Published in final edited form as:

Hum Hered. 2016 ; 81(2): 88–105. doi:10.1159/000450827.

Computational Prediction of the Global Functional Genomic Landscape: Applications, Methods and Challenges

Weiqliang Zhou¹, Ben Sherwood^{1,2}, and Hongkai Ji^{1,*}

¹Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA

Abstract

Technological advances have led to an explosive growth of high-throughput functional genomic data. Exploiting the correlation among different data types, it is possible to predict one functional genomic data type from other data types. Prediction tools are valuable in understanding the relationship among different functional genomic signals. They also provide a cost-efficient solution to inferring the unknown functional genomic profiles when experimental data are unavailable due to resource or technological constraints. The predicted data may be used for generating hypotheses, prioritizing targets, interpreting disease variants, facilitating data integration, quality control, and many other purposes. This article reviews various applications of prediction methods in functional genomics, discusses analytical challenges, and highlights some common and effective strategies used to develop prediction methods for functional genomic data.

Introduction

Since the completion of the Human Genome Project in 2003 [1, 2], the list of sequenced genomes has been steadily growing. Analyzing DNA sequences in these genomes has provided a foundation for compiling comprehensive catalogs of genes and studying genetic bases of normal and disease phenotypes. It has been a major force driving the rapid development of genomics. Today, sequence analysis remains to be a core component of genomics. However, analyzing DNA sequences alone is not enough for answering a fundamental question in biology – how genes residing in the static DNA sequences are activated and operate in a dynamic, context-dependent, and synergistic fashion to execute their functions. Functional genomics is a young research field that aims to answer this question using genome-wide experimental and computational approaches.

Elucidating the operating logic and program behind the dynamic gene activities requires one to collect and integrate multiple layers of information such as gene expression, transcription factor (TF) binding sites, epigenetic modifications, three-dimensional chromatin structure, etc., in different cellular, developmental or environmental contexts. Recent advances in high-throughput technology have made it possible to map a variety of these functional genomic signals in a genome-wide fashion. Examples include microarray [3–5] and RNA-sequencing

*To whom correspondence should be addressed: hji@jhu.edu.

²Present address: School of Business, The University of Kansas, 1654 Naismith Drive, Lawrence, KS 66045, USA

(RNA-seq) [6–10] for measuring transcriptome, chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) [11–14] for mapping TF binding sites and histone modifications, methylation microarrays (e.g. Illumina 450K array [15]) and whole-genome or reduced representation bisulfite sequencing (WGBS [16] or RRBS [17]) for analyzing DNA methylation, DNase I hypersensitive site sequencing (DNase-seq) [18] and sequencing assay for transposase-accessible chromatin (ATAC-seq) [19] for measuring chromatin accessibility, and Hi-C [20] and a number of other chromosome conformation assays for studying three-dimensional chromatin structure [21]. Using these technologies consortium projects, such as the Encyclopedia of DNA Elements (ENCODE) [22] and Roadmap of Epigenomics [23], have generated massive amounts of functional genomic data for a variety of cell lines and tissues. Additionally, millions of functional genomic profiles generated by researchers all over the world have been deposited in public repositories such as the Gene Expression Omnibus (GEO) [24] and the Sequence Read Archive (SRA) [25]. The data are continuing to grow. How to most effectively analyze these unprecedented amounts of data, determine their quality, identify truly useful signals, and utilize them to study basic biology and human diseases become a critical challenge.

Increasingly, prediction methods are used as a tool in functional genomic studies. The wealth of data opens the door to systematically exploring the correlation among different data types. This correlation provides the basis for constructing prediction models that predict one data type from the others (Fig. 1). Prediction methods can be used to fill in the missing information in the functional genomic landscape when experimental data are not available, generate new biological hypotheses, and help researchers study relationships among different types of functional genomic signals. On the other hand, globally predicting the functional genomic landscape also presents new analytical challenges because of the need to deal with large, complex and heterogeneous data. The purpose of this article is to illustrate applications of prediction methods through several examples and review common analytical issues and strategies used to develop solutions. Through such a review, we hope to facilitate the development of new prediction-based solutions in existing and new applications, and to inspire creative use of massive amounts of existing functional genomic data through prediction tools.

Functional genomics

In this article, functional genomics is broadly defined as the genome-wide study of the operating logics and programs through which static DNA sequences are turned into dynamic gene activities and functions. It encompasses studies of a variety of data types.

Consider the human body as an example. A human body is a system consisting of different organs, tissues and cell types. Different components of this system are created in a temporally and spatially ordered fashion as a human individual develops from a fertilized egg to an adult. Most cells in this system have identical DNA sequences, yet their appearances, behaviors and functions can vary substantially. How does this happen?

Dynamic gene expression is a key to understanding cells' variable phenotypes. The human genome contains 20,000–25,000 protein-coding genes. In different contexts (e.g., different

cell types, developmental stages, or locations in human body), different sets of genes are activated to produce RNAs and proteins, leading to different functions and phenotypes of cells. For this reason, it is important to obtain gene expression profiles in different contexts in order to understand how the system of human body operates.

Mapping gene expression only provides one piece of information rather than the whole picture. Gene expression is tightly controlled. There is a sophisticated program to determine when, where and at what level each gene should be expressed. To further understand how genes' activities are controlled and coordinated, additional information needs to be collected. One example is the study of genes' transcriptional regulation by transcription factors (TFs). TFs are an important class of regulatory proteins. These proteins themselves are products of genes. Humans have approximately 1500 different TFs. Each TF can bind to $10^3 - 10^5$ genomic sites to induce, repress or modulate expression of thousands of genes. Many TFs bind to specific DNA sequence patterns in the genome. These patterns are called motifs. Different TFs can recognize different motifs and regulate different sets of genes. TF binding is context-dependent. In different cell types or biological conditions, different TFs are activated to turn on or turn off different sets of genes. The binding sites of a given TF and its binding strength at each site also vary across different cell types and conditions. Therefore, decoding the gene regulatory program requires information on the genome-wide binding sites of all TFs, also called regulome, in different contexts.

Epigenetic modifications also play an important role in the regulation of gene expression. The term "epigenetics" was used to refer to the study of heritable changes in gene activity that does not depend on changes in DNA sequence [26]. In the National Institutes of Health's Roadmap Epigenomics project [23], epigenetics is more broadly defined to encompass both heritable changes and "stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable" [27]. One example of epigenetic modification is DNA methylation which is often associated with gene repression. Histone modifications and chromatin accessibility are two other examples. Both are involved in shaping cells' chromatin landscape (e.g., open or closed chromatin) and the chromatin states are closely associated with genes' transcriptional potential. Epigenetic modifications may be altered by environmental factors and hence may serve as a mediator through which environments influence gene expression [28]. Collecting genome-wide information on epigenetic modifications, or epigenome, therefore is also important for understanding how static DNA sequences are turned into dynamic activities of genes.

Three-dimensional (3D) chromatin interactions provide another layer of information for studying gene regulation. Chromosomes reside in a 3D space. Functional elements far away in the linear genome sequences can be close to each other in the 3D space due to DNA looping. One example is enhancer-promoter interactions which provide a mechanism to coordinate functions of different components of a transcriptional machinery to modulate context-dependent gene expression.

Transcriptome, regulome, epigenome, and 3D chromatin structure are only examples among many layers of information useful for studying dynamic gene activities and functions. All these data types play an important role in addressing the fundamental question of functional

genomics. Therefore, they are all viewed as part of cells' functional genomic landscape in this article.

There are many challenges in functional genomics research. For example, since gene activities and regulation are context-dependent, constructing an organism's full functional genomic landscape requires one to collect all relevant data types from different biological contexts, defined as different combinations of cell type, developmental time, environment, and other factors. Despite the rapid development of technologies, it is still not feasible to perform all types of functional genomic assays to study every biological context or every new biological sample of interest due to material and resource constraints and technological limitations. Studies of precious clinical samples and embryonic tissues, for instance, are often constrained by the limited amounts of materials. These samples often have limited numbers of cells that are not enough for all functional genomic assays. Mapping transcription factor binding sites using ChIP-seq requires high-quality antibodies which are not always available. Even if these technical constraints did not exist, the number of all possible combinations of data type and biological context is huge, and analyzing them all would be extremely costly. Moreover, data generated by many high-throughput technologies are noisy and often contain multiple sources of unwanted variation and technical artifacts. Determining which datasets are accurate and truly useful (i.e., data quality control) and detecting meaningful biological signals from the noisy data are important but nontrivial tasks. Integrating different data types and studying their relationship also pose significant challenges.

Why use prediction methods?

Prediction methods can be used for at least two different purposes. First, they provide a way to understand the relationship among different functional genomic signals. For instance, predicting gene expression using transcription factor binding data can help identify important regulators that control genes' transcriptional activities. They can also shed light on how these regulators interact to synergistically specify the regulatory program [29]. Predicting DNA methylation from DNA motifs can generate insight on how static DNA sequences may encode a dynamic epigenetic program that varies across different cell types [30]. Predicting transcription factor binding sites (TFBSs) from DNase I hypersensitivity and histone modification profiles can elucidate functional genomic signatures that distinguish functional transcription factor binding motif sites with non-functional motif sites [31].

Second, prediction methods provide a practical and cost-effective solution to estimating missing functional genomic information. As the functional genomic landscape is dynamic and changes from one biological condition to another, comprehensively mapping this landscape in all possible cell types and biological conditions is beyond the capacity of any individual laboratory and existing consortiums such as ENCODE and Roadmap of Epigenomics. By exploiting the correlation structure learned from the existing data, using prediction methods allows one to map unknown functional genomic signals based on partially observed experimental data. The predicted functional genomic signals can then be used to guide hypothesis generation, experimental target prioritization, and disease variant

interpretation. They can also be used as a bridge to integrate different data types or as a reference for data quality control [32]. This provides a cost-efficient way to study new biological systems and increases the value of existing data.

Applications of prediction methods

In this section, we illustrate applications of prediction methods using several examples (Fig. 2).

Predicting transcription factor binding sites

A critical step toward decoding gene regulatory network is to map genome-wide binding sites of all TFs in all contexts. The state-of-the-art technology for mapping TFBSs is chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) [11, 12]. For a given TF and context, ChIP-seq allows one to map genome-wide binding sites of the TF. ChIP-seq has a number of limitations. It requires high-quality antibodies that specifically recognize the TF of interest. However, ChIP-grade antibodies are not always available. Even if one could obtain all antibodies, which is currently infeasible, conducting ChIP-seq experiments for all TFs in all possible contexts would be costly and labor-intensive. Also, conventional ChIP-seq technologies require large amounts of input materials ($>10^6$ cells) which can be difficult to obtain when studying precious clinical samples or embryonic development. Although single-cell ChIP-seq (scChIP-seq) [33] has been recently reported, its signal remains highly sparse and discrete. Currently, scChIP-seq is not sensitive enough to accurately describe TF binding events at each individual genomic locus. Due to these limitations, computationally predicting TFBSs can provide a valuable complementary tool to existing experimental technologies for TFBS mapping.

Many TFs recognize specific DNA sequence motifs. Early attempts at predicting TFBSs were based on analyzing DNA motifs. Motifs for hundreds of human TFs have been determined by high-throughput technologies such as protein binding arrays [34–41]. These motifs can be computationally mapped to the genome [42, 43]. Then, binding sites of a TF can be predicted using its mapped motif sites. This approach is simple, but its prediction accuracy is low because DNA motifs are short and can be highly degenerate [44]. A motif can occur thousands to millions of times in the genome, but most of these motif sites are not bound by the TF and hence represent noise. Moreover, motif sites in a genome are static, but TFBSs are context-dependent. Predictions based solely on static motif sites cannot account for how TF binding varies across cell types and conditions.

In order to better predict TFBSs, recently numerous efforts have been made to incorporate context-dependent information such as DNase I hypersensitivity (DH) and histone modifications (CENTIPEDE [31], DNase2TF [45], PIQ [46], etc.). TFBSs are often located in regulatory regions such as promoters and enhancers. Nucleosomes in these regions carry specific histone modifications (HMs) (e.g., H3K4me1, H3K4me2 and H3K27ac for active enhancers; H3K4me2, H3K4me3 and H3K9ac for active promoters). Often, chromatin surrounding a TFBS is also sensitive to DNase I cuts. Thus, TFBSs are often associated with characteristic HM patterns and increased DH. Both HMs and DH are context-dependent and can be profiled in a genome-wide fashion using high-throughput technologies such as ChIP-

seq (HMs) and DNase-seq (DH). When coupled with DNA motif information, both data types can be used to predict TFBSs. CENTIPEDE [31], for instance, is a method that predicts TFBSs by integrating context-dependent information from DNase-seq and HM ChIP-seq with static information from DNA motifs and phylogenetic conservation using an unsupervised Bayesian mixture model. It models DNA motif sites as a mixture of bound and unbound states. The prior probability for the “bound” state depends on the static motif and conservation information via an unknown logistic function, whereas the context-dependent data are modeled using state-specific probability distributions. After fitting the model to the data, CENTIPEDE computes the posterior probability that a motif site belongs to the bound state. For the context-dependent part of the model, it was found that the spatial distributions of DNase-seq reads surrounding the motif sites under the bound and unbound states show distinct characteristics. The signals surrounding the bound motif sites tend to show a valley-shaped footprint which is not observed from the unbound motif sites. Modeling these spatial distributions of DNase-seq reads provides a key element for the model to distinguish bound and unbound motif sites. Similar to CENTIPEDE, DH and HMs are also used in numerous other methods for predicting TFBSs (Chromia [47]; Ramsey *et al.* [48]; Cuellar-Partida *et al.* [49]; Arvey *et al.* [50]; Wei *et al.* [51]; PIQ [46]; Iterative CENTIPEDE [52]). Gusmao *et al.* [53] performed a comprehensive comparison of such methods. Differential DNase-seq and HM ChIP-seq signals between two biological conditions have also been used to predict differential TF binding. For example, in the recently developed differential principal component analysis (dPCA) method [54], principal components were used to characterize main synergy patterns of differential HM ChIP-seq and DNase-seq signals in multiple datasets between two biological conditions. It was found that the first principal component was able to predict a TF’s differential binding at its motif sites. Besides DH and HMs, Xu *et al.* found that DNA methylation can also be used to predict TFBSs [55]. They found that DNA methylation levels around TFBSs and non-TF binding sites show distinct patterns. These DNA methylation patterns are TF-specific, but they are similar for the same TF across cell types. Based on these findings, they developed a method, Methylphet, that uses whole-genome bisulfite sequencing (WGBS) and TF ChIP-seq data to train random forest (RF) models for predicting TFBSs.

While TFBSs can be predicted by coupling DNA motif analysis with surrogate information such as DNase-seq, HM ChIP-seq, or WGBS, each experiment for collecting surrogate data can only analyze one biological context. Applying this approach to study all possible contexts remains difficult. Recently, Zhou *et al.* reported that genome-wide DH profile can be predicted using gene expression microarray or RNA-seq data, and the predicted DH can be coupled with DNA motif information to predict TFBSs [56, 57]. This approach creates an opportunity to use the massive amounts of gene expression data in public domains to predict TFBSs. Since public databases such as the Gene Expression Omnibus (GEO) [24] contains hundreds of thousands of gene expression samples collected by scientists worldwide from a wide variety of biological contexts, predicting TFBSs based on these data could substantially expand the existing TFBS catalog.

Predicting epigenome

Another core component of functional genomics is the study of the epigenome, which is dynamic and varies across different cell types and conditions. Epigenome provides a crucial piece of information complementary to genome sequences for understanding gene regulation, development and diseases. Different epigenetic marks are associated with different classes of functional sequence elements in the genome and different transcriptional output. For instance, histone H3 lysine 4 trimethylation (H3K4me3) is a histone modification typically found in actively transcribed promoters, whereas histone H3 lysine 4 monomethylation (H3K4me1) and H3 lysine 27 acetylation (H3K27ac) often mark enhancers. H3 lysine 27 trimethylation (H3K27me3) and DNA methylation often correlate with gene repression, and H3 lysine 36 trimethylation (H3K36me3) marks transcribed gene body. Analyzing combinatorial patterns of different epigenetic marks provides a powerful way to discover and annotate different classes of functional elements (i.e. promoters, enhancers, gene body, etc.) and understand their operating rules [32].

Genome-wide maps of histone modifications and DNA methylation can be obtained using ChIP-seq and WGBS respectively, whereas open chromatin can be profiled using DNase-seq and ATAC-seq. There are dozens of different epigenetic data types. With a finite amount of resources, most investigators can only collect data for a subset of epigenetic marks. However, different epigenetic data types are correlated. By exploiting this correlation structure, one may predict unmeasured epigenetic data types using the measured ones. In a recent study, Ernst and Kellis [32] used the massive amounts of epigenome data generated by the NIH Roadmap Epigenomics and ENCODE projects to study the correlation among different epigenetic marks. They developed ChromImpute, a computational approach based on an ensemble of regression trees to impute missing epigenomic data types based on observed ones. They have shown that the ChromImpute predicted signals can predict the true experimentally measured epigenomic signals. Using this approach, they created a comprehensive map of 127 reference epigenomes, which would not be available using experimental data alone. Interestingly, it was found that the imputed data showed better signal quality than the observed experimental data for the same epigenetic mark. Compared to the observed experimental data, the imputed data had more robust and consistent signal profiles and showed better agreement with independent functional annotations (e.g., imputed H3K4me3 signals better recover promoters than observed H3K4me3 signals). It was demonstrated that the reference epigenome map obtained after imputation can facilitate interpretation of trait-associated variants from genome-wide association studies (GWAS). For example, by analyzing the signal enrichment of the imputed H3K27ac signals at SNPs of each GWAS dataset, the ChromImpute authors found that liver was the most enriched sample for various cholesterol phenotypes and immune-related cells were the most enriched samples for various immune-related disorders. The imputed data can be useful even when the experimental data are available. For example, it was shown that the correlation between the imputed data and the experimental data can be used for data quality control (QC). Experimental data with poor quality tend to show low agreement with the imputed data. Compared with the traditional QC metrics such as read depth and proportion of reads falling in enriched regions, imputation-based QC metrics were able to capture the datasets with low data quality even when the traditional QC metrics failed. The imputed data can also help

with annotating functional elements in the genome. In the chromatin state annotation analysis based on ChromHMM [58], using the imputed data showed more consistent genome coverage across samples and better agreement with known annotations. Based on this chromatin state annotation, other researchers have developed a functional annotation-based method to increase the power to detect genes that are associated with expression quantitative trait loci (eQTLs) [59].

In another study, Zhou et al. [56] developed BIRD, a big data regression approach, for predicting open chromatin using gene expression data. Unlike ChromImpute, whose predictions are based on multiple epigenomic data types, BIRD predictions are based on gene expression data alone. Since the number of gene expression samples in public domains is orders of magnitude larger than the number of epigenomic samples, BIRD may allow one to make predictions in a much larger number of biological contexts. To demonstrate, they applied BIRD to 2000 public gene expression samples from different cell types and cell conditions in GEO. The predicted chromatin accessibility data are made available as a database PDDb (Predicted DNase I hypersensitivity database, available at <http://jilab.biostat.jhsph.edu/~bsherwo2/bird/index.php>). They showed that predicted DH in PDDb recovered known differential binding sites of MYC in human embryonic stem cells and P493–6 lymphoma cell line, and decreasing binding activities of SOX2 at its binding sites during stem cell differentiation. These results show that the predicted data in PDDb can be used to study differential regulatory element activities when experimental ChIP-seq or DNase-seq data are not available. It was also shown that by coupling known ChIP-seq data of a TF in one biological context with the predicted chromatin accessibility data in a wide variety of biological contexts in PDDb, one can classify a TF's binding sites into functionally related subclasses and explore their activities in unstudied biological contexts. For example, the BIRD authors analyzed the binding sites of MEF2A obtained from MEF2A ChIP-seq in lymphoblastoid cells (GM12878). By grouping these TFBSs into functionally related subclasses based on the predicted chromatin accessibility in the 2000 PDDb samples, they found that MEF2A binding sites can be grouped into multiple clusters which are activated in different biological contexts. As an example, one group of MEF2A binding sites was more active in neuron and brain related samples than the other samples. This was validated using independent sources of information. For instance, functional annotation analyses showed that genes regulated by these TFBSs are enriched in “regulation of neuron differentiation” and “regulation of neurogenesis” functions, consistent with the predicted increase of MEF2A binding activities in neuron and brain samples. This demonstrates how one may use the predicted data to obtain valuable information for characterizing context-dependency of TFBSs. Such analyses can be used to generate hypotheses (e.g., MEF2A may actively bind to a subset of its binding sites to regulate neurogenesis in neuron and brain) and prioritize targets for experimental validation (e.g., the binding sites predicted to be active in neuron and brain may serve as the top candidates for knock-out experiments to test the function of MEF2A in these tissue types).

Prediction methods have also been used to generate high-resolution epigenome using low-resolution data. For example, DNA methylation can be assayed by WGBS or 450K array. Although WGBS is able to measure single-site DNA methylation levels, it is expensive and difficult to perform in particular genomic regions. As a result, a large number of DNA

methylation data are measured by 450K array. Zhang *et al.* [60] developed a method to predict genome-wide DNA methylation using information obtained from methylation measured by 450K array, genome annotation and regulatory elements. They discovered that the methylation level of a CpG site can be well predicted by the methylation from its neighboring CpG sites, location in a CpG island, chromatin accessibility, TF activity (ELF1, MAZ, MXI1, and RUNX3), and histone modifications (H3K27ac, H3K4me3 and H3K9ac).

Another application is predicting epigenome for one cell type using epigenomes from other cell types within the same individual. For instance, based on the DNA methylation data from 450K array, Ma *et al.* [61] developed a method to predict DNA methylation for one cell type using methylation levels measured from another cell type. They observed linear relationships between DNA methylation levels from different cell types for each individual and utilized these relationships to make predictions.

Finally, prediction methods have also been used to study how the dynamic epigenomic program might be coded in the static DNA sequences. For example, Whitaker *et al.* [30] developed Epigram to systematically search for DNA sequence motifs associated with DNA methylation valleys [62] and histone modification marks. They then used the discovered epigenome motifs to predict histone modifications. Their study shed light on how static DNA sequence motifs, through different combinations, may help with setting up the dynamic epigenome. Their study also provided a comprehensive catalog of epigenome related DNA motifs. These motifs can be used to guide design of experiments for epigenome editing using technologies such as CRISPR–Cas9 system [63]. Predicting epigenome using DNA sequences has also been explored in [64–66].

Predicting chromatin conformation

The 3D chromatin conformation provides crucial information for elucidating genome structure and gene regulation. A number of different technologies such as Chromosome Conformation Capture (3C) [67], Circularized Chromosome Conformation Capture (4C) [68], Carbon-Copy Chromosome Conformation Capture (5C) [69], ChIA-PET [70], and Hi-C [20] have been developed to analyze 3D chromatin conformation. Among these, Hi-C is a high-throughput technology that can comprehensively survey interactions between all pairs of genomic locations. This technology has been used to discover important features of the chromatin structure such as chromatin compartment [20], topologically associated domains (TAD) [71], and interaction hub [72].

Hi-C experiments are expensive and are not routinely done. Interestingly, chromatin conformation characterized by Hi-C such as TAD boundaries and interaction hubs are often associated with specific chromatin accessibility and histone modification signatures. Using this correlation, Huang *et al.* [72] developed a method to predict chromatin interaction hubs and TADs using histone modifications and CTCF ChIP-seq data. Their study led to a hypothesis that combinatorial histone patterns may be involved in mediating chromatin interactions. However, this hypothesis still remains to be tested experimentally. Fortin and Hansen [73] developed a method to predict chromatin's A/B compartments using long-range correlation of DNA methylation or chromatin accessibility. They found that DNA methylation levels are highly correlated with Hi-C compartment signals. They have shown

that one can use DNA methylation measured by 450K array to accurately predict chromatin compartments in different cell types. Compared to Hi-C data, data for 450K arrays are more widely available, raising the possibility that one can use the existing 450K array data to expand the catalog of Hi-C profiles. However, they also reported that DNA methylation fails in predicting chromatin compartments for some cell types (e.g. whole blood). Therefore, it is crucial to further investigate in what cell types DNA methylation can be used to predict chromatin compartments before the method can be applied to a large number of cell types in the future using the widely available 450K array data.

Predicting transcriptome

With the availability of transcriptome, genome-wide TFBSs and epigenome data, models can be constructed to predict transcriptome based on other data types. These prediction models may be used to help elucidate how gene's transcriptional activity is controlled. For example, in Ouyang *et al.* [29] TF binding activities obtained from ChIP-seq were used to predict gene expression using principal component regression. For each gene, they first integrated ChIP-seq signals from a TF using a distance-based weighted summary statistic to represent the TF's binding strength. Then, they applied principal component analysis to the binding strengths of multiple TFs and used the principal components to build regression models to predict gene expression. Through this analysis, they showed that gene expression can be explained by the combinatorial activity of multiple TFs in mouse embryonic stem cells, and they provided useful information about the role of each TF in gene regulation.

Because it is difficult to obtain ChIP-seq data for a large number of TFs, Natarajan *et al.* [74] developed a method to predict cell-type-specific gene expression for 19 cell types based on chromatin accessibility obtained from DNase-seq data. Combined with DNA motifs, DNase-seq can be used to predict TFBSs for many TFs simultaneously. Taking advantage of this property, Natarajan *et al.* first calculated TF binding scores by mapping 789 known TF motifs to each DNase I hypersensitive site (DHS). They then used a logistic regression model to predict cell-type-specific gene expression based on TF binding scores. This analysis demonstrates that cell-type-specific gene expression is not only controlled by TF activity in the proximal promoter regions, but also affected by TF activity in distal regulatory regions.

Numerous efforts have also been made to use other epigenetic marks to predict gene expression. For example, histone modifications can alter the structure of the chromatin to carry out functions linked to DNA replication, DNA repair and transcription [75]. Studies that use HMs to predict gene expression have helped reveal the distinct roles of different HMs. For instance, H3K4me3, a marker commonly found in promoter regions, and H3K27ac, a marker commonly found in enhancer regions, are both found to be positively correlated with gene expression. H3K27me3, on the other hand, marks gene repression [14, 76]. In Karli *et al.* [77], ChIP-seq data from multiple histone modifications was used to predict gene expression in human CD4+ T-cells. Based on this analysis, histone modifications most informative for predicting gene expression (e.g., H3K27ac) were identified. This study also demonstrated that prediction models trained by CD4+ T-cells can be applied to predict gene expression for CD36+ and CD133+ cells. Based on a similar

approach, Dong *et al.* [78] further tested the feasibility by training a prediction model using K562 cells and applying it to predict gene expression in four other cell types (i.e. GM12878, H1-hESC, HeLa-S3 and NHEK) which resulted in good prediction accuracy. They showed that the correlation coefficient between the predicted and experimentally measured gene expression is larger than 0.8. They also found that histone modification marks H3K9ac and H3K4me3 are the best predictors for determining whether a gene is ‘on’ or ‘off’, and H3K79me2 and H3K36me3 are the best predictors for the gene expression levels. Similar analyses have also been done in other organisms [79]. These analyses have improved the understanding of the relationships between gene expression and histone modifications. They indicate that the relationships between gene expression and HMs found in these studies are not specific to one cell type, but quite general across the tested cell types. These findings suggest that with prediction models trained using existing data and the HM data in a new cell type available, one may potentially predict gene expression in that cell type when the experimental expression data are lacking. We note, however, that whether the relationship between gene expression and HMs will remain the same for all cell types and contexts still needs to be continuously tested in the future.

For the interest of space, we will not enumerate all studies and applications that globally predict one functional genomic data type based on others. However, the examples discussed above are sufficient to demonstrate the value of prediction methods in functional genomic studies. The methods discussed above are listed and summarized in Supplementary Table 1. In what follows, we will discuss some common analytical issues and strategies to develop solutions for globally predicting the functional genomic landscape.

Analytical challenges and solutions

The prediction problems discussed above involve predicting a functional genomic profile $\mathbf{y} = (y_1, \dots, y_G)'$ using other sources of information $\mathbf{X} = ((\mathbf{x}_1, \dots, \mathbf{x}_G)'$. The response $\mathbf{y}$ is a high-dimensional random vector. Examples include transcription factor binding states for all genomic loci, DNA methylation levels for all CpG sites, or gene expression levels for all genes. Elements in $\mathbf{y}$ are indexed by $g \in \{1, \dots, G\}$, and they correspond to different genomic loci (e.g., different CpG sites or genes), where G is on the order of 10^4 – 10^9 . Each locus g has a feature vector $\mathbf{x}_g = ((x_{g1}, \dots, x_{gP})'$ which serves as the predictor and P is on the order of 1 – 10^9 . The matrix $\mathbf{X}$ contains predictors for all loci. This problem presents several challenges including identifying and extracting informative features, building prediction methods, dealing with the high-dimensionality of the predictors and responses, computational efficiency, data heterogeneity, and objectively evaluating the prediction performance (Fig. 3). Below we briefly review common strategies and considerations to develop solutions to each component.

Feature extraction

Finding and constructing features predictive of the outcome are crucial for solving the prediction problem. Features can be *static* or *dynamic*. For example, TFBS can be predicted by coupling DNase-seq data with DNA sequence motif information. DNase-seq signals are dynamic since they vary across cell types and conditions. On the other hand, DNA sequence

is static since it usually does not change across conditions (Fig. 4a). Other examples of static features are phylogenetic conservation, GC contents and distance to transcription start sites. Other examples of dynamic features include histone modifications and DNA methylation.

For a particular predictor data type, information useful for prediction often can be summarized into features representing *intensity*, *shape* or both. For instance, when DNase-seq data is used to predict TFBS, one can use the total read count at each genomic locus as a feature based on the knowledge that TF binding is correlated with increased DNase I hypersensitivity. The read count is an intensity feature. Besides the read count, the spatial distribution of reads at each locus also provides information useful for prediction. While chromatin surrounding a TFBS is sensitive to DNase cuts, chromatin bound by a TF is not because it is protected by the TF protein. This leaves a footprint in the DNase-seq data, a valley without reads within the binding site but with a large number of reads in the flanking regions of the binding site [31] (Fig. 4b). This is an example of a shape feature. Two other examples of shape features are a DNA methylation valley surrounding a TFBS [55] and a bimodal distribution of histone modifications due to nucleosome displacement for TFBS prediction [80].

Features can be *local* or *global*. To predict the expression level of a gene, one can use histone modification levels, a local feature, at the gene's promoter [77]. The expression level of a gene may depend on the expression levels of TFs that control this gene, and it may also correlate with expression levels of co-regulated genes. Therefore, one can also use TF activity at promoters of other related genes as features. These are global features (Fig. 4c). Zhou et al. [56] demonstrate that global features can provide valuable information not contained in the local features for making accurate predictions. Using global features, however, may substantially increase the dimensionality of the feature vector.

Strategies for building prediction models

Depending on the structure of available data, solutions to the prediction problem can be *supervised* or *unsupervised* (Fig. 5a). In a supervised approach, one has training data for which both the response y and the predictors X are observed. The y and X may be observed in one or multiple samples, cell types or conditions, denoted using $Y = \{y^1, \dots, y^N\}'$ and $X = \{X^1, \dots, X^N\}$ respectively where N is the sample size (i.e., number of samples, cell types or conditions). Using the training data, one trains a prediction model $f: X \rightarrow y$ which can then be applied to predict y whenever one only has information on X . One simple example of the supervised approach is linear regression in which the relationship between y and X is modeled using a linear function which can be estimated using various methods such as least squares (i.e., minimizing the total squared prediction errors) or maximum likelihood estimates. The Classification and Regression Trees (CART) [81] provides another example. CART partitions the X space based on a decision tree such that responses in each partition have similar values. For each partition, the prediction of y will be based on the average response in the training data in that partition. This approach can be used to model complex non-linear relationship between y and X . Ensemble learning is also a widely used supervised learning approach that combines multiple models to improve prediction performance. One example of ensemble learning is random forests [82] which is an ensemble of CART that

uses a random subset of the predictors. These methods are only a few examples. There are a larger number of supervised learning methods [83]. Two examples of supervised learning approach being used with functional genomic data are MethyPhet which uses random forest to predict TFBS based on WGBS data [55], and ChromImpute which uses an ensemble of regression trees to impute epigenomes [32].

In an unsupervised approach, training data with y observed are not required. This approach explores patterns in X and maps those patterns to y based on existing knowledge. An example is CENTIPEDE [31], which predicts TFBSs by using a two-component mixture model to describe motif sites. The model is fitted using both dynamic ChIP-seq and DNase-seq data and static sequence information. The mixing component that has high DH intensity and the characteristic footprint for TFBSs is then used to predict TFBSs. CENTIPEDE is an unsupervised approach because the model does not use data that explicitly states whether a site is bound or not. The predictions are made after the mixture model is fitted, and they are based on comparing each mixing component with the existing knowledge about the behavior of bound sites. Another example is the differential principal component analysis (dPCA) method [54]. Principal Component Analysis (PCA) is a widely used unsupervised learning approach for analyzing multivariate data. In PCA the variables are transformed through orthogonal linear combinations such that the transformed variables (i.e. principal components) are uncorrelated and the principal components sequentially explain maximal possible data variance not explained by previous principal components. The dPCA method uses principal components to describe major differential patterns of multiple epigenomic data types between two biological conditions. It then uses the first principal component to predict differential TF binding sites. This prediction approach is unsupervised since the construction of principal components does not rely on knowing the TF binding status of each genomic locus. Predictions on differential TF binding are based on the prior knowledge that increased TF binding at a locus is associated with the increase of certain histone modifications (e.g., H3K27ac).

In general, prediction models can be built along two directions: **cross-locus** direction or **cross-sample** direction (Fig. 5b). In the cross-locus direction, one assumes that different loci are governed by a common prediction model $y_g \leftarrow f(\mathbf{x}_g)$ which can be fit by treating data from different loci as exchangeable observations. This approach is particularly useful when data are observed for only one sample, cell type or biological condition (i.e., $N=1$), or when the sample size N is small. One example is MethyPhet [55] which uses WGBS data at all known TF binding sites to learn how DNA methylation valleys and other features determine TFBSs. The learned model is then applied to predict unknown TFBSs. In the cross-sample direction, one can build a model for each locus using data observed from multiple samples. This approach requires one to have a reasonably large sample size, N . However, it allows one to build locus-specific models, $y_g \leftarrow f(\mathbf{x}_g)$, in which different loci can be governed by different prediction laws. One example is BIRD [56] which uses data from multiple cell types to train models that predict chromatin accessibility at each locus based on gene expression data. One can also use information in both directions to build prediction models. For example, ChromImpute [32] predicts genome-wide profile of an epigenetic mark using both the same sample information and the same mark information (Fig. 5c). The same mark

information represents observed signals for the epigenetic mark in question in other closely related samples. They are extracted by assuming that samples with similar profiles for other epigenetic marks have similar signal levels for the mark in question. Conceptually, this is similar to a non-parametric regression across samples. The same sample information represents features extracted from other epigenetic marks in the same sample. The same sample and the same mark information are concatenated as features and their relationship with the response is then learned using all loci across the genome.

High-dimensionality and computational efficiency

Depending on the methods for feature extraction and model building, the prediction problem can be low-dimensional or high-dimensional. For example, when predicting expression level of a gene using a few histone modification marks at its promoter region, one may build a regression model by treating all genes as independent observations. In this case, the response is univariate, and the dimension of features P is much smaller than the number of observations G . This is a prediction problem with a ***low-dimensional response and low-dimensional features***. Alternatively, one may fit a gene-specific regression using data from multiple biological samples or cell types. This requires one to fit a large number of regression models. It is a prediction problem with ***high-dimensional responses but low-dimensional features***. For predicting TFBSs, one may use DNase-seq and histone modification footprints at each motif site as predictors. Each footprint can be a vector of read counts for a continuous run of genomic bins surrounding a genomic locus. This yields a large number of features. One can build prediction models using known TFBSs and random control sites as observations. This is a prediction problem with ***low-dimensional response and high-dimensional features***. One can also use the global gene expression profile for all genes (which contains information on the transcriptional activity of TFs, their target genes and co-activated genes) as predictors to predict TFBSs [56]. Different motif sites may be controlled by different regulators. Therefore, one needs to build site-specific prediction solutions. This is a prediction problem with ***high-dimensional responses and high-dimensional features***.

As the dimension increases, one increasingly faces the problem of the curse-of-dimensionality, that is, the sample size is not big enough to support learning the complex relationship between the responses and features. For instance a least squares regression model cannot be fit when the number of features is larger than the sample size. However, solutions may be found under certain constraints. One example is the lasso [84] method which minimizes the least squares problem subject to the sum of the absolute values of the regression coefficients being below a given threshold. Other potential solutions to this problem include dimension reduction based on the correlation structure in the responses and/or features [56], using other penalized regression methods [83], and/or using ensemble learning approaches [32].

Another problem with high-dimensionality is computation. Building models for high-dimensional features can be time-consuming. For instance, the tuning parameters for regularized regression may be chosen by cross-validation which is computationally intensive. Using ensemble learning also requires one to build a large number of models

which can be computationally slow. This problem becomes worse when both the responses and features are high-dimensional [32, 56]. While there is rich literature on how to solve a problem when either the responses or the features are low-dimensional (allowing the other to be high-dimensional), the study of high-dimensional response and high-dimensional feature problems is still in its infancy. There is great demand for new solutions that are both statistically efficient (i.e., with high prediction accuracy) and computationally efficient (i.e. fast).

Data heterogeneity and batch effects

Constrained by the capacity of instruments and resources available, high-throughput data used for building prediction models often are generated in different batches and/or by different laboratories. Lab and batch effects can confound the association between features and responses, and therefore it is important to have this unwanted variation dealt with in order to avoid ill-performing prediction models. Problems and potential solutions to batch effects are reviewed in [85]. In principle, such effects may be treated in two different ways for prediction problems. In one approach, one may model such effects and remove them from the data before using the data for prediction. In the other approach, one may build prediction models for each batch of data and then combine prediction models from different batches.

Evaluation of prediction performance

Once the prediction models are built, their performance needs to be evaluated before applying them in practice. Prediction performance can be measured in two different directions [56]. The *cross-locus prediction accuracy* measures how the predicted values accurately describe the variation of true signals across different genomic loci within the same sample. The *cross-sample prediction accuracy* measures how the predicted values accurately describe the variation of true signals across different samples at the same genomic locus. For many functional genomic data types, signals exhibit strong locus-specific effects. Because of such effects, cross-locus prediction accuracy is often higher than cross-sample prediction accuracy [56]. One can also evaluate the *overall prediction accuracy* which measures how the predicted values accurately describe the signal variation across all loci and samples.

Commonly used methods for describing prediction performance include various correlation measures (e.g., Pearson or Spearman's correlation), mean squared error, receiver operating characteristic curve (i.e., sensitivity versus 1-specificity plot), and precision-recall curve amongst others. For supervised methods, performance is often evaluated using cross-validation or similar techniques so that the training data and test data are separated and the performance is evaluated using test data [83]. In real applications, the data used to train the prediction models and the data to which the model will be applied may be generated in different labs or batches. Therefore, to accurately measure the prediction accuracy, one may need to use test data with different lab/batch origins from the training data to evaluate the performance.

Discussion

In summary, prediction methods are a useful tool in functional genomics. Despite the rapid development of technologies, conducting all types of functional genomic experiments in every sample remains extremely difficult today. Computational predictions offer a cost- and time-efficient solution complementary to experimental approaches for estimating the missing functional genomic information. The value of this approach has been demonstrated by various applications such as predicting TFBSs, epigenome, chromatin structure, etc. However, these applications only involve a subset of all functional genomic data types. Much still remains to be done in order to develop solutions to predicting other data types such as non-coding RNA, protein-RNA interaction, and protein abundance. In studies of early embryos or precious clinical samples, the amount of sample material (i.e., cell number) is limited and may not be enough to support multiple functional genomic assays. Continued efforts on developing methods that use data from one or a small number of functional genomic assays to predict all other functional genomic data types will be greatly helpful. In clinical applications, it can be difficult and unethical to collect certain sample types (e.g., brain) other than those routinely collected (e.g., blood). Therefore, how to use data from easily collectable samples such as blood to predict functional genomic signals in other parts of the body remains an important problem awaiting new solutions.

Predictions can be used to generate hypotheses and prioritize research targets. For example, if a disease-associating variant discovered by GWAS is predicted to be bound by a TF in the cell type and context relevant to the disease, it would suggest that the variant may change disease susceptibility via disrupting the TF binding. This variant can then be used to guide the design of validation experiments (e.g., CRISPR–Cas9 editing) to test the hypothesis. In this way, one can use predictions to guide subsequent studies of biological mechanisms. It is important to keep in mind that predictions can be wrong, and cannot replace biological validations which are crucial for eliminating false positives. Therefore, it is important for computational biologists to collaborate closely with experimental biologists to convert predictions into discoveries.

Besides estimating missing information, predictions can also be used to help understand how different data types are connected to each other, which could provide insights into the operating logic and program behind the dynamic gene activities. Importantly, predictions can also be used as a bridge for integrating different data types and improving data analyses, such as serving as pseudo-replicates to increase the power for signal detection or providing a reference for quality control. This idea has been used by BIRD to improve DNase-seq and ChIP-seq data analysis, and by ChromImpute for quality control. In principle, one should be able to apply this idea to many other data types and applications. For example, a disease-associating variant discovered by GWAS may or may not have functional relevance to the disease. It is often difficult to identify variants that have direct functional implications due to lack of information on the functional genomic landscape in the right context. As the GEO database contains gene expression data from diverse cell types and contexts, it might be possible to use the predicted functional genomic landscape based on the rich GEO gene expression data as a prior to moderate the association tests in GWAS studies to improve detection of functionally relevant disease-associating variants. Here, gene expression data in

GEO and GWAS data are integrated together through the predicted functional genomic landscape. For each new application, how to implement the idea will require further investigation.

Building effective prediction methods for new applications requires one to develop solutions to feature extraction, prediction model building and evaluation. With the increasing volume and complexity of the data, numerous challenges still exist and new methods need to be developed for each new prediction problem. The ultra-high-dimensionality in both the responses and features requires solutions with both statistical and computational efficiency. Prediction methods provide a way to effectively utilize the massive amounts of publicly available functional genomic data, either by using them (e.g. ENCODE data) to train prediction models or by using them (e.g., gene expression data in GEO) as predictors to quickly expand the catalog of functional genomic profiles. However, data in public repositories are highly heterogeneous. Lab and batch effects are common and should be carefully considered when building and testing prediction solutions. Importantly, public data have varying quality. Successful application of prediction methods depends not only on the amount of data but also on the data quality. Therefore, it is important to account for data quality when building prediction models. It is also important to understand how data quality affects prediction accuracy. Little research has been done regarding how to systematically integrate data quality assessment with prediction model building and application to improve prediction accuracy when applying prediction methods to massive amounts of publicly available functional genomic data. Obviously, this is also an important area that warrants future investigation. As all these challenges are addressed by innovative methods, we believe that prediction tools are capable of substantially accelerating the study of human biology and disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by grants from the National Institutes of Health (R01HG006841, R01HG006282).

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860–921. [PubMed: 11237011]
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431(7011):931–945. [PubMed: 15496913]
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995 Oct 20; 270(5235):467–470. [PubMed: 7569999]
4. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*. 1996 Dec; 14(13):1675–1680. [PubMed: 9634850]
5. Kapur K, Xing Y, Ouyang Z, Wong WH. Exon arrays provide accurate assessments of gene expression. *Genome Biol*. 2007; 8(5):R82. [PubMed: 17504534]

6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature methods*. 2008; 5(7):621–628. [PubMed: 18516045]
7. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008 Jun 6; 320(5881):1344–1349. [PubMed: 18451266]
8. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008; 453(7199):1239–1243. [PubMed: 18488015]
9. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*. 2008; 5(7):613–619. [PubMed: 18516046]
10. Wang Z, Gerstein M, Snyder M. RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; 10(1):57–63.
11. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007 Jun 8; 316(5830):1497–1502. [PubMed: 17540862]
12. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*. 2007; 4(8):651–657. [PubMed: 17558387]
13. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T, Koche RP. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*. 2007; 448(7153):553–560. [PubMed: 17603471]
14. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007 May 18; 129(4):823–837. [PubMed: 17512414]
15. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011; 98(4):288–295. [PubMed: 21839163]
16. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009; 462(7271):315–322. [PubMed: 19829295]
17. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*. 2005 Sep 28; 33(18):5868–5877. [PubMed: 16224102]
18. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*. 2006; 16(1):123–131. [PubMed: 16344561]
19. 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. *Nature methods*. 2013; 10(12):1213–1218. [PubMed: 24097267]
20. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009 Oct 9; 326(5950):289–293. [PubMed: 19815776]
21. Dekker J, Marti-Renom M, Mirny LA. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nature reviews. Genetics*. 2013 May 9; 14(6):390–403.
22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489(7414):57–74. [PubMed: 22955616]
23. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518(7539):317–330. [PubMed: 25693563]
24. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002 Jan 1; 30(1):207–210. [PubMed: 11752295]

25. Leinonen R, Sugawara H, Shumway M. International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* 2011 Jan; 39(Database issue):D19–21. [PubMed: 21062823]
26. Bird A. Perceptions of epigenetics. *Nature.* 2007; 447(7143):396–398. [PubMed: 17522671]
27. Roadmap Epigenomics Consortium. Overview of the roadmap epigenomics project. <http://www.roadmapepigenomics.org/overview>
28. Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet.* 2003; 33:245–254. [PubMed: 12610534]
29. Ouyang Z, Zhou Q, Wong WH. ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A.* 2009 Dec 22; 106(51):21521–21526. [PubMed: 19995984]
30. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. *Nature methods.* 2015; 12(3):265–272. [PubMed: 25240437]
31. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* 2011 Mar; 21(3):447–455. [PubMed: 21106904]
32. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol.* 2015; 33(4):364–376. [PubMed: 25690853]
33. Rotem A, Ram O, Shores N, Sperling RA, Goren A, Weitz DA, Bernstein BE. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol.* 2015; 33(11):1165–1172. [PubMed: 26458175]
34. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol.* 2006 Nov; 24(11):1429–1435. [PubMed: 16998473]
35. Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol.* 2002 Aug; 20(8):831–835. [PubMed: 12101405]
36. Hu S, Xie Z, Onishi A, Yu X, Jiang L, Lin J, Rho HS, Woodard C, Wang H, Jeong JS, Long S, He X, Wade H, Blackshaw S, Qian J, Zhu H. Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell.* 2009 Oct 30; 139(3):610–622. [PubMed: 19879846]
37. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpää MJ, Bonke M, Palin K, Talukder S, Hughes TR, Luscombe NM, Ukkonen E, Taipale J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* 2010 Jun; 20(6):861–873. [PubMed: 20378718]
38. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004 Jan 1; 32(Database issue):D91–4. [PubMed: 14681366]
39. Wingender E, Dietze P, Karas H, Knüppel R. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Research.* 1996 Jan 1; 24(1):238–241. [PubMed: 8594589]
40. Robasky K, Bulyk ML. UniPROBE, update 2011: Expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 2011 Jan; 39(Database issue):D124–8. [PubMed: 21037262]
41. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. hPDI: A database of experimental human protein-DNA interactions. *Bioinformatics.* 2010 Jan 15; 26(2):287–289. [PubMed: 19900953]
42. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* 2009 Jul; 37(Web Server issue):W202–8. [PubMed: 19458158]
43. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol.* 2008; 26(11):1293–1300. [PubMed: 18978777]
44. Ji H, Wong WH. Computational biology: Toward deciphering gene regulatory information in mammalian genomes. *Biometrics.* 2006 Sep; 62(3):645–663. [PubMed: 16984301]

45. Sung M, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell*. 2014; 56(2):275–285. [PubMed: 25242143]
46. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol*. 2014; 32(2):171–178. [PubMed: 24441470]
47. Won KJ, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol*. 2010 Jan 22;11(1) R7-2010-11-1-r7.
48. Ramsey SA, Knijnenburg TA, Kennedy KA, Zak DE, Gilchrist M, Gold ES, Johnson CD, Lampano AE, Litvak V, Navarro G, Stolyar T, Aderem A, Shmulevich I. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*. 2010 Sep 1; 26(17):2071–2075. [PubMed: 20663846]
49. Cuellar-Partida G, Buske FA, McLeay RC, Whittington T, Noble WS, Bailey TL. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*. 2012 Jan 1; 28(1):56–62. [PubMed: 22072382]
50. Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*. 2012 Sep 1; 22(9):1723–1734. [PubMed: 22955984]
51. Wei Y, Wu G, Ji H. Global mapping of transcription factor binding sites by sequencing chromatin surrogates: A perspective on experimental design, data analysis, and open problems. *Statistics in Biosciences*. 2012 May 8; 5(1):156–178. [PubMed: 23762209]
52. Moyerbrailean GA, Kalita CA, Harvey CT, Wen X, Luca F, Pique-Regi R. Which genetics variants in DNase-seq footprints are more likely to alter binding? *PLoS Genet*. 2016 Feb 22;12(2):e1005875. [PubMed: 26901046]
53. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Meth*. 2016 Feb 22. advance online publication.
54. Ji H, Li X, Wang QF, Ning Y. Differential principal component analysis of ChIP-seq. *Proc Natl Acad Sci U S A*. 2013 Apr 23; 110(17):6789–6794. [PubMed: 23569280]
55. Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, Yao B, Zhang F, Jin P, Wu H, Qin ZS. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res*. 2015 Mar 11; 43(5):2757–2766. [PubMed: 25722376]
56. Zhou W, Sherwood B, Ji Z, Du F, Bai J, Ji H. Genome-wide prediction of DNase I hypersensitivity using gene expression. *Cold Spring Harbor Labs Journals*. 2016 bioRxiv.
57. Zhou W, Ji Z, Ji H. Global prediction of chromatin accessibility using RNA-seq from small number of cells. *Cold Spring Harbor Labs Journals*. 2016 bioRxiv.
58. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. *Nature methods*. 2012; 9(3):215–216. [PubMed: 22373907]
59. Duong D, Zou J, Hormozdiari F, Sul JH, Ernst J, Han B, Eskin E. Using genomic annotations increases statistical power to detect eGenes. *Bioinformatics*. 2016 Jun 15; 32(12):i156–i163. [PubMed: 27307612]
60. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol*. 2015 Jan 24;16:14-015-0581-9. [PubMed: 25616342]
61. Ma B, Wilker EH, Willis-Owen SA, Byun HM, Wong KC, Motta V, Baccarelli AA, Schwartz J, Cookson WO, Khabbaz K, Mittleman MA, Moffatt MF, Liang L. Predicting DNA methylation level across human tissues. *Nucleic Acids Res*. 2014 Apr; 42(6):3515–3528. [PubMed: 24445802]
62. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*. 2013; 153(5):1134–1148. [PubMed: 23664764]
63. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. *Science*. 2013 Feb 15; 339(6121):823–826. [PubMed: 23287722]

64. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.* 2006 Mar 3.2(3):e26. [PubMed: 16520826]
65. Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A.* 2006 Jul 11; 103(28):10713–10716. [PubMed: 16818882]
66. Yuan GC. Targeted recruitment of histone modifications in humans predicted by genomic sequences. *J Comput Biol.* 2009 Feb; 16(2):341–355. [PubMed: 19193151]
67. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science.* 2002 Feb 15; 295(5558):1306–1311. [PubMed: 11847345]
68. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet.* 2006 Nov; 38(11):1341–1347. [PubMed: 17033624]
69. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 2006 Oct; 16(10):1299–1309. [PubMed: 16954542]
70. Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, Goh Y, Fullwood MJ, Sung WK, Ruan X, Ruan Y. ChIA-PET analysis of transcriptional chromatin interactions. *Methods.* 2012 Nov; 58(3): 289–299. [PubMed: 22926262]
71. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012; 485(7398): 376–380. [PubMed: 22495300]
72. Huang J, Marco E, Pinello L, Yuan G. Predicting chromatin organization using histone marks. *Genome Biol.* 2015; 16(1):1–11. [PubMed: 25583448]
73. Fortin J, Hansen KD. Reconstructing A/B compartments as revealed by hi-C using long-range correlations in epigenetic data. *Genome Biol.* 2015; 16(1):1–23. [PubMed: 25583448]
74. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* 2012 Sep; 22(9):1711–1722. [PubMed: 22955983]
75. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res.* 2011; 21(3):381–395. [PubMed: 21321607]
76. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet.* 2008 Jul; 40(7):897–903. [PubMed: 18552846]
77. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A.* 2010 Feb 16; 107(7):2926–2931. [PubMed: 20133639]
78. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, Birney E. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 2012; 13(9):R53. [PubMed: 22950368]
79. Cheng C, Yan K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.* 2011; 12(2):R15. [PubMed: 21324173]
80. He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* 2012; 22(6):1015–1025. [PubMed: 22508765]
81. Breiman L., Friedman J., Stone, CJ., Olshen, RA. Classification and regression trees. CRC press; 1984.
82. Breiman L. Random forests. *Mach Learning.* 2001; 45(1):5–32.
83. Friedman J., Hastie, T., Tibshirani, R. Springer series in statistics. Springer; Berlin: 2001. The elements of statistical learning.

84. Tibshirani, R. Journal of the Royal Statistical Society. Series B (Methodological). 1996. Regression shrinkage and selection via the lasso; p. 267-288.
85. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010; 11(10):733–739.

Figure 1. Background and rationale for prediction. **(a)** Illustration of different high-throughput technologies used in assaying the transcriptome, regulome and epigenome of a biological system. The general procedure of these technologies is: first, obtain the DNA/RNA fragments of interest; then, use sequencing or microarray technologies to read out the nucleotide sequences or abundance levels of these fragments; finally, align the sequence reads to the reference genome or analyze microarray images to obtain the genomic signals (e.g. gene expression, TF binding activity, and chromatin accessibility). **(b)** Different functional genomic signals from the same genomic region (near gene MYC) in H1 human embryonic stem cell line show that these signals are intrinsically correlated. The signals show that chromatin accessibility measured by DNase-seq and H3K27ac histone modification measured by ChIP-seq are positively correlated while they are negatively correlated with DNA methylation measured by WGBS. Transcription factor E2F6 binding

activity is positively correlated with chromatin accessibility and H3K27ac histone modification in the promoter region of MYC. Gene expression measured by RNA-seq is positively correlated with TF activity, chromatin accessibility and H3K27ac histone modification.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2.
Examples of prediction methods applied to various types of functional genomic data.

Figure 3. Summary of the challenges and common solutions for a general prediction problem in functional genomics.

Figure 4.

Illustration of different types of features. **(a)** Static or dynamic feature. DNase-seq signals are dynamic features which can be different across cell types in the same genomic locus. Motif sites are static and will not change across cell types. **(b)** Intensity or shape feature. The number of reads in a DNase I hypersensitive site (DHS) is an intensity feature. Inside a DHS, a binding transcription factor will leave a footprint, a valley-like shape, which can be seen as a shape feature. **(c)** Local or global feature. To predict expression level of a gene, one can use histone modification patterns in the promoter region of the gene which are local features. Alternatively, one may use activities of regulatory elements (e.g., TF binding activities, DNase I hypersensitivity, etc.) surrounding this gene and all its related genes (e.g., the TFs that regulate this gene, and genes that co-express with this gene) which are global features.

Figure 5.

Illustration of different strategies for building prediction models. **(a)** Supervised or unsupervised approach. A supervised approach consists of a training procedure which uses existing data from both predictors (X) and responses (Y) to train prediction models. The prediction models can then be applied to new input data (X_{new}) to make predictions (Y_{pre}). An unsupervised approach explores the new input data (X_{new}) and make predictions (Y_{pre}) based on existing knowledge. **(b)** Cross-locus or cross-sample model. A cross-locus model takes data from all genomic loci in a sample and builds a prediction model for the sample. A cross-sample model takes data of a genomic locus from different samples and builds a prediction model for the locus. **(c)** Using both cross-locus and cross-sample information for prediction. ChromImpute [32] is an example which combines both cross-locus and cross-sample information to build prediction models. It combines same sample information (cross-locus) and same mark information (cross-sample) to make predictions.