frontiers in Genetics

METHODS
published: 31 May 2021
doi: 10.3389/fgene.2021.664654

Integrative Ranking of Enhancer Networks Facilitates the Discovery of Epigenetic Markers in Cancer

*Qi Wang [1,2], Yonghe Wu [3], Tim Vorberg [2], Roland Eils [1,4] and Carl Herrmann [1]**

[1] Health Data Science Unit, Medical Faculty Heidelberg and BioQuant, Heidelberg, Germany, [2] Faculty of Biosciences, Heidelberg University, Heidelberg, Germany, [3] Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany, [4] Digital Health Center, Berlin Institute of Health (BIH) and Charité, Berlin, Germany

Regulation of gene expression through multiple epigenetic components is a highly combinatorial process. Alterations in any of these layers, as is commonly found in cancer diseases, can lead to a cascade of downstream effects on tumor suppressor or oncogenes. Hence, deciphering the effects of epigenetic alterations on regulatory elements requires innovative computational approaches that can benefit from the huge amounts of epigenomic datasets that are available from multiple consortia, such as Roadmap or BluePrint. We developed a software tool named IRENE (Integrative Ranking of Epigenetic Network of Enhancers), which performs quantitative analyses on differential epigenetic modifications through an integrated, network-based approach. The method takes into account the additive effect of alterations on multiple regulatory elements of a gene. Applying this tool to well-characterized test cases, it successfully found many known cancer genes from publicly available cancer epigenome datasets.

Keywords: enhancer, epigenetics, histone modification, chromatin interaction, network analysis

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Fuhai Li,
Washington University in St. Louis,
United States
Loredana Martignetti,
INSERM U900 Cancer Et Génome
Bioinformatique, Biostatistiques Et
Épidémiologie, France

****Correspondence:***
Carl Herrmann
carl.herrmann@
bioquant.uni-heidelberg.de

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *05 February 2021*
Accepted: *29 March 2021*
Published: *31 May 2021*

Citation:
Wang Q, Wu Y, Vorberg T, Eils R and
Herrmann C (2021) Integrative
Ranking of Enhancer Networks
Facilitates the Discovery of Epigenetic
Markers in Cancer.
Front. Genet. 12:664654.
doi: 10.3389/fgene.2021.664654

INTRODUCTION

Epigenetic alterations are frequent in many cancers. In particular, DNA methylation and histone modifications are two main mechanisms that allow cancer cells to alter transcription without changing the DNA sequences, and lead to many abnormalities such as persistent activation of cell cycle control genes or deactivation of DNA repair genes. For example, promoter DNA hypomethylation accompanied by histone hyper-acetylation is frequently observed in the activation of oncogenes in cancer. Besides, aberrant activation of distal regulatory elements is often associated with the up-regulation of cancer-promoting genes. Interestingly, epigenetic modifications at proximal and distal regulatory elements often appear to be earlier events than the gene expression (Hartley et al., 2013; Ziller et al., 2014), and can hence serve as potential early markers in cancer diagnosis.

Various histone modifications on promoters have been categorized into either activation or repression effects on gene expression. Such effects can be measured by comparing histone alteration levels between tumor and their corresponding normal tissues using ChIP-Seq (Karlic et al., 2010). A number of tools, such as ChIPComp (Chen et al., 2015), ChIPDiff (Xu et al., 2008), ChIPnorm (Nair et al., 2012), csaw (Lun and Smyth, 2015), DBChIP (Liang and Keles, 2012), DiffBind (Stark and Brown, 2011), MAnorm (Shao et al., 2012), RSEG (Song and Smith, 2011) have demonstrated their usefulness in cancer studies by comparing the histone intensities between two conditions (see Steinhauser et al., 2016 for a review of these tools). However, they are limited to the comparison of a single histone mark. Furthermore, many tools such as jMOSAiCS (Zeng et al., 2013), IDEAS (Zhang et al., 2016), and ChromHMM (Ernst and Kellis, 2012) are able to perform integrative analyses across multiple epigenetic marks. However, while these tools provide an integrated description of the epigenetic characteristics at individual genome loci, they do not take into account the combined effects of these changes at multiple regulatory elements controlling a gene.

As previously mentioned, many histone modifications that potentially regulate gene expression also occur in other genomic regions besides promoters. Enhancers are distal regulatory elements that interact with gene promoters through chromosomal loops to regulate gene transcription. Most of the enhancers are located within ±1 Mb of the transcription start site (TSS) of their target genes (Maston et al., 2006). Enhancer activity is regulated through epigenetic modifications (Zentner et al., 2011), including positive regulation from histone marks, such as H3K27ac (Creyghton et al., 2010; Stasevich et al., 2014) and H3K4me1 (Heintzman et al., 2007; Calo and Wysocka, 2013), and negative regulation by H3K27me3 (Charlet et al., 2016) and H3K9me3 (Zhu et al., 2012).

Given the complexity of epigenetic regulation, novel tools are required to combine this information, and create a comprehensive overview of the differential epigenetic landscape, integrating multiple data layers. The method we developed, named IRENE (Integrative ranking with an epigenetic network of enhancers), combines a quantitative analysis on multiple differential epigenetic modifications with an integrated, network-based approach, in which we integrated two levels of epigenetic information: the signal intensity of each epigenetic mark, and the relationships between promoters and distal regulatory elements known as enhancers (**Figure 1**). In this paper, we describe the method and present the test cases. In our benchmarking tests on cancer datasets, the IRENE ranked lists have higher relevance to cancer marker genes (CMGs) than the other approaches. Being implemented as an R package, IRENE is an easy to use method allowing gene ranking between two conditions and highlighting potential cancer biomarkers.

RESULTS

IRENE: Epigenetic Ranking With an Epigenetic Network of Enhancers

IRENE analyzes epigenetic changes between two biological conditions (e.g., ChIP-seq data for histone modifications or whole-genome bisulfite sequencing for DNA methylation), and translates the differential signals at multiple regulatory elements into a unique score (**Figure 1**). Hence, IRENE performs a double integration, both across multiple epigenetic datasets and across different regulatory regions linked to a gene. To integrate multiple datasets, we use dPCA, which captures the directions of the greatest differential variance comparing two conditions, at each regulatory element (see section Materials and Methods) (Ji et al., 2013). As the goal of our method is to capture the differential signal at proximal and distal regulatory elements, we performed a dPCA analysis both at gene promoters and distal regulatory elements, which we call promoter interacting regions (PIRs) extracted from the 4DGenome database (Teng et al., 2015). Similar to standard PCA, differential PCA captures the directions of the greatest differential variance along several differential principal components (dPCs). We selected the first two dPCs, which appear to capture the differential signal both from activating and repressive epigenetic marks. The sum of the absolute values of dPC1 and dPC2 at each regulatory element was used as a score for this element. These scores are summarized as a weighted network relating regulatory elements to their target genes. The network consists of promoters and connected PIRs. Oriented edges from PIRs to promoters indicate a 3D interaction between these regulatory elements. Despite being in principle a bipartite graph (with nodes being either PIRs or promoters), we do not make a distinction between these two types of regulatory elements. A random walk based method then assigns a score to the corresponding gene. The output of the method is a ranked list of genes from the most to the least affected one, which incorporates both promoter and enhancer alterations. As a comparison, we also generated ranked lists based only on the promoter score (named promoter ranked lists in the following), discarding the contributions from distal PIR elements. This approach can be applied whenever two conditions are to be compared, for example, normal/tumor tissue, various tumor subtypes, or different developmental stages. More details are given in the Materials and Methods section. In order to benchmark our method, we used seven test cases consisting of tumor samples for seven different tumor types and normal matching samples. For each of these test cases, we compiled a list of CMGs (**Supplementary Table 2**) from the literature, and considered tissue-specific genes (TSGs) obtained from the ArchS4 database (Lachmann et al., 2018) as controls.

Cancer Marker Genes Are Scored Higher by Incorporating Enhancer in the Ranking

In our analysis, we determined that taking into account the first two dPCs is able to capture most of the differential variance for both activating and repressive epigenetic modifications (**Figures 2A,B**). After comparing the dPC1+dPC2 values between the CMGs and TSGs in each test case, we found that the scores from CMGs are generally higher than the scores of the TSGs for the enhancers, whereas the situation is less clear at promoters. This might indicate that most of the differential signal between tumor and normal occurs at distal regulatory regions. (**Figure 2C**).

Using the ranked gene lists generated by IRENE, we further computed the area under the curve (AUC) for the empirical cumulative density function (ECDF) of the high-confidence CMG ranks as a benchmarking approach, as described in the methods. First, we examined the IRENE ranks computed using the dPC1+dPC2 on gene promoters and their targeting enhancers, and found that the marker genes are ranked higher than TSGs in every test case, indicating that our approach captures the specific differential epigenetic signals at CMGs (**Figure 3A**). Moreover, both for CMGs and TSGs, the IRENE AUC values are higher than the AUC values computed using the

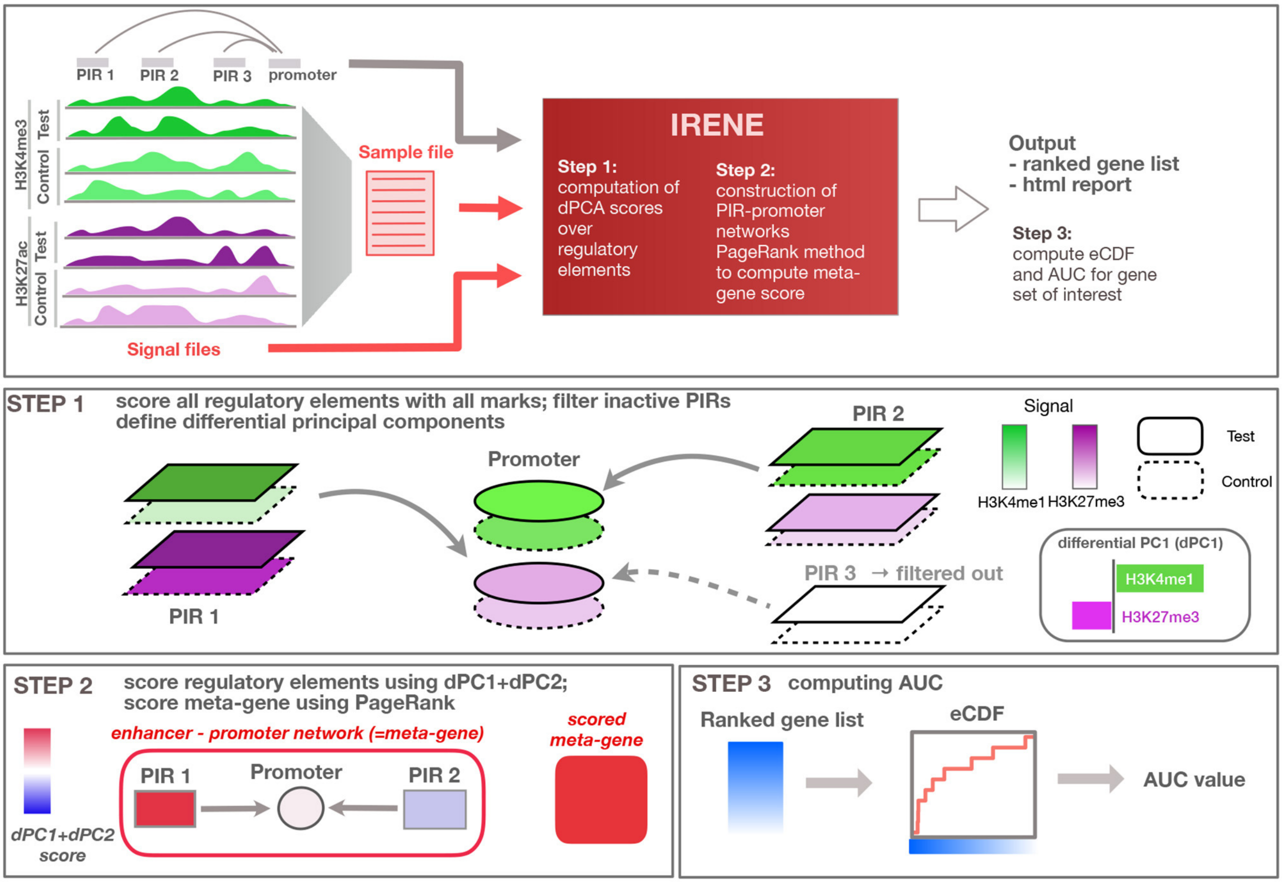
PIR 1
PIR 2
PIR 3
promoter
H3K4me3
Test
Control
H3K27ac
Test
Control
Sample file
Signal files
IRENE
Step 1:
computation of dPCA scores over regulatory elements
Step 2:
construction of PIR-promoter networks
PageRank method to compute meta-gene score
Output
- ranked gene list
- html report
Step 3:
compute eCDF and AUC for gene set of interest
STEP 1
score all regulatory elements with all marks; filter inactive PIRs
define differential principal components
Promoter
PIR 1
PIR 2
PIR 3 → filtered out
Signal
H3K4me1
H3K27me3
Test
Control
differential PC1 (dPC1)
H3K4me1
H3K27me3
STEP 2
score regulatory elements using dPC1+dPC2;
score meta-gene using PageRank
enhancer - promoter network (=meta-gene)
PIR 1
Promoter
PIR 2
scored meta-gene
dPC1+dPC2 score
STEP 3
computing AUC
Ranked gene list
eCDF
AUC value

FIGURE 1 | Overview of the method. General overview of the input/output of the IRENE method; user defined input is indicated in red, while input provided by the method (e.g., regulatory loci and interactions) are displayed in gray. Step 1: Regulatory elements are scored using all epigenetic modifications, and related to the target gene. Step 2: Epigenetic alterations are scored using the first dPC and combined using PageRank into an integrated meta-gene score. Step 3: Ranked gene lists based on the score are converted to eCDF curves showing the enrichment of a given gene set within the top-ranked genes, and corresponding area under the curve (AUC) values are computed.

dPC1+dPC2 of gene promoters only (**Figure 3A**). The fact that the genes ranked higher in IRENE suggests that a significant part of the altered epigenetic alteration arises from distal enhancer regions. We then validated these findings on the larger CMG and TSG gene sets, and we found the AUCs of CMGs are all significantly higher (one-tailed *t*-test **p**-value<0.01) than the AUCs of TSGs (**Figure 3B**).

Some genes have a much high number of linked enhancers than others. To test whether this might bias the ranks of these genes, we performed 1,000 degree-preserving random perturbations, which completely rewired the enhancer–promoter graph but maintaining the degree distribution. We used the high-confidence CMGs in the benchmarking, and the AUCs with randomly assigned enhancers dropped 5–10% on average, indicating that the higher ranks of CMGs are not explained by their higher connectivity (**Figure 4**).

We compared the target gene assignment provided by the 4DGenome database, which is based on experimental evidence, with the simpler nearest-gene assignment. As can be observed in **Figure 4**, both approaches lead to comparable results, in line with recent reports indicating that the nearest gene assignment is reasonably effective in linking enhancers with target genes (Moore et al., 2020).

As mentioned in the Introduction, several other methods have been developed to integrate multiple epigenetic marks over genomic regions. Most of these methods provide qualitative analysis in the form of discrete chromatin states. To our knowledge, none of these methods apply a network-based integration as in IRENE to summarize regulatory elements related to the same gene. In order to provide a comparison, we focused on one of the mostly used such method, ChromHMM, which integrates various histone marks into discrete chromatin states (Ernst and Kellis, 2012). We combined ChromHMM with the Chromswitch method (Jessa and Kleinman, 2018), which computes a differential score between two groups of samples over specific regions. Applying this scoring approach to promoter regions, we compared the ranked lists obtained by IRENE at promoter regions with the ChromHMM-based ranks for the

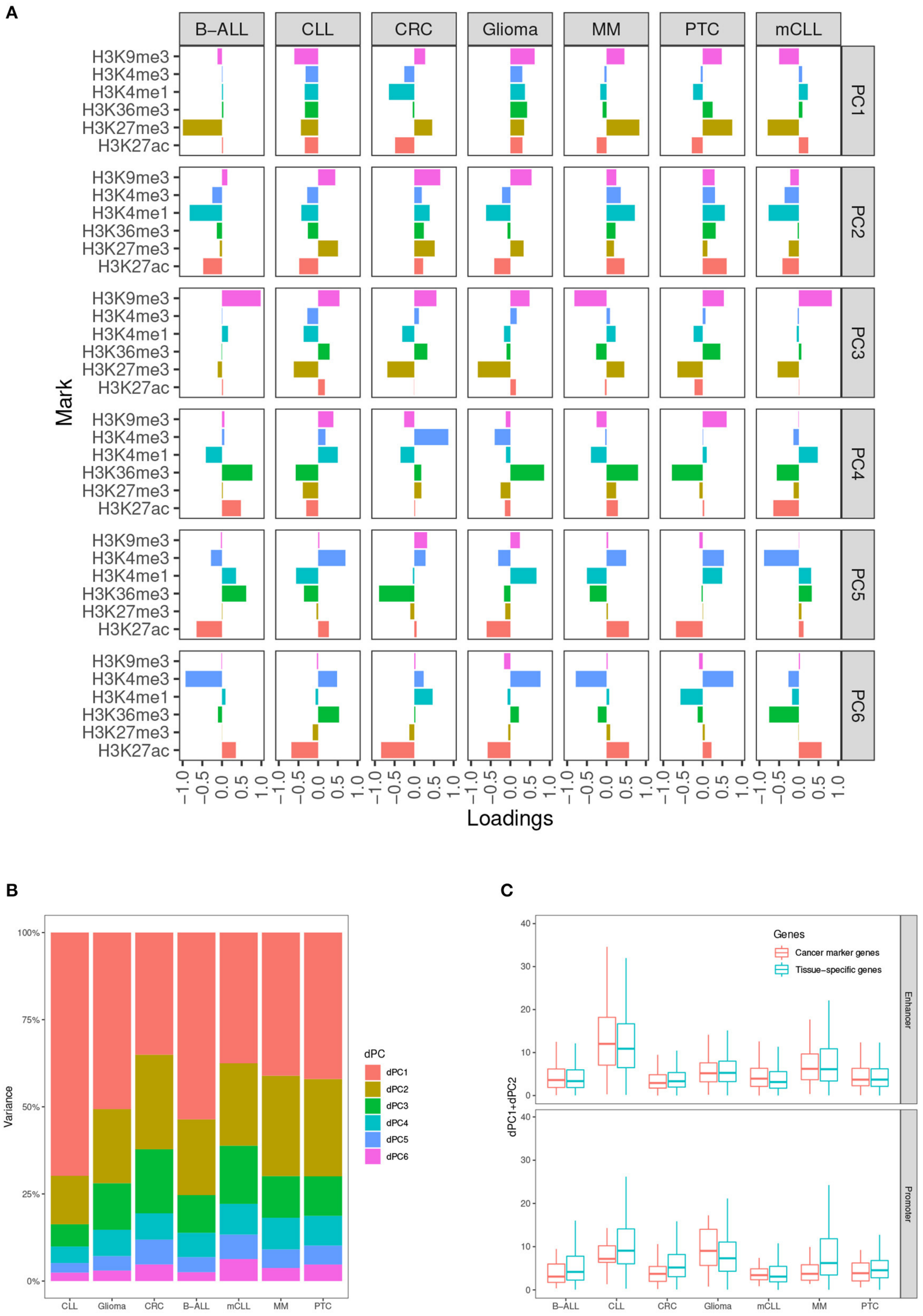

FIGURE 2 | Differential principal components. **(A)** Contributions of the six histone marks to each differential principal component (dPC). **(B)** Variances accounted for each dPC in the seven test cases. **(C)** Values of dPC1+dPC2 in the seven test cases, comparing cancer marker genes (CMGs) with TSGs, both for enhancers (top), and promoters (bottom).

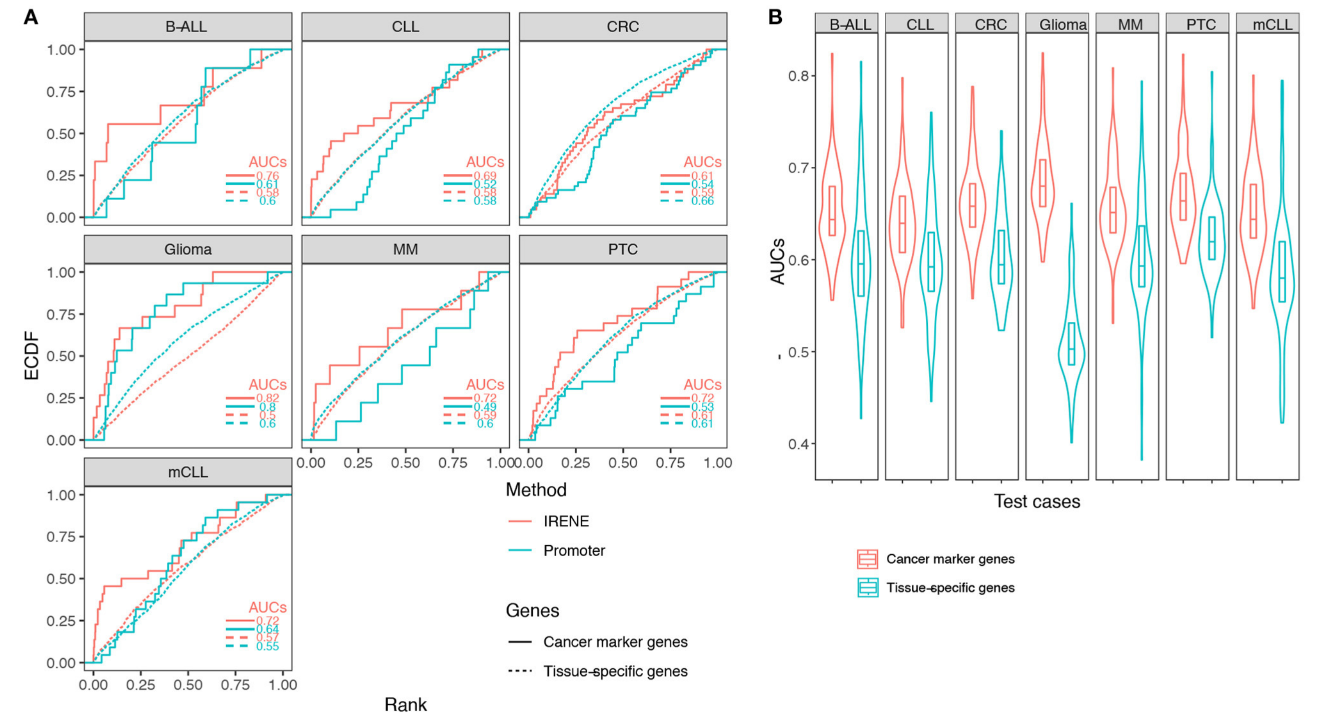

FIGURE 3 | (A) Empirical cumulative density function (ECDF) curves regarding the cancer marker genes and tissue-specific genes in seven test cases. The marker gene ranks using IRENE scores (red) are compared against their ranks using the promoter scores (cyan). **(B)** Distribution of the area under the curve (AUC) values using the cancer marker gene (CMG) sets from CancerMine in the seven test cases, compared to randomly picked tissue-specific genes to define equal size sets.

Glioma/normal brain test case, and found that the AUC values of the CMGs related to Glioma are significantly higher for the IRENE method (**Supplementary Figure 2**).

Network Analyses Characterized the Highly Ranked Genes in the IRENE and Promoter List

We downloaded 184 KEGG pathways in KGML format and loaded them as directed graphs using KEGGgraph (Zhang and Wiemann, 2009). Then we took the top 15% genes from the IRENE and promoter rank lists in each one of the seven test cases, and mapped the genes to the KEGG cancer signaling pathway (hsa05200). In total, the reference pathway contains 531 genes and 1989 interactions, and on average 208 of the 531 genes are found in the IRENE rank lists, while only 152 genes are found in the promoter rank lists. In addition, the IRENE-ranked genes differ from promoter-ranked genes in both in-degrees and out-degrees of the nodes (**Table 1**). As the IRENE nodes generally have higher in-degrees than out-degrees in the graph presentation of the reference pathway, implying the IRENE genes are more often targeted by the other regulatory genes on their enhancers as they harbor more differential enhancers. We further examined the glioma signaling pathway (hsa05214) and found 19 genes from the IRENE rank list and 10 genes from the promoter rank list in the glioma test case (**Figure 5**). One common gene, *EGFR*, is in both lists and has been reported to undergo tight control through epigenetic regulation on both promoters and enhancers (McInerney et al., 2000; Liu et al., 2015; Jameson et al., 2019). Moreover, nine genes are present only in the IRENE rank list, such as *CCND1*, which has been reported to be regulated by an estrogen-mediated enhancer (Eeckhoute et al., 2006). In conclusion, this analysis shows that the IRENE methods provide a ranked gene list, which is enriched for high-ranking, cancer-relevant genes.

DISCUSSION

From the above benchmarking on seven cancer test case studies, we showed that IRENE is a more comprehensive approach comparing to the current frequently used approaches such as separate ranking gene promoters and enhancers. This highlights the importance of epigenetic regulation through distant enhancer regions. Using IRENE, users cannot only discover the genes which show significantly epigenetic alterations on their promoters, but also the ones that are connected with strong epigenetic modifications on distal interacting enhancers, which facilitates the discovery of potential epigenetic marker genes. On the other hand, by interpreting the higher ranked genes mapped to the existing pathways, the user may also find the enhancers of interests from their differential epigenetic modifications. For example, we found the *PAX5* gene to have a significantly higher rank in the IRENE list compared to the promoter-only list in the two CLL case studies, which implies that *PAX5* is extensively regulated by enhancers. *PAX5* is a key

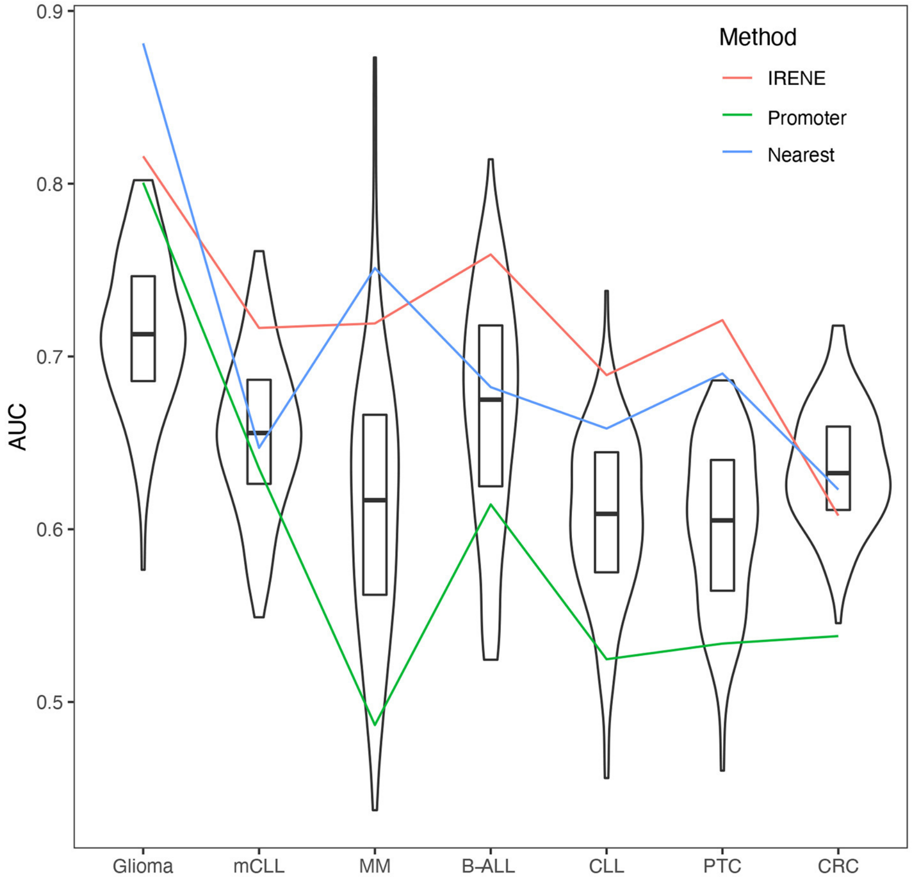

FIGURE 4 | Area under the curves (AUCs) of empirical cumulative density function (ECDF) curves of dPC1+dPC2 ranks from randomized promoter–enhancer interactions. The boxplots indicate the 25–75% quantile ranges from benchmarking each cancer marker gene set with 1,000 different rewired promoter–enhancer networks, whereas the red lines show the AUCs with the original promoter–enhancer interactions from IRENE using experimentally detected interactions (red), and interactions assigned by the nearest promoters (blue), and only promoters (green) rank lists.

TABLE 1 | Graph properties in respect of the nodes from the IRENE and promoter rank lists.

	Node number		Median in-degree		Median out-degree	
	IRENE	**Promoter**	**IRENE**	**Promoter**	**IRENE**	**Promoter**
CLL	214	167	2	2	1	3
Glioma	193	133	2	1	1	1
CRC	219	168	2	0	1	3
B-ALL	180	124	1	1	1	0
mCLL	211	168	2	0	1	3
MM	219	165	2	1	1	1
PTC	219	137	2	1	1	3

transcription factor in B-cell development, and its promoters have no significant epigenetic alterations in the CLL case studies. However, this gene is associated with several hyperacetylated and hypomethylated distal enhancers, one of which is located at 330 kilobases (kb) upstream of the *PAX5* TSS, and has been also found as extensively mutated in CLL (Puente et al., 2015) (**Figure 6**). The deletion of this enhancer resulted in a 40% reduction in the expression of *PAX5* expression and chromatin interaction of this enhancer and *PAX5* has been proven from chromosome conformation capture sequencing (4C-Seq) analysis (Puente et al., 2015). The main difficulty of this study is obtaining cell type specific enhancer–promoter interactions, as the high-resolution chromatin interaction map for the cancer cells is currently not available. We have tested two alternative approaches in this study, using either the experimentally validated chromatin interaction or distance-based interactions. The performance of the above two approaches are similar (**Figure 4**). We believe better performance can be achieved when cell type specific enhancer–promoter

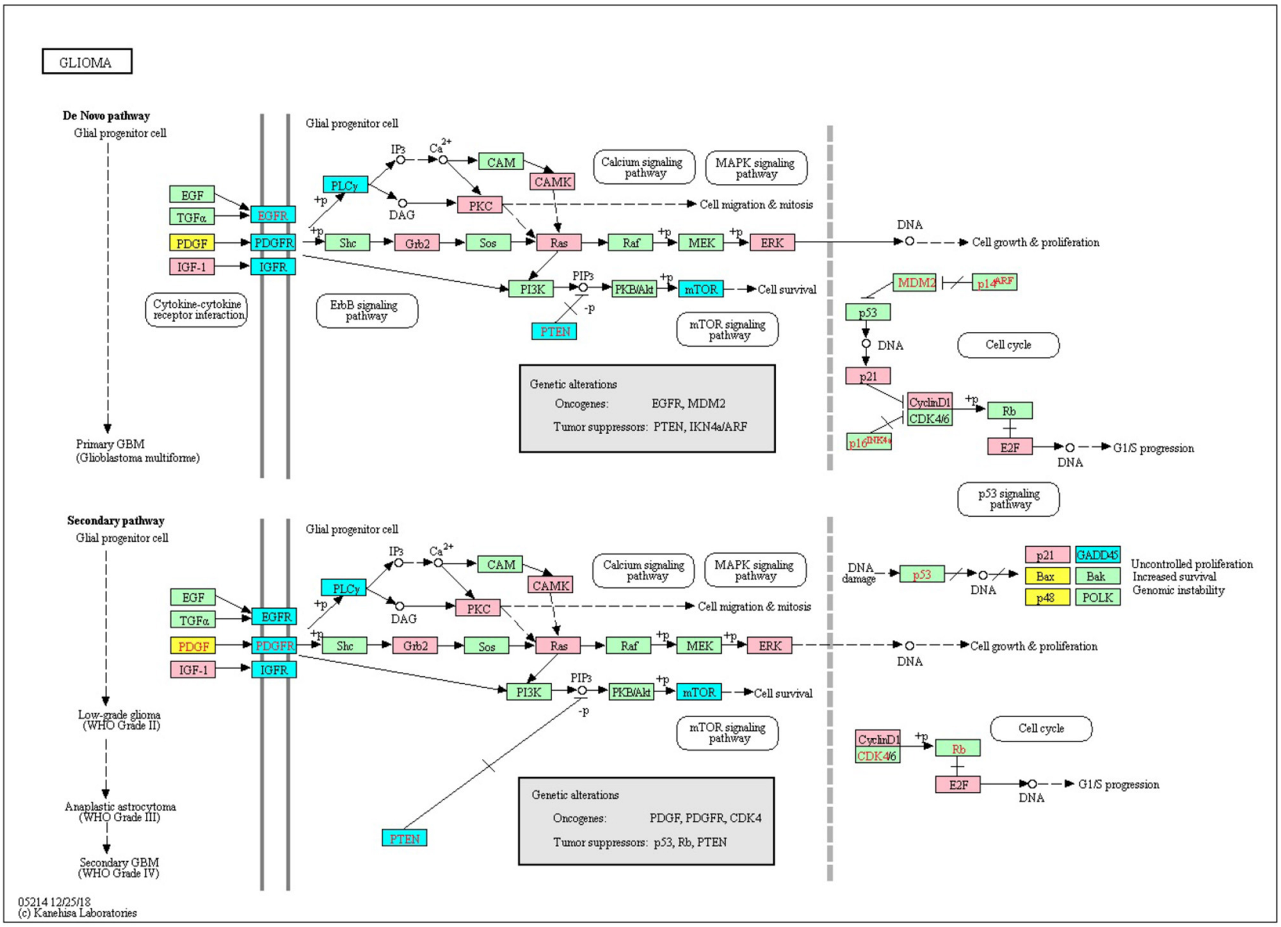

FIGURE 5 | The top 25% genes from the IRENE and promoter rank list are highlighted on the KEGG glioma signaling pathway. Pink, genes from the IRENE list; yellow, genes from the promoter list; cyan, genes from both lists.

interactions are available in the future, and using IRENE, user can replace the interaction map with a more specific one when applicable. Being a differential approach comparing two conditions, it might be affected by the possible heterogeneity of the groups being compared. If the heterogeneity is due to biological reasons (for example, different subtypes in the disease group), the comparison will be affected by the greater variance within one group. However, if the heterogeneity is of technical nature, then this noise will likely be buffered by the fact that our method integrates multiple regions to score the genes.

CONCLUSIONS

Genome-wide integrative epigenetic analysis is challenging and essential in many comparative studies. As far as we know, IRENE is the first tool that integrates quantitative and genome context information in the differential epigenetic analysis. Applying this tool to well-characterized test cases, it detects a number of candidate genes with significant epigenetic alterations, and comprehensive benchmarking validated these findings in cancer studies. As epigenomic datasets accumulate, the computational approaches employed in this study would be highly relevant in both comparative and integrative analysis of the epigenetic landscape. The discovery of novel epigenetic targets in cancers not only unfolds the fundamental mechanisms in tumorigenesis and development but also serves as an emerging resource for molecular diagnosis and treatment.

MATERIALS AND METHODS

Data Preparation

Retrieving Epigenetic Modification and Chromatin Interaction Datasets

Genome-wide ChIP-seq data are downloaded in BigWig format from NIH Roadmap Epigenomics (Bernstein et al., 2010), Blueprint (Adams et al., 2012), and the International Human Epigenome Consortium (IHEC) (Stunnenberg et al., 2016). We selected the six most frequently studied histone marks: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3. These resources allow us to investigate the histone modification differences between tumor and normal tissues (**Supplementary Table 1**). For restricting the comparisons to

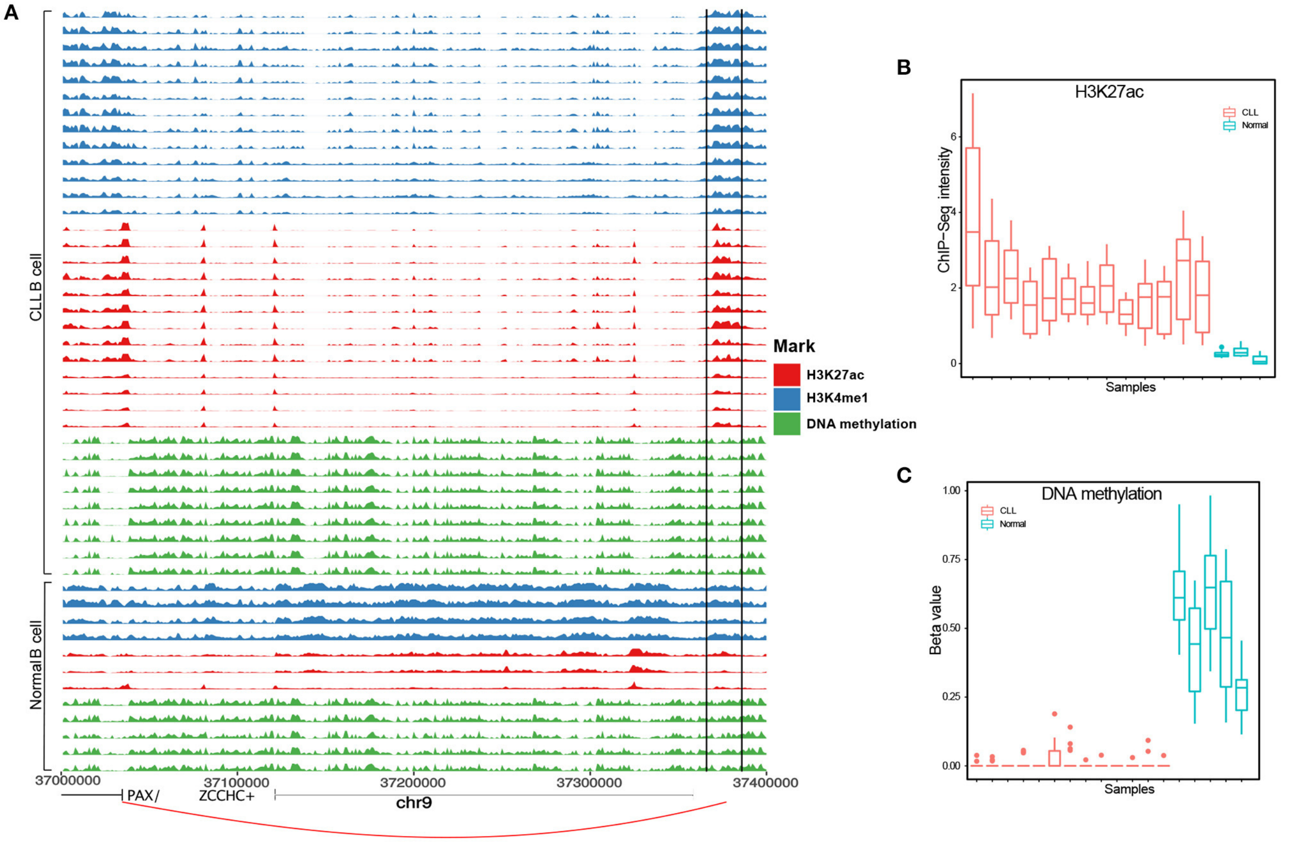

FIGURE 6 | A known PAX5 enhancer (chr9:37,370,916-37,371, and 635) in CLL exhibits hyperacetylation and hypomethylation [analysis of variance (ANOVA) *p*-value <0.05]. **(A)** PAX5 enhancer positions in each track are surrounded by two black solid lines. **(B)** H3K27ac and **(C)** DNA methylation levels between CLL and healthy samples.

the genomic loci of interests (promoters and enhancers), we downloaded the GRCh37 and GRCh38 coordinates of promoters from the eukaryotic promoter database (EPD) (Dreos et al., 2013), and the promoter interacting regions (PIRs) from the 4DGenome database (Teng et al., 2015). We treated the PIRs as potential enhancer regions, and filtered for tissue-specific enhancers by requiring the presence of H3K4me1 or H3K27ac peaks (peak calls provided in the **Supplementary Table 1**) in at least two samples from either tumor or normal tissues. By doing this, we enrich for cell type specific PIRs, which show a tissue-driven clustering (**Supplementary Figure 1**). The promoter coordinates were extended to ±1000 base pairs around the original coordinates. The sum of the numeric values from the BigWig blocks which overlap with the promoter and interacting regions are available from our project homepage. To build the relationships between and enhancers and promoters, we also download all the experimentally validated chromatin interaction datasets in various human tissues from 4DGenome.

Defining Disease and Control Datasets

We used histone modification datasets from seven cancer types in this study, i.e., B-ALL, CRC, glioma, MM, PTC, CLL, and mCLL from the Blueprint and IHEC consortia. For each cancer dataset, we paired it with the available dataset from the healthy tissue from which the cancer is most likely originated from. For example, the B-ALL, CLL, and MM were all compared against the healthy B cells in our design (see **Supplementary Table 1** for the pairs of normal/tumors used).

Definition of Cancer Marker Genes and Tissue-Specific Genes

We evaluated our algorithm on a small set of high-confidence CMGs, which is based on the tier-1 genes of the corresponding tissues from the Cancer Gene Consensus (CGC-t1) (Sondka et al., 2018) (**Supplementary Table 2**). As a negative control, we compiled a list of tissue-specific genes (TSGs) related to the tissues of interest for the tumor cases from ARCHS4_Tissues (https://maayanlab.cloud/archs4/). There are 2,318 genes for every tissue in the list. To validate our findings on independent, larger datasets of CMGs and TSGs, we compiled additional CMG lists containing 4,212 CMGs from 90 different cancer types from CancerMine (Lever et al., 2019), which incorporates the manual curated lists including the Cancer Gene Consensus (Sondka et al., 2018) and IntOGen (Gonzalez-Perez et al., 2013).

Data Processing Procedures

Combining Histone Marks

The epigenetic intensities on regulatory elements were summarized on a 1 kb scale, then power-transformed and quantile normalized. We use the dPCA (Ji et al., 2013) to decompose the matrix D representing the difference between M epigenetic datasets at G genomic loci comparing two groups of samples, into matrices B and V (1)

$$D_{G\times M} = B_{G\times R}V_{R\times M} + E \tag{1}$$

where E is the random sampling noise.

We use the first k dPCs to represent the major changes between two conditions. We implemented an R wrapper function for dPCA in our tool, which takes the mean differences of the normalized ChIP-Seq signals in each genomic locus between two biological conditions as input, and returns the dPCs from dPCA. The definition of dPCs varies between the test cases (**Figure 2A**). The largest variances of the positive and negative histone mark components are captured by dPC1 and dPC2 in our test case studies (**Figure 2B**). Therefore, we selected the sum of the absolute values of the first two dPCs for representing the overall differences of these epigenetic marks.

Promoter–Enhancer Interaction Analyses

In our approach, the enhancer–promoter relationships are described as a weighted bipartite graph, in which both enhancers and promoters are represented as vertices, and edges are directed from enhancers to their target promoters (**Figure 1** Step 1). The weights of the vertices are defined as the sum of the absolute values of the first two dPCs when combining multiple epigenetic marks, or the absolute value of the difference if a single epigenetic mark is considered. We adopt an algorithm called "PageRank," which is originally designed for evaluating the importance of web pages (Brin and Page, 1998), for ranking the magnitude of epigenetic alterations in each gene. We use the "personalized" PageRank implemented in igraph (Rye et al., 2011) to summarize the weights of one promoter and its connected enhancers into a unique meta-gene score (**Figure 1** Step 2). Since our enhancer–promoter network is a directed graph, all the enhancer weights will eventually be attributed to their target promoter using PageRank, yielding a unified score for each gene, which can be used to rank the genes. Overall, there are $\sim 251,000$ promoter interacting fragments in the promoter–enhancer interaction networks in our case studies, which is 8.5 times the number of promoters in the networks. The number of the interacting fragments targeting a gene varies from none to 227, and on average, 21 interacting fragments are targeting a promoter in the networks.

Scoring Ranked Lists

Using the gene ranks computed as described in the previous section, we can now evaluate the enrichment of a specific gene set $\mathcal{G}$ in the ranked list by computing the empirical cumulative distribution function (ECDF) obtained ranking the genes in decreasing order based on the previously described rank, and summing the indicator function

$$eCDF_{\mathcal{G}}(k) = \sum_{i=1}^{k} \delta_i \quad with \;\; \delta_i = \begin{cases} 1 \;\; if \; g_i \in \mathcal{G} \\ 0 \;\; if \; g_i \notin \mathcal{G} \end{cases} \tag{2}$$

We use the area under the curve (AUC) as a measure of the enrichment of the gene set $\mathcal{G}$, with AUC = 0.5 corresponding to a random distribution of the genes in $\mathcal{G}$ inside the ranked list.

Comparison With ChromHMM

We applied the ChromHMM method (version v1.22) to the Glioma and the healthy brain control samples (see **Supplementary Table 1**). The 6 histone marks were integrated into 10 chromatin states, of which 2 correspond to active promoter regions and one to active enhancer regions (**Supplementary Figure 2B**). The chromswitch package (Jessa and Kleinman, 2018) (v. 1.12.0) from Bioconductor was applied to the promoter and PIR regions linked to promoters for specific chromatin states. The chromswitch method determines a consensus score between changes occurring in chromatin state within a group of sample, and the labels of these samples. Hence, a maximal consensus score for a region of interest would correspond to changes in a chromatin state within the region of interest occurring only in the samples of one of the two groups. A minimal consensus score would on the opposite correspond to changes in chromatin states in the region of interest occurring in samples, which are randomly distributed over the two groups. For each gene, we compute a score by averaging the consensus score of all regulatory elements related to this gene, and use this score to rank the genes, as a comparison to the IRENE ranking.

DATA AVAILABILITY STATEMENT

The R package is available at https://github.com/hdsu-bioquant/irene. The datasets generated for this study can be found in the https://github.com/hdsu-bioquant/ irene-data. We also designed a web interface that allows users to trace back the epigenetic alterations of every enhancer and promoter, as well as every sample which is used for computing the score. We use Rmarkdown to generate static HTML pages and created a web site for presenting the results from our test case studies, which can also be found under the project home page. Users may also take advantage of this function to create a report that highlights a few genes of their interests and share the studies with the audience.

AUTHOR CONTRIBUTIONS

CH designed and supervised this project. QW and CH drafted the manuscript. QW wrote and tested the software. YW and TV participated in software testing. YW, TV, and RE revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the BMBF-funded PRECISe project (#031L0076A).

ACKNOWLEDGMENTS

The manuscript is deposited in the bioRxiv preprint server as Wang et al. (2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.664654/full#supplementary-material

REFERENCES

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechol.* 30, 224–226. doi: 10.1038/nbt.2153

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048. doi: 10.1038/nbt10 10-1045

Brin, S., and Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 107–117. doi: 10.1016/S0169-7552(98)00110-X

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: What, How, and Why? *Mol. Cell* 49, 825–837. doi: 10.1016/j.molcel.2013.01.038

Charlet, J., Duymich, C. E., Lay, F. D., Mundbjerg, K., Dalsgaard Sørensen, K., Liang, G., et al. (2016). Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol. Cell* 62, 422–431.doi: 10.1016/j.molcel.2016.03.033

Chen, L., Wang, C., Qin, Z. S., and Wu, H. (2015). A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics* 31, 1889–1896. doi: 10.1093/bioinformatics/btv094

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proce. Natl. Acad. Sci. U.S.A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

Dreos, R., Ambrosini, G., Périer, R. C., and Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* D157–D164. doi: 10.1093/nar/gks1233

Eeckhoute, J., Carroll, J. S., Geistlinger, T. R., Torres-Arzayus, M. I., and Brown, M. (2006). A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes Dev.* 20, 2513–2526. doi: 10.1101/gad.1446006

Ernst, J., and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., et al. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* 10, 1081–1084. doi: 10.1038/nmeth.2642

Hartley, I., Elkhoury, F. F., Heon Shin, J., Xie, B., Gu, X., Gao, Y., et al. (2013). Long-lasting changes in DNA methylation following short-term hypoxic exposure in primary hippocampal neuronal cultures. *PLoS ONE* 8:e77859. doi: 10.1371/journal.pone.0077859

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318. doi: 10.1038/ng1966

Jameson, N. M., Ma, J., Benitez, J., Izurieta, A., Han, J. Y., Mendez, R., et al. (2019). Intron 1-mediated regulation of EGFR expression in EGFR-dependent malignancies is mediated by AP-1 and BET proteins. *Mol. Cancer Res.* 17, 2208–2220. doi: 10.1158/1541-7786.MCR-19-0747

Jessa, S., and Kleinman, C. L. (2018). Chromswitch: a flexible method to detect chromatin state switches. *Bioinformatics* 34, 2286–2288. doi: 10.1093/bioinformatics/bty075

Ji, H., Li, X., Wang, Q.-f., and Ning, Y. (2013). Differential principal component analysis of ChIP-seq. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6789–6794. doi: 10.1073/pnas.1204398110

Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2926–2931. doi: 10.1073/pnas.0909344107

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., et al. (2018). Massive mining of publicly available rna-seq data from human and mouse. *Nat. Commun.* 9, 1–10. doi: 10.1038/s41467-018-03751-6

Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R., and Jones, S. J. M. (2019). CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* 16, 505–507. doi: 10.1038/s41592-019-0422-y

Liang, K., and Keles, S. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* 28, 121–122. doi: 10.1093/bioinformatics/btr605

Liu, F., Hon, G. C., Villa, G. R., Turner, K. M., Ikegami, S., Yang, H., Ye, Z., et al. (2015). EGFR mutation promotes glioblastoma through epigenome and transcription factor network remodeling. *Mol. Cell* 60, 307–18. doi: 10.1016/j.molcel.2015.09.002

Lun, A. T. L., and Smyth, G. K. (2015). Csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.* 44, 1–10. doi: 10.1093/nar/gkv1191

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Ann. Rev. Genomics Hum. Genet.* 7, 29–59. doi: 10.1146/annurev.genom.7.080505.115623

McInerney, J. M., Wilson, M. A., Strand, K. J., and Chrysogelos, S. A. (2000). A strong intronic enhancer element of the EGFR gene is preferentially active in high EGFR expressing breast cancer cells. *J. Cell. Biochem.* 80, 538–549. doi: 10.1002/1097-4644(20010315)80:4<538::AID-JCB1008>3.0.CO;2-2

Moore, J. E., Pratt, H. E., Purcaro, M. J., and Weng, Z. (2020). A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* 21, 17. doi: 10.1186/s13059-019-1924-8

Nair, N. U., Das Sahu, A., Bucher, P., and Moret, B. M. (2012). Chipnorm: a statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. *PLoS ONE* 7:e39573. doi: 10.1371/journal.pone.0039573

Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524. doi: 10.1038/nature1466

Rye, M. B., Sætrom, P., and Drabløs, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res.* 39:e25. doi: 10.1093/nar/gkq1187

Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H., and Waxman, D. J. (2012). MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* 13:R16. doi: 10.1186/gb-2012-13-3-r16

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696-705. doi: 10.1038/s41568-018-0060-1

Song, Q., and Smith, A. D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* 27, 870–871. doi: 10.1093/bioinformatics/btr030

Stark, R., and Brown, G. (2011). *DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data.* Bioconductor. Available online at: http://bioconductor.org/packages/release/bioc/html/DiffBind.html

Stasevich, T. J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., et al. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* 516, 272–275. doi: 10.1038/nature13714

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief. Bioinforma.* 17, 953–966. doi: 10.1093/bib/bbv110

Stunnenberg, H. G., Consortium, T. I. H. E., Hirst, M., International Human Epigenome Consortium, and Hirst, M. (2016). The International Human Epigenome Consortium: a Blueprint for Scientific Collaboration and Discovery. *Cell* 167, 1145–1149. doi: 10.1016/j.cell.2016.11.007

Teng, L., He, B., Wang, J., and Tan, K. (2015). 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics* 31, 2560–2564. doi: 10.1093/bioinformatics/btv158

Wang, Q., Wu, Y., Vorberg, T., Eils, R., and Herrmann, C. (2020). Integrative ranking of enhancer networks facilitates the discovery of epigenetic markers in cancer. *bioRxiv*. doi: 10.1101/2020.11.25.397844

Xu, H., Wei, C. L., Lin, F., and Sung, W. K. (2008). An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 24, 2344–2349. doi: 10.1093/bioinformatics/btn402

Zeng, X., Sanalkumar, R., Bresnick, E. H., Li, H., Chang, Q., and Keleş, S. (2013). jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.* 14:R38. doi: 10.1186/gb-2013-14-4-r38

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* 21, 1273–1283. doi: 10.1101/gr.122382.111

Zhang, J. D., and Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* 25, 1470–1471. doi: 10.1093/bioinformatics/btp167

Zhang, Y., An, L., Yue, F., and Hardison, R. C. (2016). Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* 44, 6721–6731. doi: 10.1093/nar/gkw278

Zhu, Y., van Essen, D., and Saccani, S. (2012). Cell-type-specific control of enhancer activity by H3K9 trimethylation. *Mol. Cell* 46, 408–423. doi: 10.1016/j.molcel.2012.05.011

Ziller, M. J., Edri, R., Yaffe, Y., Donaghey, J., Pop, R., Mallard, W., et al. (2014). Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* 518, 355–359. doi: 10.1038/nature13990

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.