Nucleic Acids Research Advance Access published November 24, 2010

Nucleic Acids Research, 2010, 1–11
doi:10.1093/nar/gkq1081

Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions

Christian Rödelsperger[1,2,3], Gao Guo[3], Mateusz Kolanczyk[2], Angelika Pletschacher[3], Sebastian Köhler[1,3], Sebastian Bauer[3], Marcel H. Schulz[2,4] and Peter N. Robinson[1,2,3,*]

[1]Berlin-Brandenburg Center for Regenerative Therapies, Charité-Universitätsmedizin Berlin, [2]Max Planck Institute for Molecular Genetics, [3]Institute for Medical Genetics, Charité-Universitätsmedizin, Berlin and [4]International Max Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany

Received June 25, 2010; Revised and Accepted October 14, 2010

ABSTRACT

Multicellular organismal development is controlled by a complex network of transcription factors, promoters and enhancers. Although reliable computational and experimental methods exist for enhancer detection, prediction of their target genes remains a major challenge. On the basis of available literature and ChIP-seq and ChIP-chip data for enhanceosome factor p300 and the transcriptional regulator Gli3, we found that genomic proximity and conserved synteny predict target genes with a relatively low recall of 12–27% within 2 Mb intervals centered at the enhancers. Here, we show that functional similarities between enhancer binding proteins and their transcriptional targets and proximity in the protein–protein interactome improve prediction of target genes. We used all four features to train random forest classifiers that predict target genes with a recall of 58% in 2 Mb intervals that may contain dozens of genes, representing a better than two-fold improvement over the performance of prediction based on single features alone. Genome-wide ChIP data is still relatively poorly understood, and it remains difficult to assign biological significance to binding events. Our study represents a first step in integrating various genomic features in order to elucidate the genomic network of long-range regulatory interactions.

INTRODUCTION

Decoding the regulatory program that controls metazoan development is a major barrier to the understanding of multicellular complexity in higher organisms. A substantial fraction of this program is likely to be encoded in gene deserts which harbor highly conserved non-coding elements (HCNEs), located up to several hundred kilobases away from the nearest gene (1,2). Many of these intergenic and intronic regions represent evolutionarily conserved enhancers and silencers, which we will refer to as 'enhancers' in the following. Enhancers coordinate tissue and developmental stage-specific expression of their target genes by inducing changes in chromatin conformation in order to bring distant regulatory elements into spatial proximity of the transcription start sites (TSS) of their target genes. Extensive experimental and computational work has been carried out on the detection of enhancer regions (3,4).

The advent of high-throughput chromatin immunoprecipitation assays (ChIP-chip and ChIP-seq) has made genome-wide *in vivo* mapping of protein–DNA interactions possible. In agreement with the observation that evolutionarily conserved regulatory elements are located primarily in intergenic regions, <10% of transcription factors have >50% of their binding sites within 2.5 kb of a transcription start site (5). Recently, Visel *et al.* (6) employed ChIP-seq to identify several thousand genomic loci in mouse embryonic tissues which were bound by the enhancer-associated p300 protein. p300 is a transcriptional coactivator (7) that is recruited by other DNA binding proteins in a tissue and cell-type specific manner to form an enhanceosome complex with regulatory activity (8). About 87% of the p300 bound loci regions showed tissue-specific enhancer activity (6).

Global correlations with expression data (6) and strong biases for HCNEs to occur in the vicinity of transcription factors and developmental genes (2) support the assumption that enhancers regulate nearby genes. To date, computational and experimental approaches for enhancer detection have employed proximity-based

*To whom correspondence should be addressed. Tel: +49 30 450566042; Fax: +49 30 450569915; Email: peter.robinson@charite.de

cutoffs on genomic distance or nearest gene assignments to associate putative enhancer regions to their target genes (4,9–11). Although the genes located closest to the enhancers are reasonable candidates for the target genes, this is not a general rule. For instance, a *Pax6* enhancer is located in an intron of a neighboring gene (12,13). Interactions between enhancers and their target genes can span large genomic distances. For instance, an enhancer of the sonic hedgehog (Shh) gene is located 1 Mb upstream of the Shh gene (14). For these reasons, enhancer targets cannot be reliably predicted by simple computational rules based on genomic proximity.

Besides genomic distance, conserved synteny is the only feature that has been considered to possess predictive power for enhancer-target gene interactions (15,16). Conserved synteny generally describes a relative order of two or more genomic loci that is conserved in more than one species (Supplementary Figure S1). This might reflect a certain pattern of co-evolution between regulatory region and target gene. The Vista enhancer browser (17) allows manual investigation of flanking genes, and some genome browsers like SynBrowse include information about conserved synteny (18), but no automated approaches exist that specifically predict the target genes of a number of predicted or known enhancers (19). Consequently, existing approaches for enhancer detection (4,20,21) remain incomplete and fail to integrate important developmental target genes into larger regulatory modules and networks that control multicellular organismal development.

One impediment to progress in this area is the paucity of experimental enhancer-target gene interaction data. Commonly used *in vivo* assays for enhancer activity that use co-injection of enhancer and minimal promoter reporter genes (2,6,22) provide evidence about the tissue specificity of the enhancer but do not indicate which genes are targets of the enhancer. On the other hand, chromosome conformation capture (3C) assays (23,24) test for physical interactions between enhancer and promoter regions, and thus can be used to identify enhancer target genes. However, no large-scale data set of enhancer-specific chromatin interactions is available with which to assess the quality of prediction methods.

To our knowledge, there has been no previous large-scale computational analysis of the prediction of enhancer targets. Ahituv *et al.* (15) mapped conserved blocks of synteny (CBSs) that were homologous among human/mouse/chicken or human/mouse/frog genomes and identified ~2000 CBSs > 200 kb for each comparison. They postulated that such CBSs were enriched for long-range regulatory interactions between enhancers and target genes because the prevalence and distribution of chromosomal aberrations leading to position effects showed a clear bias not only for mapping onto CBS but also for longer CBS size. Using a similar definition based on alignments between human and zebrafish genomes, Akalin *et al.* identified a set of genomic regulatory blocks (GRBs) located within conserved human/zebrafish-syntenic regions and predicted a set of 269 target genes of within the GRBs (25). The authors postulated that HCNEs within the GRBs are enhancers and that transcription factor genes within the GRBs are their targets, but did not develop a method for predicting target genes on a genome-wide basis or of predicting target genes of a specific enhancer protein. In this work, we present a method to predict the target genes of potential enhancers identified as bound DNA sequences in ChIP-seq and ChIP-chip experiments. We evaluated our method using published data for p300 and Gli3 in embryonic mouse tissues (6,26). Our method uses an integrative approach based on random forest analysis of a combination of genomic proximity, conserved synteny as well as distance in protein–protein interaction (PPI) networks, and Gene Ontology (GO) similarities between regulator and putative target gene. Our algorithm showed a substantially better accurracy than predictions based on any single feature in isolation.

MATERIALS AND METHODS

Genome data and alignments

We downloaded pairwise net alignments generated by blastz (27) for mouse (*Mus musculus*, mm9) against opossum (*Monodelphis domestica*, monDom4), chicken (*Gallus gallus*, galGal3), frog (*Xenopus tropicalis*, xenTro2), zebrafish (*Danio rerio*, danRer5) and fugu (*Takifugu rubripes*, fr2). We initially used data from human and dog in our analysis; however including these data sets did not improve the results (Supplementary Figure S2), and therefore these two genomes were not used for further analysis. In addition, mouse RefSeq annotations for 22 468 genes were downloaded from the UCSC Genome Browser (28). The phylogenetic distances between these species are shown in Supplementary Figure S3, whereby the branch lengths reflect the average number of subsitutions per site as calculated from genome-wide blastz alignments (29).

p300 ChIP-seq data

Visel *et al.* (6) used chromatin immunoprecipitation with the enhancer-associated protein p300 followed by massively parallel sequencing to map the *in vivo* binding sites of p300 in mouse embryonic forebrain, midbrain and limb tissue. We downloaded p300 ChIP-seq peaks and lists of upregulated genes that were identified by comparing forebrain and limb expression with E11.5 whole embryo gene expression as measured on Affymetrix GeneChip MouseGenome 430 2.0 arrays. The limb data (30) are based on E11.5 proximal hindlimb expression (GEO series GSE10516, samples GSM264689, GSM264690 and GSM264691). The ChIP-seq also includes P300 bound sites from midbrain, but we did not use this data because no set of midbrain upregulated genes were defined by Visel *et al.* We focused on upregulated genes since Visel *et al.* (6) only observed ChIP-seq peak enrichments in the vicinity of genes that are significantly upregulated in the corresponding tissue, indicating that p300 acts as a coactivator rather than as a repressor. In total 2453 ChIP-seq peaks were obtained for embryonic mouse forebrain tissue and 2105 for limb. Additionally, 1062 and 748 significantly upregulated probe sets were

obtained that correspond to 555 upregulated genes with RefSeq IDs for forebrain and 347 for limb (6). Affymetrix gene expression microarrays are not able to reliably distinguish between different transcripts of genes. Therefore, one representative transcript was chosen for each gene according to whether a transcript was in the set of differentially expressed probesets, or failing that, arbitrarily as the leftmost transcript on the Watson strand of the chromosome. This reduced the number of RefSeq IDs to 19 569.

Gli3 ChIP-chip data

We downloaded Supplementary Data sets 1 and 2 from Vokes *et al*. (26). These data sets contain 5274 Gli binding regions and 753 responsive genes that were identified using pairwise and multiple sample comparison of expression levels (Affymetrix Mouse Exon 1.0 ST arrays) for overexpressed and mutated Gli3 versus wildtype and anterior versus posterior forelimbs (26).

Genomic distances between enhancer and target gene

For each gene in a genomic window centered at the enhancer, we calculated the genomic distance between enhancer and target gene as the minimal distance between the endpoints of the enhancer region and the TSS of the candidate target genes. For the genomic distance-based predictions, the gene with the minimal distance was predicted to be the target gene.

Calculation of conserved synteny score (CSS)

We defined for each enhancer e a genomic interval in the reference species r by selecting all genes for which the genomic distance between enhancer and TSS of the gene g is less than a maximal distance threshold $d_r(e,g) < \Theta$ (Supplementary Figure S1). For each gene g in this region, we define a conserved synteny score (CSS) by testing in other species $s = 1,\ldots,k$ whether the distance $d_s(e,g)$ between aligned regions of enhancer and TSS is smaller than the threshold Θ. The CSS is then calculated as the sum of phylogenetic distances $\phi(r,s)$ (Supplementary Figure S3) between the reference r and species s, where $d_s(e,g) < \Theta$.

$$\mathrm{CSS}(e,g) = \sum_{s=1\ldots k} \delta_s(e,g) \times \phi(r,s) \quad (1)$$

$$\delta_s(e,g) = \begin{cases} 1 & \text{if } d_s(e,g) < \Theta \text{ in species } s \\ 0 & \text{otherwise} \end{cases}$$

$\delta_s(e,g)$ was also taken to be zero if an orthologous gene and enhancer could not be identified in the other species. Since some genomes in our analysis are not finished and annotations are incomplete, we identified orthologous genes on the basis of the presence of aligned sequences around the promoter region as defined by the TSS ± 1 kb. This assumes that the enhancer specifically interacts with the promoters of their target genes. This is supported by our recent finding that the occurrence of intergenic HCNEs correlates with conservation in promoter regions of nearby genes (31), which we interpret as evidence for similar evolutionary constraints acting on the enhancers as well as the promoter regions.

For the enhancer sequence, orthologous sequences were identified on the basis of aligned sequence in the other species. We note that rearrangements that disrupt collinearity are not penalized by our scoring scheme because according to our assumption, enhancers can retain their function even after chromosomal rearrangements that invert genes or change their order.

Gene Ontology similarity definition

We calculated for each GO term (t) in the ontology an information content value (IC) defined as $\mathrm{IC}(t) = -\log p_t$, where p_t is the number of genes annotated by GO term t divided by the total number of annotated genes. The similarity between two terms can be calculated as the IC of their most informative common ancestor (MICA) (32). This can be used to calculate the similarity (sim) between one set of terms, to another set of terms, each of which belongs to a particular gene (g_i, g_j):

$$\mathrm{sim}(g_i \rightarrow g_j) = avg\left[\sum_{t_1 \in g_i} \max_{t_2 \in g_j} \mathrm{IC}(\mathrm{MICA}(t_1,t_2))\right]. \quad (2)$$

Note, that $\mathrm{sim}(g_i \rightarrow g_j)$ is not necessarily equal to $\mathrm{sim}(g_j \rightarrow g_i)$. As previously described (33), we defined the similarity between two genes as the symmetric version of the formula above by calculating:

$$\mathrm{sim}(g_i,g_j) = \frac{\mathrm{sim}(g_i \rightarrow g_j)+\mathrm{sim}(g_j \rightarrow g_i)}{2}. \quad (3)$$

Distance computation for PPI networks

In order to define the similarity between two genes, we created a network based on the data of the STRING 8.2 database (34), physical and functional interactions. The network consists of 138 156 interactions including 194 direct interactions between p300 and other proteins. In a previous study, we have shown that global network similarity measures are better suited for defining functionally associated groups of genes (35). We constructed a mouse-specific adjacency matrix, which was transformed into a column-normalized adjacency matrix (A). The random walk starts at a certain node corresponding to a gene g_i at time point t_0 and randomly visits adjacent nodes. The random walk distance $p_{t+1}(g_i,g_j)$ is defined as the probability of the random walker being at node g_j at time point $t+1$ given that the walker started at g_i. For a vector of starting probabilities p_0, the state probabilities p_{t+1} can be computed iteratively:

$$p_{t+1} = (1-r)\mathbf{A} \times p_t + r \times p_0, \quad (4)$$

whereby r denotes the restart probability ($r = 0.7$). For $t \rightarrow \infty$, the state probabilities converge to a stationary distribution p_∞ that can be written as:

$$p_\infty = (\mathbf{I} - ((1-r)\mathbf{A}))^{-1} \times r \times p_0. \quad (5)$$

The matrix **I** denotes the identity matrix and the starting probabilities p_0 were set to 1 for g_i and 0 for all other genes. For two genes g_i and g_j, we define a symmetric PPI distance score by taking the average of the probabilities to encounter g_j when starting at g_i and vice versa.

Binary and discriminative random forest classifier

We first developed a binary random forest classifier (36) for the problem of deciding whether a single gene is an enhancer target based on the four features: genomic distance to an enhancer, CSS, PPI distance and GO similarity ('binary RF'). The classifier learns to predict the class from the four features and to output the ratio of trees that voted for this class. In case of missing values, we assigned the median GO similarity or PPI distance value between p300 and all other genes for the respective feature. We used an implementation of Breiman's algorithm that uses random selection of features at each node to determine a split (37) (R package 'randomForest', version 4.5-34) and to train a random forest of 1000 randomly generated decision trees. The final prediction was made by selecting among all genes in the interval the one with the highest probability (i.e. the highest number of trees voting for it).

The binary RF can yield only a yes/no decision as to whether a gene is an enhancer target or not, and is not designed to rank all the candidate genes in the interval. We therefore implemented a second classifier ('discriminative RF') that evaluates each gene pair g_i and g_j in the interval using feature values as well as pairwise rankings and then decides among the following outcomes:

(1) g_i is the target gene
(2) g_j is the target gene
(3) neither g_i nor g_j is the target

This RF takes 12 input features, corresponding to eight feature values for both genes (genomic distance, CSS, GO similarity, and PPI similarity for g_i and for g_j) and four features that assign g_i either to 'winner' (W) or 'looser' (L) or 'tied' (equal, E) in comparison with the respective feature of g_j. Since GO and PPI annotations are incomplete, we added two labels 'W?' and 'L?' to model the uncertainty that is associated with gene pairs for which one value is missing ('NA'). Then, for each of the four comparisons between genes g_i and g_j, a feature f is assigned to g_i:

$$f = \begin{cases} W & \text{if value}(g_i) > \text{value}(g_j) \\ L & \text{if value}(g_i) < \text{value}(g_j) \\ E & \text{if value}(g_i) = \text{value}(g_j) \\ W? & \text{if value}(g_i) \geq \text{median and value}(g_j) = \text{'}NA\text{'} \\ L? & \text{if value}(g_i) < \text{median and value}(g_j) = \text{'}NA\text{'} \end{cases}$$

A random forest of 1000 trees was trained using these 12 features. The output consisted of the probabilities for the three classes and the class with the majority vote. The final prediction was made by summing over all probabilities for target gene assignments in pairwise comparisons for all pairs in the interval and reporting the gene with the highest sum as the target gene. A schematic overview of both classifiers is shown in Supplementary Figure S4.

Statistical analysis

For evaluation of various values of the maximal distance parameter Θ on the forebrain and limb data, we used only the p300 enhancers with distance $<\Theta$ to a differentially upregulated gene. Depending on the distance parameter Θ, it may be that multiple differentially upregulated genes are located in a given genomic interval. In such cases, we counted the prediction as a 'correct prediction' if at least one of the upregulated genes was unambiguously predicted as a target gene by any of the prediction methods.

We calculated the precision of a method as $N_{\text{correct predictions}}/N_{\text{predictions}}$ and recall as $N_{\text{correct predictions}}/N_{\text{enhancer}}$. Precision indicates the probability that a prediction is correct and recall denotes the ratio of enhancers for which a correct prediction could be made. Precision and recall values are highly similar for most methods, only differing in cases where multiple genes are assigned the same maximal score by a method. These cases were counted as 'no prediction' in the precision and recall calculations.

For the training of the random forest classifiers, we split the enhancer sets into 80% training samples and calculated the precision and recall values on the remaining 20% validation samples. This was done 10 times, the values in Figure 4 represent the means of the different iterations. For both models, we subsampled the training set so that each possible outcome occurred an equal amount of times. The predictions in Supplementary Data S1 contain leave-one-out predictions for the $\Theta = 1000$ kb data and predictions for p300 enhancers that are >1 Mb away of an upregulated gene. For these enhancers, random forest classiers were trained on the complete data set for $\Theta = 1000$ kb.

RESULTS

Conserved synteny predictions of enhancer targets have low recall

Previously identified candidate enhancer regions have been shown to be enriched in the vicinity of transcription factors and developmental genes (2,38) and to maintain conserved synteny (15,16). However, it is not clear to what degree conserved synteny or genomic proximity can be used to predict target genes. We therefore initially compared the performance of predictions based on genomic proximity (i.e. the nearest gene is taken to be the target of an enhancer), conserved synteny and randomly choosing one of the genes in a window around the enhancer. The CSS was calculated on the basis of a conserved association between the enhancer and the promoter regions of putative target genes in related genomes (Supplementary Figure S1). We also evaluated ortholog predictions based on protein sequence similarity, but this approach showed slightly lower precision and recall values than using the genomic alignments of promoter sequences (Supplementary Figure S2).

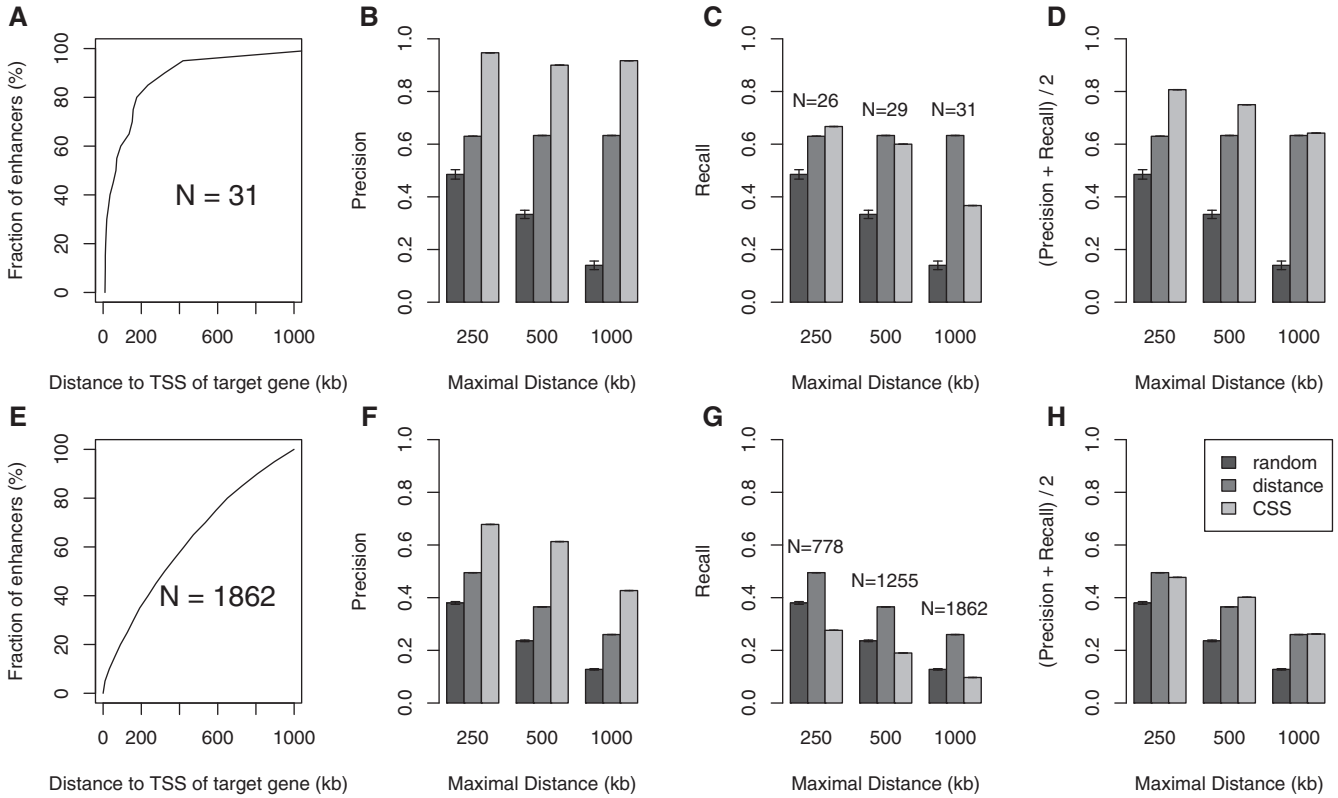

Figure 1. (**A**) Distance distribution between enhancers and target genes for 31 regulatory interactions from the literature. (**B**–**D**) Evaluation of precision (B), recall (C) and average precision and recall (D) for predictions based on conserved synteny, genomic distance or random predictions of a gene in the genomic interval defined by a maximal distance threshold $\Theta = \{250,500,1000\}$ kb around the enhancer. (**E**) Distance distribution between p300 enhancers and putative target genes (6). (**F**–**H**) Evaluation of precision (F), recall (G) and average precision and recall (H) on different sets of p300 enhancers. Although literature data and ChIP-seq data show substantially different distributions of enhancer target gene distances, conserved synteny shows the highest precision values for both data sets.

The conserved syntenies were weighted by the evolutionary distances between mouse and the target species. For each of the three methods, we evaluated the quality of predictions by calculating precision and recall for various maximal distance thresholds Θ that define a genomic window centered around the enhancer region. Several studies have outlined that HCNEs and thus putative enhancers span genomic regions of several hundreds of kilobases around their target genes (2,38). We therefore assessed the performance of predictions for $\Theta \in \{250,500,1000\}$kb. We chose an arbitrary maximal cutoff of $\Theta = 1000$ kb because the great majority of experimentally validated enhancer/target gene pairs are separated by less than this amount (c.f. Figure 1A).

At present, there is no database of enhancer targets, and information in the literature is sparse. We therefore compiled a set of 31 known enhancer-target gene interactions from the literature in order to estimate the quality of predictions that are based on genomic distance or conseved synteny. We included all interactions from either human or mouse that were identified either by observations of phenotypes due to genomic rearrangements (14), similar activation and expression pattern of enhancer and target gene (39) and 3C experiments (24). We assumed that enhancer activities are conserved in human and mouse and mapped human enhancers to the homologous sequences using blastz alignments (27) (See Supplementary Table S1 for a list of the 31 experimentally validated enhancer-target gene interactions). Synteny-based predictions showed a precision ~90% in contrast to ~61% for genomic distance (Figure 1 A–D). However, recall values for synteny-based predictions only reach a level of 69% for $\Theta = 250$ kb, 62% for $\Theta = 500$ kb and only 37% for $\Theta = 1000$ kb, probably because it is not usually possible to assign an enhancer unambiguously to a single target gene on the basis of synteny alone. In all comparisons, predictions based on CSS and genomic distance perform substantially better than random.

Binding by the transcriptional coactivator, p300 is thought to be a marker for enhancer activity. For instance, in one series of lacZ reporter gene assays in transgenic mice, 75 of 86 (87%) p300 ChIP-seq peaks showed enhancer activity (6).

In the following, we will refer to the p300 ChIP-seq peaks as 'p300 enhancers'. It should be noted that a p300 ChIP-seq peak does not necessarily represent a biologically relevant enhancer, which is a limitation of our approach. For our evaluation, we extracted 1862 enhancers from a set of about 4500 p300 enhancers identified by Visel *et al.* (6) under the assumption that

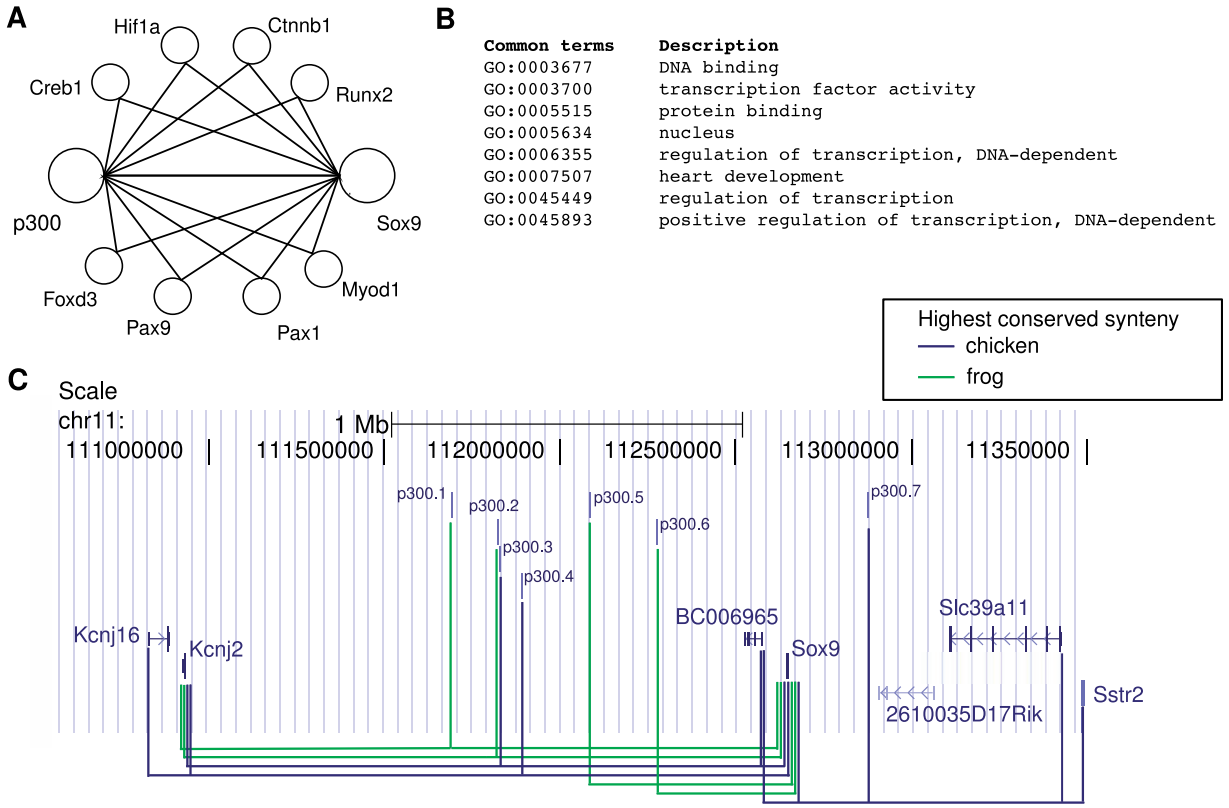

Figure 2. We hypothesize that the enhancer binding protein and its target genes show a tendency for shared functions such as 'transcription factor activity' and are located in the vicinity of one another in the protein–protein interactome. This is illustrated in the example of the potential regulation of *Sox9* by p300. (**A**) PPIs involving p300 and SOX9. p300 and SOX9 directly interact with one another (40), and also share a number of known or predicted intermediary interaction partners in the protein interaction network (34). (**B**) SOX9 has 21 GO annotations, and p300 has 35 GO annotations. The eight shared annotations are shown. (**C**) UCSC Genome Browser view on the *Sox9* locus. Seven p300 enhancers from mouse limb tissue (6) show the highest degree of conserved synteny with the *Sox9* promoter region, however, only the enhancers p300.5 and p300.6 can unambiguously be assigned to the *Sox9* promoter. For the remaining enhancers, multiple genes including *Sox9* exhibit the same degree of conserved synteny. However, high GO similarity between p300 and *Sox9* as well as their proximity in the PPI network suggest that the target gene of these enhancers is *Sox9*.

upregulated genes located in the same genomic region as p300 enhancers represent the target genes. Although the target genes of p300 enhancers are likely to also be differentially expressed, we note that this assumption may not be correct in all cases because the upregulation can be due to secondary effects. As with the gold-standard targets from the literature, we compared predictions for various maximal distance thresholds $\Theta \in \{250,500,1000\}$ kb that define a genomic window centered around the p300 enhancer, and assessed the performance of synteny-based and genomic proximity-based predictions relative to random guessing. Figure 1E–H shows precision and recall values for the predictions based on each of the two features and random guessing for the merged p300 enhancers from limb and forebrain. In agreement with our observations on the known target gene interactions, conserved synteny alone exhibits a higher precision compared with the use of distance alone. However, conserved synteny could only unambiguously assign a minority of enhancers to their target genes leading to a recall of $<20\%$ for $\Theta = 1000$ kb.

The difference in the quality of predictions for the two sets is likely to be related at least partially to the different distribution of distances between enhancer and target gene (Figure 2A and E). The results do suggest that using conserved synteny or genomic distance alone is not able to generate accurate target gene predictions for the p300 enhancers.

GO similiarity and proximity in PPI networks may be used to improve prediction of enhancer target genes

The above-mentioned results demonstrated that genomic distance and conserved synteny are of limited utility in predicting the target genes of p300 enhancers. Although CSS has reasonably high precision values, it often fails to unambiguously assign an enhancer to a target gene because multiple genes in the interval exhibit the same degree of conserved synteny. This accounts for 20–50% of p300 enhancers and thus represents a major limitation in the use of conserved synteny. For instance, seven p300 enhancers for limb tissues are located in the *Sox9* locus and might account for the upregulation of *Sox9*, observed by Visel *et al.* (6). An analysis based on genomic proximity alone would not identify Sox9 as the target, and an analysis based on conserved synteny would identify up to five additional genes in the vicinity as

potential targets. In some cases, transcription factors regulate genes with which they also physically interact, e.g. Runx2 and Dlx5 (41,42). We therefore hypothesized that p300 and its targets are located more proximal to each other in PPI networks (Supplementary Figure S5) than p300 and non-target genes. We additionally hypothesized that functional similarity between p300 and targets is greater than between p300 and non-targets. p300 has a number of GO annotations related to organ development, regulation of transcription factor activity, response to stimuli including calcium, transcription cofactor activity and others (Supplementary Table S2). GO analysis of the limb and forebrain upregulated genes from Visel *et al*. (6) revealed that both sets are significantly enriched in terms such as 'developmental process' and 'transcription factor activity' (Supplementary Table S3). These observations reflect the known role of p300 in development (43,44).

If we take the upregulation of Sox9 in the above-mentioned experiment as evidence that Sox9 is the target gene of the enhancers, then the observation that Sox9 is a direct protein interaction partner of p300, and that it shares a number of GO annotations with p300 could be used to identify Sox9, and not one of the other five genes showing conserved synteny, as the correct target gene. This observation motivates our approach (Figure 2).

We therefore tested whether GO similarity and PPI distance can be used to resolve the ambiguity in cases where CSS fails to unambiguously assign an enhancer to a target gene. Figure 3 shows that in case of ties in CSS, target genes show higher GO similarity and are closer to p300 in PPI networks than non-target genes. This observation motivated us to develop an integrative approach that combines all four features in a random forest classifier.

Accurate target gene prediction using random forest classifiers and combination of features

Decision tree induction is a supervised learning method for classifying data. During the learning phase, a tree is constructed iteratively, whereby at each node a test is derived that splits the local training set into two subsets so that the heterogeneity of the resulting subsets is minimized. Typically, the learning phase is stopped as soon as the heterogeneity falls below a certain threshold. Random forests (RFs) are an extension of decision trees to collections of trees that use randomization in the selection of features for splitting the learning sample at each node (37). The final classification is made by taking the majority vote for all trees in the forest (36). Alternatively, classification probabilities can be defined as the ratio of trees voting for a certain class.

We evaluated two random forest approaches (Supplementary Figure S4). The first was a binary RF classifier which separately calculates the probability of each gene of being a target; these probabilities can then be used to rank the k genes in a given interval according to the probability of being a target gene. The second approach involved a discriminative RF classifier which compares all gene pairs in the interval and chooses the gene that was the most frequently predicted target gene in the set of all pairs (see 'Materials and Methods' section).

In order to make the methods more easily comparable, we will use the average precision-recall (Precision+Recall/2) as performance measure, analogously to Schweikert *et al*. (45). Individual precision and recall values for limb and forebrain target gene predictions can be found in Supplementary Tables S4 and S5. Both classifiers were compared with the genomic distance-based method. Figure 4A and B shows the average precision-recall values of the three approaches for $\Theta = \{250,500,1000\}$kb. Increasing Θ leads to higher number of genes in an interval. For $\Theta = 250$ kb an average of 8.5 genes is located in the genomic window around the putative p300 from the limb data set. This number increases to 25 genes for $\Theta = 1000$ kb. Supplementary Table S6 shows the average number of genes and differentially expressed genes per interval. Binary and discriminative random forest classifiers show substantially better performance than the distance-based approach. This is true for all comparisons of random forest classifiers versus predictions based on any single feature (Supplementary Tables S4 and S5).

Since cases have been reported where the distance to the target genes exeeds 1 Mb, we also applied the classifier on a putative p300 enhancers that are located up to 2 Mb away of the nearest differentially expressed gene. The average precision-recall value stays almost constant at a level of 58% (Supplementary Tables S5).

We also applied the classifier on 1372 p300 forebrain enhancers and 1324 limb enhancers that are not in proximity of an upregulated gene ($\Theta = 1000$ kb). The predictions include previously reported *Bmp7* limb enhancer and a *Sox2* enhancer, that is active in rhombencephalon (46,47), suggesting that at least a proportion of the predictions is valid. As expected, GO enrichment analysis shows a similar pattern as the upregulated genes from Visel *et al*. (Supplementary Tables S3, S8 and S9). It is likely that the enrichment of 'transcription factor activity' and 'developmental process' is a consequence of using GO similarity in the prediction. However, both predicted target gene sets include significantly enriched terms that are unique to the respective set, such as Notch and Wnt signaling in limb and nervous system development in forebrain (Supplementary Tables S8 and S9). Supplementary Figure S6 shows the distributions of distances between p300 enhancer and predicted target genes. Forebrain enhancer-target gene pairs have a median distance of 388.2 kb with a median of two intervening genes and limb pairs have a distance of 395.3 kb spanning over three genes. A complete list of predictions for the p300 peaks is provided as Supplementary Data S1.

Prediction of Gli3 target genes in a limb ChIP-chip data set

To test whether our method might be applicable to experiments with other enhancers, we analyzed a ChIP-chip data set for Gli3 in mouse limbs (26). Gli3 is a

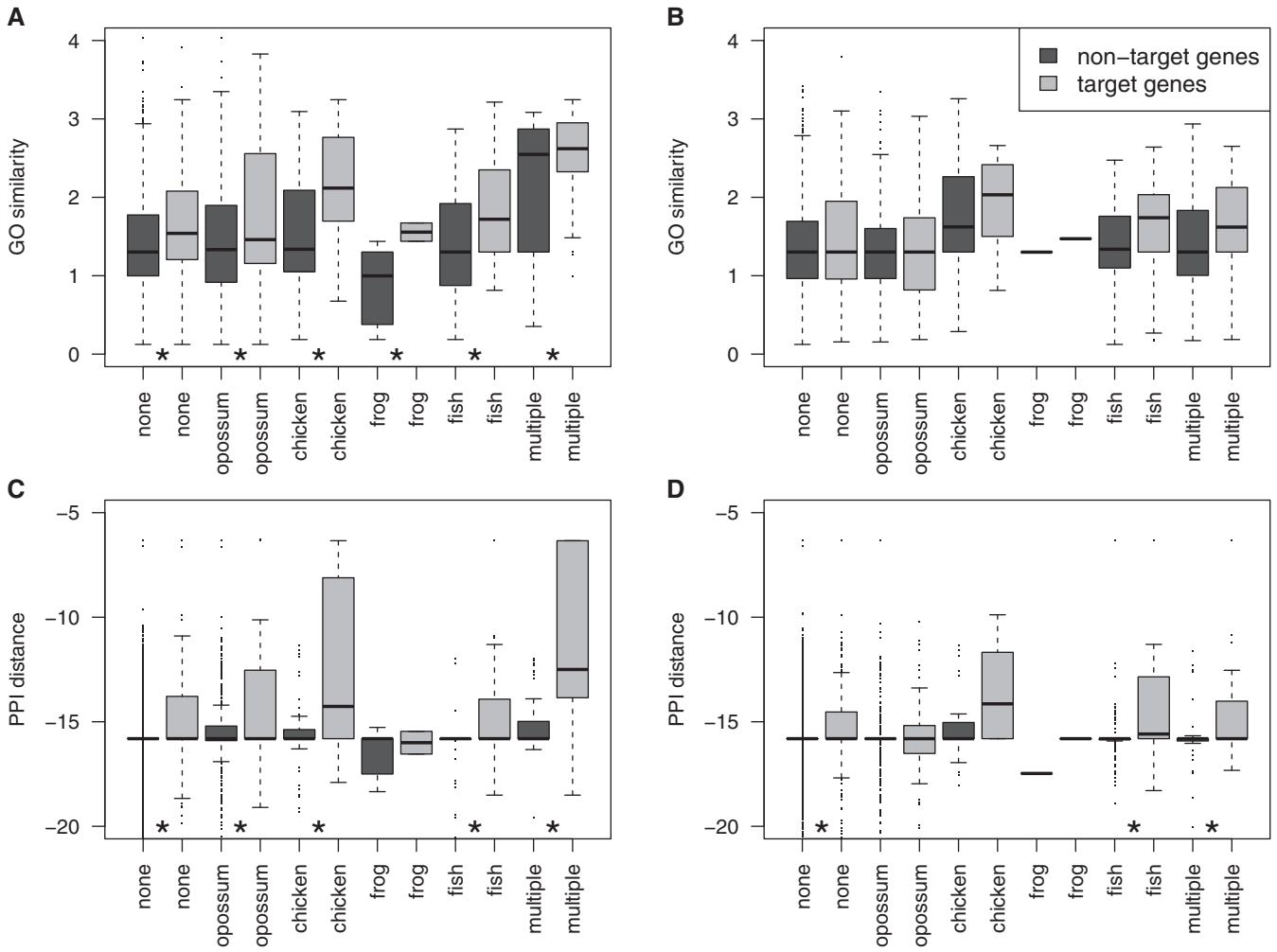

Figure 3. Since conserved synteny often fails to unambiguously predict a target gene, we tested whether GO similarity and PPI distances may help to resolve cases where multiple genes exhibit equal degrees of conserved synteny. For p300 enhancers from limb and forebrain, we identified all genes in intervals at $\Theta = 500$ kb with highest CSS but that cannot uniquely be assigned to the enhancer. We grouped this set into CSS classes that correspond to evolutionary distances from Supplementary Figure S3 with the exception that the label 'fish' indicates $1.72 < \text{CSS} \leq 2.3$] and 'multiple' corresponds to CSS>2.3. (**A** and **B**) GO similarities for target and non-target genes for p300 limb (A) and forebrain (B) enhancers. (**C** and **D**) PPI distance for target and non-target genes for p300 limb (C) and forebrain (D) enhancers. Comparison of target genes versus non-target genes within these subsets showed for all subclasses that target genes show a tendency for higher GO similarity and closer distances in PPI networks. $^*P < 0.05$, Wilcoxon test with Benjamini–Hochberg multiple testing correction.

transcription factor that is activated upon Shh signaling which specifies the anterior posterior axis in the developing limb bud and thus regulates the number of digits. Vokes *et al.* defined a high-quality set of 5274 Gli3 bound regions of which 2430 are located <1 Mb away of an Shh-repsonsive gene as identified by differential expression (26). Similar to the p300 data, conserved synteny predicts targets with higher precision but lower recall and the random forest approaches showed substantially better performance than any single feature-based prediction (Figure 4C and Supplementary Table S7).

DISCUSSION

Current research on long-range regulatory interactions is strongly focused on the computational detection and experimental validation of *cis*-regulatory elements (4,9). ChIP-seq experiments on the transcriptional coactivator p300 have proven to be a highly reliable method for experimental detection of enhancer regions in various tissues (6,8), but the identified sequences still have to be linked to their transcriptional targets.

Previous studies have postulated that evolutionarily constraints on enhancer-target gene interactions are likely to be responsible for the maintainance the conserved synteny in large genomic intervals (15,16,25). Kikuta *et al.* and Akalin *et al.* defined target genes as transcription factors with an HCNE density peak in human-zebrafish conserved-syntenic regions that were termed GRBs (16,25). This is in agreement with the observations that HCNEs are clustered around developmental genes and transcription factors (2,48), but it may not reflect the general pattern of enhancer-target gene interactions since previously defined GRBs (25) do not represent an

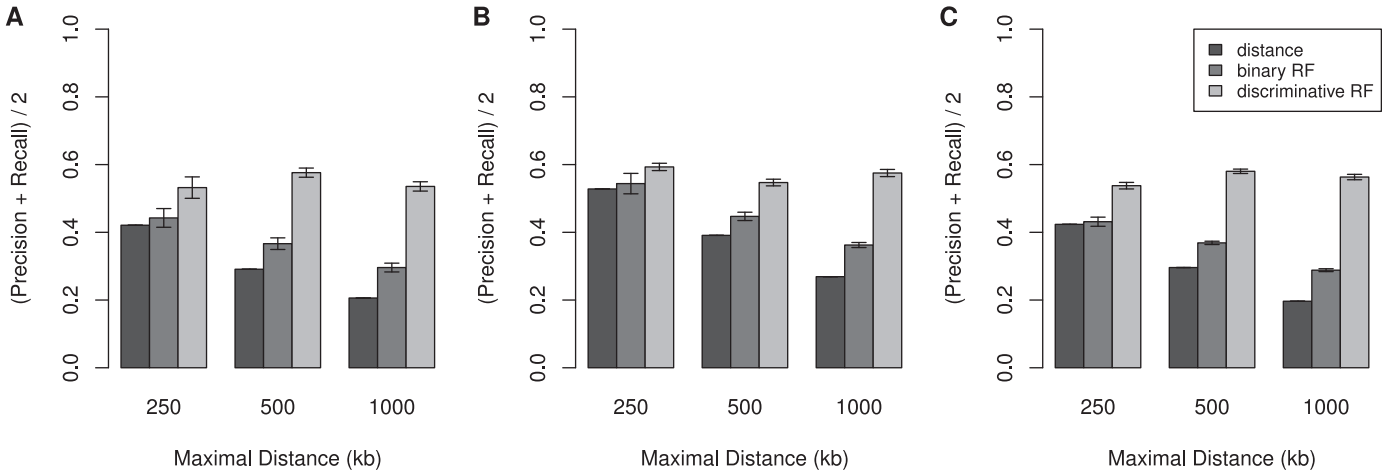

Figure 4. Evaluation of random forest classifier predictions on p300 ChIP-seq data for (**A**) limb, (**B**) forebrain and (**C**) 2430 Gli3 bound regions identified from ChIP-chip experiments (26). Average precision-recall values for predictions based on distance, and two random forest classifiers are shown. For the random forest models, the data were split into 80% training and 20% validation sets. The results shown are mean values after 10 repeated evaluations and standard errors. Combination of genomic, functional and protein interactome data allows correct target gene identification in 56–61% of cases for genomic intervals of 2 Mb.

exhaustive genome-wide collection of genes targeted by long-range regulation. For example, only 579 (12.7%) of ChIP-seq peaks from limb and forebrain overlap with aligned regions between mouse and zebrafish genomes and only 64 (7.7%) of upregulated genes in limb and forebrain overlap with the mouse orthologs of the GRB target genes. Therefore, p300 ChIP-seq data define a more general class of enhancer-target gene interactions that are less conserved and not exclusively restricted to transcription factors.

Two observations prompted us to use proximity in PPI networks and GO functional similarity as features for predicting enhancer targets. Feed forward loops and autoregulatory loops are common in gene regulatory networks (49). p300 binding in genomic regions that display conserved synteny with the Sox9 promoter suggests that it could be involved in the activation of SOX9 transcription (Figure 2). p300 also directly interacts with the SOX9 protein (40) and shares a number of known or predicted intermediary interaction partners in the protein interaction network (34). This suggested the hypothesis that the enhancer binding protein might display a relative proximity to its targets in the protein interaction network. Second, we hypothesized that the regulator would have a higher GO similarity to its targets than to non-target genes. Although GO similarity alone predicts target genes at larger distances ($\Theta > 500$ kb) with comparable recall values as genomic distance (Supplementary Table S4, S5 and S7), it cannot be utilized to predict non-transcription factor targets of very specific functions of p300 targets that are involved in cell adhesion (50) or erythropoiesis (51). The motivation of the random forest approach was therefore to exploit the complementary aspects of the four features. Our results demonstrate that the combination of features dramatically improved the prediction of target genes in genomic intervals of up to 2 Mb centered at the location of a p300 enhancer, with a recall of 58% compared with only 27% for genomic proximity and 12% for conserved synteny (Supplementary Table S5). The analysis of a second data set on Gli3 binding in embryonic mouse limbs displayed a similar advantage for the random forest predictions.

Since available data on enhancer-target gene interactions are extremely limited, we chose to interpret an upregulation of a gene in the vicinity of an enhancer to be the effect of direct regulation. This represents a limitation of our study, as the assumption that genes not found to be differentially expressed are not target genes may be incorrect, for instance because the differential expression may occur at a time point that was not measured. Another limitation is the assumption of our model that enhancers can regulate only one target gene. Enhancers may be active in various tissues (2) and multiple enhancers may coordinate the expression of one target gene (24). Nevertheless, under the assumptions of our study, we have shown that genomic distance, conserved synteny, PPI distance and functional similarity can be combined to dramatically improve predictions of the target genes.

The random forest classifiers that have been trained on limb and forebrain enhancers cannot be directly transferred to enhancer-target gene prediction in other tissues. Using the limited data now available, we have observed that the random forest classifiers are specific not only for the immunoprecipitated factor but also for the tissue (Supplementary Figure S7). However, more data will be needed to evaluate if this reflects variability between experiments or tissue specificity characteristics of regulatory interactions. With this proviso, our methodology can be applied to new ChIP-seq data to prioritize candidate enhancer-target gene interactions for validation experiments, and may also be useful for assessing the most biologically relevant hits identified by high-throughput chromosome conformation capture assays that have been developed to globally map

chromatin interactions (23,52,53). ChIP-seq is still a relatively new protocol and contains biases that are poorly understood (54); however, with more experimental data sets becoming publicly available, more detailed analyses can be performed to further evaluate how to combine functional classification of binding events and the association to target genes into an integrative downstream analysis of ChIP-seq experiments.

SUPPLEMENTARY DATA

Supplementary Data are availble at NAR Online.

FUNDING

Bundesministerium für Bildung und Forschung (BMBF, project number 0313911); Deutsche Forschungsgemeinschaft (SFB 760). Funding for open access charge: German Research Foundation (DFG).

Conflict of interest statement. None declared.

REFERENCES

1. Ovcharenko,I., Loots,G.G., Nobrega,M.A., Hardison,R.C., Miller,W. and Stubbs,L. (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Res.*, **15**, 137–145.
2. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Krys Kelly,K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
3. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
4. Pennacchio,L.M., Loots,G.G., Nobrega,M.A. and Ovcharenko,I. (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.
5. ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, **447**, 799–816.
6. 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.
7. Merika,M., Williams,A.J., Chen,G., Collins,T. and Thanos,D. (1998) Recruitment of CBP/P300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Mol. Cell*, **1**, 277–287.
8. 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.
9. Narlikar,L., Sakabe,N.J., Blanski,A.A., Arimura,F.E., Westlund,J.M., Nobrega,M.A. and Ovcharenko,I. (2010) Genome-wide discovery of human heart enhancers. *Genome Res.*, **20**, 381–392.
10. Warner,J.B., Philippakis,A.A., Jaeger,S.A., He,F.S., Lin,J. and Bulyk,M.L. (2008) Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods*, **5**, 347–353.
11. Cao,Y., Yao,Z., Sarkar,D., Lawrence,M., Sanchez,G.J., Parker,M.H., MacQuarrie,K.L., Davison,J., Morgan,M.T., Ruzzo,W.L. *et al.* (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev. Cell*, **18**, 662–674.
12. Kleinjan,D.A., Seawright,A., Elgar,G. and Heyningen,V. (2002) Characterization of a novel gene adjacent to Pax6, revealing synteny conservation with functional significance. *Mamm. Genome*, **13**, 102–107.
13. Kleinjan,D.A., Seawright,A., Mella,S., Carr,C.B., Tyas,D.A., Simpson,T.I., Mason,J.O., Price,D.J. and van Heyningen,V. (2006) Long-range downstream enhancers are essential for Pax6 expression. *Dev. Biol.*, **299**, 563–581.
14. Lettice,L.A., Heaney,S.J.H., Purdie,L.A., Li,Li., de Beer,P., Oostra,B.A., Goode,D., Elgar,G., Hill,R.E. and de Graaff,E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, **12**, 1725–1735.
15. Ahituv,N., Prabhakar,S., Poulin,F., Rubin,E.M. and Couronne,O. (2005) Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Hum. Mol. Genet.*, **14**, 3057–3063.
16. Kikuta,H., Laplante,M., Navratilova,P., Komisarczuk,A.Z., Engström,P.G., Fredman,D., Akalin,A., Caccamo,M., Sealy,I., Howe,K. *et al.* (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.*, **17**, 545–555.
17. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L. (2007) VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35(Database issue)**, D88–D92.
18. Pan,X., Stein,L. and Brendel,V. (2005) Synbrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, **21**, 3461–3468.
19. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
20. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.
21. Sinha,S., Liang,Y. and Siggia,E. (2006) Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res.*, **34(Web Server issue)**, W555–W559.
22. Müller,M., Chang,B., Albert,S., Fischer,N., Tora,L. and Strähle,U. (1999) Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. *Development*, **126**, 2103–2116.
23. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
24. D'haene,B., Attanasio,C., Beysen,D., Dostie,J., Lemire,E., Bouchard,P., Field,M., Jones,K., Lorenz,B., Menten,B. *et al.* (2009) Disease-causing 7.4 kb cis-regulatory deletion disrupting conserved non-coding sequences and their interaction with the FOXL2 promotor: implications for mutation screening. *PLoS Genet.*, **5**, e1000522.
25. Akalin,A., Fredman,D., Arner,E., Dong,X., Bryne,J., Suzuki,H., Daub,C., Hayashizaki,Y. and Lenhard,B. (2009) Transcriptional features of genomic regulatory blocks. *Genome Biol.*, **10**, R38.
26. Vokes,S.A., Ji,H., Wong,W.H. and McMahon,A.P. (2008) A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev.*, **22**, 2651–2663.
27. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with blastz. *Genome Res.*, **13**, 103–107.
28. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.*, **36(Database issue)**, D773–D779.
29. Miller,W., Rosenbloom,K., Hardison,R.C., Hou,M., Taylor,J., Raney,B., Burhans,R., King,D.C., Baertsch,R., Blankenberg,D. *et al.* (2007) 28-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Res.*, **17**, 1797–1808.
30. Krawchuk,D. and Kania,A. (2008) Identification of genes controlled by LMX1B in the developing mouse limb bud. *Dev. Dyn.*, **237**, 1183–1192.
31. Rödelsperger,C., Köhler,S., Schulz,M.H., Manke,T., Bauer,S. and Robinson,P.N. (2009) Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics*, **94**, 308–316.

32. Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., pp. 448–453.
33. Köhler,S., Schulz,M.H., Krawitz,P., Bauer,S., Dölken,S., Ott,C.E., Mundlos,C., Horn,D., Mundlos,S. and Robinson,P.N. (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, **85**, 457–64.
34. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37(Database issue)**, D412–416.
35. Köhler,S., Bauer,S., Horn,S. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–58.
36. Geurts,P., Irrthum,A. and Wehenkel,L. (2009) Supervised learning with decision tree-based methods in computational and systems biology. *Mol. Biosyst.*, **5**, 1593–1605.
37. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
38. Sandelin,A., Bailey,P., Bruce,S., Engström,P.G., Klos,J.M., Wasserman,W.W., Ericson,J. and Lenhard,B. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.
39. Bejerano,G., Lowe,C.B., Ahituv,N., King,B., Siepel,A., Salama,S.R., Rubin,E.M., Kent,W.J. and Haussler,D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, **441**, 87–90.
40. Furumatsu,T., Tsuda,M., Yoshida,K., Taniguchi,N., Ito,T., Hashimoto,M., Ito,T. and Asahara,H. (2005) Sox9 and p300 cooperatively regulate chromatin-mediated transcription. *J. Biol. Chem.*, **280**, 35203–35208.
41. Roca,H., Phimphilai,M., Gopalakrishnan,R., Xiao,G. and Franceschi,R.T. (2005) Cooperative interactions between RUNX2 and homeodomain protein-binding sites are critical for the osteoblast-specific expression of the bone sialoprotein gene. *J. Biol. Chem.*, **280**, 30845–30855.
42. Holleville,N., Matéos,S., Bontoux,M., Bollerot,K. and Monsoro-Burq,A. (2007) Dlx5 drives Runx2 expression and osteogenic differentiation in developing cranial suture mesenchyme. *Dev. Biol.*, **304**, 860–874.
43. Ghosh,A.K. and Varga,J. (2007) The transcriptional coactivator and acetyltransferase p300 in fibroblast biology and fibrosis. *J. Cell Physiol.*, **213**, 663–671.
44. Shikama,N., Lutz,W., Kretzschmar,R., Sauter,N., Roth,J., Marino,S., Wittwer,J., Scheidweiler,A. and Eckner,R. (2003) Essential function of p300 acetyltransferase activity in heart, lung and small intestine formation. *EMBO J.*, **22**, 5175–5185.
45. Schweikert,G., Zien,A., Zeller,G., Behr,J., Dieterich,C., Ong,C.S., Philips,P., de Bona,F., Hartmann,L., Bohlen,A. *et al.* (2009) mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.*, **19**, 2133–2143.
46. Adams,D., Karolak,M., Robertson,E. and Oxburgh,L. (2007) Control of kidney, eye and limb expression of Bmp7 by an enhancer element highly conserved between species. *Dev. Biol.*, **311**, 679–690.
47. Uchikawa,M., Ishida,Y., Takemoto,T., Kamachi,Y. and Kondoh,H. (2003) Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev Cell*, **4**, 509–519.
48. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
49. Kielbasa,S.M. and Martin Vingron,M. (2008) Transcriptional autoregulatory loops are highly conserved in vertebrate evolution. *PLoS ONE*, **3**, e3210.
50. Kim,Y., Lee,S., Ye,S. and Lee,J.W. (2007) Epigenetic regulation of integrin-linked kinase expression depending on adhesion of gastric carcinoma cells. *Am. J. Physiol Cell Physiol.*, **292**, C857–C866.
51. Engel,J.D. and Tanimoto,K. (2000) Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. *Cell*, **100**, 499–502.
52. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Bin Mohamed,Y., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
53. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
54. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, **9**, R137.