Published in final edited form as:

Methods Mol Biol. 2021 ; 2190: 229–248. doi:10.1007/978-1-0716-0826-5_11.

Leverage large-scale biological networks to decipher the genetic basis of human diseases using machine learning

Hao Wang, Jiaxin Yang, Jianrong Wang

Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824

Abstract

A fundamental question in precision medicine is to quantitatively decode the genetic basis of complex human diseases, which will enable the development of predictive models of disease risks based on personal genome sequences. To account for the complex systems within different cellular contexts, large-scale regulatory networks are critical components to be integrated into the analysis. Based on the fast accumulation of multi-omics and disease genetics data, advanced machine learning algorithms and efficient computational tools are becoming the driving force in predicting phenotypes from genotypes, identifying potential causal genetic variants, and revealing disease mechanisms. Here, we review the state of the art methods for this topic, and describe a computational pipeline that assembles a series of algorithms together to achieve improved disease genetics prediction through the delineation of regulatory circuitry step by step.

Keywords

Regulatory networks; multi-omics; genetic variants; disease mechanisms; fine-mapping

1. Introduction

Being able to precisely predict disease risks based on personal genomic sequences is one of the central questions in biomedical and bioinformatics research. To address this problem, we need to estimate mapping functions ($f: \text{genetic variant} \rightarrow \text{disease status}$) that can link individual genetic variants to final disease outputs, and assign a probabilistic score to quantify the risk of observing a disease-related phenotype when specific genetic variants occur (Figure 1), i.e. $P(\text{phenotype} | \text{genetic variants})$.

Due to the complex nature of the human genome, the mapping functions for different genetic variants may be substantially different, depending on the biological roles of the genomic locations around specific genetic variants and how the genomic location interacts with other genomic loci in diverse biological processes. Therefore, accurate re-construction of large-scale regulatory networks and network-based genetics modeling are two major tasks for advanced prediction of disease risks and characterization of disease mechanisms.

1.1 Traditional genetics approaches to identify disease-associated genetic variants

Genome-wide association studies (GWAS) have been the standard approach to look for genetic variants that are statistically correlated with specific complex diseases or traits [1,2]. In a typical GWAS analysis for a specific disease (e.g. Alzheimer's Disease), genome-wide genetic variants are profiled for a group of cases (i.e. patients with Alzheimer's disease phenotypes) and a group of controls (people without Alzheimer's disease phenotypes). In this case-control setting, each genetic variant (mainly refers to single nucleotide polymorphisms: SNP) is evaluated in turn. The frequency of occurrence of each genetic variant in cases is compared to its frequency of occurrence in controls. If the frequency in cases is significantly higher than the frequency in controls, based on the p-value calculated by Chi-squared test, then the specific genetic variant is predicted to be correlated with the disease. After testing all the genetic variants profiled in the genome, a set of significantly disease-associated SNPs will finally be identified.

One of the goals of GWAS analysis is to lay a quantitative foundation to predict disease risks so that people can build models (e.g. logistic regression or random forest models) where disease-associated SNPs are used as features and disease phenotypes are treated as response variables. Given the genomic sequence of a new patient, the model is expected to predict the risk of observing disease phenotypes. But there are several fundamental limitations of the GWAS based approach, which have been widely recognized in the field [1–3]. The three major limitations are: 1) Low statistical power. Due to the huge number of genetic variants across the human genome (i.e. the number of hypotheses to test), many real disease-associated SNPs are evaluated to be not significant enough and are therefore missed in traditional GWAS (false negatives). 2) Unable to identify causal genetic variants. Due to the correlative nature of GWAS, the disease-associated SNPs may not be the actual causal variants that lead to disease phenotypes. This problem is further complicated by the fact that neighboring SNPs are usually correlated with each other across human populations (i.e. linkage disequilibrium). Among hundreds of SNPs located close to each other, it is computationally challenging to pinpoint which SNP is the causal one. 3) Lack of mechanistic understanding of disease. The GWAS approaches basically treat the underlying disease mechanisms as black box and are not able to determine why certain SNPs are correlated with disease phenotypes. But it is highly useful for new therapeutic development to know how the disruptive effects of specific SNPs at DNA sequence propagate to the organism level phenotypes. For example, since the majority (>97%) of the human genome are not protein-coding gene exons, i.e. are non-coding regions, significant SNPs from GWAS are usually not located in exons. Knowing the molecular mechanisms of these non-coding SNPs, such as which transcription factors bind there and which genes are regulated by the SNPs, can substantially improve our capacity to understand and prioritize disease-associated non-coding genetic variants. In the meantime, knowledge of disease mechanisms can also help to address the first two problems (low statistical power and identification of causal variants) by enabling sophisticated generative probabilistic models with prior knowledge flexibly incorporated, instead of black-box discriminative models.

1.2 Whole-genome annotation of regulatory elements

In order to overcome these limitations, the genome sequence itself, along with the basic gene annotations, does not provide enough information. Additional functional genomics data are needed to provide a better annotation of the human genome, such as genomic locations of specific transcription factor binding sites, enhancers, insulators, chromatin accessible sites, and DNA methylated sites [4–9]. These diverse functional sites are collectively defined as regulatory elements, which can regulate tissue-specific gene expression at transcriptional level. Although not directly producing proteins, they are critical genomic units that can control the levels of specific proteins in different cellular contexts (e.g. tissues or cell-types). Therefore, the functional annotation of regulatory elements is highly useful for people to understand the non-coding portion of the human genome (>97%). Among different families of regulatory elements, enhancers are the largest family and are characterized with complex features [10–13], such as combinatorial transcription factor (TF) binding within enhancers, multiple associated epigenetic signatures (e.g. histone modifications) and high levels of chromatin accessibility (Figure 1).

Intuitively, assuming the genomic locations of tissue-specific enhancers are known along the whole human genome and across different tissues, a non-coding genetic variant located within a brain-specific enhancer may disrupt a specific gene's expression in the brain and may have higher prior probability to be a causal variant in Alzheimer's disease, compared to genetic variants located in genomic regions that are neither genes nor enhancers. Similarly, a genetic variant located within a T cell specific enhancer may be more likely to be a causal variant for autoimmune diseases.

Therefore, as the first step in going beyond traditional GWAS analysis, algorithms to predict the genomic locations of regulatory elements (especially enhancers) in non-coding regions are urgently needed. Fortunately, next-generation sequencing-based techniques have produced large amounts of genome-wide omics data, such as ChIP-seq data of TF bindings and histone modifications, DNase-seq or ATAC-seq for chromatin accessibility, and bisulfite sequencing for DNA methylation, that can be used as features to predict enhancers in different tissues or cell-types [11,12,14].

1.3 Regulatory networks are needed to reveal the mechanisms

Enhancers are functionally important, because they are genomic hubs of combinatorial TF bindings and are responsible for precise regulation of specific target gene expression. Therefore, as the major components of tissue-specific transcriptional regulation, the complex interactions among upstream TFs, enhancers, and downstream target genes can be viewed as regulatory networks (Figure 1). In general, multiple TFs bind to a specific enhancer, which is also tightly related with the epigenetic signal landscape around the enhancer, and turn on the enhancer's regulatory activity. Multiple enhancers may collaborate and regulate the transcription level of a specific target gene, and one enhancer may participate in the regulation of multiple genes. And the regulatory relationships are often tissue-specific. Another layer of complexity of tissue-specific regulatory networks comes from the long-range regulation of enhancers. For a specific enhancer element, the nearest gene in the genome may not be the target gene (Figure 1). Considering the human genome as an article

and a specific enhancer as a word, the precise interpretation of the word depends not only on the texts of local neighborhood but also distal contexts. Biologically, the long-range regulation of target genes by enhancers is mediated by 3D chromatin interactions or loops [15,16]. Although the target gene's promoter may be located far away from the specific enhancer in the 1D genome, they may be located in proximity with each other in 3D space, based on chromatin interactions.

Knowing the structure of tissue-specific regulatory networks is apparently useful for genotype-based phenotype prediction. Based on the regulatory network, for a specific non-coding SNP in an enhancer, it is feasible to predict which TF's binding is disrupted, whether the enhancer's regulatory activity is turned off, and which specific target gene's expression may be affected (Figure 1). For a group of potential disease-associated SNPs, it is also useful to identify the group of target genes whose expression regulation may be disrupted and whether they are enriched in specific biological pathways. In addition, the regulatory networks also provide a foundation to predict the common TF binding sites that are disrupted by a group of disease-associated SNPs and further infer the master TF regulators underlying specific diseases. All these regulatory network-based predictions will lead to more in-depth understandings of disease mechanisms and how the disruptive effects of genetic variants propagate through regulatory networks.

To further illustrate the complexity of tissue-specific regulatory networks, we can consider genes, the most important genomic units that may be directly related to diseases, as bulbs. And we can consider enhancers as switches that can control the bulbs (turn the bulbs on or off), and TFs as workers that send instructions to control the switches (Figure 2). Multiple switches can control one bulb, and in the meantime, one switch may also control multiple bulbs. Similarly, there is also a multiple-to-multiple mapping from workers to switches. In addition, due to long-range control, the switches may be located in New York while the bulbs are in Chicago (Figure 2). Furthermore, considering the scale of this system (i.e. >20k genes, ~1M enhancers, ~1k TFs), it is a significant challenging problem in machine learning to delineate the specific links in the network.

1.4 Machine learning developments to integrate regulatory networks for advanced GWAS

Given the multiple benefits of network-based approaches to improve disease genetic analysis, a variety of machine learning algorithms are needed for each step, from identifying the specific genomic locations of regulatory elements to discovering the downstream dysregulated genes or upstream master TFs. Different modeling frameworks are needed, depending on the goal of each specific step. In this chapter, we present an algorithmic pipeline that assembles a series of machine learning algorithms to achieve an improved prediction of disease associated genetic variants and their related mechanisms. Along with the introduction of this pipeline, we provide systematic reviews of the machine learning algorithms included in the pipeline. Overall, there are three major steps (Figure 3): 1) Predict enhancer locations in the human genome (i.e. identify 'nodes' for the networks), 2) Identify upstream TFs disrupted by SNPs (i.e. identify TF-enhancer edges in the networks), and 3) Identify downstream target genes dysregulated by SNPs (i.e. identify enhancer-gene edges in the networks). For each step, there are 2 major tasks (Figure 3): 1) Network re-

construction and 2) Integration of networks with genetic variants. These algorithms not only help to discover SNPs underlying diseases more accurately, but also help to prioritize potential causal genetic variants and the perturbed regulatory pathways.

1.4.1 Predict regulatory elements and prioritize disease-associated SNPs.—

A tissue-specific regulatory network is a graph consisting of nodes and edges, where nodes are mainly non-coding regulatory elements or genes and edges represent regulatory links. Therefore, the first step of regulatory network re-construction is to identify the nodes, *i.e.* we need to know where they are in the human genome (Figure 3). Since genes are already largely well annotated for the human genome, the major computational task is to identify the genomic locations of non-coding regulatory elements, *e.g.* enhancers. Traditionally, computational algorithms mainly rely on DNA sequence features to predict regulatory elements, such as sequence conservation across different species or TF motif scanning. Since the DNA sequence features are not tissue-specific, the prediction accuracy for functional regulatory elements, which are highly tissue-specific, is low. Based on recent development of next-generation sequencing based techniques, large amounts of ‘omics’ datasets have been profiled by several large consortia, such as ENCODE and Roadmap Epigenomics projects [17,18] (Table 1).

These ‘omics’ datasets provide tissue-specific information of the functional states of potential regulatory elements along the human genome, which are not directly provided by the DNA sequences alone. The major goal for this step is to explore and identify the combinatorial patterns of multiple ‘omics’ signal tracks that are predictive of regulatory elements, as for each specific genomic location it is no longer a single-valued signal (*i.e.* DNA sequence) but a high-dimensional array of ‘omics’ signals from different histone modifications, DNA methylation, chromatin accessibility and TF binding.

Generating an accurate genome-wide prediction of non-coding regulatory elements in this step is already a substantial achievement, since the predicted regulatory elements can be integrated with genetic variants (Table 1) and help to prioritize the real disease-associated SNPs (Figure 3), compared to traditional GWAS analysis. From the probabilistic modeling perspective, predicted regulatory elements can help to set more biologically meaningful prior probabilities for genetic variants to be associated with diseases. Intuitively, SNPs located in regulatory elements are expected to have higher prior probabilities to induce diseases because they may cause dysregulation of gene expression, compared to other SNPs located in the genomic background. Instead of assuming uninformative priors across all SNPs in the genome (*i.e.* equal priors), the regulatory element annotation-based priors can help address the challenges of low statistical powers and fine-mapping of causal variants as we discussed above.

1.4.2 Predict TF-enhancer links and identify master regulators disrupted by SNPs—The second step in network re-construction is to link upstream TFs to specific enhancers, and the TF-enhancer links are a major type of regulatory links in the network (*i.e.* network edges). As we mentioned before, enhancers are functional in gene expression regulation because they are bound by specific combinations of different TFs. The binding of specific TFs can activate the enhancers to further regulate genes. Different enhancers are

usually bound by different TFs (Figure 3). And there are groups of enhancers, i.e. co-regulated enhancer modules, which are usually bound by the same combinations of TFs. If a specific TF can bind to a specific enhancer, then there is a regulatory link between the node of the TF and the node of the enhancer, in the re-constructed networks. The basic approach to link TFs to specific enhancers is by scanning TF binding motifs in predicted enhancer regions. For example, if there is a strong sequence hit of GATA1 motif in a specific enhancer, it is assumed that GATA1 may have the potential to bind to the enhancer element. Due to the large false positives of sequence motif-based methods, genome-wide ChIP-seq data of specific TF binding is important for predicting TF-enhancer links, as TF ChIP-seq data represent real biological events of TF binding in-vivo. But due to the complexity of antibody preparation and cost, ChIP-seq data of TF binding is largely missing for different cell-types or tissues. Therefore, an accurate computational algorithm to impute ChIP-seq signals of specific TF binding in specific cell-types along the human genome, based on features of DNA sequences and tissue-specific chromatin accessibility, is the major focus in this step of network re-construction.

The importance of imputed tissue-specific TF binding in enhancer regions, when combined with genetic variants, is to facilitate the identification of master TF regulators associated with diseases and reveal mechanistic insights (Figure 3). The imputed tissue-specific TF binding profiles will enable researchers to investigate whether disease-associated SNPs are enriched in certain TF's binding sites, and the enriched TFs are probably the master regulators to maintain normal cellular states. If a specific TF's binding sites are significantly disrupted by disease-associated SNPs, it indicates that the TF and its regulatory circuitry may be candidate targets for new therapeutic techniques. In the meantime, the imputed TF binding signals within enhancers can further help to prioritize the causal SNPs as SNPs located within TF binding sites are more likely to induce dysregulation of gene expression.

1.4.3 Predict enhancer-gene links and discover dysregulated genes by SNPs

—The third step in regulatory network re-construction is to link specific enhancers to their downstream target genes, which is another major type of regulatory links in the network (Figure 3). Enhancers are key non-coding genomic units because they play critical roles in precise tissue-specific regulation of gene expression. Knowing the downstream target genes that are regulated by specific enhancers provides a bridge to map non-coding SNPs to their effects in coding regions (i.e. gene expression changes). Although some enhancers regulate the nearest gene located in the genome, many of them participate in long-range regulation and their specific target genes may be located far away from them in the 1D genome. The long-range regulation of enhancers is biologically mediated by 3D chromatin interactions, where the distal target genes are placed in proximity to the corresponding enhancers by chromatin loops in 3D space. It is therefore one of the most challenging computational problems in bioinformatics. In order to predict the long-range target genes for specific enhancers, multiple 'omics' datasets across diverse panels of cell-types or tissues are needed.

The benefit of integrating the predicted enhancer-genes links with disease-associated non-coding SNPs is apparent: it enables the mechanistic interpretation of why certain non-coding genetic variants are associated with diseases (Figure 3). Considering that the majority of

disease-associated SNPs are in non-coding regions and are enriched in enhancers, the predicted enhancer-gene links will expand our capability to understand how the disruptive effects of SNPs propagate through target gene expression perturbations and can also help to identify genes that are significantly associated with diseases. Furthermore, based on the set of genes dysregulated by non-coding SNPs in enhancers, pathway enrichment analysis can be carried out to look for the affected biological processes underlying human diseases.

In order to address the machine learning problems listed above, the fast accumulation of multi-omics data and large cohorts of human disease genetics studies (Table 1) provides unique opportunities nowadays to develop novel computational algorithms and facilitate their wide implementation. This is a highly active research area with many new machine learning algorithms and types of computational software being developed [19,20]. Based on our own testing and practical applications, we have selected a series of software and algorithms to use as a pipeline (Figure 3) which can re-construct the regulatory networks step-by-step and leverage the network predictions to improve disease genetics studies. In the rest of this chapter, we will review the selected software through this pipeline, introduce their computational model design, and demonstrate their usage.

2. Software and data resources

The pipeline to be introduced mainly includes three steps for regulatory network re-construction. For each step, the predicted network results are integrated with genetic variants or DNA sequence features for improved disease analysis (Figure 3). The six algorithms that we review and recommend here are: ChromHMM [21] (<http://compbio.mit.edu/ChromHMM/>), bfGWAS [22] (<https://github.com/yjingj/bfGWAS>), Catchitt [23] (<http://jstacs.de/index.php/Catchitt>), deltaSVM [24] (<http://www.beerlab.org/deltasvm/>), IM-PET [25] (<http://tanlab4generegulation.org/IM-PET.html>) and ARVIN [26] (<https://github.com/gaolong/arvin>). For software applications, there are large amounts of multi-omics datasets, SNP dataset and TF motif datasets, which were generated by several consortia [17,18,27–29] and are summarized in Table 1.

3. Methods

We describe the machine learning algorithms and software by following the three major steps in the pipeline. In each step, two algorithms will be introduced: one on regulatory network re-construction and one on integration with disease genetics.

3.1 Regulatory element prediction and improved fine-mapping of disease SNPs

3.1.1 ChromHMM algorithm.—ChromHMM [21] is one of the most widely used algorithms to predict regulatory elements based on multiple epigenetic features. As we discussed in the introduction, traditional approaches based on DNA sequence features alone are not accurate enough, with many false positives, and are not sufficient to predict tissue-specific functional enhancers. ChromHMM was designed to leverage the tissue-specific epigenetic features, such as ChIP-seq signals of different histone modifications, because prior biological studies have shown that specific histone modification combinations are associated with different functional genomic units (e.g. enhancers, promoters, transcription

regions and repressive domains). One of the major advantages of ChromHMM compared to other algorithms is the robustness to signal noise and a lower requirement of user-defined hyper-parameters.

Model formulation of ChromHMM.: ChromHMM considers the regulatory element problem as a segmentation task: the human genome is composed with consecutive latent segments that are not directly observed, different segments have different generation probabilities of specific histone modifications, and there are spatial dependencies between segments. Thus, ChromHMM employs hidden Markov model (HMM) as the modeling framework to segment the human genome into different chromatin states, a subset of which correspond to regulatory elements. In the hidden Markov model, the hidden states are different chromatin states and each chromatin state has different emission probabilities to generate observable signals of different histone modifications. There is also a transition probability matrix to characterize the spatial relationship among different chromatin states. The observed data along the human genome are multiple signal tracks for different histone modifications, which are assumed to be generated by the underlying hidden chromatin state path in the model. The goal is to infer the optimal hidden chromatin state path which can maximize the probability of observing the given multi-tracks of histone modifications along the genome in the specific cell-type or tissue. The well-known Baum-Welch algorithm (a special case of the EM algorithm) is used to infer the final segmentation of the genome into chromatin states, along with the emission and transition probabilities. To assign the biological meanings to the learned hidden states, the segments of each state are compared with other genomic annotations (e.g. TSS, gene bodies, p300 binding sites) and biological labels are assigned based on the enriched annotations. For example, chromatin states that are enriched with H3K4me1 and H3K27ac histone modification peaks or p300 binding peaks are labeled as regulatory enhancer elements.

Software usage and key parameters of ChromHMM.: Based on the instructions of ChromHMM [21], a typical command to run is: `java -jar ChromHMM.jar` plus settings of several parameters, including memory, number of processors, the directory of binarized input data (i.e. histone modification ChIP-seq signal files), the directory for output and also the specific human genome assembly. The only major parameter that the user needs to decide is the number of hidden states for the hidden Markov model. In practice, it's suggested to try several different numbers (common choices include 5, 10, 15 and 18 states) and check the final results. The key outcomes from ChromHMM will be the genomic locations of different chromatin states. For our purpose, the locations of regulatory elements (especially enhancers) will be used for other steps.

3.1.2 bfGWAS algorithm.—bfGWAS [22] can be used to incorporate the predicted regulatory element locations into GWAS analysis to fine-map the disease-associated SNPs and, in the meantime, increase the statistical power (i.e. reduce the number of false negatives). Since the regulatory element information is integrated for improved estimation of prior probabilities of SNPs to be associated with diseases, bfGWAS outperforms traditional GWAS studies that rely on genotype and phenotype data alone.

Model formulation of bfGWAS.: bfGWAS considers the identification of disease-associated SNPs as a variable selection problem and employs a Bayesian variable selection regression model as the inference approach. In the regression model, the disease phenotype of different individuals (including patients and controls) is treated as a responsible variable, which is a function of genotypes (i.e. the SNP profiles across different individuals). Each SNP is associated with an effect size, which follows a spike-and-slab prior. For the majority of SNPs, the effect size is near zero, while the disease-associated SNPs are associated with effect sizes that are significantly non-zero. Because of the huge number of SNPs in the human genome, the statistical power to identify SNPs with non-zero effect sizes is extremely low. bfGWAS borrows information from regulatory element annotations and incorporates the biological observation that potentially causal SNPs are enriched in regulatory element regions (e.g. enhancers, promoters, or general TF binding sites) compared to other genomic regions. Basically, for SNPs located in regulatory regions, the prior that has a non-zero effect size is expected to be higher, which is the key novelty of bfGWAS compared to basic GWAS. bfGWAS employs a Markov chain Monte Carlo (MCMC) strategy to simultaneously infer the priors for different types of genomic locations and the effect sizes of SNPs. And a parallel inference algorithm is used to scale up the computation.

Software usage and key parameters of bfGWAS.: bfGWAS can be run using the generated analysis Makefile [22]. The key inputs include the genotype file in VCF, PLINK or Tab-Delimited formats, the phenotype file in Tab-Delimited format, and the genomic annotation file (e.g. annotation locations of genes and regulatory elements) in Tab-Delimited format. The memory usage can also be specified by the user. The major outcomes will include the prior probabilities of being disease-associated for different types of genomic elements, along with the prioritized SNPs associated with the diseases.

3.2 TF binding imputation and incorporation with genetic variants

3.2.1 Catchitt algorithm.—Given the predicted regulatory elements along the human genome, specific TF binding sites within regulatory elements are needed to link TFs to regulatory elements in the regulatory networks. Since ChIP-seq data for most TFs in different cell-types are missing, imputed signals of TF binding are important to further help identify disease-associated non-coding SNPs. Catchitt is one of the top-scoring imputation algorithm for TF binding [23], which mainly relies on features of TF motif hits in DNA sequences and chromatin accessibility signals in the specific cell-type (assayed by DNA-seq or ATAC-seq).

Model formulation of Catchitt.: To impute TF binding signals for specific cell types, Catchitt takes the combined advantages of DNA sequence motifs and chromatin accessibility. DNA sequence motifs for specific TFs are easy to obtain and can be calculated based on the genomic sequence alone. Chromatin accessibilities are highly cell-type specific and can indicate cell-type specific in-vivo TF binding sites and reduce the number of false positives. To do this, Catchitt first uses the available ChIP-seq data for a specific TF in a certain cell-type, and then trains a logistic regression model of the TF binding ChIP-seq signal, based on features of sequence motif hits, chromatin accessibility, and distances to nearby genes. Based on the trained logistic regression for the specific TF, Catchitt is able to

impute the binding signals of the same TF in other cell-types, as long as chromatin accessibility data is available in other cell-types. To further improve the accuracy, Catchitt built an iterative training procedure to fine-tune the coefficients in the logistic regression model along with modifying the imputed signals of TF binding.

Software usage and key parameters of Catchitt.: To run Catchitt, a series of commands need to be run using `java -jar Catchitt.jar <tool name>`, where the commands identify TF binding peaks of the available ChIP-seq for training, calculate chromatin accessibility signals, score motif matches along the human genome, and iteratively train the logistic regression model. The ChIP-seq file of TF binding can be either narrowPeak or broadPeak formats. The chromatin accessibility file (e.g. DNase-seq) need to be in bigwig format. The TF motifs PWM are from the HOCOMOCO database [30] by default. The major output is the probabilistic scores of TF binding along the human genome in a given cell-type, and therefore we can predict the links between specific TFs and specific enhancers in the re-constructed networks.

3.2.2 deltaSVM algorithm.—The benefits of incorporating imputed TF binding signals into disease analysis come in two folds. First, it helps to further prioritize the potentially causal non-coding SNPs that are not only located within regulatory elements but also may disrupt TF binding. Second, based on the aggregated analysis of TF binding for a group of SNPs, it can indicate the potential master TF regulators that are enriched in disease-associated SNPs, providing new insights on disease mechanisms. We recommend to use deltaSVM [24] in the pipeline for this step, an algorithm which was originally developed to discover SNPs that may alter chromatin accessibility levels. In our practice, we found it is also a good tool to identify SNPs that may disrupt TF binding.

Model formulation of deltaSVM.: deltaSVM aims to identify the discriminative sequence features in regulatory elements that are informative to classify regions with high chromatin accessibility levels (or TF binding) versus regions with low levels, by training a support vector machine (SVM). Given a training ChIP-seq dataset of a TF's genome-wide binding in a specific tissue, deltaSVM first generates genomic regions with high signal levels as positive training samples and regions with low signal levels as negative training samples. Each sample region is then represented as a high-dimensional vector, where each element is a gapped k-mer (i.e. DNA sequence features). Based on the implementation of SVM, deltaSVM is able to find the important k-mers that are predictive for high-levels of TF binding, and furthermore, to predict the level of TF binding for an arbitrary DNA sequence. In this way, deltaSVM predicts the level of TF binding for a specific genomic location based on the reference genomic sequence first, and then predicts the level based on the altered sequence with a specific SNP. The changed predicted TF binding level ('delta') induced by the SNP is then used as a metric to prioritize SNPs that may disrupt specific TF binding in regulatory elements, which have higher probability to be real disease-associated genetic variants. Furthermore, for a group of SNPs from GWAS studies, enrichment analysis can be used to discover specific TFs that are recurrently disrupted by GWAS SNPs, suggesting potential master TF regulators whose regulatory circuitry is substantially affected in diseases.

Software usage and key parameters of deltaSVM.: deltaSVM can be run using the R package of gkmSVM with commands specifying gkmsvm_delta. The major inputs include the reference genomic sequences and the altered sequences with genetic variants, both in FASTA formats.

3.3 Enhancer-gene link prediction and identify disease-associated genes

3.3.1 IM-PET algorithm.—Based on the predicted enhancers along the human genome and the TF-enhancer links, the last step in regulatory network re-construction is to identify the downstream target genes regulated by enhancers. As we discussed, enhancer regulation is highly cell-type specific and involves long-range enhancer-promoter interactions mediated by 3D chromatin loops. Therefore, nearest genes in the 1D genome may not be good estimates as target genes for specific enhancers. We recommend using the algorithm IM-PET [25] to link specific enhancers to their distal target genes, based on combined features of enhancer activity, gene expression, TF activity and genomic distances.

Model formulation of IM-PET.: IM-PET is a supervised algorithm using experimental 3D chromatin map data for training. For a specific cell type, IM-PET considers enhancer-gene pairs that have 3D chromatin interactions as positive training samples and other random enhancer-gene pairs as negative samples. The four features for the prediction are: activity correlations between the enhancer and the gene across multiple cell-types, the correlation of expression between the TF (which has binding sites within the enhancer) and the gene, the co-evolution signal between the sequences of enhancer and gene, and the separation distance between enhancer and gene along the genome. The activity correlation between the enhancer and gene is the most important feature, since the target gene regulated by an enhancer is expected to show coordinated expression dynamics. The feature of distance is also critical, as experimental 3D chromatin interactions show a highly skewed distribution with respect to the distance. IM-PET employs the random forest method to make the predictions, considering the potential non-linear relations among the four features.

Software usage and key parameters of IM-PET.: IM-PET was written in Perl and the basic command to run it is: perl IM-PET.pl, with several settings of the input files. The input files mainly include the genomic locations of enhancers, enhancer activity values and gene expression levels across different cell-types. The outcome of IM-PET is the re-constructed regulatory network including the long-range links between specific enhancers and genes.

3.3.2 ARVIN algorithm.—Based on the re-constructed regulatory networks linking non-coding enhancers to genes, we are finally able to assign the functional interpretations of disease-associated SNPs with respect to their disruptive effects on gene expression and downstream biological pathways. By analyzing the biological functions of the target genes and their protein products (which have largely been annotated for the human genome), information can be obtained about the mechanisms of the non-coding SNPs associated with disease phenotypes. Moreover, aggregating non-coding SNPs located in enhancers can help to provide additional genes that may be associated with diseases, which are usually missed by traditional protocols that focus only on coding region SNPs. We recommend using a recent algorithm ARVIN [26] to integrate enhancer-gene links with disease genetics.

Model formulation of ARVIN.: ARVIN is an integrative model which incorporates a series of regulatory or functional features of non-coding SNPs and topological signatures of the regulatory networks and gene-gene networks. The key feature in the model is the predicted enhancer-gene links which facilitate the information flow between the non-coding SNPs and the coding genes. One of the basic model assumptions is that if a gene is frequently regulated by enhancers which contain disease-associated SNPs, the gene is more likely to be disease associated also, even if there is no strong evidence from coding SNPs. For features from the local genomic contexts of non-coding SNPs, ARVIN leverages TF-motif disrupting scores induced by the SNPs, the epigenetic and chromatin accessibility signals and other basic genomic sequence information. For the features on the gene side, ARVIN innovatively leverages the gene-gene networks which characterize the functional similarity or association among genes. The assumption is that if a gene is located close to a known disease-gene in the graph neighborhood of the gene-gene network, it is also possibly associated with the disease. ARVIN employs random forest as the modeling framework to combine all these features together to predict novel disease-associated genes and non-coding regulatory SNPs.

Software usage and key parameters of ARVIN.: To run ARVIN, based on its instructions, the major preparation is to collect all the necessary features for the integrative predictions. The input files include genotype file with SNP information, the predicted enhancer-gene links from IM-PET, the TF-motif disrupting scores, epigenetic signals, chromatin accessibility signals and other genomic sequence features, along with the differential gene expression dataset. The outcomes include the prioritized disease-associated genes and the corresponding non-coding SNPs.

3.4 Examples illustrating results from the pipeline

Based on the three steps described above, the pipeline will be able to re-construct large-scale networks of cell-type specific regulatory elements, upstream TF to enhancer links and enhancer to downstream target gene links. In the meantime, the pipeline will also effectively integrate different components of the predicted network with genotype/phenotype datasets to reveal in-depth understandings of disease mechanisms. We illustrate what can be learned by running the software introduced earlier with two interesting examples based on our analysis (Figure 4).

In the first example (Figure 4A), we prioritized a non-coding somatic mutation as a potentially functional genetic variant, because this SNP is located in a predicted non-coding enhancer element and, more specifically, within the binding peaks of MAX within the enhancer, which indicates that the SNP may disrupt MAX binding and inactivate the enhancer element. MAX is a well-known TF involved in blood cancer. Furthermore, the predicted long-range (>100kb) target gene of the enhancer element is BHLHE40, which has also been shown to be related with cancer. In support of this prediction, the prioritized SNP is indeed a significant genetic variant associated with leukemia from TCGA studies. Similarly, in the second example (Figure 4B), we prioritized a SNP (rs113859809) which is also located in a predicted enhancer element within an intron of the gene GATS and the predicted target gene is PVRIG, whose expression across diverse cell-types shows highly specific high expression in immune-related cells (T cell and B cells). More specifically, this

SNP is located with the peak of EBF1, suggesting that this variant may disrupt EBF1 binding. EBF1 (Early B cell Factor 1) is an important transcription factor in B cell development. As support of this prediction, this SNP is indeed a significant eQTL in blood tissues based on GTEx project.

These two examples demonstrate that we can substantially expand our view of the underlying biological processes associated with genetic variants in different diseases, including not only the prioritized non-coding variants but also the regulatory mechanisms. These predicted mechanisms will also provide valuable information to guide new experimental designs to further interrogate the genetic basis of complex human diseases.

4. Notes

4.1 Epigenetic features to predict enhancers.

As suggested by the Roadmap Epigenomics project and the ENCODE project, the major histone modifications that can be helpful to predict enhancers and other regulatory elements are: H3K4me1, H3K27me3, H3K4me3, H3K9me3 and H3K36me3. These five histone marks have differential enrichments in enhancers, promoters, actively transcribed gene bodies and repressive domains. And these five marks have the largest coverage across many cell-types or tissues and are widely available to use. Additional useful histone modifications include H3K27ac, which is very informative in predicting cell-type specific active enhancers. Other epigenetic features that are suggested to use are DNA methylations and chromatin accessibility, if their data are available for the specific cell-type or tissue under study.

4.2 Training dataset to predict enhancer-gene links.

IM-PET is a type of supervised machine learning software and relies on training data of 3D chromatin interactions. The preferred experimental 3D interaction data to train the model is ChIA-PET datasets, which have better resolution and lower noise levels than other experimental data of 3D chromatin. Considering that ChIA-PET data is not always available for different cell-types or tissues, an optional solution is to leverage Hi-C chromatin interactions data and combine the Hi-C interactions with histone modifications. Hi-C chromatin interactions are more stable across cell-types but they have limitations such as low resolution, high noise levels, and less cell-type specificity. These limitations can be mitigated by overlapping with cell-type specific histone modifications to improve the resolution of interacting fragments (e.g. enhancers), reduce noise, and substantially increase cell-type specificity.

4.3 Selection of the appropriate cell-type or tissue for the study of a specific disease.

As the final goal is to delineate the genetic basis of specific diseases, which are usually associated with multiple cell-types, the training data and machine learning predictions need to be obtained or generated in the matching cell-types. But it is sometimes not clear based on prior biological knowledge which cell-types are associated with the specific disease under study. One strategy of cell-type selection, as suggested by the Roadmap Epigenomics project, is to use the basic GWAS significant SNPs of the specific disease and check the

enrichment of overlaps of those SNPs with active enhancers or promoters. If the GWAS SNPs of the disease are globally enriched in enhancers of a subset of cell-types, those cell-types are more likely to be the associated cell-types and can be used in the described pipeline.

Acknowledgements

Hao Wang, Jiaxin Yang and Jianrong Wang were supported by NIH R01GM131398. The authors would like to thank iCER at MSU for providing the high-performance computing facilities.

References

1. Lander ES (2011) Initial impact of the sequencing of the human genome. *Nature* 470 (7333):187–197. doi:10.1038/nature09792 [PubMed: 21307931]
2. Visscher PM et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101 (1):5–22. doi:10.1016/j.ajhg.2017.06.005 [PubMed: 28686856]
3. Tam V et al. (2019) Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 20 (8):467–484. doi:10.1038/s41576-019-0127-1 [PubMed: 31068683]
4. Do C et al. (2017) Genetic-epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biol* 18. doi:ARTN 120 10.1186/s13059-017-1250-y
5. Gallagher MD, Chen-Plotkin AS (2018) The Post-GWAS Era: From Association to Function. *Am J Hum Genet* 102 (5):717–730. doi:10.1016/j.ajhg.2018.04.002 [PubMed: 29727686]
6. Hawkins RD et al. (2010) Next-generation genomics: an integrative approach. *Nat Rev Genet* 11 (7):476–486. doi:10.1038/nrg2795 [PubMed: 20531367]
7. Deplancke B et al. (2016) The Genetics of Transcription Factor DNA Binding Variation. *Cell* 166 (3):538–554. doi:10.1016/j.cell.2016.07.012 [PubMed: 27471964]
8. Watanabe K et al. (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 8. doi:ARTN 1826 10.1038/s41467-017-01261-5
9. Schaub MA et al. (2012) Linking disease associations with regulatory information in the human genome. *Genome Res* 22 (9):1748–1759. doi:10.1101/gr.136127.111 [PubMed: 22955986]
10. Shlyueva D et al. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* 15 (4):272–286. doi:10.1038/nrg3682 [PubMed: 24614317]
11. Creighton MP et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *P Natl Acad Sci USA* 107 (50):21931–21936. doi:10.1073/pnas.1016071107
12. Kimura H (2013) Histone modifications for human epigenome analysis. *J Hum Genet* 58 (7):439–445. doi:10.1038/jhg.2013.66 [PubMed: 23739122]
13. Lister R et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462 (7271):315–322. doi:10.1038/nature08514 [PubMed: 19829295]
14. Zhou VW et al. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 12 (1):7–18. doi:10.1038/nrg2905 [PubMed: 21116306]
15. Schoenfelder S, Fraser P (2019) Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* 20 (8):437–455. doi:10.1038/s41576-019-0128-0 [PubMed: 31086298]
16. Heintzman ND et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459 (7243):108–112. doi:10.1038/nature07829 [PubMed: 19295514]
17. Dunham I et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414):57–74. doi:10.1038/nature11247 [PubMed: 22955616]
18. Kundaje A et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518 (7539):317–330. doi:10.1038/nature14248 [PubMed: 25693563]
19. Valencia AM, Kadoch C (2019) Chromatin regulatory mechanisms and therapeutic opportunities in cancer. *Nat Cell Biol* 21 (2):152–161. doi:10.1038/s41556-018-0258-1 [PubMed: 30602726]

20. Kim K et al. (2016) Chromatin structure-based prediction of recurrent noncoding mutations in cancer. *Nat Genet* 48 (11):1321–1326. doi:10.1038/ng.3682 [PubMed: 27723759]
21. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 9 (3):215–216. doi:10.1038/nmeth.1906 [PubMed: 22373907]
22. Yang JJ et al. (2017) A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. *Am J Hum Genet* 101 (3):404–416. doi:10.1016/j.ajhg.2017.08.002 [PubMed: 28844487]
23. Keilwagen J et al. (2019) Accurate prediction of cell type-specific transcription factor binding. *Genome Biol* 20. doi:ARTN 9 10.1186/s13059-018-1614-y
24. Lee D et al. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 47 (8):955–+. doi:10.1038/ng.3331 [PubMed: 26075791]
25. He B et al. (2014) Global view of enhancer-promoter interactome in human cells. *P Natl Acad Sci USA* 111 (21):E2191–E2199. doi:10.1073/pnas.1320308111
26. Gao L et al. (2018) Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat Commun* 9. doi:ARTN 702 10.1038/s41467-018-03133-y
27. Lonsdale J et al. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45 (6):580–585. doi:10.1038/ng.2653 [PubMed: 23715323]
28. Kheradpour P, Kellis M (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* 42 (5):2976–2987. doi:10.1093/nar/gkt1249 [PubMed: 24335146]
29. Wang YL et al. (2018) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* 19. doi:ARTN 151 10.1186/s13059-018-1519-9
30. Kulakovskiy IV et al. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* 41 (D1):D195–D202. doi:10.1093/nar/gks1089 [PubMed: 23175603]

Figure 1. Large-scale regulatory networks associated with disease genetics.

The disruptive effects of non-coding genetic variants propagate through complex regulatory networks (TF-enhancer-gene links) and finally induce complex disease phenotypes.

Figure 2. Illustration of the regulatory network complexity.

Figure 3. Flowchart of the suggested pipeline.

The pipeline includes three steps and each step involves two algorithms.

Methods Mol Biol. Author manuscript; available in PMC 2021 January 01.

Table 1.

Large-scale data resources of multi-omics and genetic variants.

Data resource names	Data types	Scalability	Website
ENCODE	DNase data	~200 DNase-Seq datasets in ~140 cell-types	https://www.encodeproject.org/
	TF binding data (ChIP-Seq)	~2000 ChIP-Seq datasets in ~100 cell-types	
	Gene expression data	100 total RNA-seq in 58 cell-types	
	Chromatin interaction	37 Hi-C datasets in 24 cell-types 52 ChIA-PET datasets in 27 cell types	
Roadmap Epigenomics	Epigenomic data	~100 H3K27ac/ H3K9me3/ H3M4me1 datasets in ~25 functional groups of cell-types	http://www.roadmapepigenomics.org/
	DNase data	~70 DNase datasets in 44 functional groups of cell-types	
GTEx	SNPs	54 tissues	https://gtexportal.org/home/
Motif browser	Motif instances	Genome-wide binding of ~500 TFs	http://compbio.mit.edu/encode-motifs/
3D genome browser	Hi-C	>50 cell-types	http://promoter.bx.psu.edu/hi-c/
	ChIA-PET	6 cell-types	
	Capture Hi-C	>45 cell-types	