Inference of transcriptional regulation in cancers

Peng Jiang*, Matthew L. Freedmanb,c,d, Jun S. Lius, and Xiaole Shirley Liu*a,1

*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA 02215; bDepartment of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215; cCenter for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02215; dProgram in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142; and eDepartment of Statistics, Harvard University, Cambridge, MA 02138

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved April 20, 2015 (received for review December 18, 2014)

Despite the rapid accumulation of tumor-profiling data and transcription factor (TF) ChIP-seq profiles, efforts integrating TF binding with the tumor-profiling data to understand how TFs regulate tumor gene expression are still limited. To systematically search for cancer-associated TFs, we comprehensively integrated 686 ENCODE ChIP-seq profiles representing 150 TFs with 7484 TCGA tumor data in 18 cancer types. For efficient and accurate inference on gene regulatory rules across a large number and variety of datasets, we developed an algorithm, RABIT (regression analysis with background integration). In each tumor sample, RABIT tests whether the TF target genes from ChIP-seq show strong differential regulation after controlling for background effect from copy number alteration and DNA methylation. When multiple ChIP-seq profiles are available for a TF, RABIT prioritizes the most relevant ChIP-seq profile in each tumor. In each cancer type, RABIT further tests whether the TF expression and somatic mutation variations are correlated with differential expression patterns of its target genes across tumors. Our predicted TF impact on tumor gene expression is highly consistent with the knowledge from cancer-related gene databases and reveals many previously unidentified aspects of transcriptional regulation in tumor progression. We also applied RABIT on RNA-binding protein motifs and found that some alternative splicing factors could affect tumor-specific gene expression by binding to target gene 3′UTR regions. Thus, RABIT (rabit.dfci.harvard.edu) is a general platform for predicting the oncogenic role of gene expression regulators.

Results
Landscape of Transcriptional Regulation in Cancer. To systematically search for TFs that drive tumor-specific gene expression patterns, we developed an integration framework, RABIT (rabit.dfci.harvard.edu). As a key distinction from other algorithms integrating gene expression and TF ChIP-seq data (15–18), RABIT better captures the properties of cancer cells, such as CNA and DNA methylation, that shape tumor gene expression independently from TF regulation. Additionally, somatic mutations of the TF-coding region can perturb transcriptional regulation. As another difficulty of transcriptional regulation analysis in cancer, most public ChIP-seq datasets were generated under experimental conditions distinct from those in cancers. To model and control the above confounding factors, RABIT uses three steps to identify TFs that drive tumor-
specific gene expression patterns at both the individual tumor level and the whole cancer-type level (Fig. 1). In step 1, RABIT screens for TFs that significantly affect the gene expression patterns in each tumor and select the most relevant ChIP-seq profile if multiple profiles exist for the same TF. In step 2, RABIT further selected a subset of TFs among those screened in step 1 to achieve an optimized model error. In step 3, RABIT investigates how well the public ChIP-seq profiles can capture the active TF targets in each cancer type and clean up insignificant TFs.

We first collected 686 ChIP-seq profiles from the ENCODE project, representing 150 TFs and 90 cell types (7). For a given TF ChIP-seq dataset, candidate target genes are identified by weighting the number of binding sites by their distance to the transcription start site (TSS) of each gene, using the BETA method we developed (19). Then, in each tumor sample, RABIT tests whether the putative target genes of a TF show significant differential expression compared with the normal controls (step 1 in Fig. 1). To correct for the influence of gene CNA, promoter DNA methylation, promoter CpG content, and promoter degree (total number of ChIP-seq peaks near the gene TSS) on gene expression (SI Appendix, Fig. S1), RABIT uses multivariate linear regression (Table 1 and SI Appendix, Fig. S2A). Because this linear regression needs to be conducted against many ChIP-seq profiles and in a large collection of tumor samples, RABIT uses the efficient Frisch–Waugh–Lovell (FWL) method for regression (20). FWL separates factors that are invariant in each tumor, such as CNA and promoter degree, and only regresses against each variable ChIP-seq profile to speed up the calculation (Fig. 1B).

After running the regression in each tumor, a regulatory activity score can be defined for each TF, which is the t value of the linear regression coefficient t test (coefficient/SE). When one TF has several ChIP-seq profiles from different conditions, RABIT only keeps the profile that gives the largest statistical effect of regulatory activity on target genes. In step 2, RABIT further selects a subset of TFs among those screened in step 1 by stepwise forward selection to achieve an optimized model error.

Using linear regression, the effect of TF regulatory potential on target gene expression is evaluated after controlling for background effects of promoter degree, promoter CpG content, gene CNA, and promoter methylation. In this example, the ENCODE MYC ChIP-seq profile in the MCF-7 cell line is analyzed together with TCGA data of breast tumor TCGA-AO-A03P-01A. The significance of the regression coefficient is evaluated by the t test, and the TF regulatory activity score is defined as the t value (regression coefficient/SE) of the TF regulatory potential, which is shown in the bold text row.

Table 1. Multivariate linear regression for TF regulatory activity

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>SE</th>
<th>t value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter degree</td>
<td>0.0013</td>
<td>0.0003</td>
<td>4.99</td>
<td>6.14e-07</td>
</tr>
<tr>
<td>CpG content</td>
<td>0.1823</td>
<td>0.0384</td>
<td>4.75</td>
<td>2.05e-06</td>
</tr>
<tr>
<td>Gene CNA</td>
<td>0.7281</td>
<td>0.0211</td>
<td>34.55</td>
<td>7.92e-252</td>
</tr>
<tr>
<td>Promoter methylation</td>
<td>-0.6742</td>
<td>0.0744</td>
<td>-9.07</td>
<td>1.37e-19</td>
</tr>
<tr>
<td>Regulatory potential</td>
<td>0.6359</td>
<td>0.0417</td>
<td>15.24</td>
<td>4.79e-52</td>
</tr>
</tbody>
</table>

Fig. 1. Search transcription factors driving tumor-specific gene expression patterns. (A) The input to RABIT framework includes TF ChIP-seq profiles, recognition motifs, and tumor-profiling datasets. RABIT uses three steps to identify TFs that drive tumor-specific gene expression patterns at both the individual tumor level and the whole cancer-type level. In steps 1 and 2, for each tumor sample, RABIT tests whether the TF target genes are significantly up-regulated or down-regulated compared with the normal controls. In step 3, for each cancer type, RABIT tests whether the TF gene expression and somatic mutation are correlated with the scores of TF regulatory activity on target genes across all tumors and cleans up TFs with poor correlation. (B) In step 1, the efficient Frisch–Waugh–Lovell method of linear regression is applied to test the impact of TFs on target gene regulation after controlling for background factors. A set of TFs with significant regulatory activity is screened. If one TF has several ChIP-seq profiles from different conditions, RABIT only keeps the profile that gives the largest absolute value of regulatory activity on target genes. In step 2, RABIT further selects a subset of TFs among those screened in step 1 by stepwise forward selection to achieve an optimized model error.
After modeling the TF activity on individual tumor level, RABIT investigates how well the public ChIP-seq profiles used can capture the active TF targets in each cancer type (step 3 in Fig. 1). RABIT tests whether the TF gene expression and somatic mutations are correlated with the differential expression of TF target genes across all tumors in one cancer type. As an example, the GATA3 expression level is positively associated with its target gene differential expression across breast tumors, and the presence of GATA3 somatic mutation is negatively associated with its target gene differential expression (Table 2 and SI Appendix, Fig. S2B). Those TFs with insignificant cross-tumor correlation are removed from the results.

With the RABIT framework, we integrated 686 ENCODE ChIP-seq profiles with 7484 TCGA tumor profiles over 18 cancer types (Fig. 2 and SI Appendix, Table S1) (7, 10). The impact of TFs on tumor gene expression predicted by our framework is highly consistent with previous knowledge. For example, RABIT predicted the target genes of MYC to be significantly up-regulated in numerous cancers (star in Fig. 2), consistent with the known role of MYC as an oncogenic TF (21). FOXM1 also up-regulates its target genes in many cancer types (star in Fig. 2) and could become a potential therapeutic target for many cancer types (5). For MYC and FOXM1, they are clustered together with several other TFs by their regulatory similarity across cancer types (cluster 1 in SI Appendix, Fig. S4). The target genes of TFs in this cluster are preferentially up-regulated in most cancer types, which indicates these TFs as pervasive oncogenic regulators. Another example, RAD21, a member of the cohesion complex with important roles in chromosome maintenance (22), has its target genes repressed in breast cancer (BRCA) (Fig. 2 and SI Appendix, Fig. S3). One recent study reported a high level of RAD21 expression to be indicative of poor prognosis and resistance to chemotherapy in breast cancers (23), which is consistent with our findings. We found RAD21 is clustered with a set of TFs, whose target genes are generally repressed in most cancer types (cluster 3 in SI Appendix, Fig. S4).

Besides capturing knowledge from previous studies, our analysis also predicted putative TF functions in cancer. For example, the target genes of SPI1 are significantly up-regulated in Glioblastoma (GBM) and kidney renal clear cell carcinoma (KIRC), as predicted by RABIT (Fig. 2). SPI1 is clustered with several other TFs, featured by significant up-regulation of target genes in both GBM and KIRC (cluster 4 in SI Appendix, Fig. S4). We found that high SPI1 expression is associated with poor patient survival in GBM and KIRC (SI Appendix, Fig. S6A–D). To our knowledge, the role of SPI1 is not well studied in these cancers. Our result indicates SPI1 as a promising target for further study.

As another example of predictions from RABIT, we explored our results on breast cancer, which is an intensively studied cancer type in the past decade. We found that 80 out of 150 TFs analyzed have support from the National Cancer Institute (NCI) cancer gene index or Google search with gene name to be related to breast cancer. Among the rest, there are 21 TFs whose target genes show significant differential expression in TCGA and METABRIC cohorts (SI Appendix, Fig. S7), but little is known about their role in breast cancer from the literature. For example,
the histone demethylase PHF8 target genes are up-regulated in more than 60% of breast tumors (SI Appendix, Fig. S7). There has been no study implicating PHF8 in breast cancer to the best of our knowledge, but PHF8 is known oncogene and potential therapeutic target for esophageal squamous cell carcinoma and prostate cancer (24, 25). Our prediction suggests that PHF8 might also be an oncogene in breast cancer.

**Quality Assessment of RABIT Result and Method Comparison.** Encouraged by the consistency of RABIT output with a few previous studies, we set out to systematically check the quality of our computational predictions. To measure the cancer relevance level of a TF, we computed the percentage of tumors with the TF target genes differentially regulated and averaged across all TCGA cancer types. We tested whether the cancer relevance levels defined above are consistent with cancer gene databases. The NCI cancer gene index project assigned a number for each gene as the count of sentences from MEDLINE abstracts in which the gene name and a cancer term cooccurred (26). We also included databases with annotations of cancer-related genes, including the Bushman Laboratory cancer driver gene list (27, 28), the COSMIC somatic mutation catalog (29), and the CCGD mouse cancer driver genes (30). We found TFs with higher percentage of tumors showing target differential expression are associated with the cancer gene annotations in all databases (Fig. 3A and SI Appendix, Fig. S8A).

Because our predicted TF cancer relevance is highly consistent with the annotations from cancer gene databases, we use the knowledge from these databases as the gold standard to compare RABIT performance with several other methods. The gold-standard positive set is defined as TFs annotated as cancer-associated in at least two out of four cancer gene databases described above, and the negative set is defined as the rest of the TFs. Using receiver-operating characteristic (ROC) and precision-recall (PR) curve, we compared the ability of predicting cancer-associated TFs among several methods (details provided in SI Appendix, SI Methods). RABIT has the largest area under curve (AUC) among all methods (Fig. 3B and SI Appendix, Fig. S8B and C). The second-best methods are LAR and LASSO, which are very popular regression-based feature selection algorithms (31, 32). Because the first two steps of RABIT framework (without step 3) also composed a general feature selection algorithm (Fig. 1B), we compared among RABIT, LAR, and LASSO on the performance of feature selection. Using each algorithm, we select the top 10 most significant TFs to predict the tumor gene expression patterns. RABIT achieved better cross-validation error and shorter running time than LAR and LASSO in all cases tested (SI Appendix, Fig. S9).

To check whether RABIT can accurately identify important TFs in a condition studied, we compared the RABIT results with the growth phenotypes after TF knockout in cell lines. There are two recent works of genome-wide CRISPR screening in K562 cell and HL60 cell (33, 34). In a screening experiment, each gene is assigned a score to represent whether the cell growth rate is negatively correlated with the TF gene CRISPR screening scores (Fig. 3C and D). This means if RABIT assigns a TF as highly active, knocking out that TF will significantly slow down the cell growth.

To check whether our results on TCGA datasets are consistent with other tumor-profiling cohorts, we applied RABIT on METABRIC breast cancer data (35), Rembrandt glioma data (36), Gravendeel glioma data (37), and Genotype-Tissue Expression (GTEx) normal tissue data (38) (SI Appendix, Fig. S10 A and B). As expected, the TCGA results show positive correlation with other tumor cohort results but negative correlation with GTEx normal tissue cohorts (SI Appendix, Fig. S10 C and D). For example, onecogenes in breast cancer (such as MYC, FOXM1, and RAD21) and oncogenes in GBM (such as SPI1) show the same direction of up- or down-regulation of their target genes in both TCGA and other tumor cohorts, but no regulatory activity is observed in GTEx normal tissue (SI Appendix, Fig. S10 A and B).

Besides ChIP-seq data, TF recognition motifs can also model TF-binding specificity. To check whether our findings based on ChIP-seq data are consistent with the TF regulatory motifs’ result, we collected recognition motifs of 505 TFs from several studies (9, 39, 40). We searched the matches of TF motifs near gene transcription start sites and applied RABIT to characterize TF activity in regulating tumor gene expression (SI Appendix, Fig. S11A). Between ChIP-seq results and regulatory motif results of a TF in each tumor, we computed the correlation of TF regulatory activity scores (Fig. S11B), indicating consistency between ChIP-seq and recognition...
motif analysis. This result is not surprising because ChIP-seq-binding peaks and TF motif-binding sites significantly correlate with each other (SI Appendix, Fig. S11C). Thus, our result of TF activity correlation could be derived from the similarity of TF target genes.

**Landscape of Posttranscriptional Regulation in Cancer.** Besides TFs, RBPs can also control gene expression through posttranscriptional regulation. Recent years saw increasing studies demonstrating RBPs as important players in tumorigenesis (41), although the function and targets of the vast majority of RBPs are still uncharacterized. To this end, we collected 172 recognition motifs for 133 RBPs (42) and predicted the putative targets of each by searching the recognition motifs over gene 3′UTR regions (43). We then analyzed the regulatory activity of RBPs in the same way as TFs, using RABIT.

As an example of RBP regulatory activity in cancer, there are four RBPs (RBFOX1, RBFOX2, RBFOX3, and EIF2S1) with almost the same binding preference of GCAUG sequence, and their motifs are clustered together (Fig. 4A). The gene targets of this motif cluster are strongly down-regulated in most cancer types, especially in GBM (Fig. 4B). We found that a higher level of motif targets down-regulation indicated worse patient survival in GBM (Fig. 4C and SI Appendix, Fig. S6 E and F). Because there are several RBP members in this cluster, we applied the forward selection algorithm and found RBFOX1, RBFOX2, and RBFOX3, but not EIF2S1, are the relevant factors driving regulatory activity (SI Appendix, Table S2). RBFOX1, RBFOX2, and RBFOX3 are known as evolutionarily conserved tissue-specific alternative splicing regulators in metazoans (44). A previous study showed that RBFOX1 suppresses malignancy in glioma by regulating the alternative splicing of TPM1 (45). Our analysis suggests that besides regulating alternative splicing, RBFOX1 and its homologs could bind the gene 3′UTR regions and increase mRNA stability (SI Appendix, Table S2). The loss of RBFOX1 target stabilization is universal among GBM tumors and serves as an indicator of poor patient survival (Fig. 4).

Besides the RBFOX1 regulatory motif cluster, there are many other RBP motifs showing significant regulatory impact in cancer. For example, cluster 1, which is composed of nine RBPs, has its target genes significantly up-regulated in most cancer types (Fig. 4B). Among cluster 1 members, HnrN (ELAV1) is a well-known oncogenic RBP that promotes tumor proliferation and malignancy in many human cancers (46). Besides these examples of cancer-associated RBPs, the RABIT framework provides a regulatory map between 133 RBPs and 18 cancer types, which will facilitate further exploration of posttranscriptional regulation in cancers.

**Discussion**

This study comprehensively integrated TF ChIP-seq and binding motifs with TCGA tumor-profiling data for systematic identification of cancer-associated TFs. RABIT has shown superior performance compared with other state-of-the-art methods. We also applied RABIT to identify a set of RBPs that might play important roles in shaping gene expression in tumors, demonstrating RABIT as a versatile framework for finding cancer-associated gene expression regulators. Notably, there are abundant previous works that integrate ChIP-seq and gene expression data to understand gene regulatory mechanisms (15). For example, ChIP-seq profiles of 12 TFs and RNA-seq expression profiles in mouse embryonic stem cells have been analyzed together, using the regression method (16–18). However, these previous studies were conducted when ChIP-seq and expression-profiling data were generated in the same condition, without further requirement of removing any background confounding effect. Thus, compared to previous works, RABIT is specially designed for large-scale regulatory analysis across diverse cancer types.

As a limitation of RABIT, the linear model used assumes a TF either up-regulates or down-regulates its target genes, which can be represented by one regression coefficient with a positive or negative sign. However, certain TFs may up-regulate and down-regulate their target genes, which can be represented by one regression coefficient with a positive or negative sign. Certain TFs may up-regulate and down-regulate their target genes, which can be represented by one regression coefficient with a positive or negative sign. Certain TFs may up-regulate and down-regulate their target genes, which can be represented by one regression coefficient with a positive or negative sign.

With the development of high-throughput sequencing technologies, high-quality transcriptome and mutation profiles of tumor samples have been rapidly generated for diverse cancer types. However, a big gap still exists between getting the tumor profiles and understanding the molecular mechanism of tumorigenesis. Public resources such as ChIP-seq and recognition motifs provide a rich base for bridging this gap and understanding how cancer genes are regulated. Our study provides a cost-effective and systematic framework for integrating regulatory genomics resources with tumor-profiling data to better understand gene regulation in cancers.

**Methods**

**Background Factors of Tumor Gene Expression.** A large portion of tumor gene expression variation is derived from the gene CNA and promoter DNA methylation, which are not direct effects of TF regulation (14). We also found promoter degree (total number of ChIP-seq peaks near the gene TSS) is positively correlated with tumor expression patterns in most cancer types (SI Appendix, Fig. S1A). The promoter Cpg content, defined as (CpG dinucleotide frequency)/C frequency × G frequency) 1kb around gene TSS, also has a strong positive correlation with gene expression in cancer types.
such as lung squamous cell carcinoma (LUSC) (SI Appendix, Fig. S18). When testing the TF regulatory impact on gene expression, RABIT controls the effect from these background factors (gene CNA, promoter DNA methylation, promoter degree, and CpG content).

Search TFS Driving Tumor Gene Expression Patterns. RABIT uses three steps to identify TFS that drive tumor-specific gene expression patterns at both the individual tumor level and the whole cancer-type level. In step 1, RABIT screens for TFS that significantly affect the gene expression patterns in each tumor and selects the most relevant ChIP-seq profile, if multiple profiles exist for the same TF. In step 2, RABIT further selects a subset of TFS among those screened in step 1 to achieve an optimized model error. In step 3, RABIT investigates how well the public ChIP-seq profiles can capture the active TF targets in each cancer type and cleans up insignificant TFS.

In step 1, RABIT runs a TF screening by testing the regulatory impact for each individual TF with a linear regression (Fig. 1B). The regression units are human genes, and the response variable is gene expression difference between tumor and normal sample (SI Appendix, Fig. S2A). The regression covariates include the regulatory posttransductional scores of an individual TF over gene promoters and four background factors defined in the section above (five covariates in total). The regression coefficients and their SEs are estimated by the least squares method. We defined the t value (coefficient/SE) as the regulatory activity score for each TF and assessed its significance by the t test (Table 1). The P values are converted to false discovery rates (FDR) by the Benjamini–Hochberg procedure. We screen a set of significant TFS with an FDR threshold of 0.05. If several ChIP-seq profiles exist for the same TF, we select the profile with the highest absolute value of TF regulatory activity score (Fig. 1B and SI Appendix, Fig. S3).

In step 2, to achieve an optimized error in predicting tumor gene expression patterns, we apply stepwise forward selection to find a subset of TFS among those screened in step 1 (Fig. 1F). We use Mallows’s Cp as a model selection criterion to decide which covariates (screened TFS) should be included in the model (47). For each tumor, we start with four covariates of gene expression background factors (promoter degree, CpG content, gene CNA, promoter methylation), and search through the TFS screened in step 1. At each round, one TF is selected from the candidate set to best minimize the Mallow’s Cp. The process is repeated until no TFS can be added to further reduce the Mallow’s Cp.

In step 3, RABIT cleans up cases where the target genes identified from public ChIP-seq profiles cannot represent active TF target genes in a cancer type because most public ChIP-seq data were generated under experimental conditions distinct from those in cancers. For each TF, RABIT computed its regulatory activity score in each tumor in step 1, which measures the level of the TF’s control of gene expression’s differential expression (Table 1). We regressed the TF regulatory activity score against the two covariates, TF gene expression value and TF somatic mutation count, across all tumor samples in the same cancer type (Table 2 and SI Appendix, Fig. S2B). The significance level of each covariate is assessed by the t test, and the P values of all covariates are grouped together and converted to FDRs by the Benjamini–Hochberg procedure. Those TFS with insignificant coefficients on both covariates (gene expression and somatic mutation) are excluded from results (FDR threshold, 0.05).

Acknowledgments. We thank Myles Brown, David Louis, Mario Suva, and Yang Cao for helpful discussions. The work was supported by NIH Grants HG5R U1A1 HG7000, NC1 U01 CA108990, R01 GM131262-01, and NSF DBI 1357771.