Published online 24 October 2016 Nucleic Acids Research, 2017, Vol. 45, No. 4 e19
doi: 10.1093/nar/gkw956

CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data

Timothy O'Connor[1,2], Mikael Bodén[2] and Timothy L. Bailey[3,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia, [2]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane 4072, Australia and [3]Department of Pharmacology, University of Nevada School of Medicine, Reno, NV 89557-0357, USA

Received March 08, 2016; Revised September 30, 2016; Editorial Decision October 06, 2016; Accepted October 10, 2016

ABSTRACT

Identifying the genomic regions and regulatory factors that control the transcription of genes is an important, unsolved problem. The current method of choice predicts transcription factor (TF) binding sites using chromatin immunoprecipitation followed by sequencing (ChIP-seq), and then links the binding sites to putative target genes solely on the basis of the genomic distance between them. Evidence from chromatin conformation capture experiments shows that this approach is inadequate due to long-distance regulation via chromatin looping. We present CisMapper, which predicts the regulatory targets of a TF using the correlation between a histone mark at the TF's bound sites and the expression of each gene across a panel of tissues. Using both chromatin conformation capture and differential expression data, we show that CisMapper is more accurate at predicting the target genes of a TF than the distance-based approaches currently used, and is particularly advantageous for predicting the long-range regulatory interactions typical of tissue-specific gene expression. CisMapper also predicts which TF binding sites regulate a given gene more accurately than using genomic distance. Unlike distance-based methods, CisMapper can predict which transcription start site of a gene is regulated by a particular binding site of the TF.

INTRODUCTION

Transcription factors regulate gene transcription by binding to specific regions of DNA called regulatory elements. This binding then activates or inhibits the action of transcriptional machinery at the transcription start site (TSS) of each gene it regulates. Particular TF binding sites are often unique to a specific cell type, condition, developmental stage or tissue (for brevity hereinafter referred to as a 'tissue'), and defective binding due to mutations in the bound region (e.g. 'regulatory SNPs' (1)) or in the TF itself (2) can cause dysregulation of genes and pathological phenotypes. Thus, two key questions are (i) which genes does a given TF regulate in a particular tissue, and, for a given gene, (ii) which binding sites of the TF affect its expression?

The current preferred method for determining the regulatory actions of a TF begins with predicting where it binds the genome in a given tissue using a chromatin immunoprecipitation followed by sequencing (ChIP-seq) assay (3). The next step usually assumes that each such predicted TF binding site (TFBS) regulates the closest gene, or that each gene is regulated by the closest TFBS, where distance is measured in bases (b) along the chromosome between a TSS of the gene and the TFBS.

This 'nearest neighbor' assumption works fairly well in practice for predicting the gene targets of a TF, since many TFs regulate by binding in the promoter of the target gene. However, a good deal of regulation is via distal enhancer regions and involves chromatin looping (4,5), which causes these distance-based methods to make incorrect predictions. In one human cell line (GM12878), fully 41% of chromatin loops connecting a non-promoter region to a promoter skip one or more intervening promoters (6), violating the 'closest gene' assumption. Similarly, if the target gene has multiple TSSs, distance-based methods cannot tell which TSS is the actual target of a TF bound at a nearby enhancer. Finally, if a TF binds at multiple locations near a gene, there is no guarantee that the closest site actually regulates the gene, as the 'closest TFBS' method assumes.

A number of methods for linking *regulatory elements* (such as enhancers) to target genes have previously been proposed that are not based on distance alone, but none have been tested with TFBSs predicted by TF ChIP-seq. The method of Ernst *et al*. (7) uses distance plus data for three histone modifications (H3K4me1, H3K4me2 and H3K27ac) and gene expression in a panel of tissues. It requires a supervised learning training step, and was not tested with regulatory elements predicted in a tissue not included in the panel. Similarly, Thurman *et al*. (8) showed that cross-tissue correlation of DNaseI hypersensitivity

*To whom correspondence should be addressed. Tel: +1 775 784 4651; Fax: +1 775 784 1620; Email: timothybailey@unr.edu

(DHS) between DHS regions overlapping promoters and DHS regions not overlapping promoters can predict regulatory relationships, but it is not clear how to extend their approach to linking TFBSs to promoters. DHS data are also available in far fewer organisms than histone modification data, restricting the applicability of that approach. The PreSTIGE algorithm (9) uses cross-tissue correlation of H3K4me1 and expression, but it was designed for linking enhancers (not TFBSs) to genes, requires CTCF binding data and only predicts links when both the H3K4me1 and expression signals are specifically enriched in a given tissue. He *et al*. (10) and Roy *et al*. (11) also proposed methods for training predictors of regulatory links between regulatory elements and genes using a large number of input features (e.g. histone modifications, DHS and TF ChIP-seq). These predictors are more accurate than the simple correlation-based approaches like PreSTIGE, but require data from many assays in order to make predictions in a tissue of interest.

We previously described a method for predicting links between enhancers and genes using cross-tissue correlation between histone modifications and gene expression (12), and in the current work we extend and validate that approach for TFBS-gene links. Our primary goal is to provide a method for analyzing peaks from TF ChIP-seq experiments that is as easy to use as distance-based methods, but is substantially more accurate. We propose a method we call CisMapper that, like distance-based methods, only requires the user to provide the genomic locations of predicted TFBSs. Rather than using distance, CisMapper infers regulatory links from the correlation between the presence of a selected histone modification (typically H3K27ac) at the TFBS and the expression of a gene across a panel of tissues in the same organism. We make available for free download the CisMapper software (suitable for OS X, Linux or Unix) and panels of histone and expression data for human (13) and mouse (14) from ENCODE, and for human from the Roadmap Epigenomics Project (15). We show that CisMapper is substantially more accurate than distance-based methods for predicting regulatory links between a TF's binding sites and specific TSSs, that the target tissue need not be present in the tissue panel, and that the target TF need not be expressed in all the panel tissues. We also show that accuracy increases with the number of tissues in the panel, and that CisMapper predictions can improve gene enrichment analyses.

MATERIALS AND METHODS

The CisMapper algorithm

Given a set of ChIP-seq peaks for a TF in some tissue along with auxiliary information in the form of expression and histone modification data for each of a panel of tissues in the same organism, CisMapper computes a score for a (peak, TSS) link using the correlation of expression at the TSS and the presence of the histone modification at the peak across the panel of tissues (Figure 1). Specifically, the score of a (peak, TSS) link is the *p*-value of the Pearson correlation coefficient between the log of the histone modification signal at peak and the log of the expression at the TSS. (Details are given in the Supplementary Material). We also tested using the Spearman rank correlation coefficient, but found it to give worse results (data not shown).

Here, we study using the active enhancer mark H3K27ac (16), the poised enhancer marks H3K27me3 and Zentner2011 (17), and the active promoter mark H3K4me3 (18), but in principle any histone mark could be used with CisMapper. (Note that CisMapper only uses data for a single histone mark at a time.) Using the *P*-value of the correlation as the score normalizes for panel size, allowing us to compare the effect of the score threshold across experiments with varying panel sizes. Although the correlation of a histone mark a ChIP-seq peak with expression at a TSS can be positive or negative, with positive correlation implying that the mark increases expression, since we are using histone marks indicative of active enhancers and promoters we restrict our analyses here to positive correlations.

CisMapper generates four ranked lists of predictions from the set of scored (peak, TSS) links. Two 'target' lists rank TSSs and genes, respectively, as potential targets of the ChIP-ed TF. The target score for a TSS is the minimum (best) score of any of its links. The target score for a gene is the minimum (best) target score of any of its TSSs. Two 'element' lists rank TF ChIP-seq peaks as potential regulators of TSSs and genes, respectively. The regulatory element lists group all the links for a given TSS or gene together, and sort within each group in increasing order by link score. Details of list creation are given in the Supplementary Material.

For practical reasons, it is necessary to restrict the set of possible (peak, TSS) links for which CisMapper computes link scores. First, in this work we restrict CisMapper to links where the TF ChIP-seq peak and the TSS are on the same chromosome and separated by at most 500 Kb. We do this to reduce the required compute time as well as to reduce the number of possible links with low (good) link scores merely due to chance. We note that previous studies that predicted enhancer-promoter links also chose to limit the maximum link length considered for similar reasons (e.g. 125 Kb in Ernst *et al*. (7), 500 Kb in Thurman *et al*. (8) and 2 Mb in He *et al*. (10)). Second, following related work by (19), CisMapper only computes scores for links where there is non-zero variation in the histone level at the peak and the variation in expression at the TSS meets certain criteria. (See Supplementary Methods for details.) Subject to the above caveats, CisMapper computes link scores for all possible (peak, TSS) pairs, so each peak can be linked to multiple TSSs, and vice-versa.

Validating predictions using chromatin contacts

We look for direct evidence of physical contact between CisMapper high-scoring (peak, TSS) pairs from promoter capture Hi-C (CHiC) data. We use these data to study (i) the coverage and accuracy of CisMapper predictions, (ii) the necessity of the target (ChIP-ed) tissue in CisMapper's panel and (iii) whether the ChIP-ed TF needs to be expressed in the panel tissues. The chromatin contact data we use are for GM12878 cells (6), which was the highest resolution data available when this study was conducted. To measure accuracy, we use the positive predictive value (PPV), which is equal to one minus the false discovery rate

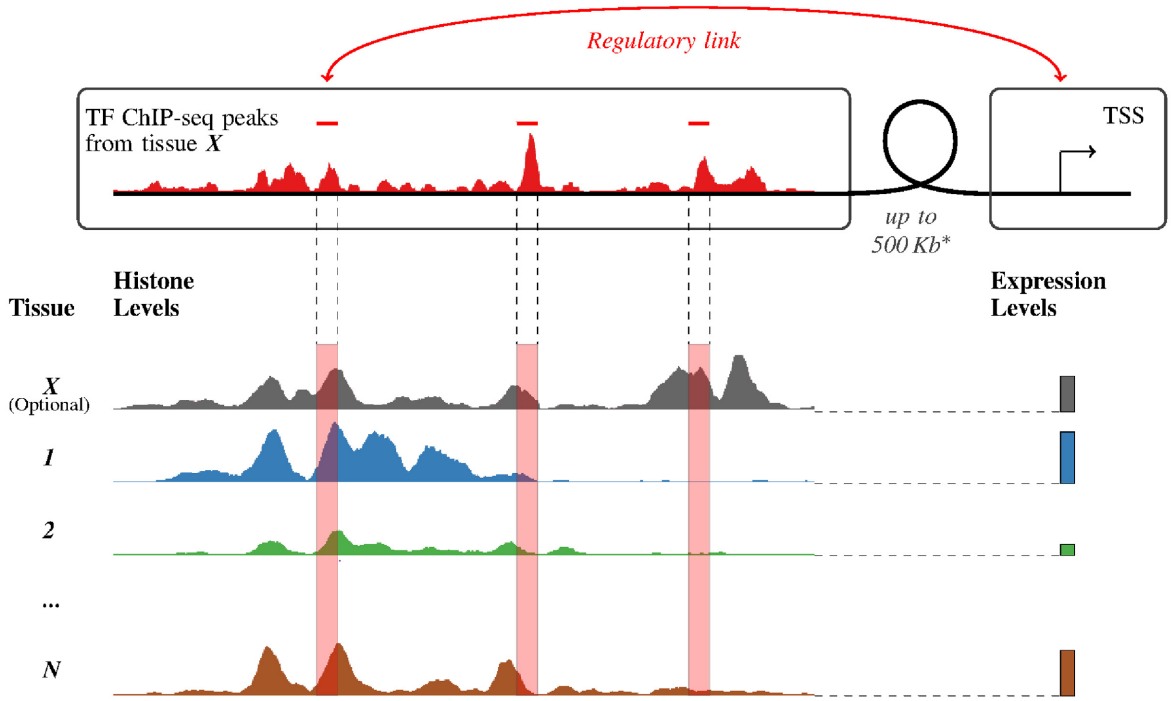

Figure 1. Schematic of the CISMAPPER method. CISMAPPER predicts regulatory links in tissue *X* between TF ChIP-seq peaks (red) and TSSs of genes by measuring the correlation of histone levels (shown as colored tracks) that overlap peaks (highlighted in red) with expression levels across a set ('panel') of tissues. Tissue *X* need not be present in the panel. *Distance limit is user configurable with 500 Kb chosen for this work.

(1−FDR), where a predicted link is confirmed if its two ends overlap the two ends of a promoter-other chromatin contact in the Mifsud *et al.* (6) data. The CISMAPPER panel consists of eight tissues—GM12878, Ag04450, H1-hESC, HeLa-S3, HepG2, HUVEK, K562 and NHEK—and the histone (H3K27ac) and expression data (CAGE) come from ENCODE (Supplementary Table S2 lists data sources). TF ChIP-seq peaks are for the 19 TFs in Supplementary Table S1 with ENCODE ChIP-seq data in GM12878 cells. Further details are given in Supplementary Methods.

Validating predictions using differential TF activity

Sikora-Wohlfeld *et al.* (20) developed the 'differential TF activity' evaluation method and used it to evaluate a large number of distance-based predictors of regulatory interactions from TF ChIP-seq data. This evaluation method uses sets of TSSs that are differentially expressed in two tissues in which the ChIP-ed TF is active. They reasoned that if a TF is active in both tissues, some of the changes in gene expression between those tissues should be due to changes in activity of the TF. Hence, the top 500 differentially-expressed TSSs should be enriched for direct targets of the ChIP-ed TF. The figure of merit is the size of the overlap between the top 500 differentially-expressed TSSs and the top 500 predictions of predictor being evaluated, minus size of overlap expected if the predictor guessed randomly. Sikora-Wohlfeld *et al.* (20) found that the differential TF activity evaluation method gave results consistent with other evaluation methods that use TF perturbation data, functional homogeneity of target genes or consistency of target gene predictions across multiple ChIP-seq data sets, respectively. Note that although we use the evaluation method of Sikora-Wohlfeld *et al.* (20), we do not use their data or results. A diagram (Supplementary Figure S2) and further details are given in Supplementary Methods.

Validating predictions using gene enrichment analysis

We analyze the enrichment of genes predicted by CISMAPPER or GREAT (21) to be associated with TF ChIP-seq peaks for p300 in embryonic (E14.5) mouse neocortical tissue from Table S1 of Wenger *et al.* (22). For CISMAPPER we use Mouse ENCODE histone (H3K27ac) and expression (long polyA+) for a panel of 22 mouse tissues listed in Supplementary Table S11 and a distance limit of 500 Kb. We use the target gene list produced by CISMAPPER with a link score threshold of 0.01. We then apply the DAVID (23,24) on-line gene enrichment tool to the gene targets predicted by CISMAPPER to determine enriched Gene Ontology (25) terms. For comparison, we perform enrichment analysis on the same TF peaks using GREAT with its default region-gene association rule. This associates each peak with every gene whose 'genomic region' it overlaps. GREAT defines the genomic region of a gene as a basal domain of −5 Kb to +1 Kb around its TSS, which it then extends that up to 1 Mb in either direction, stopping if it encounters another gene's basal domain.

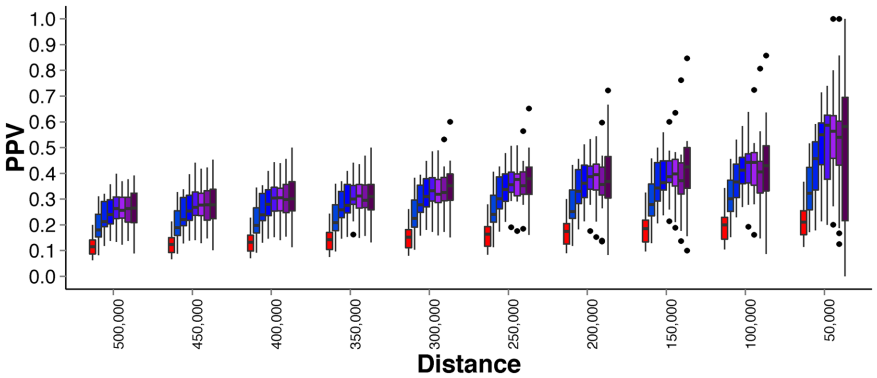

Figure 2. Distribution of the accuracy of peak-TSS links of different maximum lengths predicted by CISMAPPER for 19 TFs in GM12878 cells. The plot shows the distribution of the accuracy (PPV) of predicted links with lengths less than a given distance and CISMAPPER link scores less than or equal to 1, 0.1, 0.01, 0.001, 10^{-4}, 10^{-5}, 10^{-10} or 10^{-20} (blue to purple boxplots, from left to right for each distance). Links are validated using CHiC contact data. The red box plots (score ≤ 1) correspond to predicting that TF peaks regulate every TSS within the given distance from them. All CHiC and TF ChIP-seq data are from GM12878 cells. CISMAPPER links were scored using H3K27ac histone and CAGE expression data from a panel of eight tissues including GM12878, and only positive correlations are considered. The box plots summarize the results for 19 sets of TF ChIP-seq peaks; boxes show the range of the middle quartiles with a line at the median, and dots are outliers further than 1.5 times the interquartile range (the whiskers) from the median.

RESULTS

CISMAPPER accurately predicts contacts between promoters and TF-bound regions

We first demonstrate that CISMAPPER can accurately predict the long-distance contacts between TF-bound regions and promoters to be expected when a distal TFBS regulates a gene. For validation we use CHiC chromatin contact data (see Materials and Methods), and observe that CISMAPPER predicted (peak, TSS) links are greatly enriched for chromatin contacts compared with links predicted by distance. Using a panel of eight tissues and TF ChIP-seq peaks for 19 TFs in GM12878 cells, the potential regulatory links predicted by CISMAPPER with link scores less than 0.01 are at least 73% more likely to be confirmed by CHiC chromatin contact data than all links of the same length (Figure 2). High-confidence CISMAPPER links (score $< 10^{-5}$) shorter than 50 Kb have a median PPV of 0.57 across 19 TF ChIP-seq data sets, whereas all potential (peak, TSS) links shorter than 50 Kb have a median accuracy of only 0.21 (2.7-fold improvement).

As shown in Figure 2, the median accuracy of CISMAPPER-predicted links is higher than that of all similar length links for all tested score thresholds (from 0.1 to 10^{-20}) and for all tested link lengths (50–500 Kb). The maximum improvement in accuracy is seen for short links ($d < 50$ Kb) and score thresholds below 0.001 (2.7-fold improvement in median PPV). Prediction accuracy increases with decreasing link length and increasing score stringency, with a maximum median PPV of 59% for links shorter than 50 Kb and a score threshold of 10^{-4} or lower. The higher accuracy of CISMAPPER predictions relative to distance-based predictions is consistent across the 19 TF ChIP-seq data sets analyzed here (Supplementary Figure S4C). The PPV of all CISMAPPER links predicted at a score threshold of $<10^{-5}$ ranges from a high of 37% for RXRA to a low of 12% for ZBTB33. For all 19 of the TFs studied in this experiment, the PPV of CISMAPPER predictions is higher than that of links predicted using a distance threshold yielding a similar length distribution (350 Kb).

CISMAPPER's approach is clearly superior to using distance alone for predicting specific regulatory interactions between a bound TFs and TSSs. What is more, predicted links can easily be thresholded on both link score and link length (as done in Figure 2) to select links with high probability (>50%) of corresponding to contacts between promoters and TF-bound regions (Supplementary Figure S6). In this experiment, prediction accuracy for links predicted using a CISMAPPER score threshold of 0.01 drops below 10% (see Supplementary Figure S5) for the subset of links with lengths in the range 450–500 Kb. While this level of accuracy is still nearly twice as high as using a distance threshold alone, the 500 Kb limit on link length we have chosen here may be a reasonable value in practice.

Although the coverage (recall) of CISMAPPER is relatively low compared to using a simple distance threshold (Supplementary Figure S4A and B), we would argue that this is a reasonable trade-off in circumstances where a set of predicted regulatory links is desired for further examination. Higher PPV means lower FDR, so if predictions will be tested via expensive wet-lab experimentation, a smaller set of predicted links of higher precision may be preferable to a larger set of links that contains a higher proportion of false positives.

The target tissue need not be present in the panel

We wondered if CISMAPPER could successfully predict potential regulatory interactions using TF ChIP-seq data from a tissue not included in its panel. If true, this would greatly expand its utility. To examine this question we repeated the CHiC validation experiment after removing the target tissue (GM12878) from CISMAPPER's panel. As seen in Figure 3, using a score threshold of 10^{-5} CISMAPPER's predictions are still substantially more accurate at all distance thresholds than distance alone. On the other hand, including the target tissue in the panel does increase accuracy, especially for links shorter than 50 Kb. It is clear, therefore,

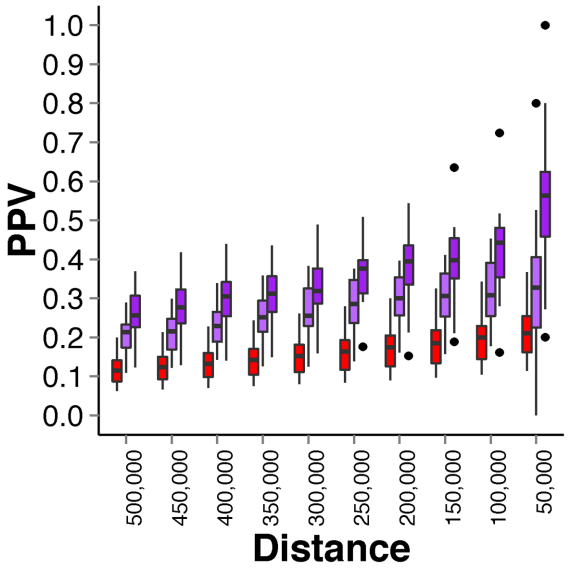

Figure 3. Including the ChIP-ed tissue in CISMAPPER's panel improves accuracy. The plot shows the distribution of the accuracy (PPV, *y*-axis) of predicted links with lengths less than a given distance (*x*-axis) and CISMAPPER link scores less than or equal to 1 (red, distance-only) or 10^{-5} when we exclude (purple) or include (dark purple) the ChIP-ed tissue (GM12878) in the tissue panel. The data and methods are the same as in Figure 2.

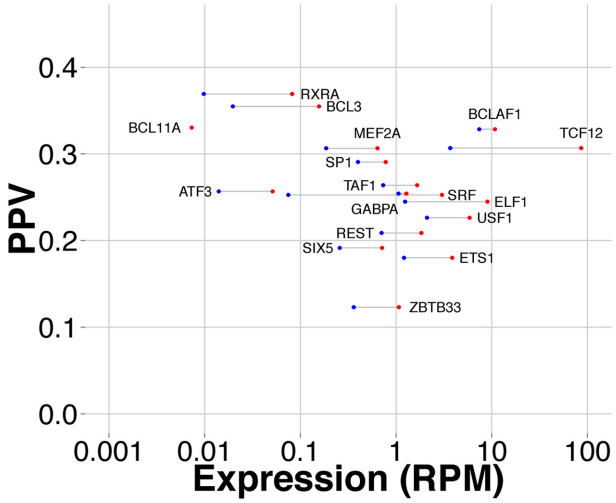

Figure 4. Accuracy does not depend strongly on expression of the ChIP-ed TF in the tissue panel. Each point shows the accuracy (PPV) and either the maximum (red) or median (blue) level of expression (reads-per-million, RPM) of a single TF across the panel of eight tissues used by CISMAPPER to predict regulatory links (score $< 10^{-5}$) between ChIP-seq peaks for the TF and TSSs in GM12878 cells. The TFs are labeled and their points are connected with a gray line. The data are from the same experiments as in Figure 2. (The median expression of BCL11A is zero and is not plotted).

that CISMAPPER is useful for analyzing TF ChIP-seq peaks from tissue types not included in its tissue panel, but accuracy will be better if the panel includes the tissue in which the TF was ChIP-ed.

The ChIP-ed TF need not be expressed in all panel tissues

We also wondered if the ChIP-ed TF needs to be expressed across CISMAPPER's tissue panel. Consequently we examined the relationship between the accuracy of predicted regulatory links and the level of expression of the TF across the panel for the CHiC validation experiments. As can be seen in Figure 4, there is no discernible relationship between accuracy (PPV) and the expression of the ChIP-ed TF across the panel. For example, the median expression of a single TF varies by four orders of magnitude (from 0.01 to 100 reads-per-million, Figure 4, blue), but this has no consistent effect on the accuracy of CISMAPPER's predictions. The TF for which CISMAPPER's predictions are most accurate is RXRA, which has the smallest median and third smallest maximum of expression across the panel of tissues used by CISMAPPER (data not shown). In fact, RXRA has no measurable expression (according to the ENCODE CAGE data used here) in two of the eight tissues, including in GM12878, the tissue in which it was ChIP-ed. Two other TFs have no measurable expression in *five out of eight* tissues (data not shown), yet they rank third (BCL11A, PPV = 0.33) and eighth (PU.1, PPV=0.27) in accuracy among the 19 TFs tested here (Supplementary Figure S4C).

Some TFs show highly tissue specific expression, so we wondered if CISMAPPER could predict regulatory links for them even if they were not expressed in *any* tissue included in its panel. We therefore repeated our validation using chromatin contacts after removing any tissue from the panel where the ChIP-ed TF showed measurable expression. In this new experiment, we selected five additional TFs (RUNX3, PAX5, IRF4, IKZF1 and BATF) with ENCODE ChIP-seq peaks in GM12878 because these TFs have measurable expression in GM12878 and at most two other panel tissues. When we exclude these tissues, each panel contains at least five of the original eight panel tissues (but the number and identities of the tissues varies depending on the TF). CISMAPPER's predictions are still more accurate at all distance thresholds than distance alone (Supplementary Figure S8) for these five 'tissue specific' (with respect to the panel) TFs. The ability to make predictions for a TF not expressed in any tissue in the histone/expression panel is likely due to the fact that the TF binds in enhancer regions that are active (and varying) across the panel.

CISMAPPER is more accurate than distance-based methods

We next explore how CISMAPPER accuracy compares with distance-based approaches. A recent survey of distance-based methods for linking TF ChIP-seq peaks to genes studied six methods and found two—LINEAR and CLOSESTGENE—to be consistently superior to the others they tested (20). The window-based LINEAR method simply adds a value between 0 and 1 to a gene's score for each peak within 10 Kb of the gene's TSS, where the value added decreases linearly with the peak-TSS distance. The CLOSESTGENE method assigns each peak to the nearest gene, then scores the peak based on how well the distance fits the observed distribution of peak-TSS distances, and finally sums all the peak scores for each gene. We applied CISMAPPER, LINEAR and CLOSESTGENE to ChIP-seq data for 27 TFs in a variety of tissues (Supplementary Table S1), and estimated the accuracy of the predictions using Sikora-Wohlfeld *et al.*

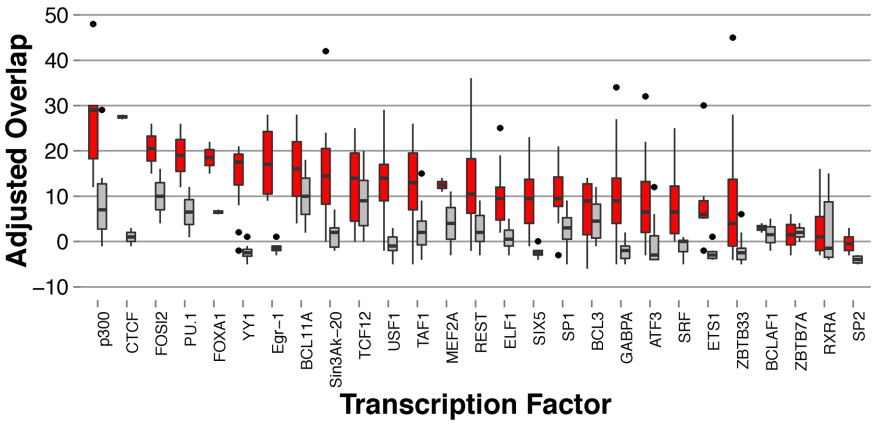

Figure 5. Validation of CISMAPPER target TSS predictions using differential TF activity. The figure shows the accuracy ('Adjusted Overlap') of TSS target predictions from CISMAPPER (red) or CLOSESTGENE (grey) for 27 different TFs (*X*-axis). Each prediction method is allowed to predict 500 TSS targets for a TF, and the adjusted overlap is the number of predicted TSS targets that are also among 500 most differentially expressed TSSs between the ChIP-ed tissue and another tissue, minus the expected size of the overlap by chance. The box-and-whisker plots summarize the results of between 2 and 30 experiments involving ChIP-seq peaks for the given TF. Outliers that are further than 1.5 times the interquartile range from the median are shown as black dots in the plots. All CISMAPPER maps are built using H3K27ac histone data and CAGE expression data, and the differentially expressed TSS sets are also based on CAGE data.

(20)'s 'differential TF activity' evaluation method (see Materials and Methods).

Overall, CISMAPPER predictions are substantially more accurate than those made by CLOSESTGENE (Figure 5) or LINEAR (Supplementary Figure S9). The median accuracy of the TSS target predictions made by CISMAPPER is higher than that of CLOSESTGENE for 26 out of 27 TFs tested ($P < 10^{-6}$, sign test), and higher than that of LINEAR for 25 of 27 TFs tested ($P < 10^{-5}$, sign test). For 26 out of 27 TFs, CISMAPPER correctly identifies between 1.5 and 26.5 more TSS targets than CLOSESTGENE, and correctly identifies 10 times more TSS targets on average (Supplementary Table S4). CISMAPPER is also more accurate than CLOSESTGENE for predicting gene (rather than TSS) targets for 20 of 27 TFs (Supplementary Figure S10, $P < 0.01$, sign test). Here the CISMAPPER panel of tissues draws from six of the eight following tissues: Ag04450, GM12878, H1-hESC, HeLa-S3, HepG2, HUVEC, K562 and NHEK; the CAGE expression and H3K27ac histone data is from the ENCODE sources listed in Supplementary Table S2 (see Supplementary Methods for details).

To check the consistency of our two evaluation methods, we looked at how they ranked the accuracy of CISMAPPER predictions on the 19 TF ChIP-seq data sets that we evaluated using both methods. In both these evaluations, CISMAPPER based its predictions on an enhancer mark (H3K27Ac), so we divided the 19 TFs into two groups according to their preference for binding in enhancer regions, based on data from Ernst *et al.* (26). Supplementary Table S5 shows that for six of the seven TFs that bind preferentially in enhancer regions, CISMAPPER predictions are ranked highly by both evaluation methods. The notable exception is that the two evaluation methods disagree strongly on the accuracy of the CISMAPPER predictions for the RXRA ChIP-seq data set. This anomaly may be due to poor quality of the RXRA ChIP-seq data set. There is no significant enrichment of any of the known motifs for RXRA from the JASPAR database (27) in the RXRA ChIP-seq peaks based on a CentriMo (28) motif enrichment analysis (data not shown). The high PPV of the links predicted by CISMAPPER in the RXRA data set according to the chromatin contact evaluation method suggests that those ChIP-seq peaks frequently contain regions in contact with neighboring genes. The low accuracy according to the differential TF activity evaluation is not surprising given the lack of evidence of actual RXRA binding in the peaks. Thus, with the exception of the RXRA data set, both evaluation methods estimate the accuracy of CISMAPPER predictions based on an enhancer mark to be generally highest for TFs binding primarily in enhancer regions, as would be expected.

CISMAPPER can use a variety of histone marks

Thus far we have only presented results based on using the active enhancer histone mark H3K27ac in CISMAPPER's tissue panel. When we repeat the TSS target prediction experiment above using histone data for the active promoter histone mark H3K4me3 in place of the H3K27ac data used above, CISMAPPER is more accurate than CLOSESTGENE, although the comparative advantage is smaller than when using H3K27ac (Supplementary Figure S11). For 21 of 27 TFs, the median accuracy of CISMAPPER predictions is higher than that of CLOSESTGENE ($P < 0.003$, sign test), compared with 26 of 27 TFs when CISMAPPER uses H3K27ac data (Figure 5). We also examined using histone marks H3K27me3, associated with poised enhancers (29) and H3K36me3, associated with active enhancers and transcribed genes (17). We found that the accuracy of predicted links was somewhat lower using these two marks (data not shown). These results suggests that CISMAPPER can be used effectively with ChIP-seq data for histone marks other than H3K27ac should data for that mark not be available for enough tissues to build a panel (see next section).

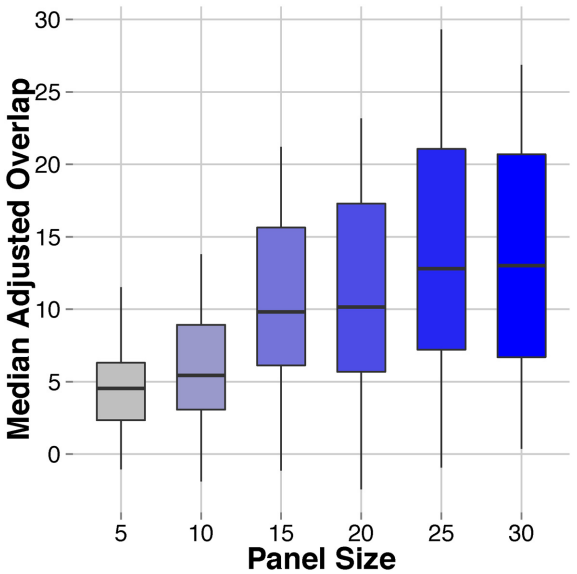

Figure 6. Accuracy increases with panel size. The plot summarizes the distribution of the median adjusted overlap score (y-axis) for the top 500 gene target predictions of 19 TFs as the number of tissues used by CISMAPPER ('Panel Size', x-axis) increases. The TFs used are those with three or more ChIP-seq peak sets in Supplementary Table S4. The expression data used for validation comes from CAGE expression in ENCODE tissues. CISMAPPER gene target predictions use histone (H3K4me3) and expression (long polyA+) data from subsets of the 38 tissues from the Roadmap Epigenomics Project. Smaller panels are always a subset of the next larger panel, and values in the figures are averages over 15 independent nested panel sets.

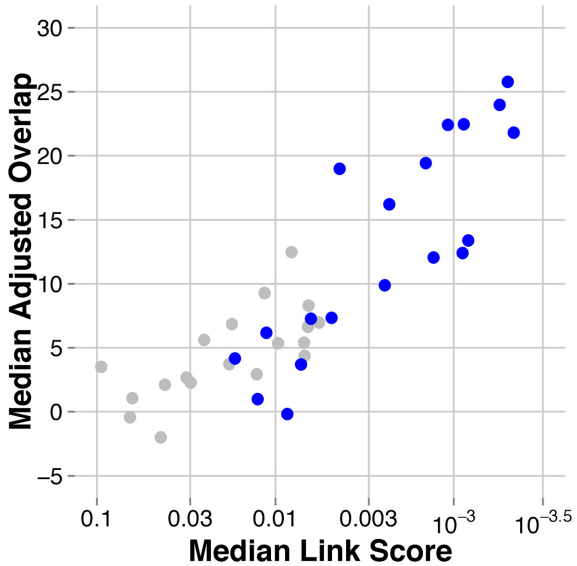

Figure 7. Score calibration does not depend on panel size. Each point represents a single TF, and shows the median link score of the 500th ranked target (x-axis) versus the median accuracy of CISMAPPER of those 500 target gene predictions (y-axis), averaged across all ChIP-seq peak sets for the TF, with tissue panels of size 5 (grey) or 30 (blue). The data are from the same experiments as in Figure 6.

Increasing panel size improves CISMAPPER coverage and accuracy

We assumed that CISMAPPER coverage and accuracy should increase with the size of the panel of tissues it uses for computing peak-TSS correlations. To test this we again used the differential TF activity method, but switched to data from the more extensive Roadmap Epigenomics Project (15) to allow us to create panels of from 5 to 30 tissues using histone ChIP-seq data for H3K4me3, and polyA+ RNA-seq expression data. Since RNA-seq data does not identify the TSS as accurately as CAGE data, we use the gene target list output by CISMAPPER rather than its TSS target list in this evaluation. (See Supplementary Methods for details.)

The accuracy of CISMAPPER target predictions increases with the panel size (Figure 6). The median of the adjusted overlap score almost triples over the range of panel sizes we tested (5–30). What is more, the coverage of CISMAPPER target predictions increases with the panel size (Supplementary Figure S12A), as might be expected due to the increased statistical power of larger panels. A similar increase in accuracy between tissue panels of size 5 and 30 is seen for each of the 19 individual TFs we tested (Supplementary Figure S12B). Although we observe a plateau in the accuracy of CISMAPPER gene target predictions when the panel size reaches 25 tissues (Figure 6), for most of the 19 TFs we tested, the number of gene targets predicted by CISMAPPER at a link score threshold of 0.001 more than doubles. This plateau is probably due to limitations in the available data reducing the diversity of any additional tissues added to the panel beyond 25. (See Supplementary Methods for further discussion of this issue.)

CISMAPPER scores are calibrated

Using data from the previous section, we checked that CISMAPPER scores are 'calibrated' in the sense that a given score corresponds to the same accuracy regardless of panel size. This is evidenced by the scatter plot in Figure 7, which shows the accuracy (y-axis) of gene target predictions using the link score threshold given on the x-axis. Each point represents the median CISMAPPER results for one TF ChIP-seq data set, averaged over the different tissue subset panels, as described above. The X-value of each point is the median of the link score of the 500th gene in the target list, and the Y-value is the median accuracy (adjusted overlap scores).

Two things are clear from Figure 7. First, there is a very strong correlation between the CISMAPPER link score threshold and gene target prediction accuracy. Secondly, the slope of this correlation is essentially unchanged when CISMAPPER uses a panel of five tissues (grey points) or 30 tissues (blue points). This implies that the prediction accuracy when using a given link score threshold does not depend strongly on panel size. Therefore, a reasonable choice of link score threshold will remain so regardless of how many paired histone-expression data sets are provided as input to CISMAPPER. Thus, the main effect of increasing panel size is to increase the coverage (number of predictions) at a given link score, while maintaining the accuracy of those predictions.

CISMAPPER predictions can improve gene enrichment analyses

Perhaps the most common downstream analysis applied to TF target gene predictions is gene enrichment analysis, and we wondered if this type of analysis would benefit from the improved accuracy of CISMAPPER predictions. To address this, we compare gene enrichment analysis of gene targets predicted by CISMAPPER with a similar analysis using the distance-based enrichment analysis tool GREAT (21). The TF ChIP-seq peaks are for p300 in embryonic (E14.5) mouse neocortical tissue (22). Given the tissue and stage of neocortical development, we expect p300-bound regions to regulate many neural-development related functions.

In this example, the gene enrichment analysis based on the CISMAPPER predicted targets appears more informative than analysis based on distance-based target prediction (see Supplementary Tables S8, S9 and S10). Although the GREAT tool identifies many neural-related biological processes and molecular functions enriched among its predicted 4676 gene targets (22), the 938 gene targets predicted by CISMAPPER are enriched for important neural-related processes and functions that are not identified by GREAT. For example, only the CISMAPPER-predicted targets are enriched for genes involved in the neural projection biological process (Supplementary Table S8), a critical process in neuron formation within the cortex (30). CISMAPPER also scores a key regulator of neural projection in neuron development, *Fezf2* (31,32), as a top target.

Furthermore, CISMAPPER predictions identify genes primarily enriched in ion transport and charge potentiation molecular functions (Supplementary Table S9), crucial to the excitatory function of pyramidal neurons in the neocortex (33). These are missing from the GREAT predictions, which mainly identify transcription-related functions.

Finally, there are no enriched cellular component terms among the GREAT-predicted gene targets, whereas terms highly relevant to neocortical neurons such as 'neural projection', 'plasma membrane' (the location of ion channels), 'axon' and 'synapse' are enriched among the CISMAPPER-predicted gene targets (Supplementary Table S10).

DISCUSSION

Several previous studies have sought methods for accurately identifying the gene targets of regulatory regions (7–11,34) using auxiliary data on gene expression, TF binding, DNaseI hypersensitivity and histone modifications. Although demonstrably more accurate, these methods have not supplanted simple distance-based association of TF ChIP-seq peaks with putative target genes in practice. This is probably due mainly to the relative simplicity of distance-based methods, as well as to the fact that the more advanced methods have not been explicitly validated on regulatory regions defined by TF ChIP-seq peaks. We developed CISMAPPER to provide a method that is more accurate than simple distance-based methods, but that places a minimum burden on the user to provide auxiliary data. CISMAPPER uses only data for a single histone modification and gene expression across a small panel of tissues, requires no training step and has been extensively evaluated here as an alternative to distance-based methods for analyzing TF ChIP-seq peaks.

CISMAPPER can analyze TF ChIP-seq peaks to predict regulatory links between TF binding sites and the TSSs of genes. It predicts these links using cross-tissue correlation between histone marks overlapping the TF binding site and expression at the TSS. The target lists output by CISMAPPER can be used to predict either which TSSs or which genes a given TF regulates. Similarly, the regulatory element lists it outputs can be used to predict which specific TF binding sites are most likely to regulate a given TSS or gene.

We have shown that the regulatory links predicted by CISMAPPER coincide with chromatin contacts at a higher rate than links predicted based on the distance between the binding site and the TSS, the current method of choice. Direct chromatin contact between a bound TF and a TSS is highly suggestive of a possible regulatory interaction, which is what CISMAPPER is intended to predict. We also report experiments using the differential TF activity evaluation method to show that CISMAPPER's lists of the gene and TSS targets of a TF have higher accuracy than predictions made by distance-based methods. We have also shown that CISMAPPER is especially accurate for predicting long-distance regulatory links that are beyond the reach of distance-based prediction methods, and that as more histone and expression data become available across a larger number of tissues, the accuracy of CISMAPPER's regulatory predictions will improve. Based on these results, we believe that CISMAPPER is a valuable addition to the standard bioinformatic toolkit for analyzing TF ChIP-seq data.

Importantly, we have shown that CISMAPPER requires neither histone nor expression data from the tissue of interest, only the genomic loci of the ChIP-seq peaks for a TF in that tissue. However, if such histone and expression data are available, it can and should be included in CISMAPPER's input, as we expect it to improve prediction accuracy.

We have also shown that CISMAPPER does not require the TF to be expressed in *any* of the panel tissues to accurately predict regulatory links to its TFBS. This suggests that even if a TF's expression is tissue specific, CISMAPPER can still detect when it binds to enhancers showing varying activity across CISMAPPER's tissue panel.

Suitable compendia of histone mark and expression data currently exist for using CISMAPPER to analyze TF ChIP-seq data from human, mouse, fly and worm. For analyzing human data, extensive histone and expression data are available from the Roadmap Epigenomics Project (15), from ENCODE (13) and from FANTOM5 ((35); expression data only). Data for mouse are available from the mouse ENCODE project (14), and a mouse blood-specific compendium has been published recently (36). The modENCODE project provides data for both fly (37) and worm (38). Each of these compendia contain matched histone and expression data from seven to over 100 tissues, and our results show that CISMAPPER can make useful regulatory predictions when provided with such data for as few as five tissues in the organism of interest.

While we have shown that CISMAPPER predictions are more accurate than distance-based predictions, the coverage of CISMAPPER and distance-based methods is quite distinct. On the one hand, distance-based methods are confounded when chromatin looping causes a TF binding site to regulate a TSS other than the nearest one. On the other

hand, CISMAPPER can only predict a regulatory link between a TF binding site and a TSS when there is variation in their histone mark and expression, respectively, across the tissues in the histone/expression compendia provided to CISMAPPER. Consequently, the regulatory predictions made by CISMAPPER are somewhat complementary to those made by distance-based methods.

Due to the complementarity of the distance- and correlation-based approaches to regulatory interaction prediction, a future version of CISMAPPER will integrate genomic distance directly with histone-expression correlation in calculating the link score. We anticipate this will improve CISMAPPER's coverage. In the mean time, we recommend analyzing TF ChIP-seq peaks with both CISMAPPER and a distance-based method. The CISMAPPER predictions will provide a higher quality set of predicted targets and regulatory binding sites, and the union of those predictions with the distance-based predictions will provide a higher-coverage, albeit less-accurate, set.

CISMAPPER predictions of regulatory links are also complementary to those inferred from chromatin conformation capture (CCC) data because they are based on completely different types of evidence. Specifically, the link score that CISMAPPER calculates for a pair of loci indicates how related histone and expression levels are between the loci, whereas, the read count for a pair of loci produced by a conformation capture assay, after conversion to a score that corrects for distance-dependent and other biases, can be used to infer if the two loci are in contact. Thus, CISMAPPER and chromatin conformation capture assays (e.g. 3C (39), 4C (40), 5C (41), Hi-C (42), ChIA-PET (43) or CHiC (6)) provide scores that are independent predictions of regulatory interactions between pairs of genomic loci. This independence suggests that intersecting the sets of loci pairs predicted by CISMAPPER with those predicted by CCC in the same tissue should yield an even more accurate set of predicted regulatory interactions.

Analyses of the regulation of expression by a transcription factor should benefit from CISMAPPER's more accurate and highly specific predictions of regulatory links between its binding sites and particular TSSs. For example, when searching for regulatory SNPs, it is reasonable to assume that those contained in TF binding sites predicted by CISMAPPER to be regulatory are more likely to be important biologically. (Note that we assume that the binding *sites* can be identified within the TF-bound regions predicted by CISMAPPER via standard motif-based methods (44).) Likewise, gene ontology analysis (25) performed using the more accurately predicted target gene set provided by CISMAPPER should better elucidate the biological roles of the ChIP-ed TF. Finally, when validating predicted regulatory binding sites via genome editing (e.g. using CRISPR/Cas (25)), CISMAPPER's ability to associate specific binding sites with a gene and to rank them by regulatory potential should prove invaluable.

The use of CISMAPPER need not be restricted to the analysis of TF ChIP-seq data. CISMAPPER can take as input *any* set of loci (expressed as a BED file) from the genome of interest, and will generate lists of the TSSs and genes that those regions may regulate. Previously we showed that the cross-tissue histone-expression correlation approach used by CISMAPPER can predict regulatory links between enhancers and TSSs (12), providing the first validation of this idea (19,45). As noted above, distance-based methods cannot reliably distinguish which TSS might be regulated by a given locus due to the possibility of chromatin looping. This ability to make TSS-specific predictions of regulation by arbitrary genomic loci is a novel feature of CISMAPPER.

A second novel feature of CISMAPPER is that it can utilize data for any type of histone mark in making its predictions, and the regulatory links it predicts will depend on the histone mark chosen (e.g. H3K27ac or H3K4me3). By contrast, distance-based methods do not make predictions that take into account the histone state of the predicted regulatory loci. In future work we will explore running CISMAPPER using a series of distinct histone marks in order to classify links according to their 'histone profiles'—the set of histone marks that identify the given link. This may allow us to group regulatory links into biologically relevant classes (e.g. activating, repressing, promoter-specific, enhancer-specific, etc.) in a way analogous to previous work that uses histone profiles to assign genomic loci to classes such as promoter, enhancer, insulator, etc. (46,47). In principle, this link-profiling approach might be used to classify links predicted by CISMAPPER from TF binding sites (ChIP-seq peaks), enhancers, disease-associated SNPs or chromatin conformation contact data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health [R01 GM103544 to T.L.B]. Funding for open access charge: NIH [R01 GM103544 to T.L.B].
Conflict of interest statement. None declared.

REFERENCES

1. Macintyre,G., Bailey,J., Haviv,I. and Kowalczyk,A. (2010) is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**, i524–i530.
2. Heruth,D.P., Hawkins,T., Logsdon,D.P., Gibson,M.I., Sokolovsky,I.V., Nsumu,N.N., Major,S.L., Fegley,B., Woods,G.M., Lewing,K.B. *et al.* (2010) Mutation in erythroid specific transcription factor KLF1 causes Hereditary Spherocytosis in the Nan hemolytic anemia mouse model. *Genomics*, **96**, 303–307.
3. Bailey,T., Krajewski,P., Ladunga,I., Lefebvre,C., Li,Q., Liu,T., Madrigal,P., Taslim,C. and Zhang,J. (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.
4. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
5. Bulger,M. and Groudine,M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, 327–339.
6. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.
7. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

8. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
9. Corradin,O., Saiakhova,A., Akhtar-Zaidi,B., Myeroff,L., Willis,J., Cowper-Sal lari,R., Lupien,M., Markowitz,S. and Scacheri,P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.
10. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E2191–E2199.
11. Roy,S., Siahpirani,A.F., Chasman,D., Knaack,S., Ay,F., Stewart,R., Wilson,M. and Sridharan,R. (2015) A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.*, **43**, 8694–8712.
12. O'Connor,T.R. and Bailey,T.L. (2014) Creating and validating cis-regulatory maps of tissue-specific gene expression regulation. *Nucleic Acids Res.*, **42**, 11000–11010.
13. ENCODE Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
14. Mouse ENCODE Consortium, Stamatoyannopoulos,J.A., Snyder,M., Hardison,R., Ren,B., Gingeras,T., Gilbert,D.M., Groudine,M., Bender,M., Kaul,R. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.*, **13**, 418.
15. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
16. 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.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
17. Zentner,G.E., Tesar,P.J. and Scacheri,P.C. (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.*, **21**, 1273–1283.
18. Voigt,P., Tee,W.-W. and Reinberg,D. (2013) A double take on bivalent promoters. *Genes Dev.*, **27**, 1318–1338.
19. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
20. Sikora-Wohlfeld,W., Ackermann,M., Christodoulou,E.G., Singaravelu,K. and Beyer,A. (2013) Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003342.
21. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
22. Wenger,A.M., Clarke,S.L., Notwell,J.H., Chung,T., Tuteja,G., Guturu,H., Schaar,B.T. and Bejerano,G. (2013) The enhancer landscape during early neocortical development reveals patterns of dense regulation and co-option. *PLoS Genet.*, **9**, e1003728.
23. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
24. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
26. Ernst,J. and Kellis,M. (2013) Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.*, **23**, 1142–1154.
27. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
28. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
29. Rada-Iglesias,A. and Wysocka,J. (2011) Epigenomics of human embryonic stem cells and induced pluripotent stem cells: insights into pluripotency and implications for disease. *Genome Med.*, **3**, 36.
30. Parnavelas,J.G. (2000) The origin and migration of cortical neurones: new vistas. *Trends Neurosci.*, **23**, 126–131.
31. Chen,J.-G., Rasin,M.-R., Kwan,K.Y. and Sestan,N. (2005) Zfp312 is required for subcortical axonal projections and dendritic morphology of deep-layer pyramidal neurons of the cerebral cortex. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 17792–17797.
32. Lodato,S., Rouaux,C., Quast,K.B., Jantrachotechatchawan,C., Studer,M., Hensch,T.K. and Arlotta,P. (2011) Excitatory projection neuron subtypes control the distribution of local inhibitory interneurons in the cerebral cortex. *Neuron*, **69**, 763–779.
33. Magee,J., Hoffman,D., Colbert,C. and Johnston,D. (1998) Electrical and calcium signaling in dendrites of hippocampal pyramidal neurons. *Annu. Rev. Physiol.*, **60**, 327–346.
34. Zhu,Y., Chen,Z., Zhang,K., Wang,M., Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, **7**, 10812.
35. FANTOM Consortium the RIKEN PMI (DGT), C., Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M. J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
36. Lara-Astiaso,D., Weiner,A., Lorenzo-Vivas,E., Zaretsky,I., Jaitin,D.A., David,E., Keren-Shaul,H., Mildner,A., Winter,D., Jung,S. *et al.* (2014) Immunogenetics. Chromatin state dynamics during blood formation. *Science*, **345**, 943–949.
37. modENCODE Consortium, Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.
38. Gerstein,M.B., Lu,Z.J., Nostrand,E.L.V., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.
39. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
40. Simonis,M., Klous,P., Splinter,E., Moshkin,Y., Willemsen,R., de Wit,E., van Steensel,B. and de Laat,W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.
41. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
42. 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.
43. 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.
44. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.
45. Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.
46. 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.
47. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.