OPENANNO: annotating genomic regions with chromatin accessibility

Shengquan Chen[1,a], Yong Wang[2,3,*,b], Rui Jiang[1,*,c]

[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic and Systems Biology; Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China

[2] CEMS, NCMIS, MDIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.

[*] Corresponding authors.

Abstract

Chromatin accessibility, as a powerful marker of active DNA regulatory elements, provides rich information to understand the regulatory mechanism. The revolution in high-throughput methods has accumulated massive chromatin accessibility profiles in public repositories as a valuable resource for machine learning and integrative studies. Nevertheless, utilization of these data is often hampered by the cumbersome and time-consuming collection, processing, and annotation of the chromatin accessibility information. Motivated by the above understanding, we developed a web server, named OPENANNO, to **anno**tate the **open**ness of genomic regions across diverse cell lines, tissues, and systems. The annotation is based on 871 DNase-seq experiments across 199 cell lines, 48 tissues, and 11 systems from ENCODE, and openness values rigorously defined by four statistical strategies. Particularly, we designed a parallel program to allow efficient annotation and visualization of the openness of a vast amount of genomic regions. OPENANNO will help users extract and download formulated data in a batch follow-up analysis. Besides, we illustrate the valuable information provided by OPENANNO using an enhancer of blood vessels from VISTA Enhancer Browser as an example. Furthermore, we demonstrate three applications of OPENANNO in regulatory mechanism and association studies. We believe that OPENANNO will serve as a comprehensive and user-friendly web server to facilitate methodology development and biological insights discovery, specifically to explore the biological questions and model the regulatory landscape of genome. OPENANNO is freely available at http://health.tsinghua.edu.cn/openness/anno/.

KEYWORDS: Chromatin accessibility; Openness; Annotation; Visualization; Web server

Introduction

The era of the personal genome is arriving with the widespread sequencing technologies and the ultimate promise for precision medicine. However, it remains distant in interpreting the context of variations in non-coding DNA sequence associated with disease and other phenotypes, deciphering their biological functions in gene regulation, and further understanding disease mechanism and dynamic response to treatment [1, 2]. Chromatin accessibility is a measure of the ability of nuclear macromolecules to physically contact DNA [3], and plays the role of a powerful marker of active regulatory genomic regions, which have a wide range of effects on the transcription, DNA repair, recombination, and replication [4, 5]. The recent revolution in high-throughput, genome-wide methods invented several biological assays for extracting open chromatin, such as DNase-seq (deoxyribonuclease), FAIRE-seq (formaldehyde-assisted isolation of regulatory elements), ATAC-seq (assay for transposase-accessible chromatin) and MNase-seq (micrococcal nuclease), and thus open a new door for us to make extensive use of chromatin accessibility [6-11]. For example, accessible genomic regions are regarded as the primary positions of regulatory elements [12], and thus provide a great opportunity to study transcription factor binding sites, DNA methylation sites, histone modification markers, gene regulation, and regulatory network [13, 14]. In addition, changes in chromatin accessibility have been implicated with different perspectives of human health as a result of the alterations of nucleosome positioning affected by mutations in chromatin remodelers [15-17].

The development of high-throughput sequencing techniques has accumulated a vast amount of chromatin profiles across a variety of cell lines. Large collaborative projects, such as Encyclopedia of DNA Elements (ENCODE) [18], have become a part of the major effort. The Roadmap Epigenomics project provides another similar resource for human stem cells and tissues [19]. Nevertheless, many experimental biologists may lack the bioinformatics expertise to make full use of these valuable resources efficiently. Cistrome DB, a data portal for ChIP-Seq and chromatin accessibility data, although comprises species, factors, biological source, publication, and other information for their collected ChIP-seq and DNase-seq data [20], limited to containing only a part of currently available transcription factors and histone marks. Therefore, it is still very cumbersome and time-consuming to collect, process, and incorporate the chromatin accessibility information of arbitrary genomic regions into bioinformatics and

epigenetics studies, which thus makes it difficult to full use of the vast amount of chromatin profiles.

We noticed that the epigenome consists of signals from chemical modifications of histones, DNA methylation, non-coding RNA expression, and transcription factors that work in concert to determine the accessibility of the regulatory regions, so-called open region. Then the open regulatory regions can work together with transcription factors, RNA polymerases, and other cellular regulatory machines and produce the final gene expression pattern. In this sense, the chromatin 'openness', *i.e.*, the accessibility of genomic regions, bridges the epigenome and transcriptome and plays an important role in understanding the regulatory mechanism. Motivated by the above demand, we built a web server, named OPENANNO, to **anno**tate the **open**ness of genomic regions across diverse types of cell lines, tissues, and systems. We downloaded the raw sequencing data of 871 DNase-seq experiments across 199 cell lines, 48 tissues and 11 systems from ENCODE data portal [18] and processed by a uniform pipeline. We defined the openness of genomic regions by four statistical strategies, including foreground read count, raw read openness, narrow peak openness, and broad peak openness. Furthermore, we designed a parallel program to enable OPENANNO to efficiently annotate and visualize the openness of a vast amount of genomic regions. We finally demonstrate three applications of OPENANNO in regulatory mechanism and association studies. We believe that this web server will help both the computational and experimental community to facilitate developing methods and discovering important insights, explore the basic biology and various applications, and open a new door to model the regulatory landscape of genome.

Web server content and usage

Overall design of OPENANNO

As illustrated in **Figure 1**, the diagram for constructing the OPENANNO web server consists of three main parts, *i.e.*, the source, host, and web interface. In the source part, we deposited the meta-information, raw data, and uniformly processed data of 871 human DNase-seq experiments from the ENCODE project [18]. It also includes the datasets of regulatory elements that have been experimentally validated from FANTOM[21] and VISTA[22]. The second part, calculation node in the host, monitors the file path that contains annotation tasks, and will automatically calculate and store

the results in a specific file path once the server node creates a new task. The server node bridges the host and web interfaces by achieving tasks from users and extracting results or other information for visualization and downloading. The third part, web interface, was developed in a concise and easy-to-use mode. The 'Annotate' page provides a service to annotate the openness of a vast amount of genomic regions, and is built on a specially designed infrastructure to extract and visualize big table. The 'Browse' page enables users to study the openness of a particular genomic region more intuitively. The 'Statistics' page provides detailed information for all the 871 DNase-seq experiments, and an intuitive comparison of the number of experiments in different cell lines, tissues or biological systems. The 'Download' page endows users with the ability to directly download the openness of collected experimentally validated regulatory elements. The 'Help' page provides other commonly used information to improve the usability of the web server.

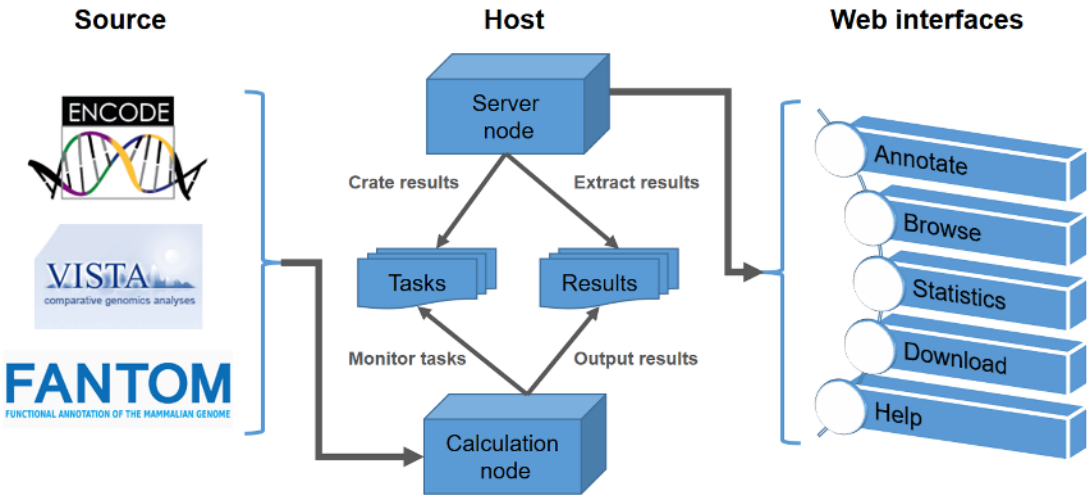

Figure 1 The diagram for constructing the OPENANNO web server

Web interfaces for annotating

The 'Annotate' page, as the major site of OPENANNO, can annotate the openness of genomic regions in batches. As illustrated in **Figure 2**, there are five major steps for this workflow. First of all, we provide a concise task submission approach, which avoids the confusion caused by redundant information to users. By clicking the 'Browse' button, users can upload a bed or bed.gz file (uncompressed or compressed in gzip format, *e.g.*, ENCFF001WKF.bed.gz). The parallel program in our calculation node will extract the first three columns and the sixth column (the chromosomes, starting sites, terminating sites and strands, respectively) separated by tabs for calculating the

openness. Note that sorting input in advance using other toolkits, such as bedtools (https://bedtools.readthedocs.io), is preferred for speeding up the calculation. In the current web server release, we provide the service to annotate the openness of genomic regions in human reference genome GRCh37 (hg19). We will provide the option of other genomes or species in future releases. Users can choose to calculate the openness of these genomic regions in a particular cell line, or directly calculate the openness in all the 871 cell lines. After saving the data as local files, the users can compare the openness in different cell lines or perform advanced analysis such as the co-openness analysis, which will be demoed in the following section. Furthermore, users can enable the option of Per-base pair annotation to calculate the openness of each base-pair of the genomic regions for some bioinformatics analysis such as machine learning tasks which will be demoed in the following section.

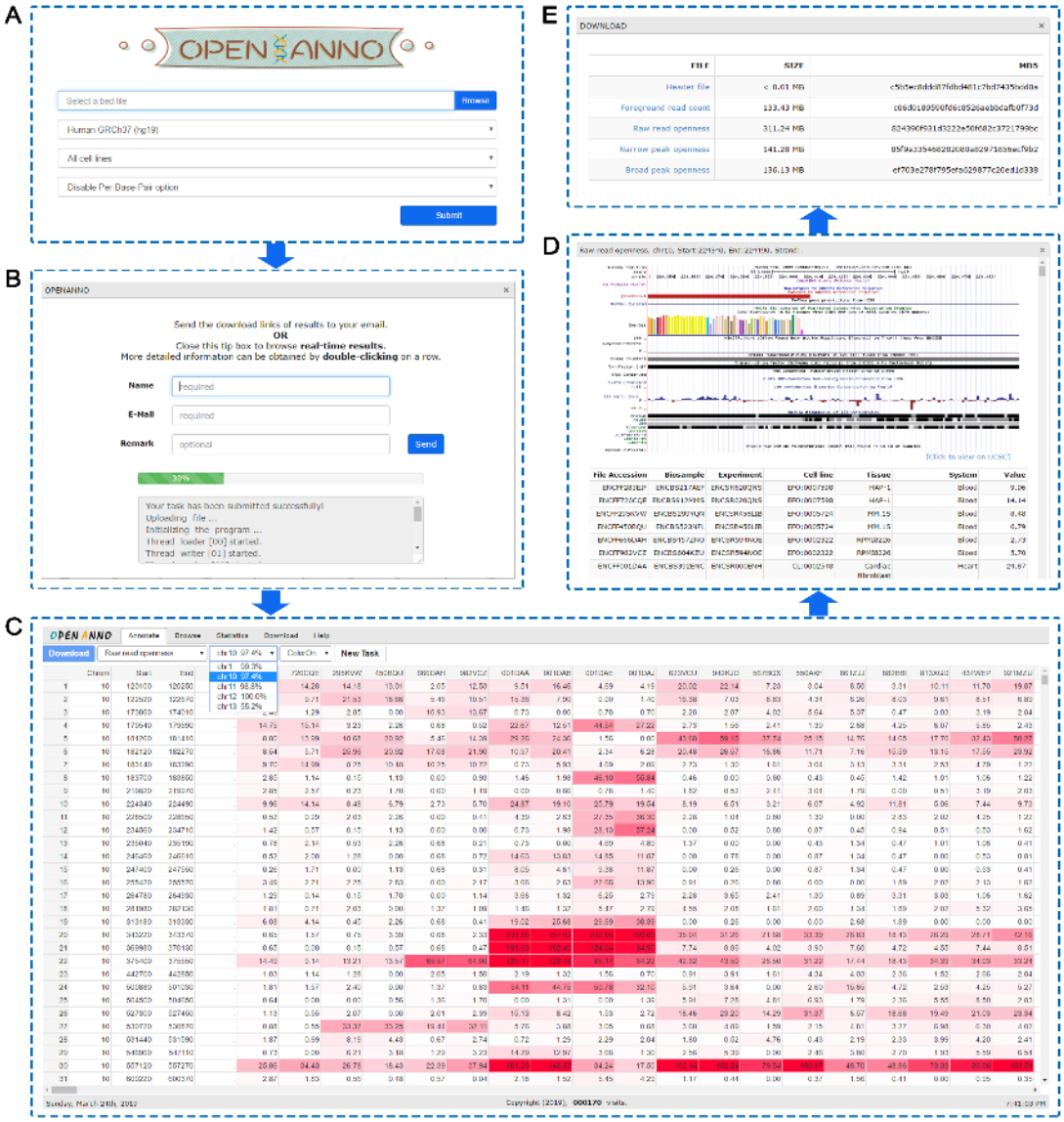

Figure 2 Web interfaces for annotating the openness of genomic regions in batches

A. The interface for submitting a new annotation task. **B.** Display of total calculation progress, and the interface for sending remarks and download links of results by email. **C.** Display of real-time annotation results. Users can browse part of a big table and scroll to any row and any column smoothly. **D.** More detailed information about the experiments and the visualization in UCSC Genome Browser of a specific genomic region. **E.** The interface for downloading the result plain text dump files.

After submitting the annotation task, users can follow the links of results sent to their email, and download the results after the calculation is completed. Furthermore, users can directly browse real-time results of each region. Here we demonstrate a seamless integration of hardware and software to hold the big result tables which may contain billions of rows and hundreds of columns. Users can observe part of a big table and scroll to any row and any column smoothly. The computational status in different chromosomes will be updated in real time. Users can arbitrarily switch between different chromosomes or the four different types of openness values. By enabling the Color option, each sample in the table will be colored according to the score of openness for intuitively comparing the openness of different genomic regions in different experiments. More detailed information about the experiments and the visualization of a specific genomic region in UCSC Genome Browser [23] can be obtained by double-clicking on a row. After the calculation is completed, users can extract the result files through the 'Download' button. By clicking the 'New Task' button, users can submit a new task on a new tab in the browser.

Web interfaces for browse

On the 'Browse' page, users can study the openness of a particular genomic region more intuitively. As illustrated in **Figure 3**, we take an enhancer (chr10: 94,513,996-94,517,989) of the tissue of blood vessels provided by the VISTA Enhancer Browser [22] as an example to demonstrate the service in this page. After submitting the form that contains the chromosome, starting site, terminating site, and strand of a particular genomic region, the web server provides information from three perspectives, including (1) the average openness scores of this genomic region in different biological systems,

tissues, and cell lines, (2) the visualization of this genomic region in UCSC Genome Browser [23], and (3) openness scores and details of the 871 DNase-seq experiments.

Figure 3 The interface for studying the openness of a particular genomic region more intuitively

First of all, the average openness scores enable users to compare the openness across different biological systems, tissues, and cell lines more intuitively. For example, as shown in **Figure 3**, the average openness score of this enhancer in the vessel tissue is obviously higher than that in other tissues. We calculated z-scores, standard deviations from the mean, the average openness scores of this enhancer in the vessel tissue, and found that the z-scores in all the 4 different types of openness definition

192 beyond the three-sigma limits (3.18 of foreground read count, 4.07 of raw read
193 openness, 3.29 of narrow peak openness and 4.37 of broad peak openness), which
194 demonstrates that all the 4 different types of openness of the enhancer in the vessel
195 tissue are significantly higher than that in other tissues from the statistical perspective.
196 This coincides with the fact that this genomic region is an enhancer in blood vessels
197 according to VISTA annotation. Second, the visualization in UCSC Genome Browser
198 provides rapid and reliable displaying of the requested genomic region at any scale,
199 together with dozens of aligned annotation tracks (http://genome.ucsc.edu/) [23]. We
200 also provide a hyperlink to UCSC Genome Browser to facilitate users to achieve more
201 concrete information of the particular genomic region. Finally, openness scores and
202 detailed information of the 871 DNase-seq experiments are filled in a table with
203 advanced features. Users can sort the table according to the openness to find out in
204 which experiment the genomic region has higher openness. Users can also sort the table
205 on the basis of different columns according to their own needs. We also provide a
206 convenient service for searching in the table. Users can quickly query information they
207 are interested in. Furthermore, users can directly copy the table or download tables
208 stored in different formats for other requirements.
209
210 **Web interfaces for statistics, downloading, and help**
211 On the 'Statistics' page, as shown in **Figure 4**, users can intuitively compare the number
212 of experiments across different cell lines, tissues, or biological systems. A table with
213 advanced features also provides detailed information, including file accessions,
214 biosamples, experiments, cell lines, tissues, and systems, of all the 871 DNase-seq
215 experiments. We collected regulatory element datasets from FANTOM[21] and
216 VISTA[22], including 184,476 FANTOM5 human promoters, 32,693 FANTOM5
217 human enhancers, and 979 VISTA human enhancers. We calculated the openness of
218 these regulatory elements in advance, and provide a download service on the
219 'Download' page to allow users to directly download plain text dump files in
220 compressed (gzip) format. We will continue to provide the openness of other public
221 and validated regulatory elements. To improve the usability, we provide a 'Help' page
222 with other commonly used information of the web server, including frequently asked
223 questions, news about the releases of OPENANNO, tutorials of each web interface,
224 citation information, and contact information for help and feedback.

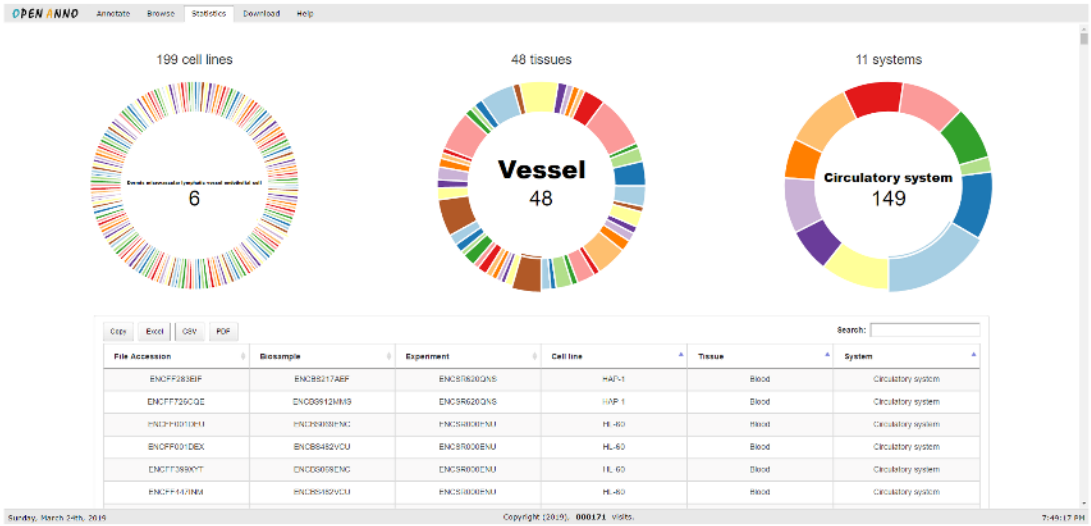

Figure 4 The interface for intuitively comparing the number of experiments across different cell lines, tissues, or biological systems, and achieving details of all the 871 DNase-seq experiments

OPENANNO facilitates regulatory mechanism studies

The chromatin 'openness', *i.e.*, the accessibility of genomic regions, calculated using our method has been widely applied to various studies of regulatory mechanism. Here, we show two examples to demonstrate the output of OPENANNO contains valuable information. For example, a model named DeepTACT has been proposed to integrate DNA sequences and chromatin accessibility data for the prediction of chromatin contacts between regulatory elements [24]. Briefly, DeepTACT first performs a one-hot encoding strategy and calculates the raw read openness of each site for characterizing a given genomic region. DeepTACT takes the sequences of two one-hot encoded regulatory elements, and their chromatin openness scores derived from OPENANNO of a given cell type as input. The output of DeepTACT is the predictive score that represents the probability the two regulatory elements have 3D contact. In the deep neural network of DeepTACT, a sequence module is used to extract features from DNA sequences, an openness module is adopted to learn epigenomic features from chromatin openness scores, and an integration module merges outputs of these two modules and extracts high-level features with an attention-based recurrent neural network to predict the probability that the two regulatory elements have 3D contact.

Using sequence features and the raw read openness of genomic regions, DeepTACT, as a bootstrapping deep learning model, outperforms existing methods on the task of

249 inferring both promoter-enhancer and promoter-promoter interactions. In more detail,
250 with same test sets, DeepTACT achieves a mean auPRC (the area under the precision-
251 recall curve) score of 0.89 for inferring promoter-promoter interactions compared with
252 0.76 of SPEID [25] and 0.23 of Rambutan [26]. For inferring promoter-enhancer
253 interactions, DeepTACT achieves a mean auPRC of 0.82 compared with 0.67 of SPEID
254 and 0.36 of Rambutan.

255 Besides, DeepTACT provides a finer mapping of promoter-enhancer and promoter-
256 promoter interactions from high-quality promoter capture Hi-C data. Furthermore, the
257 class of hub promoters identified by DeepTACT, and the integrative analysis of existing
258 GWAS data and chromatin contacts predicted by DeepTACT demonstrate the openness
259 calculated by OPENANNO bridges the epigenome and transcriptome and plays an
260 important role in understanding the regulatory mechanism.

261 In addition to DeepTACT, a model named DeepCAPE is proposed to predict
262 enhancers via the integration of DNA sequences and DNase-seq data [27] with the
263 understanding that DNase I hypersensitivity has been shown to be important to identify
264 active cis-regulatory elements including enhancers, promoters, silencers, insulators,
265 and locus control regions [28]. Briefly, DeepCAPE uses the raw read openness of each
266 site of a genomic region as the information of chromatin accessibility to greatly improve
267 the performance of predicting enhancers. In more detail, when the ratio of positive and
268 negative samples is 1:20, the auROC (the area under the receiver operating
269 characteristic curve) and auPRC scores of DeepCAPE are on average 0.151 and 0.590
270 higher than gkmSVM [29], 0.151 and 0.598 higher than DeepSEA [30], and 0.150 and
271 0.588 higher than DeepEnhancer [31], respectively. One-sided paired-sample
272 Wilcoxon signed rank tests consistently suggest that DeepCAPE consistently achieves
273 higher auPRC scores (p-values < 2.2e-16 for all the other three baseline methods), and
274 higher auROC scores than a baseline method (p-values < 2.2e-16 for all the other three
275 baseline methods).

276 In the model ablation analysis for evaluating the contributions of DNA sequences
277 and DNase-seq data, DeepCAPE illustrates that DNase-seq data provides more
278 information than DNA sequences to greatly improve the performance of prediction.
279 Besides, the information provided by DNA sequences also plays an important role in
280 promoting the performance and making the performance more stable. Because the
281 number of DNase-seq experiments varies between cell lines, the dimensionality of input
282 data varies between cell lines and prevents the use of convolutional neural networks in

283 the cross cell line prediction. DeepCAPE therefore adopts a neural network designed
284 for unsupervised learning of efficient encodings [32], named auto-encoder, to embed
285 chromatin openness scores of a DNA fragment derived from OPENANNO into a vector
286 of fixed length in a low-dimensional latent space. Comparing to the model without an
287 auto-encoder, and other two strategies that average the replicates or randomly select a
288 single replicate, DeepCAPE with an auto-encoder not only makes cross cell line
289 prediction possible, but also maintains superior performance even if the dimensionality
290 of the openness data is reduced. In addition, with a collective scoring strategy,
291 DeepCAPE achieves an average auROC of 0.971 and an average auPRC of 0.862 in
292 the cross cell-line prediction when the ratio of positive and negative samples is 1:20,
293 and thus establishes a landscape of potential enhancers specific to a cell line that still
294 lacks systematic exploration of enhancers.
295 To sum up, DeepCAPE not only achieves superior prediction performance in a cell
296 line-specific manner, but also makes accurate cross cell line predictions possible with
297 the openness scores calculated by OPENANNO. With this understanding, analogous
298 machine learning frameworks can possibly be adapted for the prediction of other
299 functional elements in the genome, including but not limited to promoters, silencers,
300 insulators, repressors, and locus control regions. In addition, the strategy that integrates
301 DNA sequences and chromatin openness can also be generalized for the prioritization
302 of candidate variants in whole-genome sequencing studies, and thus facilitate the
303 regulatory mechanism studies.
304
305 **OPENANNO facilitates association studies**
306 Network-based functional studies play an important role in the identification of disease-
307 associated genes and the interpretation of disease mechanism. The functional
308 relationship of a pair of genes is influenced not only by the co-activation of transcripts,
309 but the regulation mechanism [33]. Besides, the consistence of gene chromatin
310 accessibility indicates the tendency of genes being co-regulated. With this
311 understanding, a gene co-opening network has been constructed based on the raw read
312 openness of genes [34]. Briefly, they take alternative promoters of genes into account
313 when calculating the correlation (absolute value of *Pearson*'s correlation coefficient)
314 of the openness scores between two genes, considering the prevalence of the alternative
315 splicing phenomena. By calculating a co-opening score for every pair of genes, they
316 obtain a co-opening matrix for all genes to facilitate the downstream analysis.

The results demonstrate that the co-opening network contains new information different from co-expression networks and protein-protein interactions networks. In addition, the genes related to a specific biological process or a specific disease has been demonstrated to tend to be clustered together in the co-opening network, which facilitates detecting functional clusters in the network and predicting new functions for genes. Particularly, through integrative analysis with fruitful genome-wide association studies (GWAS) data, the co-opening network provides a new perspective to the discovery of genes associated with complex diseases, and thus benefits elucidating gene associations and the deciphering of disease mechanisms. For example, by simulating a random walk process on the co-opening network, they use the steady state probability assigned to a gene as a score to measure the likelihood that the gene is associated with the disease under investigation. Applying this strategy to a complex disease named *Psoriasis*, a potentially disfiguring immune-mediated inflammatory disease of skin, they discovered *TNFSF14* (TNF super family member 14, a biomarker of *Psoriasis* [35]), which was ranked second by the random walk model while cannot be identified by GWAS (p-value = 0.1616, ranked 1259 based on the p-value). In general, the co-opening network is ready to serve as a useful resource complementary to the widely used co-expression network, and thus shed light on the studies in system biology.

Perspectives and concluding remarks

Chromatin accessibility, which bridges the epigenome and transcriptome, is a very valuable resource for interpreting non-coding genomic region and understanding the regulatory mechanism. In this study, we downloaded raw sequencing data of 871 DNase-seq experiments across 199 cell lines, 48 tissues and 11 biological systems from ENCODE, and defined the openness of genomic regions from four perspectives. In addition, we take an enhancer of the tissue of blood vessels provided by the VISTA Enhancer Browser as an example to illustrate the valuable information provided by all the four different types of openness from statistical perspective. Furthermore, we designed a parallel program to endow OPENANNO with the ability to efficiently annotate and visualize openness for a vast amount of genomic regions. Finally, we introduced three examples to demonstrate the output of OPENANNO serves as valuable input for follow-up regulatory mechanism and association studies.

Our web server has four main application scenarios. First, one can use our web server to annotate openness of genomic regions, and then integrate the information of openness to a machine learning model for superior performance. Second, one can use our web server to visualize the openness of a specific genomic region to intuitively understand this region has higher openness in which cell lines, tissues, or systems, and thus contribute to the study of functional implications of this genomic region. Third, our web server offers a new opportunity to reinterpret abundant data cumulated by genome-wide association studies, and thus one can characterize variants by integrating upstream openness annotated with our web server and downstream gene expression. Finally, one can use the openness annotated with our web server to construct gene co-opening networks which provide a new perspective to association studies.

To better serve the academic community, we will continue to collect public data and update OPENANNO regularly in the future. Our next plan is to provide the option of other genomes or species, and the option of annotating using other chromatin accessibility data, such as ATAC-seq data. We will continue to provide the openness of other public and validated regulatory elements for downloading directly. According to users' feedbacks, we will continue to improve the interfaces and performance of OPENANNO. We believe that OPENANNO would serve as a useful tool for both bench scientists and computational biologists, and shed light on studies including but not limited to bioinformatics and system biology.

Materials and methods

Data collection

We first parsed a total of 41,418 JSON files from ENCODE to obtain detailed information about experiments, biosamples, cell lines, tissues, and systems of the DNase-seq data provided by the ENCODE project [18]. Under the constraint that each experiment contains both narrow peaks and broad peaks, we downloaded raw sequencing data of 891 DNase-seq experiments in human reference genome GRCh37 (hg19), and then identified experiments corresponding to 199 cell lines, 48 tissues and 11 biological systems. We collected datasets of regulatory elements from FANTOM [21] and VISTA[22] that have been experimentally validated, including 184,476 FANTOM5 human promoters, 32,693 FANTOM5 human enhancers, and 979 VISTA human enhancers.

382

383 **Definition of openness**

384 We defined the openness of given genomic regions from four perspectives, including
385 foreground read count, raw read openness, narrow peak openness, and broad peak
386 openness. Specifically, given the raw sequencing data of a DNase-seq experiment, we
387 provided the number of reads (*N*), *i.e.*, foreground read count, falling at a specific
388 genomic region to facilitate special applications that may use raw read counts of a
389 DNase-seq experiment directly. To remove the effect of sequencing depth, we defined
390 the raw read openness (*S*) of a genomic region as the average foreground read count
391 divided by the average number of reads falling into a background region of size *W*
392 surrounding the given region. The raw read openness (*S*) can be calculated as

$$S = \frac{N/L}{K/W}$$

393 where *L* denotes the size of the specific genomic region, and *K* the number of reads
394 falling into the background region of size *W*. The size of a background region *W* is set
395 to 1 M base pairs.

396 Analogously, we defined the narrow peak openness (and broad peak openness) of a
397 genomic region as the average number of narrow peaks (or broad peaks) overlapping
398 with the genomic region, divided by the average number of narrow peaks (or broad
399 peaks) overlapping with a position in a background region of size *W* surrounding the
400 given genomic region.

401

402 **Parallel computing and real-time browsing**

403 To facilitate querying reads and peaks at high frequencies, and the massive demand for
404 annotating openness of a large number of genomic regions, we designed a parallel
405 strategy that endows OPENANNO with an ability to efficiently annotate openness of a
406 vast amount of genomic regions. We used C++, a programming language known for its
407 high efficiency, to develop a multithreaded program that consists of a reading module,
408 a calculating module, and a writing module. As illustrated in **Figure 5**, the reading
409 module packages the query regions in the input file into data blocks and pushes them
410 into the input pipeline *I_pipe*. Each data block contains a portion of the query regions,
411 and header information that contains the number and indexes of these query regions in
412 the input file. When a thread in the calculating module is idle, the data block is
413 automatically extracted from the input pipeline for calculation. The results are pushed

into the output pipeline *O_pipe* in a similar form of data blocks, which are then popped out and written into the disks by the writing module. Through the communication pipeline *C_pipe*, the reading module can respond to the working state of the writing module. When the writing speed is lower than the reading speed, the reading module pauses pushing the data block into *I_pipe* to effectively save memory resources. With the parallel program, OPENANNO can calculate the four types of openness across all the 871 DNase-seq experiments of one thousand genomic regions within 1 second.

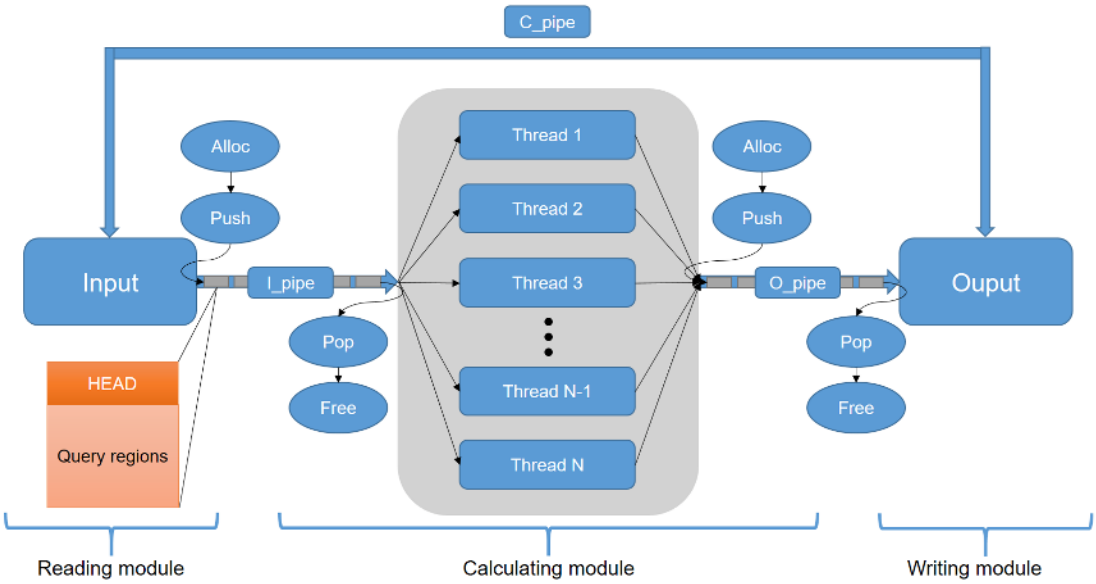

Figure 5 The multithreaded program for efficiently annotating openness of a vast amount of genomic regions

Users may need to annotate a large number of genomic regions, and thus results in very large tables of openness, especially when the per-base option is enabled. Loading all the results directly to the front-end of OPENANNO will take a long time and result in an unsmooth experience, and thus is an unrealistic choice. It is, therefore, necessary to provide a highly efficient front-end web application to browse these tables. To accomplish this, we used WebSocket (https://www.websocket.org/), a computer communications protocol, to provide full-duplex communication channels over a single socket over the web and the remote host, realize the real-time browsing of calculation results, and thus users can observe part of a large table and scroll to any row and any column smoothly.

436 **Web server implementation**

437 The whole design of the OPENANNO is shown in Figure 1, with the possible jumps
438 among web pages illustrated. OPENANNO is freely available to all users without a
439 login requirement. The current version of OPENANNO was deployed on a calculation
440 node and a server node of a high-performance computer cluster. The calculation node,
441 whose operating system is CentOS 7.5 (one of the most popular Linux distributions;
442 https://www.centos.org), has 56 hyper-threaded processors to perform efficient parallel
443 computing, RAM of 775 GB to support a large number of reading and writing
444 operations, and storage space of 321 TB to store the vast amounts of data. Incron
445 (http://inotify.aiken.cz/?section=incron) is used to monitor filesystem events and
446 executes predefined commands. Once the server node creates a new task in a specific
447 file path, the parallel program automatically runs and stores the results in another
448 specific file path for the website to read and users to download.

449 In the server node, a Linux-based Nginx server (https://www.nginx.com/), which
450 uses dramatically less memory than Apache (https://www.apache.org/) and can handle
451 roughly four times more requests per second, is deployed to improve the performance,
452 reliability, and security of OPENANNO. PHP (http://www.php.net/) is used for server-
453 side scripting. Bootstrap (a popular toolkit for developing with HTML, CSS and
454 JavaScript; https://getbootstrap.com/) and jQuery (a fast and feature-rich library
455 designed to simplify the JavaScript programming; https://jquery.com/) are adopted for
456 the front-end, *i.e.*, the interactive and responsive user interface, of OPENANNO. The
457 user interface of OPENANNO can automatically respond to the devices and browsers
458 with different screen resolutions, and change its structure and shape according to the
459 resolution in order to optimize the visualization. DataTables (https://datatables.net/), as
460 a plug-in for the jQuery and JavaScript library, is used to add advanced features to the
461 tables. The visualizations of bar charts and pie charts are implemented with JavaScript
462 libraries named CanvasJS (https://canvasjs.com/) and morris.js
463 (https://morrisjs.github.io/morris.js/index.html), respectively. To obtain the
464 visualization in UCSC Genome Browser of a query region, we put the information of
465 chromosome, start position, end position and reference genome to the link
466 http://genome.ucsc.edu/cgi-
467 bin/hgTracks?db=(reference_genome)&position=(chromosome):(start_position)-
468 (end_position) to finish UCSC link construction.

All codes are developed using Vim (a highly flexible text editor that supports any kind of text; https://www.vim.org/). The performance of OPENANNO has been tested in Chrome, Firefox, Opera and Microsoft Edge on Windows 10, Ubuntu 16.04 and MacOS 10.12. We hope users could feedback their comments and suggestions through the contact page on our website to help us improve OPENANNO.

Authors' contributions

RJ and YW designed the project. SC and RJ collected data and implemented the web server. SC, YW and RJ wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China (Grant No. 2018YFC0910404), the National Natural Science Foundation of China (Grant Nos. 61873141, 61721003, 61573207, 11871463, and 61671444), and the Tsinghua-Fuzhou Institute for Data Technology. Rui Jiang is a RONG professor at the Institute for Data Science, Tsinghua University.

Reference

[1] Frazer KA. Decoding the human genome. Genome Res 2012;22(9):1599-601.

[2] Sosnay PR, Cutting GR. Interpretation of genetic variants. Thorax 2014;69(3):295-7.

[3] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019;20(4):207-20.

[4] Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin 2014;7(1):33.

[5] Radman-Livaja M, Rando OJ. Nucleosome positioning: how is it established, and why does it matter? Dev Biol 2010;339(2):258-66.

[6] Rizzo JM, Sinha S. Analyzing the global chromatin structure of keratinocytes by MNase-seq. Methods Mol Biol 2014;1195:49-59.

[7] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;10(12):1213-8.

[8] Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. Methods Mol Biol 2012;833:413-9.

[9] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17(6):877-85.

510 [10] John S, Sabo PJ, Canfield TK, Lee K, Vong S, Weaver M, et al. Genome-scale
511 mapping of DNase I hypersensitivity. Curr Protoc Mol Biol 2013;Chapter 27:Unit 21
512 7.
513 [11] Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of
514 regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc
515 2012;7(2):256-67.
516 [12] John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, et al. Chromatin
517 accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet
518 2011;43(3):264-8.
519 [13] Ward LD, Kellis M. Evidence of abundant purifying selection in humans for
520 recently acquired regulatory functions. Science 2012;337(6102):1675-8.
521 [14] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general
522 framework for estimating the relative pathogenicity of human genetic variants. Nat
523 Genet 2014;46(3):310-5.
524 [15] Gaspar-Maia A, Alajem A, Polesso F, Sridharan R, Mason MJ, Heidersbach A, et
525 al. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. Nature
526 2009;460(7257):863-8.
527 [16] Hargreaves DC, Crabtree GR. ATP-dependent chromatin remodeling: genetics,
528 genomics and mechanisms. Cell Res 2011;21(3):396-420.
529 [17] Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, Jacob K, et al.
530 Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric
531 glioblastoma. Nature 2012;482(7384):226-31.
532 [18] Consortium EP. An integrated encyclopedia of DNA elements in the human
533 genome. Nature 2012;489(7414):57-74.
534 [19] Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A,
535 Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat
536 Biotechnol 2010;28(10):1045-8.
537 [20] Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a
538 data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic
539 Acids Res 2017;45(D1):D658-D62.
540 [21] Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al.
541 Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol
542 2015;16:22.
543 [22] Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a
544 database of tissue-specific human enhancers. Nucleic Acids Res 2007;35(Database
545 issue):D88-92.
546 [23] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The
547 human genome browser at UCSC. Genome Res 2002;12(6):996-1006.
548 [24] Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via
549 bootstrapping deep learning. Nucleic Acids Res 2019.
550 [25] Singh S, Yang Y, Póczos B, Ma J. Predicting Enhancer-Promoter Interaction from
551 Genomic Sequence with Deep Neural Networks. bioRxiv 2018:085241.

[26] Schreiber J, Libbrecht M, Bilmes J, Noble W. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv 2018:103614.
[27] Chen S, Gan M, Lv H, Jiang R. DeepCAPE: a deep convolutional neural network for the accurate prediction of enhancers. bioRxiv 2018:398115.
[28] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature 2012;489(7414):75-82.
[29] Lee D. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics 2016;32(14):2196-8.
[30] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12(10):931-4.
[31] Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics 2017;18(Suppl 13):478.
[32] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504-7.
[33] Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. Biosystems 2009;96(1):86-103.
[34] Li W, Wang M, Sun J, Wang Y, Jiang R. Gene co-opening network deciphers gene functional relationships. Mol Biosyst 2017;13(11):2428-39.
[35] Chandran V, Cook RJ, Edwin J, Shen H, Pellett FJ, Shanmugarajah S, et al. Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. Rheumatology (Oxford) 2010;49(7):1399-405.

575 ## Figure legends

576 **Figure 1 The diagram for constructing the OPENANNO web server**

577

578 **Figure 2 Web interfaces for annotating the openness of genomic regions in**
579 **batches**

580 **A.** The interface for submitting a new annotation task. **B.** Display of total calculation
581 progress, and the interface for sending remarks and download links of results to email.
582 **C.** Display of real-time annotation results. Users can observe part of a big table and
583 scroll to any row and any column smoothly. **D.** More detailed information about the
584 experiments and the visualization in UCSC Genome Browser of a specific genomic
585 region. **E.** The interface for downloading the result plain text dump files.

586

587 **Figure 3 The interface for studying the openness of a particular genomic region**
588 **more intuitively**

589

590 **Figure 4 The interface for intuitively comparing the number of experiments**
591 **across different cell lines, tissues or systems, and achieving detail information of**
592 **all the 871 DNase-seq experiments**

593

594 **Figure 5 The multithreaded program for efficiently annotating openness of a**
595 **vast amount of Genomic regions**