Published online 18 June 2015

Nucleic Acids Research, 2015, Vol. 43, No. 19 ***e125***
doi: 10.1093/nar/gkv607

By the company they keep: interaction networks define the binding ability of transcription factors

Davide Cirillo[1,2], Teresa Botta-Orfila[1,2] and Gian Gaetano Tartaglia[1,2,3,*]

[1]Centre for Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain and [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluís Companys, 08010 Barcelona, Spain

Received January 15, 2015; Revised May 25, 2015; Accepted May 28, 2015

ABSTRACT

Access to genome-wide data provides the opportunity to address questions concerning the ability of transcription factors (TFs) to assemble in distinct macromolecular complexes. Here, we introduce the PAnDA (Protein And DNA Associations) approach to characterize DNA associations with human TFs using expression profiles, protein–protein interactions and recognition motifs. Our method predicts TF binding events with >0.80 accuracy revealing cell-specific regulatory patterns that can be exploited for future investigations. Even when the precise DNA-binding motifs of a specific TF are not available, the information derived from protein-protein networks is sufficient to perform high-confidence predictions (area under the ROC curve of 0.89). PAnDA is freely available at http://service.tartaglialab.com/new_submission/panda.

INTRODUCTION

Recent chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) experiments have provided genome-wide details of transcription factors binding sites (TFBS), revealing important information on transcription factors (TFs) activities in human cells (1). Although computational analyses are extremely useful to identify DNA motifs associated with individual TFs (2), the complexity of regulatory networks requires advanced methods to capture cell-specific TF interactions with DNA regions (3). TFs assemble in hetero-complexes (4), which indicates that protein–protein interactions (PPI) could play an important role in TFBS recognition. Indeed, recent reports showed that TFs are present at different concentrations in human cells (5) and form highly dynamic complexes (6), suggesting that their DNA-binding abilities are influenced by the way components of PPI networks assemble together (7). Here, we investigated how the information contained in components of PPI networks can be exploited to perform accurate predictions of TFBS. The opportunity to study the role of PPI in TF-DNA associations is offered by details of ChIP-seq protocols, which involve crosslinking of genomic regions with TFs and their protein partners (8). Our analysis of ChIP-seq data sheds light on combinatorial associations of TFs and reveals new properties of regulatory networks (9).

From analysis of experimental data, we formulated a novel approach, PAnDA (Protein And DNA Associations), to predict TF interactions with DNA regions. Our approach will be useful for the designing of new applications for biotechnological research, including somatic stem cell reprogramming to pluripotency (10) and genome engineering with transcription activator-like effectors (11).

MATERIALS AND METHODS

We developed the PAnDA algorithm to predict DNA targets of TFs using the information contained in PPI networks. For each TF (*ChIP-seq datasets*), we performed a literature search to retrieve PPI networks (*Interaction networks*) as well as regulatory motifs of DNA-binding components (*Regulatory motifs*). We then selected PPI elements active in specific cell-lines using expression levels (*Expression levels*) and combined their regulatory motifs using a machine learning approach that distinguishes between low- and high-affinity interactions with genomic regions (*Trend analysis and models selection*). The algorithm is freely available as a web server (*PAnDA web server*).

ChIP-seq datasets

We retrieved 607 datasets from ENCODE Transcription Factor Binding Sites by ChIP-seq tracks [June 2012 uniform processing (8)]. Datasets containing <2000 peaks (i.e. 60 datasets) or sequences smaller than 20 nt in length [i.e. comparable with the average length of a DNA motif (12)] were filtered out. In the training phase (*Trend analysis and models selection*), we used 361 datasets whose target TF have at least one motif annotated in JASPAR CORE (13), Jolma (14) and UniPROBE (15). For independent testing,

*To whom correspondence should be addressed. Tel: +34 933160116; Fax: +34 933969983; Email: gian.tartaglia@crg.es

we used 186 datasets without annotated motifs for target TFs. The filter for availability of expression data (*Expression levels*) slightly reduced the number of cases to 404 ChIP-seq datasets (134 target TFs): 275 datasets (86 target TFs, 12 cell lines) for model selection and 147 datasets (48 target TFs, 9 cell lines) for independent testing (i.e. target TFs not present in the training phase).

Interaction networks

Within each PPI network, we classify proteins as: (i) target TFs (layer 1); (ii) cofactors (layer 2); (iii) mediated cofactors (layer 3) [Figure 1a]. To retrieve protein interactions we used the STRING database (16) that contains physical and functional interactions (total of 5 214 234 associations) collected from different publicly available sources including BioGRID [822 889 interactions](17), MINT and IntACT [523 070 interactions](18,19), DIP [11 000 interactions](20) and PID [16 823 interactions](21). In our analysis, we discarded interactions with STRING scores < 0.8 because they are frequently linked to text-mining and co-expression studies that cannot be easily linked to physical evidence (16). Using confidence scores >0.8, we compared high peaks versus low peaks motif counts and found strong enrichments in second and third layers of PPI networks (total of 1759 unique pairwise associations; layer 2 enrichment: *P*-value = 1.23e-21; layer 3 enrichment: *P*-value = 2.69e-10; Wilcoxon's test). We note that using STRING scores > 0.9, the number of interacting partners is sensibly reduced (total of 1 235 pairwise associations) and the enrichments were not significant (layer 2 enrichment: *P*-value = 0.07; layer 3 enrichment: *P*-value = 0.955; Wilcoxon's test).

As a number of physical interactions are missing at STRING scores > 0.8, we proceeded with their reintegration using BioGRID [physical interactions: 'FRET', 'Two-hybrid', 'Co-localization', 'Co-purification'; total of 2047 unique pairwise associations]. The combination of BioGRID and STRING led to a substantial increase in the enrichments significance (Figure 1b; layer 2 enrichment: *P*-value = 7.43e-34; layer 3 enrichment: *P*-value = 3.43e-66; Wilcoxon's test). We also investigated cases linked to stronger evidence for physical interaction in BioGRID ('FRET' and 'Two-hybrid'; total of 1976 unique pairwise associations), but we did not observe sensible changes in discriminative performances (layer 2 enrichment: *P*-value = 3.16e-25; layer 3 enrichment: *P*-value = 7.58e-16; Wilcoxon's test). Indeed, fold-change enrichments are higher when STRING is combined with BIOGRID classes 'FRET', 'Two-hybrid', 'Co-localization', 'Co-purification' (layer 2 fold enrichment: 2.08; layer 3 fold enrichment: 1.55; median comparison between distributions) than 'FRET' and 'Two-hybrid' (layer 2 fold enrichment: 1.68; layer 3 fold enrichment: 1.12; median comparison between distributions).

In summary, we retrieved 1 093 841 (STRING) and 225 943 (BioGRID) unique protein–protein associations for our PPI networks. Upon selection of DNA-binding proteins with annotated sequence motifs (*Regulatory motifs*), we counted a total of 13 785 (STRING), 1 529 (BioGRID) and 20 800 (STRING + BioGRID) associations. Finally, the application of expression level thresholds (*Expression levels*) reduced the set to 3 581 (STRING), 372 (BioGRID) and 5 333 (STRING + BioGRID) interactions.

Regulatory motifs

DNA motifs of PPI networks were retrieved from JASPAR CORE [205 motifs of 212 TFs](13), Jolma [738 motifs of 432 TFs](14) and UniPROBE [318 motifs of 224 TFs; Supplementary Table S1](15). Notably, motifs contained in JASPAR and Jolma databases were determined through SELEX experiments, while UniPROBE reports motifs identified via universal protein binding microarray (PBM) technology. Calculation of motif occurrences was performed using FIMO software [default *P*-value threshold of 1e-4] (22).

We cross-validated our results using databases of motifs derived from ENCODE ChIP-seq experiments [Wang database (23): 82 motifs of 65 TFs; SeAMotE database (24): 95 motifs of 95 TFs]. Overall, we collected 1 438 distinct motifs assigned to 570 human TFs using NCBI HomoloGene database (http://www.ncbi.nlm.nih.gov/homologene). Motif frequencies were computed as fraction of DNA sequences (ChIP-seq peaks) containing ≥ 1 motifs recognized by one or more factors in the experiments from the same cell line.

To select the DNA-motif database that better describes the binding sites of a specific PPI network, we introduced the concept of *mappability*:

$$\text{mappability} = \frac{\text{number of mapped motifs}}{\text{number of available motifs}} \times \text{motifs size}$$

The *mappability* score is a measure of the number of motifs covered by cofactors and mediated cofactors present in each TF network (*number of mapped motifs*/*number of available motifs*) weighted by their average length (*motifs size*). For instance, in the case of EP300 the *mappability* scores are 8.24 [Wang (23)], 7.17 [UniPROBE (15)], 5.32 [JASPAR (13)], 4.27 [Jolma (14)] and 2.87 [SeAMotE (24)]. Importantly, the *mappability* score correlates with testing performances on each database (*PAnDA web server* and also 'Results and Discussion': *Stability of PAnDA models*). Cofactors play a predominant role in TFBS classification compared with mediated cofactors, which results in stronger correlation between their *mappability* and predictive performances.

Expression levels

We estimated protein abundances using expression levels from ENCODE experiments (25), that provide a consistent and homogeneous source of information for our comparative analysis. Although protein and mRNA levels can differ owing to regulatory mechanisms and post-translational modifications (26), recent studies indicate that they are significantly correlated in human cells (27,28). In agreement with these findings, we recently showed that predictions of protein interactions can be performed using mRNA levels instead of protein abundances (29), which is also in line with our previous observations (30,31).

In our algorithm, we employed 12 cell types matching both ChIP-seq and RNA-seq data (A549, AG04450, BJ, GM12878, H1-hESC, HUVEC, HeLa-S3, HepG2, K562, MCF-7, NHEK, SK-N-SH_RA) representing the main

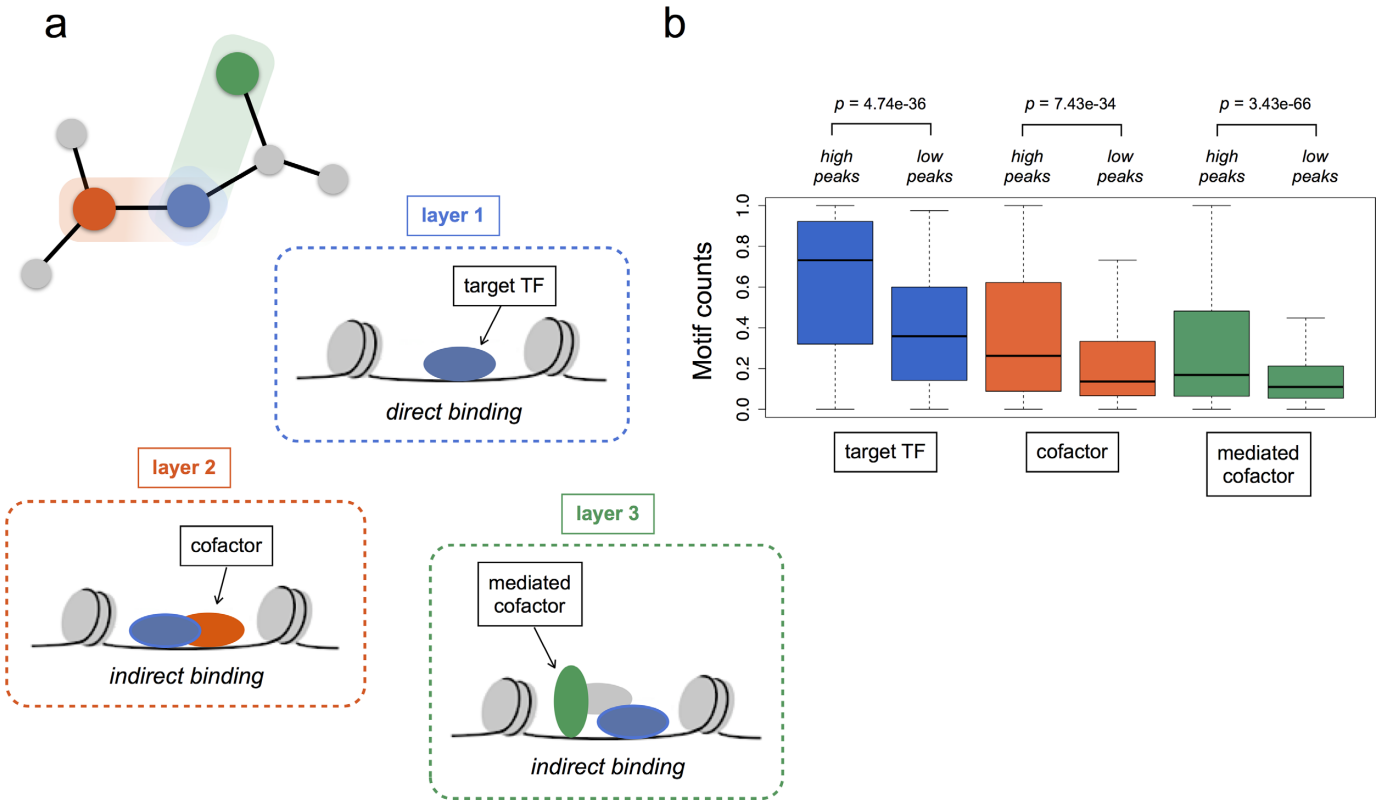

Figure 1. Trends in PPI networks. (**a**) Graphical representation of TF binding modes. In addition to direct binding (layer 1 in the interaction network—blue dots), we take into account the contribution of cofactors (layer 2—red dots) and mediated cofactors (layer 3—green dots). (**b**) Each network layer shows significant difference in frequencies of binding motifs associated with high and low ChIP-seq peaks (motifs were retrieved from a number of open-source databases; *Online methods: TFBSs databases*).

cell types in ENCODE ChIP-seq experiments. For each gene, we computed overall expression levels as sum of RPKM (Reads Per kb per Million reads) values averaged across replicates with np-IDR (non-parametric Irreproducible Discovery Rate) lower than 0.1.

We used quantile normalization to analyze expression levels from different cell lines because it allows comparison between statistical distributions without prior knowledge of their intrinsic features (e.g. 0.1 quantile = bottom 10% of all observations; 0.5 quantile = 50% of all observations = median). Thus, the approach introduces an absolute criterion to select components of PPI networks using optimal thresholds (from 0.1 to 0.5: threshold = 1 – quantile; *Trend analysis and models selection*). For each motif database, thresholds were derived in the training phase (Supplementary Table S1) and tested on the independent validation set.

Trend analysis and models selection

We employed state-of-the-art machine-learning approaches (32) to evaluate the contribution of PPI networks to DNA recognition: K-nearest neighbors (KNN), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM) and Random Forest (RF) [see the *Online Tutorial* for further details]. Each classifier provides a classification score for protein–DNA interactions (interactions are considered positive when the score is >0.5). Using expression levels information to select DNA-binding proteins (*Expression levels*), we built three distinct classifiers (Supplementary Figure S1) for the first (i.e. target TFs), second (i.e. target TFs and cofactors) and third layer (i.e. target TFs, cofactors and mediated cofactors) of PPI networks. An additional algorithm (model 4) was trained without information on target TFs (Supplementary Figure S1).

In each model, cell-specific expression levels (Supplementary Table S1) were used to select cofactors and mediated cofactors associated with individual target TFs. Motif frequencies for each set of interest (*Regulatory motifs*) were employed as input for the machine-learning approach (model 1: target TF motifs; model 2/4: target TF and cofactor motifs/cofactor and mediated cofactor motifs; model 3: target TF, cofactor and mediated cofactor motifs). We evaluated performances by measuring accuracies with leave-one-out cross-validation (LOOCV) on 275 datasets (86 target TFs, 12 cell lines; LOOCV has been performed on individual TF networks; Supplementary Figures S1-S4). For each network, positive and negative non-redundant instances were balanced by randomly under-sampling the more populated class (i.e. using the same number of positive and negative cases).

We estimated performances of PAnDA models using slopes (ΔPerformance) obtained from a linear regression on LOOCV accuracies ('layer 1 → layer 2' and 'layer 2 → layer 3'; Figure 2; Supplementary Figure S1). Expression thresholds (*Expression levels*; Supplementary Table S1) were se-

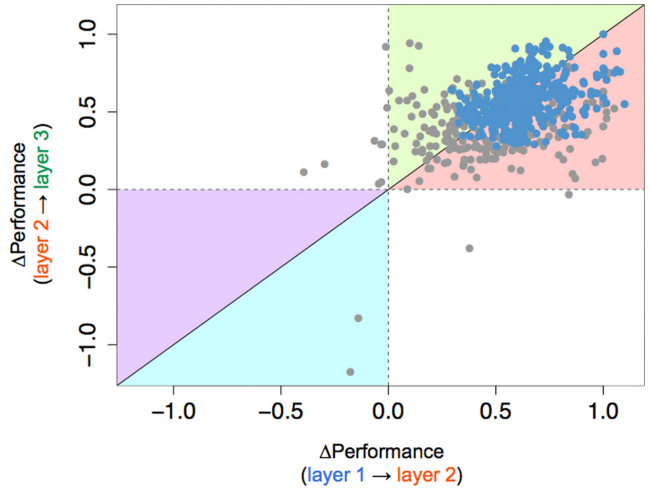

Figure 2. Training the PAnDA approach. Network layers and algorithm performances. In 275 datasets (8), we observed a consistent increase in cross-validation accuracies (ΔPerformance >0; *Online Methods: trend analysis and models selection*) upon layers integration (layer 1 → layer 2; layer 2 → layer 3; *Online Methods: TFBSs databases*; blue dots indicate ΔPerformance with *P*-value < 0.01). Colors highlight specific trends: light green and pink indicate that addition of layers 2 and 3 is associated with an increase in predictive power (light green: layer 3 has stronger signal than layer 2; pink: *vice versa*), while light blue and purple indicate decrease (light blue: layer 3 has lower signal than layer 2; purple: *vice versa*).

lected by evaluating classifiers agreement on the criterion ΔPerformances >0 (Figure 2; Supplementary Figures S2 and S3). Performances on independent test sets (*ChIP-seq dataset*) were evaluated using the area under the ROC curve (AUROC), which is independent of positive and negative datasets size.

PAnDA web server

One or more protein and DNA sequences in FASTA format are used as input. Query protein sequences are searched for homologous sequences in PAnDA TFs database using Pfam (33) DNA binding domain families (version 27.0) and filtering BLAST (34) hits (*E*-value ≤ 0.01; sequence identity ≥ 95%). Interaction networks of homologous TFs are generated as described in *Interaction networks* section. As for query DNA sequences, FIMO (22) software is used to search motifs occurrences in PWMs databases. Motif frequencies are calculated as described in the *Regulatory motifs* section. Expression thresholds are calculated as in the *Expression levels* section. If the 'Default mode' option is selected, all the motifs repositories in our database are employed along with their optimal expression thresholds and optimal classifiers identified in LOOCV (*Trend analysis and models selection*; Supplementary Table S1). If no motifs are found for layer 1, optimal model 4 (*Trend analysis and models selection*) is used. Binding propensities are listed using the ranking provided by the *mappability* score that correlates with predictive performances (*Regulatory motifs;* Supplementary Table S1 and Supplementary Figure S5). If the 'Expert mode' option if selected, the user can choose motif databases, expression thresholds and classifiers. By default we provide components of PPI networks derived from STRING (confidence score > 0.8) and BioGRID ('FRET', 'Two-hybrid', 'Co-localization', 'Co-purificational') and allow users to decide combinations depending on their specific analyses. We note that all combinations of BIOGRID and STRING sets showed similar AUROCs on the test set (model 4) in the range of 0.88–0.90, with exception of STRING confidence score > 0.9 that has AUROC of 0.84 due to the low number of elements in PPI networks (*Interaction networks*).

RESULTS AND DISCUSSION

In this study we analyzed TFBS derived from ENCODE ChIP-seq datasets (8) ('Materials and Methods' section: *ChIP-seq databases*) using PPI networks ('Materials and Methods' section *Interaction networks*), regulatory motifs ('Materials and Methods' section: *Regulatory motifs*) and expression levels of DNA-binding proteins ('Materials and Methods' section: *Expression levels*). For each TF (i.e. layer 1 in Figure 1a), we retrieved high-confidence functional partners (i.e. cofactors: layer 2 in Figure 1a) as well as their interactions (i.e. mediated cofactors: layer 3 in Figure 1a), thus covering up to two 'degrees of separation' (35) of functionally related layers of PPI networks.

Trends in PPI networks

Comparing high-affinity (top 500 ChIP-peaks) with low-affinity (bottom 500 ChIP-peaks) regions, we found that not only target TFs but also cofactors and mediated cofactors are significantly enriched in DNA-binding motifs (target TFs: *P*-value = 4.74e-36, fold-change = 2.08; cofactors: *P*-value = 7.43e-34, fold-change = 2.09; mediated cofactors: *P*-value = 3.43e-66, fold-change = 1.55; Wilcoxon's test and median comparison between distributions; Figure 1b). Such findings suggest that the PPI network contains relevant information that can be exploited to identify TFBS.

Training the PAnDA approach

The ability of PPI networks to identify TFBS was explored using state-of-the-art algorithms ('Materials and Methods' section: *Trend analysis and models selection*). In our calculations, we selected TF interactions considering cell-specific expression levels ('Materials and Methods' section: *Expression levels*; Supplementary Tables S1) and measured performances of classifiers trained on DNA-binding motif counts of PPI networks (Figure 2; 'Materials and Methods' section: *Regulatory motifs*; Supplementary Figures S1–S3). With respect to standard approaches restricted to the first layer of PPI networks (i.e. layer 1 or DNA motifs of target TF), we observed a consistent increase in performances when second and third layers were taken into account (Figure 2; Supplementary Figures S1–S3; 'Materials and Methods' section: *Trend analysis and models selection*). Using leave-one-out cross-validation (LOOCV), we identified an ensemble of models with high predictive power (average accuracy of 0.81 and area under the ROC curve, AUROC, of 0.80 on LOOCV per TF network; 'Materials and Methods' section: *Trend analysis and models selection*; Supplementary Figure S1a; Supplementary Table S1).

To illustrate PAnDA performances with an example, TFBS of T-cell acute lymphocytic leukemia protein 1

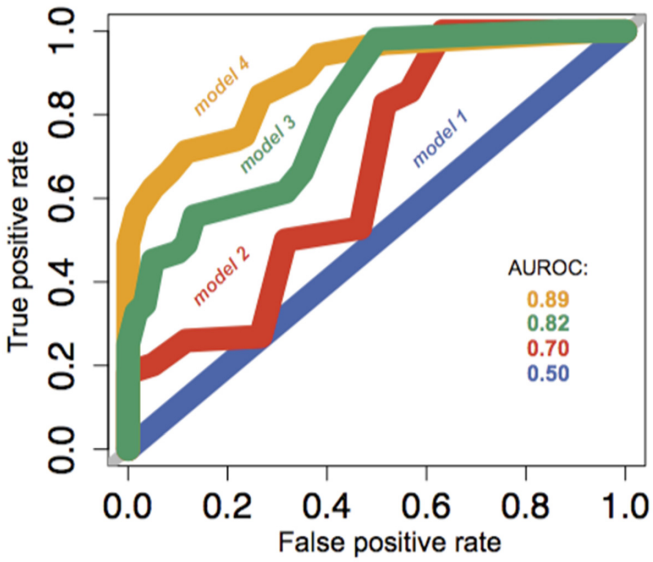

Figure 3. Testing the PAnDA approach. Performances on independent test sets [147 datasets (8); 'Material and Methods' section: *ChIP-seq datasets*]. Four models based on different network layers (model 1: layer 1; model 2: layers 1 and 2; model 3: layers 1,2 and 3; model 4: layers 2 and 3) have been applied to cases without annotated target TF motifs (Supplementary Figure S1c). Areas Under the ROC curve (AUROCs) show that interaction network information (layers 2 and 3) provides accurate description of binding events.

(TAL1) were predicted with accuracies of 0.50 (layer 1), 0.75 (layers 1 and 2) and 0.83 (layers 1, 2 and 3) in K562 cells. Similarly for proto-oncogene c-Fos (FOS), TFBS were identified with accuracies of 0.65 (layer 1), 0.71 (layers 1 and 2) and 0.77 (layers 1, 2 and 3) in 4 cell lines (GM12878, HUVEC, HeLa-S3 and K562). Interestingly, DNA-binding activity of both TAL1 and FOS have been recently investigated using innovative biotechnologies, including epigenome editing tools and single-cell high-throughput sequencing (36,37). We speculate that our predictions could be interfaced with novel experimental platforms to investigate and engineer TF interaction networks.

Testing the PAnDA approach

We tested PAnDA on 147 never-seen-before ChIP datasets of target TFs for which DNA-binding motifs are not reported in literature ('Materials and Methods' section: *Motif frequencies*). High predictive performances were observed (AUROC = 0.82; Figure 3; Supplementary Figure S1c; Supplementary Table S1; 'Materials and Methods' section: *ChIP-seq datasets*), indicating that PPI networks contain relevant information to identify TFBS even in absence of target TF motifs. For instance, in the case of reprogramming factor homeobox protein NANOG (38), regulatory signals are not available (layer 1) but DNA motifs exist for several cofactors and mediated cofactors. Comparing our calculations with ChIP-seq data in H1-hESC cells, we found strong agreement when TF network information is taken into account (AUROCs of 0.50, 0.60 and 0.98 using respectively one, two or three PPI layers; Figure 4). Similarly, DNA interactions of paired amphipathic helix protein Sin3a (SIN3A) were predicted with high confidence in A549, GM12878, H1-hESC, HepG2 cells (AUROCs of 0.50, 0.51 and 0.88 upon addition of PPI layers). In accordance with experimental evidence, we found that SIN3A, MYC-associated factor X (MAX), homeobox TGIF1 and mothers against decapentaplegic homolog SMAD3 have the ability to co-bind DNA sequences (39). Furthermore, we note that the association SMAD3-TGIF1 is particularly relevant for the activity of histone acetyltransferase EP300 (40), whose genomic interactions were also predicted with high confidence in seven cell lines (AUROCs of 0.52, 0.69 and 0.82 increasing the number of layers from 1 to 3). The finding that PPI networks provide accurate description of DNA-binding events is further corroborated by the high performances (AUROC of 0.89; 0.96 for NANOG; Figure 4) of a model trained without information on target TF motifs (i.e. only using cofactors and mediated cofactors motif frequencies; Figure 4; Supplementary Figure S1; 'Materials and Methods' section: *Trend analysis and models selection*). Importantly, recalibration of the algorithm with information contained in second and third layers allows better weighting of the contribution coming from cofactors and mediate-cofactors, which are intrinsically associated with lower signals (Figure 1b; see also *Stability of PAnDA models*).

Specificity of PAnDA models

To test the sequence-specificity of PPI networks for DNA regions, we built 10 models using random associations between proteins and their regulatory motifs. We observed a substantial decrease in both training [average accuracy = 0.56 with maximum of 0.60 on motif dataset by Wang *et*

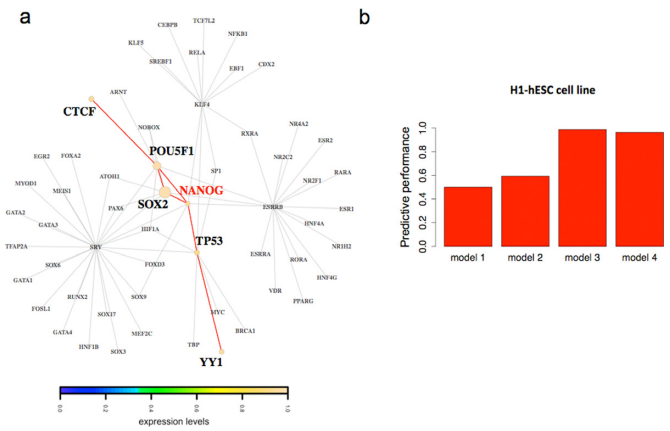

Figure 4. Example of PPI networks used in PAnDA calculations. (**a**) Components of PPI networks selected for predictions of NANOG interactions [H1-hESC cell line] (8); (**b**) Performances based on DNA-binding motifs of target TF (model 1: NANOG; AUROC = 0.50), target TF and cofactors (model 2: NANOG, TP53, SOX2 and POU5F1; AUROC = 0.60), target TF, cofactors and mediated cofactors (model 3: NANOG, TP53, SOX2, POU5F1, CTCF and YY1; AUROC = 0.98) and cofactors and mediated cofactors (model 4: TP53, SOX2, POU5F1, CTCF and YY1; AUROC = 0.96). The network is represented using squares for target TF (NANOG) and circles for other proteins (cofactors and mediated cofactors). The color palette refers to quantiles of expression levels (increasing from blue to yellow). Factors predicted to be not relevant for the binding of target TF are colored in gray.

al. that was trained on both low and high peaks of ChIP-seq data (23)] and testing sets (Figure 5a), which indicates that optimal performances require recognition of precise regulatory motifs. We also investigated the cell-specificity of PAnDA predictions by shuffling expression levels of DNA-binding proteins (10 random models). In this test, elements of PPI networks have been removed or added considering randomized expression levels with respect to thresholds identified in the training phase. Also in this case, we observed poor performances on both training (average accuracy = 0.52 with maximum of 0.59 on the motif dataset by Wang *et al*. 2012) and testing sets (Figure 5b), indicating that our absolute criterion based on quantile normalization is key to identify factors and mediated cofactors participating in specific cell lines ('Materials and Methods' section: *Expression levels*).

Stability of PAnDA models

In an additional analysis, we modified PPI networks and DNA sequences to evaluate their impact on PAnDA predictions. In this test, we used the UniPROBE dataset as it shows the highest performances in the testing phase (AUROC of 0.89). Elimination of one cofactor per TF network significantly reduced performances (AUROC of 0.67; *P*-value = 0.0006, Student's *t*-test; 'Materials and Methods' section: *Models stability*), while removal of mediated cofactors showed less dramatic effects (AUROC of 0.82 upon elimination of one mediated cofactor; *P*-value = 0.01; Figure 6a). Thus, our findings suggest that cofactors play a predominant role in TF binding (see also Supplementary Figure S5), in agreement with previous reports (41). Similarly, modifications of DNA sequences dramatically affected PAnDA predictive power (Figure 6b). We found that

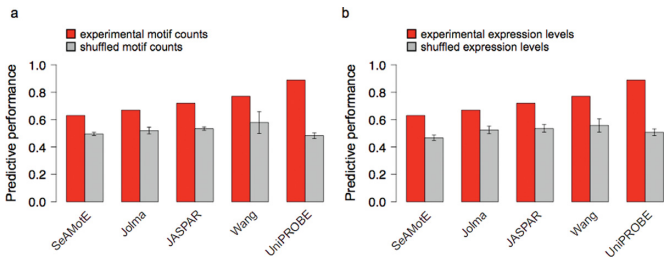

Figure 5. Specificity of PAnDA models. (**a**) Randomization of regulatory motifs. We built 10 independent models using shuffled associations between regulatory motifs and DNA-binding proteins present in the following databases: SeAMotE (24), Jolma (14), JASPAR CORE (13), Wang (23) and UniPROBE (15). Compared to PAnDA performances (red bars), the random models (gray bars) show negligible predictive power (AUROCs ~ 0.50) on the test set, indicating that regulatory motifs are specific for DNA targets. We note that the regulatory motifs generated with the SeAMotE approach (24) are of smaller size [6 nucleic acids on average] than those present in Jolma (14) [12 nucleic acids], JASPAR CORE (13) [12 nucleic acids], Wang (23) [16 nucleic acids] and UniPROBE (15) [16 nucleic acids], which results in poorer performances. (**b**) Randomization of expression levels. For each PPI network, selection of cofactors and mediated cofactors is based on cell-line abundances. Shuffling the expression levels of all DNA-binding proteins, we built 10 models (gray bars) with randomized PPI networks. On the test set, the models have poorer predictive power (AUROCs ~ 0.50) than PAnDA (red bars), which suggests that components of PPI network are highly specific for the cell line of interest. In both plots, AUROC averages and standard deviations are shown.

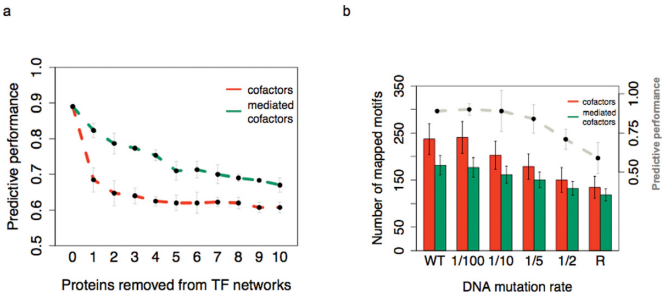

Figure 6. Stability of PAnDA models. (**a**) Interaction network destabilization. We found a significant decrease in predictive performance (AUROC; averages and standard deviations shown) upon removal of cofactors and mediated cofactors (model 4; *Online Methods: Models stability*). (**b**) Mutations of DNA sequences. From low (1/100 or 1 mutation in 100 nt) to high (R or 1 mutation each nucleotide) mutation rates, motifs mapped by cofactors and mediated cofactors are sensibly reduced (500 sequences per ChIP dataset; model 4; *Online Methods: Models stability*), which affects predictive performances (AUROC; averages and standard deviations shown).

performances were significantly reduced at high mutation rates (AUROC of 0.59; *P*-value = 0.003, Student's *t*-test; *Online Methods: Models stability*), although a number of binding sites could be still identified due to degeneration of consensus motifs (Figure 6b).

Use of the PAnDA approach

Given a pool of TFs and DNA sequences, PAnDA retrieves components of PPI networks from publicly available databases ('Materials and Methods' section: *Interaction networks*) selecting proteins that have expression levels compatible with the cell-line of interest ('Materials and Methods' section: *Expression levels*). Once the binding motifs ('Materials and Methods' section: *Regulatory motifs*) are mapped onto the DNA sequences, three independent classifiers are employed to predict the binding propensities of TF, their cofactors and mediated-cofactors (Figure 7). As binding motifs are collected from various sources, the *mappability* score is used to select the set with highest information content ('Materials and Methods' section: *Regulatory motifs*). In the web server implementation, the algorithm stops if the signal in the submitted datasets is comparable with random submissions (*Specificity of PAnDA models*). Otherwise, PAnDA reports DNA sequences and the interacting PPI networks for further analysis.

CONCLUSIONS

PAnDA is a powerful method to explore protein–DNA networks and can be used to design experiments targeting genomic regions such as promoters, enhancers as well as other functional elements. The key ingredients of our approach rely on intrinsic aspects of experimental measurements. Indeed, due to the formaldehyde fixation step during immunoprecipitation (8), binding regions reported in ENCODE ChIP-seq data involve multiple protein interactions. Starting from this observation, we found that PPI network information substantially improves the ability to classify TFBS (Figure 3; *Online Methods: Interaction networks*). Our results are in agreement with previous statistical analyses re-

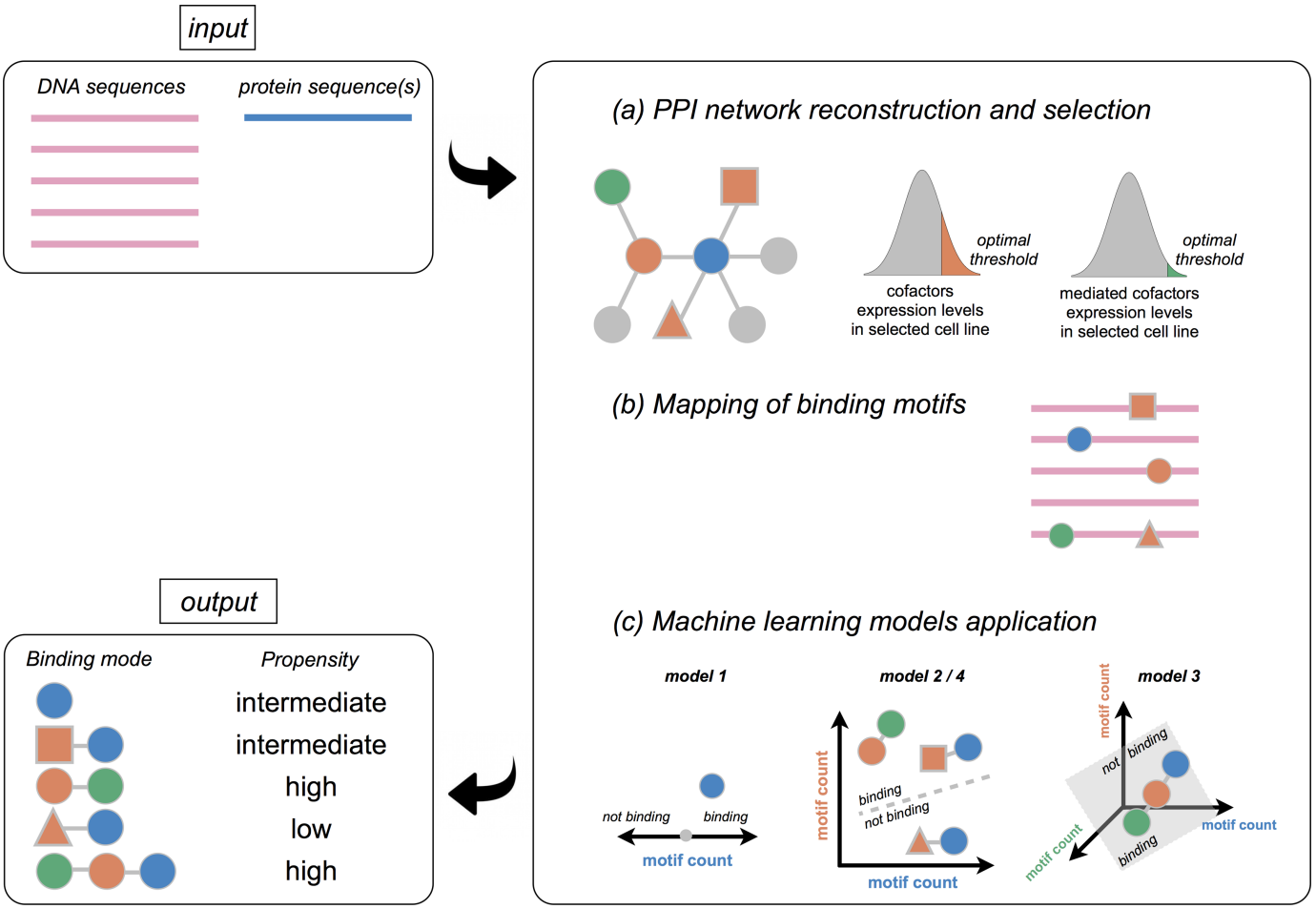

Figure 7. Using the PAnDA approach. Once DNA and TF sequences are submitted to the PAnDA web server, (a) PPI networks are selected from publicly available databases using expression levels to retrieve components of PPI networks that are active in specific cell-lines; (b) Regulatory motifs of DNA-binding proteins are mapped onto DNA sequences and reported in a table; (c) Three algorithms predict protein-DNA interactions exploiting first (TF), second (TF and cofactors) and third (TF, cofactors and mediated cofactors) layers of PPI networks. If DNA motifs of input TFs are missing, an alternative model (model 4) based on motifs of cofactors and mediated cofactors is employed. Each protein association is scored with a value for the propensity of the interaction to occur (see also *Online Tutorial*).

porting that regulatory motifs of cofactors are significantly enriched in proximity of transcription factors binding sites (42). Moreover, our algorithm very well complements recent catalogues of TF interactions (43), providing a tool to predict combinatorial associations in large-scale studies. It should be mentioned that our approach is an attempt toward the development of a multi-body potential for molecular interactions, which could overcome limitations of binary predictors (44). Implementation of new algorithms based on combinatorial features will impact performances of existing methods such as for instance *cat*RAPID for protein–RNA interactions (45).

In conclusion, while binding site identification based on nucleic acid motifs of individual proteins provides low-accuracy predictions, integrative approaches such as the one presented here will facilitate the discovery of complex functionalities based on combinatorial associations of proteins, leading to a better understanding of phenomena that govern genome evolution and stability (46). We envisage that PAnDA will be extremely useful to investigate and manipulate regulatory networks in future engineering studies (47).

Further details about the assessment of PAnDA models are available in the Supplementary Data. PAnDA web-service is freely available at http://service.tartaglialab.com/new_submission/panda.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank R. Guigó, P. Cosma, B. Lehner, F. Agostini and A. Breschi for stimulating discussions.

Author contributions: G.G.T. conceived this study. D.C. and G.G.T. designed the *in silico* experiments. D.C. performed the computational analysis. D.C. and G.G.T. analyzed the data. D.C., T.B.O. and G.G.T. wrote the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement RIBOMYLOME_309545; the Fundació La Marató de TV3 (20142731); and the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208). Funding for Open Access charge: European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement RIBOMYLOME_309545 and Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208). This work reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein.
Conflict of interest statement. None declared.

REFERENCES

1. Villar,D., Flicek,P. and Odom,D.T. (2014) Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat. Rev. Genet.*, **15**, 221–233.
2. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotech.*, **31**, 126–134.
3. Weingarten-Gabbay,S. and Segal,E. (2014) The grammar of transcriptional regulation. *Hum. Genet.*, **133**, 701–711.
4. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.-K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
5. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
6. Voss,T.C. and Hager,G.L. (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.*, **15**, 69–81.
7. Balaji,S., Babu,M.M., Iyer,L.M., Luscombe,N.M. and Aravind,L. (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.
8. ENCODE Project Consortium, Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
9. Luscombe,N.M., Madan Babu,M., Yu,H., Snyder,M., Teichmann,S.A. and Gerstein,M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
10. Dunn,S.-J., Martello,G., Yordanov,B., Emmott,S. and Smith,A.G. (2014) Defining an essential transcription factor program for naïve pluripotency. *Science*, **344**, 1156–1160.
11. Perez-Pinera,P., Ousterout,D.G., Brunger,J.M., Farin,A.M., Glass,K.A., Guilak,F., Crawford,G.E., Hartemink,A.J. and Gersbach,C.A. (2013) Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat. Meth.*, **10**, 239–242.
12. Sela,I. and Lukatsky,D.B. (2011) DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.*, **101**, 160–166.
13. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C., 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.
14. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
15. Robasky,K. and Bulyk,M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **39**, D124–D128.
16. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
17. Stark,C., Breitkreutz,B.-J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
18. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
19. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
20. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
21. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
22. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
23. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
24. Agostini,F., Cirillo,D., Ponti,R.D. and Tartaglia,G.G. (2014) SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC Genomics*, **15**, 925–933.
25. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
26. Vogel,C. and Marcotte,E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.*, **13**, 227–232.
27. Schwanhäusser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
28. Wilhelm,M., Schlegl,J., Hahne,H., Gholami,A.M., Lieberenz,M., Savitski,M.M., Ziegler,E., Butzmann,L., Gessulat,S., Marx,H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.
29. Cirillo,D., Marchese,D., Agostini,F., Livi,C.M., Botta-Orfila,T. and Tartaglia,G.G. (2014) Constitutive patterns of gene expression regulated by RNA-binding proteins. *Genome Biol.*, **15**, R13, 1–12.
30. Tartaglia,G.G., Dobson,C.M., Hartl,F.U. and Vendruscolo,M. (2010) Physicochemical determinants of chaperone requirements. *J. Mol. Biol.*, **400**, 579–588.
31. Tartaglia,G.G. and Vendruscolo,M. (2009) Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol. Biosyst.*, **5**, 1873–1876.
32. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
33. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
34. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
35. Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
36. Mendenhall,E.M., Williamson,K.E., Reyon,D., Zou,J.Y., Ram,O., Joung,J.K. and Bernstein,B.E. (2013) Locus-specific editing of histone modifications at endogenous enhancers. *Nat. Biotech.*, **31**, 1133–1136.

37. Pollen,A.A., Nowakowski,T.J., Shuga,J., Wang,X., Leyrat,A.A., Lui,J.H., Li,N., Szpankowski,L., Fowler,B., Chen,P. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotech.*, **32**, 1053–1058.
38. Mitsui,K., Tokuzawa,Y., Itoh,H., Segawa,K., Murakami,M., Takahashi,K., Maruyama,M., Maeda,M. and Yamanaka,S. (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, **113**, 631–642.
39. Liberati,N.T., Moniwa,M., Borton,A.J., Davie,J.R. and Wang,X.F. (2001) An essential role for Mad homology domain 1 in the association of Smad3 with histone deacetylase activity*. *J. Biol. Chem.*, **276**, 22595–22603.
40. Wotton,D., Lo,R.S., Swaby,L.A. and Massagué,J. (1999) Multiple modes of repression by the Smad transcriptional corepressor TGIF. *J. Biol. Chem.*, **274**, 37105–37110.
41. Haynes,B.C., Maier,E.J., Kramer,M.H., Wang,P.I., Brown,H. and Brent,M.R. (2013) Mapping functional transcription factor networks from gene expression data. *Genome Res.*, **23**, 1319–1328.
42. Liu,F. and Miranda-Saavedra,D. (2014) rTRM-web: a web tool for predicting transcriptional regulatory modules for ChIP-seq-ed transcription factors. *Gene*, **546**, 417–420.
43. Griffon,A., Barbier,Q., Dalino,J., van Helden,J., Spicuglia,S. and Ballester,B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.
44. Pahikkala,T., Airola,A., Pietilä,S., Shakyawar,S., Szwajda,A., Tang,J. and Aittokallio,T. (2015) Toward more realistic drug–target interaction predictions. *Brief. Bioinform.*, **16**, 325–337.
45. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
46. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
47. Gaj,T., Gersbach,C.A. and Barbas,C.F. (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.*, **31**, 397–405.