PLOS ONE

RESEARCH ARTICLE

iFORM: Incorporating Find Occurrence of Regulatory Motifs

Chao Ren☯, Hebing Chen☯, Bite Yang, Feng Liu¤, Zhangyi Ouyang, Xiaochen Bo*, Wenjie Shu*

Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing, China

☯ These authors contributed equally to this work.
¤ Current address: Department of Information, the 188th Hospital of Chaozhou, Chaozhou, China
* shuwj@bmi.ac.cn (WS); boxc@bmi.ac.cn (XB)

Citation: Ren C, Chen H, Yang B, Liu F, Ouyang Z, Bo X, et al. (2016) iFORM: Incorporating Find Occurrence of Regulatory Motifs. PLoS ONE 11 (12): e0168607. doi:10.1371/journal.pone.0168607

Editor: Denis Dupuy, INSERM U869, FRANCE

Received: August 3, 2016

Accepted: December 2, 2016

Published: December 19, 2016

Data Availability Statement: Both the source and binary codes for iFORM can be freely accessed at https://github.com/wenjiegroup/iFORM. The identified TF binding sites across human cell and tissue types using iFORM have been deposited in the Gene Expression Omnibus under the accession ID GSE53962.

Funding: This work was supported by grants from the Major Research plan of the National Natural Science Foundation of China (No. U1435222), the Program of International S&T Cooperation (No. 2014DFB30020) and the National High Technology Research and Development Program of China (No. 2015AA020108). The funders had no role in study

Abstract

Accurately identifying the binding sites of transcription factors (TFs) is crucial to understanding the mechanisms of transcriptional regulation and human disease. We present incorporating Find Occurrence of Regulatory Motifs (iFORM), an easy-to-use and efficient tool for scanning DNA sequences with TF motifs described as position weight matrices (PWMs). Both performance assessment with a receiver operating characteristic (ROC) curve and a correlation-based approach demonstrated that iFORM achieves higher accuracy and sensitivity by integrating five classical motif discovery programs using Fisher's combined probability test. We have used iFORM to provide accurate results on a variety of data in the ENCODE Project and the NIH Roadmap Epigenomics Project, and the tool has demonstrated its utility in further elucidating individual roles of functional elements. Both the source and binary codes for iFORM can be freely accessed at https://github.com/wenjiegroup/iFORM. The identified TF binding sites across human cell and tissue types using iFORM have been deposited in the Gene Expression Omnibus under the accession ID GSE53962.

Introduction

Gene regulation is co-ordinately regulated by interactions of many transcription factors (TFs), many of which bind promoter and enhancer DNA preferentially at characteristic sequence 'motifs'. Motifs are short patterns described as position weight matrices (PWMs) that tend to be conserved by purifying selection. Identifying and understanding these TF motifs can provide critical insight into the mechanisms of transcriptional regulation and human disease. However, accurate identification of these binding site motifs is still challenging because a single TF will often recognize a variety of similar sequences.

Over the past several decades, many classical algorithms have been developed to discover DNA regulatory motifs. FIMO [1], Consensus [2, 3], and STORM [4] function solely as motif scanners, whereas RSAT [5] and HOMER [6] provide multiple functions for analysing regulatory sequences in general. S1 Table summarizes the features of iFORM and these five algorithms as motif scanners. Each of these methods has its own merits for identifying potential TF binding; however, it is still a major challenge to integrate superiorities and to preclude

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

inferiorities of these complementary methods. Several studies have demonstrated that higher accuracy and sensitivity can be achieved by incorporating multiple motif discovery programs [7–9]; however, this approach will result in considerable computational overhead for scoring the large number of discovered motif instances obtained with different programs.

We describe incorporating Find Occurrence of Regulatory Motifs (iFORM), a software tool for scanning DNA sequences with TF motifs by integrating five classical motif discovery methods. We used Fisher's combined probability test to convert the resulting *p*-values obtained from the five methods to a χ^2 statistic, which follows a chi-squared distribution. We then applied false discovery rate analysis to estimate a *q*-value for each motif instance. By systematically assessing the accuracy of prediction performance, iFORM achieves higher accuracy and sensitivity relative to the five classical algorithms. Although iFORM integrates five methods, it is efficient, allowing for scanning sequences at a rate of 3.5 Mb/s on a single CPU. Additionally, we have illustrated the use of iFORM by producing high-quality genome-wide maps of transcription factor binding sites (TFBSs) for 542 TFs within DNaseI hypersensitive sites (DHSs) of 133 human cell and tissue types that were generated by the ENCODE Project [10] and the NIH Roadmap Epigenomics Mapping Consortium [11] in our recent studies. We also demonstrated the utility of iFORM to further reveal individual roles of functional elements in gene regulation and diseases based on these maps.

Materials and Methods

Data sets

DNaseI Hypersensitivity by Digital DNaseI data were obtained from both the ENCODE Project [10] and the NIH Roadmap Epigenomics Mapping Consortium [11]. TFs identified by ChIP-seq data were obtained from the ENCODE Project [10]. The uniform processing pipeline of the ENCODE Integrative Analysis Consortium was used to generate uniform peaks from both the DNase-seq and ChIP-seq data. The use of these data adheres strictly to the ENCODE and Roadmap Epigenomics Consortium Data Release Policy. PhastCons were extracted from the hg19 conservation track of the UCSC Genome Browser [12].

iFORM

iFORM integrates the TF binding sites identified by five classic methods: FIMO [1], Consensus [2, 3], STORM [4], RSAT [5] and HOMER [6]. For each discovered TFBS, a combined *p*-value obtained from these five methods was calculated using Fisher's combined probability test. First, a test statistic was calculated using the following formula:

$$\chi^2 = -2\sum_{i=1}^{5} log_e\,(p_i) \sim \chi^2(10)$$

where p_i is the *p*-value calculated from the five methods for each TFBS and χ^2 follows a chi-squared distribution $\chi^2(10)$ when p_i is independent. Thus, a combined *p*-value can be assigned to this test statistic. Then, we used a bootstrap method [13] to estimate false-discovery rates (FDRs) and report for each *p*-value a corresponding *q*-value. Before applying Fisher's combined probability test to our iFORM, we collected different sets of p_i values derived from the five methods by scanning diverse TF motifs in diverse cells/tissues in the human genome to test the independence of p_i values derived from the five methods. A Hilbert-Schmidt independence criterion (HSIC) method was used to test the independence jointly [14]. The threshold and HSIC score were calculated using *d*-variable HSIC (dHSIC) R-packages [15], with alpha level 0.05. The independent hypothesis was rejected if and only if threshold was equal or less

than HSIC score. Our result demonstrated that the thresholds were strictly greater than HSIC scores in all the collected p_i value sets, suggesting the joint independence of the p_i values derived from the five methods (S2 Table). Thus, Fisher's combined probability test can be applied to our iFORM. iFORM can be freely accessed at https://github.com/wenjiegroup/iFORM.

Generation of "gold-standard" data

To validate the predicted TFBSs using the ROC (receiver operating characteristic) approach, we applied a method similar to that presented in a previous study [16] to generate "gold-standard" data for a few TFs. Briefly, we scanned the genomic sequences under the uniform DHSs in the hg19 genome using the five methods incorporated by iFORM with default parameters for each TF motif in the "gold-standard" set separately. Then, for each TF, all motif instances (p-value $< 10^{-9}$) located within the corresponding TF ChIP-seq peaks were considered the set of TFBS positives. The set of TFBS negatives was defined as all motif instances (p-value $< 10^{-9}$) that did not overlap with a ChIP-seq peak and had a greater fraction of total mapped reads from the "Control" compared to the ChIP-seq treatment. In total, there were 103 sets of "gold-standard" data for 12 TFs from 51 cells/tissues (S3 Table).

Assessing the performance of iFORM

The prediction performance of iFORM was assessed using three methods, including ROC curve analysis, precision-recall curve analysis and a correlation-based approach. First, we used ROC curves and the area under the curve (AUC) to assess the accuracy of the prediction performance of iFORM. The ROC approach assumes that there are "gold-standard" data, including TFBS positives and TFBS negatives, that can be used for comparison to confidently evaluate our predictions of TF binding for a subset of the candidate motif sites. Furthermore, we also adopted precision-recall curves to assess the imbalance of the "gold-standard" data sets.

In addition to the ROC curve analysis, which classifies the motif instances as either TFBS positive or TFBS negative, we adopted a correlation approach that takes into account all locations except those in or near repetitive regions. The correlation approach assumes that the larger the correlation between the predicted value and the ChIP-seq signal and the smaller the correlation with the "control" background noise, the better the prediction accuracy. We extracted the ChIP-seq reads and the "control" reads around each motif site, and we used the Pearson correlation to measure the association between the total number of reads and the scores reported by iFORM and the five classical methods.

Results and Discussion

Implementation of iFORM

iFORM integrated the motif instances identified by the five classical algorithms, FIMO [1], Consensus [2, 3], STORM [4], RSAT [5] and HOMER [6], based on Fisher's method. We chose these five methods due to their great popularity in applications and the great generality of their scoring methods as motif scanners. iFORM was built using C, and an overview of the workflow is illustrated in Fig 1. To improve efficiency, we extracted the core source code of the motif discovery functions of these five algorithms and integrated them into the framework of iFORM instead of simply combining the resulting p-values obtained from these algorithms.

iFORM accepts DNA sequences in FASTA format or genomic coordinates in BED or GFF formats, and it takes as an input one or more TF motifs (represented as PWMs) that can be

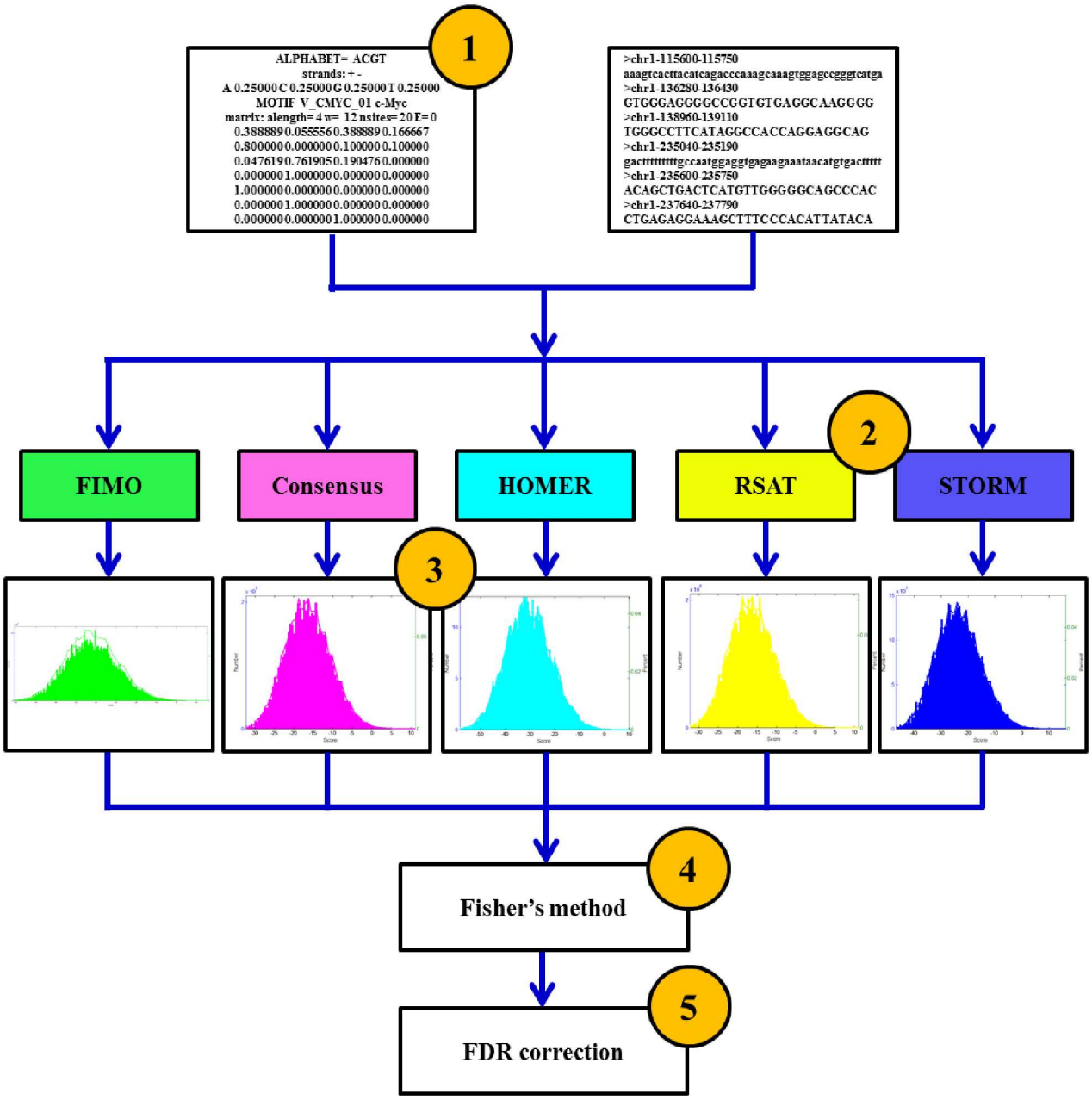

Fig 1. The workflow of iFORM. Overview of the iFORM workflow. (1) Assemble input data. The results may be improved by restricting the input to high-confidence sequences. Some programs achieve improved performance using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains. (2) Choose several motif discovery programs for the analysis. (3) Test the statistical significance of the resulting motifs. Use control calculations to estimate the empirical distribution of scores produced by each program on random data. (4) For each identified TFBS, a combined *p*-value obtained from these five methods was calculated using Fisher's combined probability test. (5) FDR correction of the combined *p*-values obtained from Fisher's method.

doi:10.1371/journal.pone.0168607.g001

collected from an existing motif database, generated the MEME algorithm, or even be defined by the user. For each motif instance, iFORM computes a χ^2 statistic that was obtained by combining *p*-values resulting from the five algorithms using Fisher's method, and it converts these statistics to *p*-values using the chi-squared distribution. Finally, iFORM uses a bootstrap

method [13] to estimate FDRs and reports for each p-value a corresponding q-value, which is defined as the minimal FDR threshold at which the p-value is deemed significant [17].

iFORM outputs a ranked list of motif instances, each with an associated χ^2 score, p-value and q-value. The list is illustrated as an HTML report, as an XML file in CisML format [18], as a plain text file and as tab-delimited files in gff formats suitable for input in the UCSC Genome Browser [12].

Performance assessment of iFORM using "gold-standard" data

Because iFORM integrates five classical algorithms based on Fisher's method, it is expected that iFORM can achieve higher accuracy and sensitivity compared with the five methods. To test this expectation, we first examined the CTCF motif instances identified by these six methods separately. We found that many of the CTCF binding sites, which are well annotated by DNase-seq, DNaseI digital genomic footprint (DGF), and corresponding TF ChIP-Seq data in H1 cells, can only be discovered by iFORM, not the other five classical methods (Figure A in S1 File). To systematically assess the accuracy of prediction performance for each motif instance, we used ROC curves and the corresponding AUC and precision-recall curves on the "gold-standard" data of six TF ChIP-seq data in GM12878 cells provided in a previous study [16] (Fig 2A and Figure C and D in S1 File). Our results suggest that our iFORM method showed higher AUC values than those of the other five classical methods.

To further validate the predicted TF binding sites, we used the method presented by Roger Pique-Regi et al. [16] to generate "gold-standard" data for many TF ChIP-seq sequences that were produced from different laboratories and from different cells/tissues (S3 Table). Using our newly generated "gold-standard" data of CTCF in GM12878 cells, we obtained ROC curves and precision-recall curves similar to those presented previously (Fig 2B and Figure C in S1 File). In addition, similar ROC curves and precision-recall curves were also obtained for the newly generated "gold-standard" data of CTCF in GM12878 produced from different labs (Fig 2C) and for the newly generated "gold-standard" data of CTCF in other cells (Figure C and D in S1 File). Furthermore, we assessed the prediction performance using new "gold-standard" data from 103 sets of TF ChIP-seq data produced in the ENCODE project (S3 Table). Across these "gold-standard" datasets, our iFORM achieves significantly higher AUCs than the other five algorithms (Fig 2D and S3 Table, p-value $< 10^{-10}$, Kolmogorov–Smirnov test). These results suggest that iFORM manifested superior performance consistency in different laboratories and different cells/tissues across multiple TFs, and it demonstrated higher accuracy and sensitivity compared with the five classical methods.

Furthermore, we examined whether the performance of iFORM was superior to that of integrating only two, three, or four methods based on "gold-standard" data in GM12878 cells provided in a previous study [13] (Fig 3A) and the "gold-standard" data generated in this study (Fig 3B). Again, our iFORM achieved significantly higher AUCs compared with integrating only two, three, or four methods across these "gold-standard" datasets (Fig 3C and S3 Table, p-value $< 10^{-9}$, two-sample Kolmogorov–Smirnov test). Taken together, our results convincingly demonstrate that iFORM presents superior performance compared with the five classical methods as well as integrating only two, three, or four methods.

Performance assessment of iFORM using a correlation approach

Thus far, we have validated iFORM with ROC analysis, which requires a binary classification of TFBSs as positive/negative using ChIP-seq as a "gold standard". However, it is becoming increasingly evident that TF binding is not perfectly dichotomous (bound or unbound); in many cases, quantitative partial binding occurs, which corresponds to the fraction of cells that

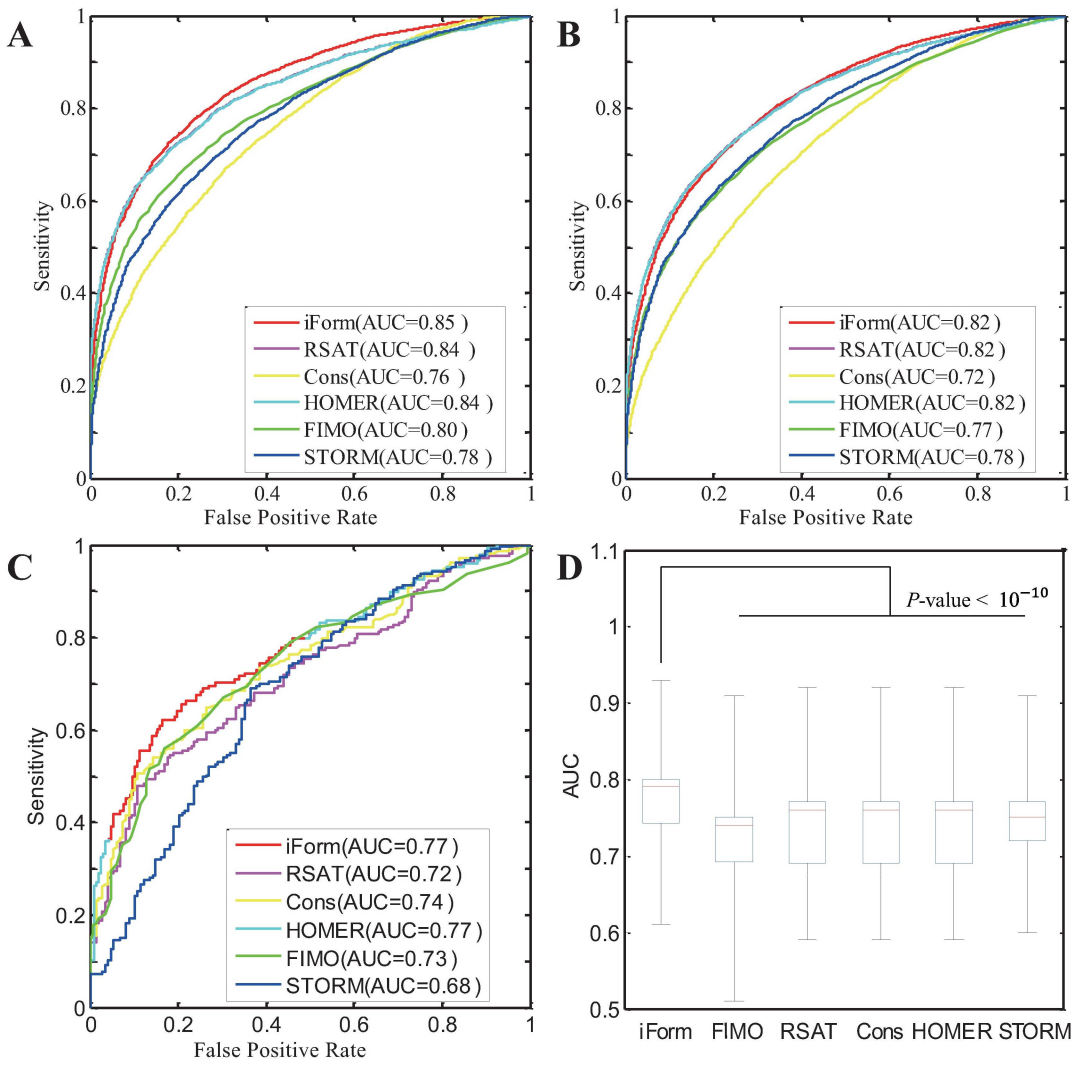

Fig 2. Performance assessment of iFORM using ROC curves. (A–C) ROC curves of multiple algorithms using "gold-standard" data from (A) CTCF in GM12878 cells provided in a previous study; (B) CTCF in GM12878 cells that was generated in our study. (C) REST in H1 cells. (D) Boxplot distributions of AUC for diverse TFs in multiple cells/tissues.

doi:10.1371/journal.pone.0168607.g002

have TF binding at a particular site at any given time [19, 20]. Indeed, our iFORM method is formulated as a combined χ^2 statistic, which could quantitatively reflect the level of TF occupancy to some extent. Thus, we presumed that correlations of the statistical scores with the ChIP-seq signal and with the "control" background noise could be used to validate iFORM. The larger the former value and the smaller the latter value, the better the prediction accuracy of iFORM. Applying this procedure to the other five classical methods, we found that iFORM presented substantially superior performance relative to the other methods (Fig 4A and 4B). Across the multiple TFs that we tested, iFORM showed consistently superior accuracy compared with the other methods (Fig 4C and 4D and S5 Fig, p-value < 0.005, two-sample Kolmogorov–Smirnov test). Taken together, these findings suggest that iFORM provides state-of-the-art performance relative to the five existing methods.

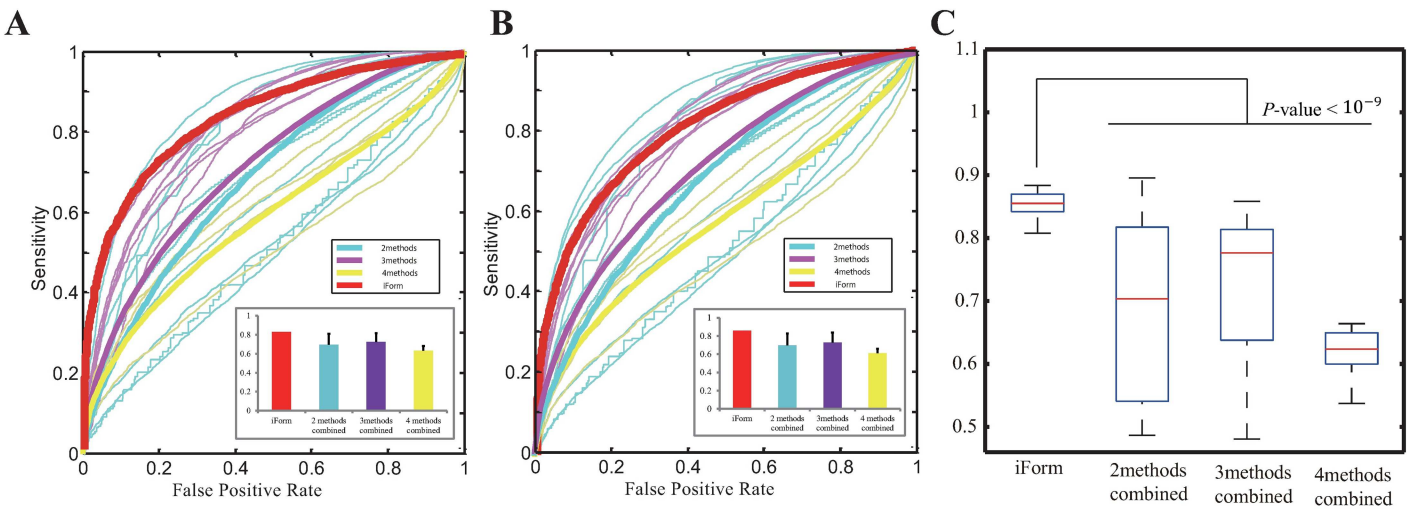

Fig 3. Performance comparisons between iFORM and integrating methods. (A-B) Performance comparisons between iFORM and integrating methods. ROC curves of iFORM and that of integrating only two, three, or four methods based on "gold-standard" data CTCF in GM12878 cells provided in a previous study (A) and the "gold-standard" data CTCF in GM12878 cells generated in this study (B). (C) Boxplot distributions of AUC of iFORM and of integrating only two, three, or four methods for diverse TFs in multiple cells/tissues.

doi:10.1371/journal.pone.0168607.g003

Application of iFORM for identifying TFBSs across multiple cells/tissues

We collected PWMs of the 542 TFs corresponding to 796 motif models from the TRANSFAC [21], JASPAR [22], and UniPROBE [23] databases as described in our previous study [24–26]. We applied iFORM to produce high-quality genome-wide maps of the TFBSs for the 542 TFs within DHSs of 133 human cell and tissue types that were generated by the ENCODE Project [10] and the NIH Roadmap Epigenomics Mapping Consortium [11] (p-value $< 10^{-18}$). On average, scanning DHSs of each cell/tissue took 30 min and 10 s of wall clock time on an Intel Xeon 2.2 GHz CPU, which is equivalent to scanning at a speed of 3.5 Mp/s.

Based on these data, we found that TFBSs were clustered together in the human genome, and we report the first comprehensive map of TFBS cluster regions across human cell and tissue types. An integrative analysis of these regions revealed a novel transcriptional regulation model of the accessible chromatin landscape [24]. Additionally, we investigated the HOT (high-occupancy target) regions, which were defined as TFBS cluster regions with extremely high TFBS complexity. We found that HOT regions play key roles in human cell development and differentiation [25]. Furthermore, we explored the prevalence of single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) in HOT regions, which demonstrated key roles of HOT regions in human disease and cancer [26]. These findings represent a critical step towards further understanding disease biology, diagnosis, and therapy.

Conclusions

The iFORM tool presented in this study was designed to incorporate five classical regulatory motif discovery methods using Fisher's method. iFORM is an easy-to-use and efficient motif discovery tool that achieves higher accuracy and sensitivity by integrating the results from multiple motif discovery programs. iFORM has provided accurate results using a variety of data from the ENCODE Project and the NIH Roadmap Epigenomics Project in our recent

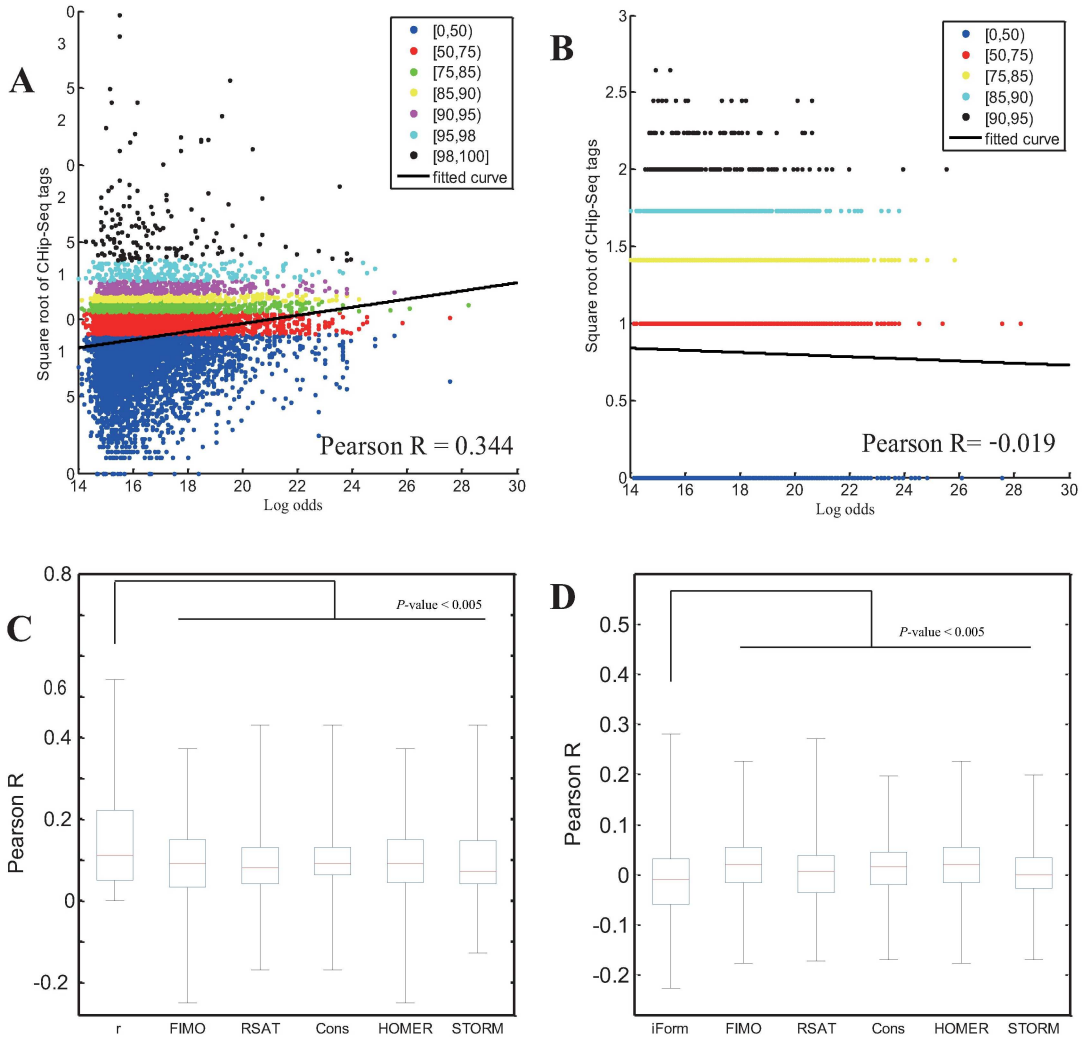

Fig 4. Performance assessment of iFORM using correlation methodology. (A-B) Scatter plot of the correlation between the log of odds from iFORM and the square-root transformed count of ChIP-seq (A) and "control" (B) reads in the 400 bp region surrounding each CTCF motif instance. (C–D) Performance comparison between iFORM and currently available methods. Boxplot distribution of Pearson correlations between the log of odds from iFORM and the square-root transformed count of ChIP-seq (A) and "control" (B) reads for diverse TFs in multiple cells/tissues.

doi:10.1371/journal.pone.0168607.g004

studies, and it has demonstrated its utility to further elucidate individual roles of functional elements in the mechanisms of transcriptional regulation and human disease.

Supporting Information

S1 File. This file contains all Supplementary figures (A-E).
(PDF)

S1 Table. Summaries of the five motif scanners.
(DOCX)

S2 Table. Independence test of p_i derived from the five methods.
(XLSX)

S3 Table. Performance assessment of iFORM.
(XLSX)

Acknowledgments

We wish to thank the ENCODE Project Consortium and the Roadmap Epigenomics Project Consortium for making their data publicly available.

Author Contributions

Conceptualization: WS.

Data curation: CR.

Formal analysis: BY FL ZO.

Funding acquisition: WS XB.

Investigation: CR HC.

Methodology: WS XB.

Project administration: WS.

Resources: WS XB.

Software: CR HC.

Supervision: WS.

Validation: CR.

Visualization: CR.

Writing – original draft: WS.

Writing – review & editing: WS.

References

1. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27(7):1017–8. Epub 2011/02/19. doi: 10.1093/bioinformatics/btr064 PMID: 21330290
2. Stormo GD, Hartzell GW 3rd. Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences of the United States of America. 1989; 86(4):1183–7. Epub 1989/02/01. PMID: 2919167
3. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics (Oxford, England). 1999; 15(7–8):563–77. Epub 1999/09/17.
4. Schones DE, Smith AD, Zhang MQ. Statistical significance of cis-regulatory modules. BMC bioinformatics. 2007; 8:19. Epub 2007/01/24. doi: 10.1186/1471-2105-8-19 PMID: 17241466
5. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, et al. RSAT 2011: regulatory sequence analysis tools. Nucleic acids research. 2011; 39(Web Server issue):W86–91. Epub 2011/08/17. doi: 10.1093/nar/gkr377 PMID: 21715389
6. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38(4):576–89. Epub 2010/06/02. doi: 10.1016/j.molcel.2010.05.004 PMID: 20513432
7. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431(7004):99–104. doi: 10.1038/nature02800 PMID: 15343339

8. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nature biotechnology. 2005; 23(1):137–44. doi: 10.1038/nbt1053 PMID: 15637633
9. MacIsaac KD, Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. PLoS computational biology. 2006; 2(4):e36. doi: 10.1371/journal.pcbi.0020036 PMID: 16683017
10. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. doi: 10.1038/nature11247 PMID: 22955616
11. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology. 2010; 28(10):1045–8. Epub 2010/10/15. doi: 10.1038/nbt1010-1045 PMID: 20944595
12. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. Nucleic acids research. 2016; 44(D1):D717–25. doi: 10.1093/nar/gkv1275 PMID: 26590259
13. Storey JD. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol. 2002; 64:479–98.
14. Gretton A, Fukumizu K, Choon HT, Song L, Sch\ lkopf B, Alex JS. A Kernel Statistical Test of Independence. In: Platt JC, Koller D, Singer Y, Roweis ST, editors. Advances in Neural Information Processing Systems 20 (NIPS 2007): Curran Associates, Inc.; 2007. p. 585–92.
15. Niklas Pfister PB, Bernhard Schölkopf, Jonas Peters. Kernel-based Tests for Joint Independence. ArXiv e-prints. 2016: 67.
16. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome research. 2011; 21(3):447–55. Epub 2010/11/26. doi: 10.1101/gr.112623.110 PMID: 21106904
17. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat. 2003; 31(6):2013–35.
18. Haverty PM, Weng Z. CisML: an XML-based format for sequence motif detection software. Bioinformatics (Oxford, England). 2004; 20(11):1815–7.
19. MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Bio. 2009; 10(7):R80. Epub 2009/07/25.
20. Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS biology. 2010; 8(3):e1000343. doi: 10.1371/journal.pbio.1000343 PMID: 20351773
21. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic acids research. 2006; 34(Database issue):D108–10. Epub 2005/12/31. doi: 10.1093/nar/gkj143 PMID: 16381825
22. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic acids research. 2010; 38(Database issue):D105–10. Epub 2009/11/13. doi: 10.1093/nar/gkp950 PMID: 19906716
23. Robasky K, Bulyk ML. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic acids research. 2011; 39(Database issue):D124–8. Epub 2010/11/03. doi: 10.1093/nar/gkq992 PMID: 21037262
24. Chen H, Li H, Liu F, Zheng X, Wang S, Bo X, et al. An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. Scientific reports. 2015; 5:8465. Epub 2015/02/17. doi: 10.1038/srep08465 PMID: 25682954
25. Li H, Liu F, Ren C, Bo X, Shu W. Genome-wide identification and characterisation of HOT regions in the human genome. BMC genomics. 2016; 17(1):733. doi: 10.1186/s12864-016-3077-4 PMID: 27633377
26. Li H, Chen H, Liu F, Ren C, Wang S, Bo X, et al. Functional annotation of HOT regions in the human genome: implications for human disease and cancer. Scientific reports. 2015; 5:11633. doi: 10.1038/srep11633 PMID: 26113264