740 random genomic sites have been selected. EMdeCODE scored a PPV of 96.9% ± 1.1 with an optimal window of 300 bp. Notably, this window is considerably smaller than 2000 bp reported elsewhere (12,13,20), probably owing to the higher resolution provided by ChIPSeq data. Comparison with previous methods is reported in Table 1.

Enhancer identification

EMDdeCODE has also been used to reconstruct the p300 binding sites distribution in CD4+ T cells, as a strong

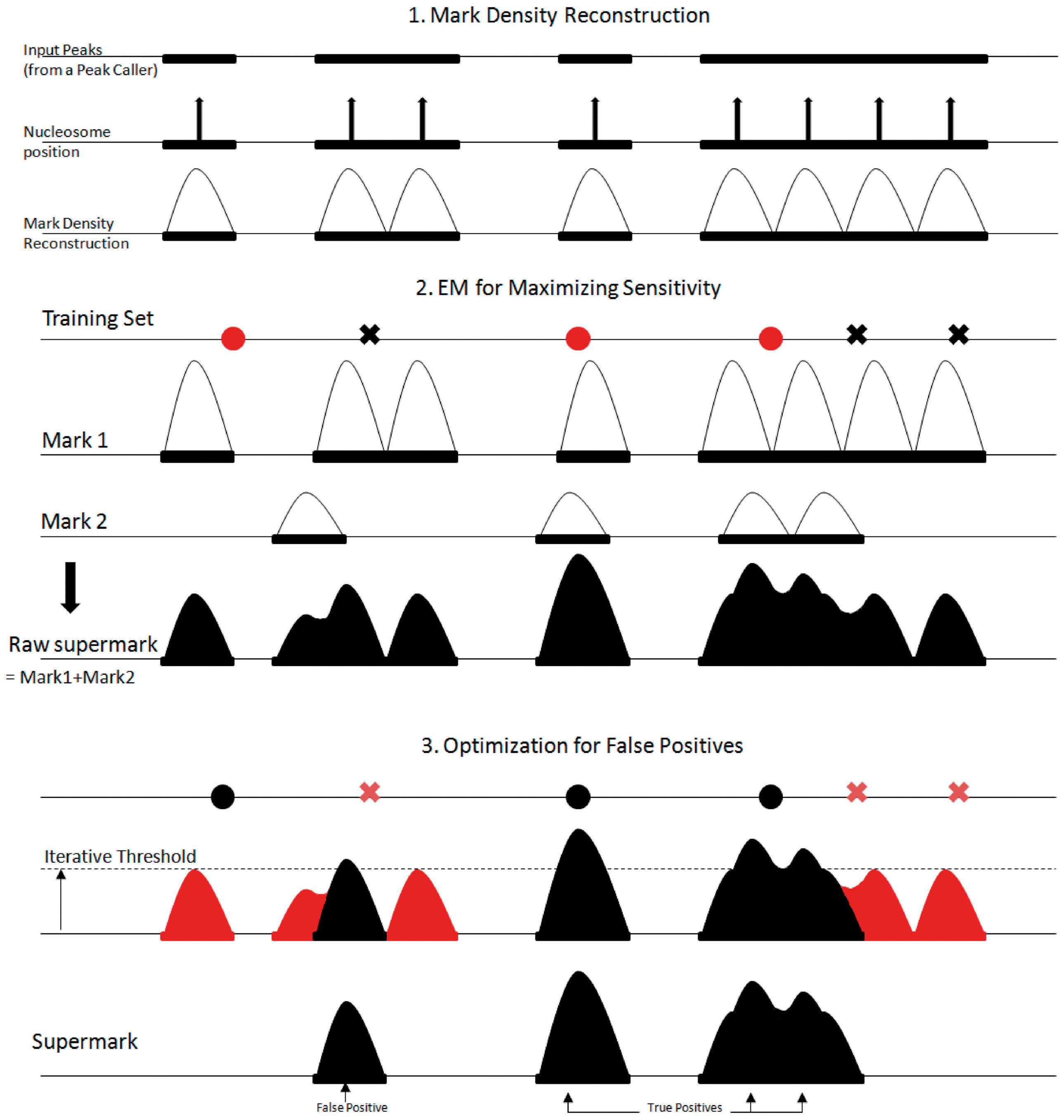

Figure 1. (1) EMdeCODE reconstructs mark density distributions by placing the appropriate number of nucleosomes into each peak region and modeling nucleosome occupancy by a Gaussian distribution with variance 200 bp. (2) Expectation Maximization evaluates the sensitivity of each mark to true positives (filled round dots). Marks are weighted accordingly and linearly combined to generate the raw supermark. (3) Peaks of the supermark accounting for random loci (crosses: false positives) are trimmed off the supermark to optimize the *F* score.

marker for enhancer identification (22). Thirty-nine ChIPSeq histone modifications (acetylation and methylation) profiles have been used as input (see 'Materials and Methods' section). Similarly to the analysis proposed by Firpi *et al.* (12), EMdeCODE was trained with 213 p300 distal binding sites [at least 2.5 Kbp from a known Transcription Starting Site (TSS)] overlapping PreMod (23) predicted enhancers and 2130 random genomic sites. Here, two loci are considered as overlapping if their genomic distance is <2 Kbp.

EMdeCODE generated 22 878 whole genome putative enhancer sites, 53% overlapping (12 165) with the 36 796 sites generated by CSI-ANN [(12), data in Supplementary Material]. To validate this prediction, the EMdeCODE generated enhancer dataset has been compared with the whole genome distal p300-enriched regions [>2.5 Kbp from TSS, calculated by SICER (24)]. In all, 74% (6069/8245) of p300 regions overlap with EMdeCODE predictions, 29% more than CSI-ANN (45%, 3740/8245). When compared with a background dataset to evaluate the number of false positives (70 000 random selected genomic sites), EMdeCODE predictions obtained an F score of 0.91 (Precision = PPV = 96%, Recall = Sensitivity = 74%), whereas the CSI-ANN F score was 0.74 (Precision = PPV = 89%, Recall = Sensitivity = 45%). This is somehow expected because EMdeCODE aims to maximize the F score. A visual comparison between EMdeCODE and CSI-ANN is reported in Figure 2 by Chromosome Projection Mandalas (5). It is interesting to observe that, consequently, EMdeCODE improves both PPV and Sensitivity. Previously it has been shown that enhancers are enriched with DNase I sensitive sites (13,22,25). Indeed, 80% of EMdeCODE predictions is supported by at least one among p300, DNase sites (26) and PreMod predicted functional sites, 20% more than CSI-ANN (60%). As additional comparison, 49 746 p300 distal binding sites were obtained by peak calling from the original p300 ChIPSeq data. In all, 18 759 (82%) EMdeCODE-generated sites were supported by p300 peaks, 45% more than CSI-ANN (37%). Together, these results indicate that EMdeCODE is a more effective algorithm to approximate the p300 distribution and a superior enhancer predictor.

Enhancer 'Code'

The composition of the enhancer 'code' generated by EMdeCODE (reported in Supplementary Figure S1A) is dominated by H3K36ac, an highly conserved histone modification already reported to be strongly enriched at enhancer sites (27,28) followed by H2BK120 and H3K4ac, two marks associated with active enhancers and promoters (29). Notably, EMdeCODE enhancers are also enriched in H3K4me3 (76%) and H3K27ac (70%) and depleted in H3K27me3 (<1%), indicating that most of them are likely active enhancers (30,31).

Table 1. Positive Predicted Value comparison among enhancer predicting algorithms in a 5-fold cross-validation test

Method (H3K4me1, H3K4me3)	Enhancer PPV(±SD)[a]
EMdeCODE	96.9% ± 1.10
CSI-ANN (12)	96.22% ± 2.14
HMM (20)	94.1% ± 0.89
Heintzmann *et al.* (13)	85%

[a] Where available.

Interestingly, this particular composition depends on modeling ChIP-Seq profiles as probability mass distributions and specifically on the normalization $\int_{-\infty}^{+\infty} p(x)dx = 1$. When min-max normalization is applied (i.e. $C = \max \sum_{p\in\Gamma} e^{\frac{-(i-p)^2}{2\sigma^2}}$ in eq. 0), EMdeCODE generates 45 616 putative enhancers, 16 961 (out of 22 878, 75%) overlapping with those ones produced by standard normalization. In this case the composition of this supermark is dominated by the well known marks H2AZ, H3K9me1, H3K4me1 and H3K4me3 (Supplementary Figure S1B). Notably, the performance of this prediction is slightly better in Sensitivity (77%, 6374/8244 of p300 regions overlap with this prediction) but not in Precision (89%), leading to a lower F score = 0.86. For this reason, the standard normalization has been preferred in EMdeCODE implementation.

Retroviral integration site prediction

EMdeCODE has been used to generate new virtual marks by combining 39 different histone marks previously extracted from CD4+ T cells [methylation (32) and acetylations (29)] with 6 histone marks from HeLa cells (21). These new marks have been tested for association with several gammaretroviral integration datasets with respect to randomly generated matched control datasets [similarly to (5)].

Indeed, EMdeCODE improved the prediction of retroviral integration sites with respect to (5). The Sensitivity (Recall) increases 5% (to 10%) explaining up to 85% of integration sites in CD4+ T cells. Accordingly, the related F score increases 3% (to 7%, Figure 3). The mixture coefficient vector Θ, calculated for each chromosome, stores the contribution of each mark to the final association. As expected, the optimal receipt includes all marks related to active transcription and open chromatin, whereas

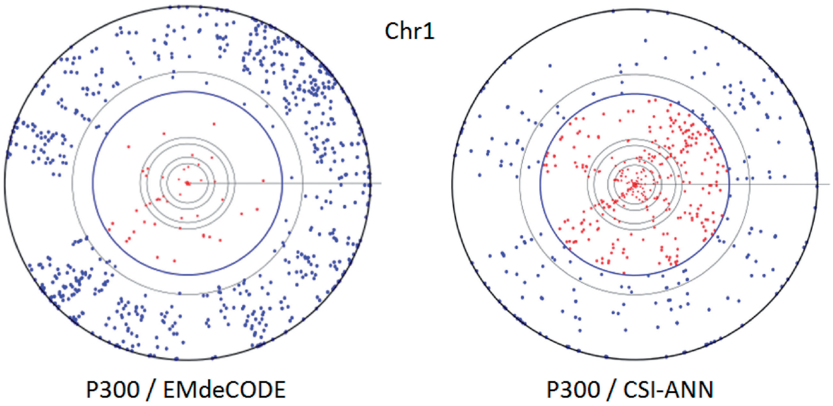

Figure 2. Chromosome projection mandalas calculated within 2 Kbp for EMdeCODE and CSI-ANN enhancer predictions versus p300 enriched regions on chr1. Each dot on the mandala represents the center of a p300 region in polar coordinates. Angular distance corresponds to genomic location, and radial distance from the outer circle indicates the log-scaled distance in nucleotides from the closest predicted enhancer. Inner circles mark 1, 2 (in blue), 10, 20 Kbp and so on. Blue filled circles represent p300 sites within 2 Kbp from the nearest predicted enhancer.

H3K27me3 and other heterochromatin-related marks give no contribution (Supplementary Figure S2). All results have been checked using a 5-fold cross-validation strategy similar to what presented in (5).

Discriminating between coding and non-coding DNA

Another possible application of EMdeCODE is the identification and the quantification of differential epigenetic profiles between two distinct genomic features. Here, the algorithm has been used to identify epigenetic differences that could characterize coding from non-coding transcripts as, in this case, long non-coding RNA (lncRNA). It is believed that non-coding mRNAs transcribed by the PolII transcriptional machine have the same epigenetic landscape found in promoter regions specific of coding mRNA (33). To investigate this hypothesis, a total of 182 lncRNA, 84 not overlapping with any known expressed transcript within 6 Kbp upstream of the TSS and 98 not overlapping within 6 Kbp downstream of the 3′-end of transcription regions (EoT), all actively transcribed in CD4+ T cells [accordingly to publicly available RNASeq data (28,34)], with a clear PolII peak in the promoter region [data from (32)] and more than 400 nt long have been selected. As a matched control, 84 + 98 coding PolII transcribed RefSeq genes with the same 6 Kbp non-overlapping condition as selected lncRNA and with comparable transcriptional level and length have been randomly chosen (Supplementary Figure S3). EMdeCODE was then fed with the above-mentioned 39 markers to identify possible discriminating factors over promoter regions, gene bodies (Tx) and the 3′-EoT. A comparison with 2000 randomly selected genomic regions has been performed as reference.

Diagrams in Figure 4A report F score curves calculated for coding versus non-coding genes (Gene versus ncRNA), coding versus random (Gene versus Rnd) and non-coding versus random (ncRNA versus Rnd), considering gene bodies and increasing regions from TSS and 3′ EoT

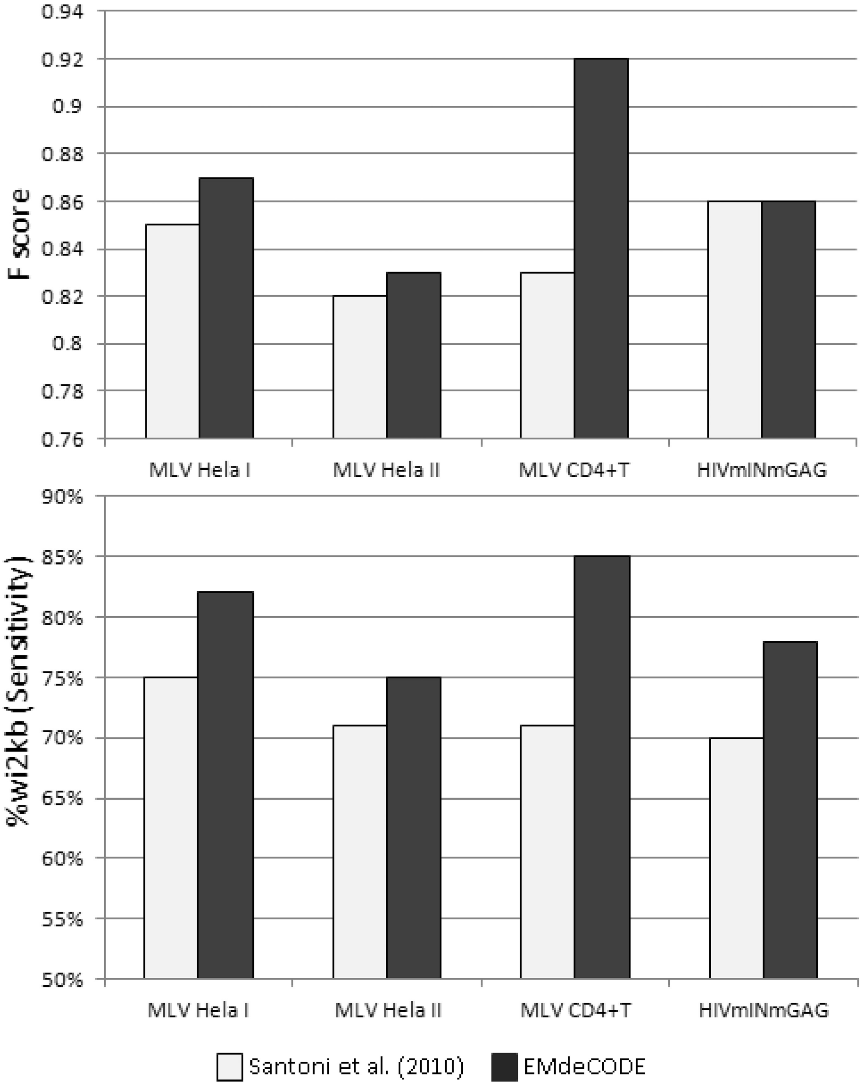

Figure 3. Histograms of the *F* score (upper panel) and the percentage of associated proviruses wi2Kbp of the EMdeCODE-generated supermark (lower panel) with respect to MLV (I and II) proviruses (41,42) and HIVmINmGAG chimera (41) integrated in HeLa in comparison with the 'supermarker' previously reported by Santoni *et al.*

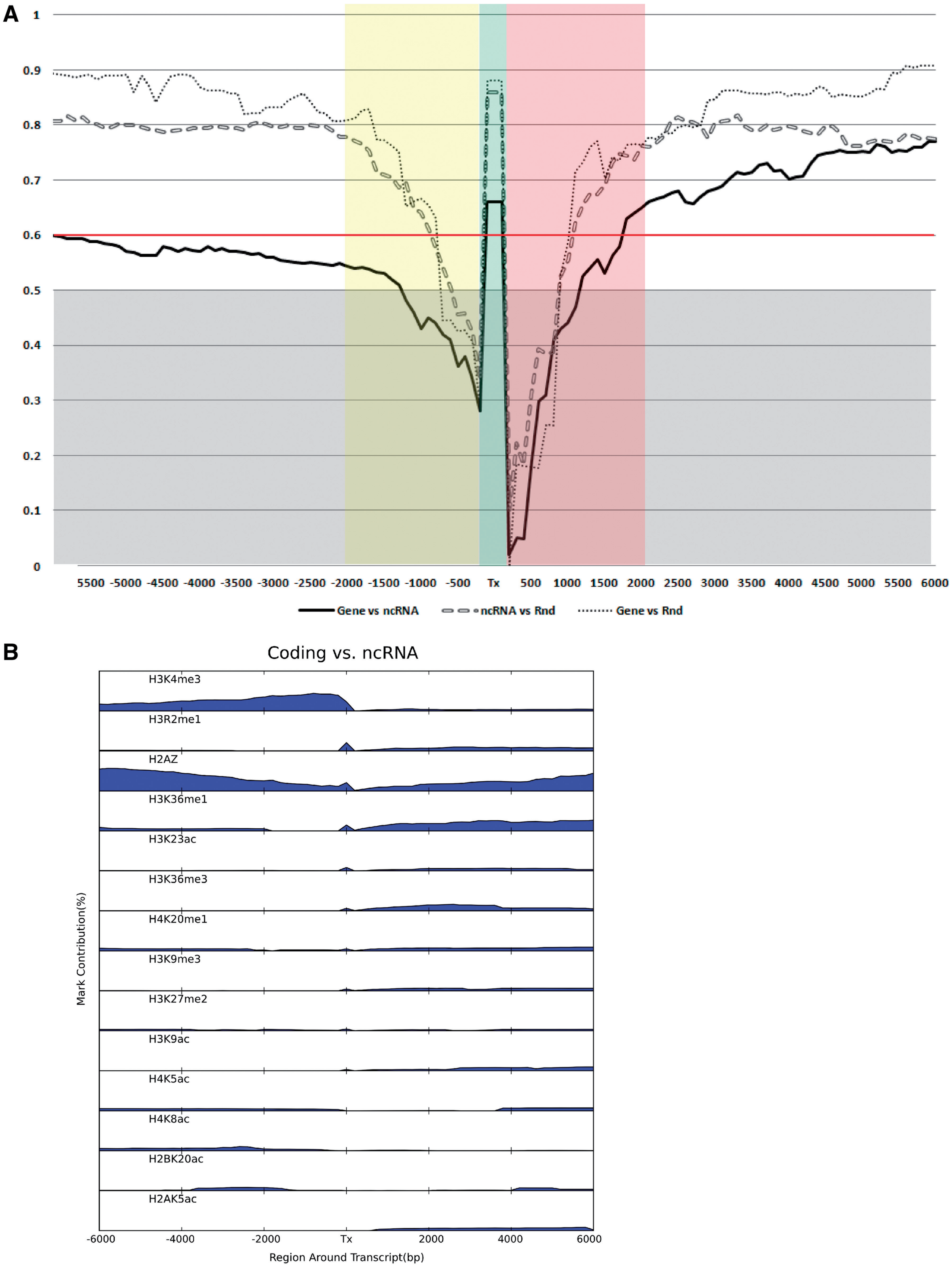

Figure 4. Coding (Gene), ncRNA and Rnd comparisons. (**A**) Supermarks and corresponding F scores are calculated for gene bodies (Tx) and for genomic regions considering regions extending upstream from TSS (negative sizes on x axis) and downstream of 3′-end of transcription (EoT, positive sizes). Association is considered significant if F score >0.6. (**B**) Single-mark contributions (mixture coefficient vector Θ) to supermark discriminating Gene versus lncRNA ordered by contribution in gene bodies. Only marks contributing >1% are reported.

respectively. As expected, both coding genes and non-coding RNA (ncRNA) can easily distinguished from random sites in gene bodies (Tx, F score = 0.87 and 0.86), whereas it is more difficult but still possible to discriminate between them (Tx, F score = 0.67). PolII transcribed lncRNA and coding genes do not differ in the epigenetic profile of the promoter region (F score <0.6), accordingly to what previously reported, but they are clearly distinct up to 6 Kbp downstream the EoT (*F* score = 0.78). Figure 4B reports the contribution of each mark to the final supermark (i.e. the mixture coefficient vector Θ defined in eq. 2) sorted by their percentage in gene bodies. The leading histone mark here is H3K4me3, unexpectedly followed by H3R2me1 that is also the fourth contributor in downstream region after the well-known marks H2AZ, H3K36me1 and H3K36me3 and in comparison with random sites (Gene versus Rnd, Supplementary Figure S4). No functions in human cells have been associated to this mark so far. Notably, H3R2me1 does not show up in lncRNA versus Rnd analysis (Supplementary Figure S5), indicating that this mark might be specific to coding regions. To support this hypothesis, I calculated H3R2me1 and PolII density plots extending 6 Kbp 5′ and 3′ gene bodies of the two matched dataset. H3R2me1 appears to be indeed enriched in coding gene bodies with respect to expressed lncRNA. Its density increases drastically in the promoter region reaching its maximum immediately after TSS and it decreases smoothly but peaking again around the 3′ EoT region (Figure 5A, black line). Conversely, H3R2me1 is almost flat in lncRNA regions (Figure 5A, red line). This different behavior cannot be explained by differences of length, expression level or by PolII-mediated activation as clearly shown in Figure 5B. Moreover, correlation between level of expression of coding genes and lncRNA with H3R2me1 density in T cell data is rather low: 0.25 and 0.22, respectively. Interestingly, coding PolII density has a ‘bump’ at 3′ EoT [interpreted as a slow release of PolII from DNA (35)], absent in lncRNA and that overlap with H3R2me1 increased density in the same region (Figure 5A and B).

DISCUSSION

Understanding the combinatorial complexity of histone post-translational modifications is important to elucidate the mechanisms of gene expression regulation as well as protein–DNA interactions. In this perspective, EMdeCODE recreates the probability mass distribution of observing a specific event on a specific DNA location by aggregating ChIPSeq histone marks profiles into a new associated supermark. The algorithm is based on classical Expectation Maximization approach and mixture modeling with statistical selection. The goal is to maximize the association in term of F score, thus reaching an optimal tradeoff between Precision and Sensitivity.

Compared with previous methods, EMdeCODE significantly improves the identification of putative enhancers, as demonstrated by the good approximation of the experimental p300 distribution in CD4+ T cells. One of the reasons for this increased performance can be ascribed to the choice of reconstructing a minimal profile by considering only significant peaks from ChIPSeq data, thus consistently reducing the associated noise and defining a common background for all the available marks. Moreover, the interpretation of ChIPSeq profiles as probability mass densities and the choice of the F0.5 score as cost function increase the contribution of less abundant but significantly associated marks and, at the same time, reduce the number of false positives.

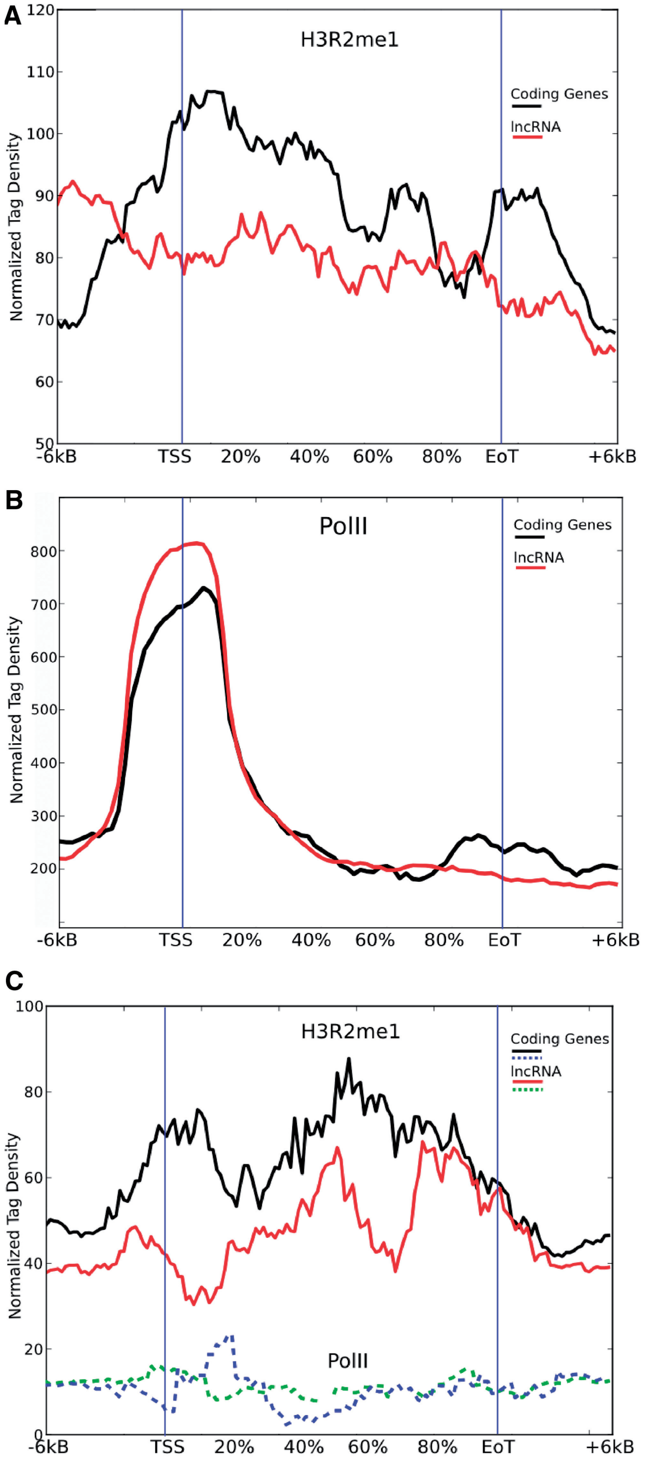

Figure 5. H3R2me1 (**A**) and PolII (**B**) read density plots of coding (black) and non-coding (red) gene bodies normalized by length and divided in 50 non-overlapping bins +15 bins for ±6 Kbp extensions. (**C**) Density plots in untranscribed genes.

When applied to the prediction of retroviral integration sites, EMdeCODE outperforms the heuristic method reported in (5) and generates a supermark localized within 2 Kbp of 75–85% of gammaretroviral proviruses, while increasing the corresponding F score. It is worth to observe that the best performance is obtained in CD4+ T cells where the highest number of marks is also available. This again demonstrates that EMdeCODE exploits and properly balances even minor contributions that are underrated by the heuristic approach.

EMdeCODE can also be used to determine whether some marks can be used to discriminate among genomic events. Indeed, it has been capable of determining supermarks which distinguish coding from non-coding genes, despite substantial epigenetic homogeneity of transcriptional regions having been previously reported (36). This seems to be true only for the promoter region up to 6 Kbp (upstream from the TSS), where no significant differences can be observed. Regions downstream (from 2 to 6 Kbp) of coding genes, instead, are particularly enriched in transcription associated marks, such as H2AZ, H3K4me3 and H3K36me1-3 among others, likely accounting for a complex regulatory structure that is less pronounced for most of non-coding transcripts. The gene body, also, is clearly enriched of H3K36me3 and H3K36me1 in coding genes, especially in comparison with random sites (Supplementary Figure S4), in line with what has been previously reported (37). Perhaps less expected is the strong influence of H3R2me1 in coding gene bodies with respect to lncRNA and random regions as shown in Figure 4B and Supplementary Figure S4. The specificity of this mark for coding regions is confirmed by its absence in lncRNA compared with random regions even if they can be clearly discriminated (F score >0.75, Figure 4A and Supplementary Figure S5). H3R2me1 significantly contributes to coding genes characterization in 3′ EoT regions and 6 Kbp downstream. H3R2me1 read density profile shows a peculiar shape in coding regions of the matched dataset by markedly outlining the gene body. Conversely, H3R2me1 is considerably depleted in non-coding gene bodies, explaining why it has been picked up by EMdeCODE (Figure 5A). H3R2 is methylated by PRMT1 and CARM1 that subsequently coactivate nuclear hormone receptor-mediated transcription, suggesting that the H3R2 methylation may be involved in transcriptional activity (38), as recently observed in Drosophila (39). On the other hand, H3R2me1 is not particularly enriched in active promoter regions of human DNA (32) and it is not strongly correlated with expression levels [see Results and (32)]. Interestingly, a recent study proposed that symmetrical dimethylation of H3R2 may have a role in euchromatin maintenance by mediating histone H3K4 methylation and H3 and H4 acetylation (40). Figure 5C reports PolII and H3R2me1 densities of 150 untranscribed 6 Kbp non-overlapping coding and non-coding genes. Again, monomethylated arginin of histone 3 is denser in coding regions even in complete absence of transcriptional activity (flat PolII density). These observations together with the other results obtained by EMdeCODE support the hypothesis that H3R2 is a key amino acid residue epigenetically involved in the maintenance and the protection of coding regions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–5.

ACKNOWLEDGEMENTS

The author thanks Massimo Pizzato and Stylianos Antonarakis for stimulating discussions and suggestions.

FUNDING

European Research Council [ERC-HuCNC-249968 to F.A.S.]. Funding for open access charge: University of Geneva, Department of Genetic Medicine and Development.

Conflict of interest statement. None declared.

REFERENCES

1. Bird,A. (2007) Perceptions of epigenetics. *Nature*, **447**, 396–398.
2. Jirtle,R.L. and Skinner,M.K. (2007) Environmental epigenomics and disease susceptibility. *Nat. Rev. Genet.*, **8**, 253–262.
3. Spivakov,M. and Fisher,A.G. (2007) Epigenetic signatures of stem-cell identity. *Nat. Rev. Genet.*, **8**, 263–271.
4. Nightingale,K.P., O'Neill,L.P. and Turner,B.M. (2006) Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr. Opin. Genet. Dev.*, **16**, 125–136.
5. Santoni,F.A., Hartley,O. and Luban,J. (2010) Deciphering the code for retroviral integration target site selection. *PLoS Comput. Biol.*, **6**, e1001008.
6. Turner,B.M. (2007) Defining an epigenetic code. *Nat. Cell Biol.*, **9**, 2–6.
7. Huff,J.T., Plocik,A.M., Guthrie,C. and Yamamoto,K.R. (2010) Reciprocal intronic and exonic histone modification regions in humans. *Nat. Struct. Mol. Biol.*, **17**, 1495–1499.
8. Hon,G., Ren,B. and Wang,W. (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.*, **4**, e1000201.
9. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
10. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
11. van Steensel,B., Braunschweig,U., Filion,G.J., Chen,M., van Bemmel,J.G. and Ideker,T. (2010) Bayesian network analysis of targeting interactions in chromatin. *Genome Res.*, **20**, 190–200.
12. Firpi,H.A., Ucar,D. and Tan,K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**, 1579–1586.
13. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
14. Micsinai,M., Parisi,F., Strino,F., Asp,P., Dynlacht,B.D. and Kluger,Y. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.*, **40**, e70.
15. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
16. Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.
17. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.

18. Boyle,A.P., Guinney,J., Crawford,G.E. and Furey,T.S. (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.
19. Bilmes,J.A. (1997) A gentle tutorial on the EM algorithm and application to Gaussian Mixtures and Baum-Welch. Technical Report TR-97-021, ICSI.
20. Won,K.J., Chepelev,I., Ren,B. and Wang,W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.
21. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
22. 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.
23. Ferretti,V., Poitras,C., Bergeron,D., Coulombe,B., Robert,F. and Blanchette,M. (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.*, **35**, D122–D126.
24. Wang,Z., Zang,C., Cui,K., Schones,D.E., Barski,A., Peng,W. and Zhao,K. (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019–1031.
25. 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.
26. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
27. Morris,S.A., Rao,B., Garcia,B.A., Hake,S.B., Diaz,R.L., Shabanowitz,J., Hunt,D.F., Allis,C.D., Lieb,J.D. and Strahl,B.D. (2007) Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification. *J. Biol. Chem.*, **282**, 7632–7640.
28. Valouev,A., Johnson,S.M., Boyd,S.D., Smith,C.L., Fire,A.Z. and Sidow,A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516–520.
29. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.
30. Rada-Iglesias,A., Bajpai,R., Swigut,T., Brugmann,S.A., Flynn,R.A. and Wysocka,J. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283.
31. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2011) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA*, **107**, 21931–21936.
32. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
33. Ugarkovic,D. (ed.), (2011) *Long Non-Coding RNAs.* Springer-Verlag, New York.
34. Barski,A., Chepelev,I., Liko,D., Cuddapah,S., Fleming,A.B., Birch,J., Cui,K., White,R.J. and Zhao,K. (2010) Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol.*, **17**, 629–634.
35. Rahl,P.B., Lin,C.Y., Seila,A.C., Flynn,R.A., McCuine,S., Burge,C.B., Sharp,P.A. and Young,R.A. (2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432–445.
36. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
37. Tanaka,Y., Katagiri,Z., Kawahashi,K., Kioussis,D. and Kitajima,S. (2007) Trithorax-group protein ASH1 methylates histone H3 lysine 36. *Gene*, **397**, 161–168.
38. An,W., Kim,J. and Roeder,R.G. (2004) Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53. *Cell*, **117**, 735–748.
39. Kirmizis,A., Santos-Rosa,H., Penkett,C.J., Singer,M.A., Green,R.D. and Kouzarides,T. (2009) Distinct transcriptional outputs associated with mono- and dimethylated histone H3 arginine 2. *Nat. Struct. Mol. Biol.*, **16**, 449–451.
40. Migliori,V., Muller,J., Phalke,S., Low,D., Bezzi,M., Mok,W.C., Sahu,S.K., Gunaratne,J., Capasso,P., Bassi,C. *et al.* (2012) Symmetric dimethylation of H3R2 is a newly identified histone mark that supports euchromatin maintenance. *Nat. Struct. Mol. Biol.*, **19**, 136–144.
41. Lewinski,M.K., Yamashita,M., Emerman,M., Ciuffi,A., Marshall,H., Crawford,G., Collins,F., Shinn,P., Leipzig,J., Hannenhalli,S. *et al.* (2006) Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog.*, **2**, e60.
42. Wu,X., Li,Y., Crise,B. and Burgess,S.M. (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300**, 1749–1751.