SCIENTIFIC REPORTS

OPEN

A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data

Received: 16 December 2014
Accepted: 20 April 2015
Published: 27 May 2015

Qiongshi Lu[1], Yiming Hu[1], Jiehuan Sun[1], Yuwei Cheng[2], Kei-Hoi Cheung[2,3,4,5] & Hongyu Zhao[1,2,5]

Identifying functional regions in the human genome is a major goal in human genetics. Great efforts have been made to functionally annotate the human genome either through computational predictions, such as genomic conservation, or high-throughput experiments, such as the ENCODE project. These efforts have resulted in a rich collection of functional annotation data of diverse types that need to be jointly analyzed for integrated interpretation and annotation. Here we present GenoCanyon, a whole-genome annotation method that performs unsupervised statistical learning using 22 computational and experimental annotations thereby inferring the functional potential of each position in the human genome. With GenoCanyon, we are able to predict many of the known functional regions. The ability of predicting functional regions as well as its generalizable statistical framework makes GenoCanyon a unique and powerful tool for whole-genome annotation. The GenoCanyon web server is available at http://genocanyon.med.yale.edu

Annotating functional elements in the human genome is a major goal in human genetics. Despite years of efforts from both experimental and computational scientists, functional annotation remains challenging, especially in the non-protein-coding regions. It is estimated that approximately 98% of the human genome is non-protein-coding[1]. Because of the apparent importance of coding regions, many computational tools have been developed to annotate DNA variants in the coding regions[2–4]. Although the non-coding regions were considered "junk DNA" for many years, much has been learned on the potential roles of these regions in the last decade. First, extensive comparative genomic studies have shown that the majority of mammalian-conserved regions consist of non-coding elements[5]. Second, results from genome-wide association studies show that close to 90% of the significant variants associated with human diseases reside outside of the coding regions[6], only slightly less underrepresented among all the variants in the human genome, where about 95% of known variants are from the non-coding regions. Third, high-throughput experiments, e.g. the ENCODE project[7], also suggest that a large fraction of the human genome are functionally relevant. All of this evidence suggests the importance and need for extending the annotation tools from the coding regions to the entire human genome.

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA. [2]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [3]Yale Center for Medical Informatics, Yale School of Medicine, New Haven, CT, USA. [4]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, USA. [5]VA Connecticut Healthcare System, West Haven, CT, USA. Correspondence and requests for materials should be addressed to H.Z. (email: hongyu.zhao@yale.edu)

Despite the increasing need to functionally annotate the human genome, there is no universal definition of genomic function[8,9], which differs among geneticists, evolutionary biologists, and molecular biologists. The experimental approaches and analysis techniques of detecting functional genomic elements among these scientists also vary greatly. Extensive work in some genomic regions such as the β-globin gene complex has shown that no single approach is sufficient to identify all the regulatory activities in the non-coding regions[8,10]. In order to obtain a comprehensive picture of the genomic functional structure, all the valuable information acquired through different approaches needs to be combined using appropriate statistical learning techniques.

Several annotation tools focusing on the non-coding regions have been established recently[11–15]. Similar to the long list of deleteriousness prediction tools developed for the coding regions, most of these new methods aim to distinguish tolerable variants from the deleterious ones. Though important, prediction of deleteriousness does not cover every aspect of functional annotation. The potential of these variant classifiers in understanding the genomic architecture on a large scale and in detecting regulatory elements such as cis-regulatory modules remains to be thoroughly investigated. Moreover, scientists now routinely analyze different cell types[7], and even single cells[16]. In order to keep up with these technological advances, it is critical to develop a functional annotation framework that can be generalized to different species, cell types, and single cells. Such a generalizable framework can be achieved through biologically-motivated and statistically-justified models. As for choosing between a supervised approach, where some gold standard datasets are needed to train the model, and an unsupervised approach, where no labeled data are used, we focus on developing an unsupervised learning method in this article. This is because current supervised-learning-based annotation tools suffer from highly biased training data, which is largely due to our limited knowledge of non-coding regions. This may become less of an issue after we have gained a deeper understanding of non-coding functional mechanisms. However, at such an early stage, we think unsupervised learning techniques would be advantageous.

In this paper, we present GenoCanyon (inspired by the canyon-like plots it generates), a whole-genome annotation tool based on unsupervised statistical learning. From a collection of the comparative genomic conservation scores and biochemical signals obtained from the ENCODE project[17], the posterior probability of a genomic position being functional is used as the prediction score. Compared to existing methods, GenoCanyon not only measures the deleteriousness of variants, but also the functional potential of each genomic location. Its flexible and generalizable statistical framework could also benefit future applications.

Results

Estimating the Proportion of Functional Regions in the Human Genome. Genetic approaches that focus on studying the consequences of genetic perturbations are often referred to as a gold standard for defining function[8]. Such a genetic definition is also directly related to causal inference, which is at the core of developmental biology and disease research[9]. In this study, we also adopt this genetically meaningful definition of genomic function. On the other hand, we treat the conservation measures and the biochemical signals as consequences of genomic function (Fig. 1A). For a specific location in the human genome, define Z to be the latent indicator of function. We collected 22 different annotations, denoted as $\boldsymbol{A}$ (Supplementary Table 1). We also assumed that the 22 annotations are conditionally independent given Z (Fig. 1B). Then, the posterior probability $P(Z = 1|\boldsymbol{A})$ serves as the prediction score of the functional potential at this location (See Methods, Fig. 1C).

We have pre-calculated the prediction scores for the entire human genome (hg19). Overall, when using 0.5 as the cutoff for defining functionality, 33.3% of the human genome was predicted to be functional. The proportion of functional elements is mostly stable across chromosomes (Supplementary Table 2; Supplementary Figure 1). We note that the functional proportion of the human genome has been estimated using many different approaches[8,18–22] and results differed drastically. Comparative genomic analysis of multiple mammals revealed that constrained elements consist of approximately 4.5% of the human genome[18,19]. At the other extreme, the ENCODE project found that 80% of the human genome has detectable biochemical activities in at least one cell line[7]. However, it has been discussed recently that several corrected constraint estimations would each suggest two to three times increase to the original estimate of 4.5%[20–22]. Also, it still remains non-trivial to distinguish real biochemical signals from biological noises in the ENCODE data[8]. The large amount of observed biochemical activities have also been criticized to be more like an "effect" rather than "function"[9]. Our prediction falls in the middle of these highly diverging estimates of functional regions in the literature. It is worth noting that the GenoCanyon functional prediction represents a mixed probability involving multiple tissues. A smaller proportion of the human genome would be expected to be functional for a particular tissue.

Prediction for cis-regulatory Modules in the HBB Gene Complex. The intensively studied β-globin (HBB) gene complex on chromosome 11 contains embryonically expressed HBE1, fetally expressed HBG1 and HBG2, and adult globin genes HBD and HBB, along with a pseudogene HBBP1. This locus is known to provide a paradigm for developmental gene expression and regulation[23,24]. A large number of cis-regulatory modules (CRMs) that control both the developmental timing and the spatial pattern of gene expression have been discovered in the HBB gene complex[10]. More interestingly, the

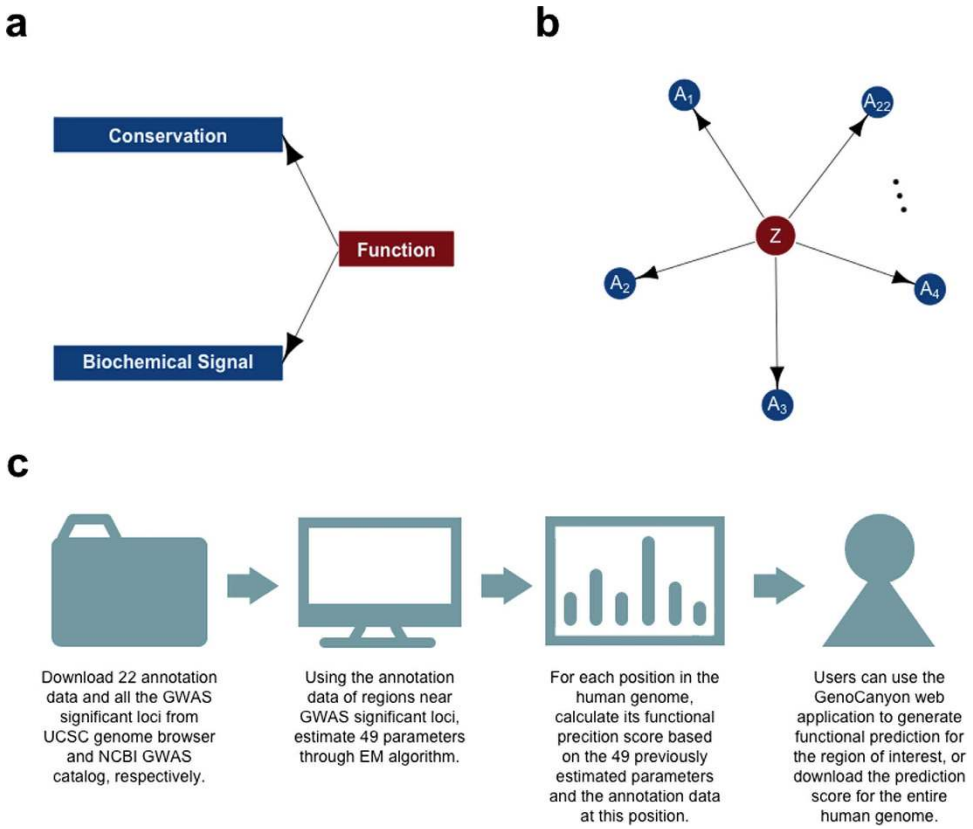

Figure 1. Modeling of causal relationship among variables. (**a**) We adopt the biologically meaningful definition of function, and treat conservation measures and biochemical signals as consequences. (**b**) The latent functional indicator Z is modeled as the parental variable and all the 22 annotations are treated as consequences. Also, we assume there is no direct causal relationship between any two annotations. Therefore the annotations are conditionally independent given Z. (**c**) Workflow of GenoCanyon functional prediction.

epigenetic and evolutionary signals at these CRMs differ substantially[8]. Therefore, the HBB gene complex provides a perfect example to test if GenoCanyon could effectively combine different sources of signals and successfully predict the functional segments.

We analyzed the prediction results in the HBB gene complex. On the entire chromosome 11, 32.2% of the DNAs were predicted to be functional. Strong enrichment of signals was observed at this locus. Using 0.5 as the cutoff, 62.2% of HBB gene complex and 97.0% of the CRMs were predicted as functional (Fig. 2A). Remarkably, a cluster of five DNase I hypersensitive CRMs upstream of the HBB gene complex, known as the locus control region (LCR)[25], showed strong functional signals as a whole (Fig. 2B). The 3'HS1 enhancer blocker (chr11: 5226013-5226493; hg19) downstream of the HBB gene complex was also successfully predicted with high resolution. Interestingly, these CRMs showed highly variable patterns of annotations (Fig. 2C). This proved that GenoCanyon could effectively combine different sources of information. Recent research revealed several new regulatory elements at this locus, including one in the intergenic region between HBBP1 and HBG1[24], and another one upstream of HBD[23]. These elements also reside in the highly scored regions. Moreover, it is worth noting that the understanding of CRMs is still incomplete even in a relatively well-studied region such as the HBB complex. Some of the apparent false positives might actually be regulatory elements not yet discovered. The functional regions provided by our method could potentially offer a guideline for further studies.

Among the 23 CRMs being reviewed[10], only the promoter of HBB did not get the perfect score (Table 1). Therefore, we analyzed the HBB gene and its promoter in more details (Fig. 2D). Within the HBB gene, the 600 bp segment near the 3'UTR was predicted to be functional. 77 pathogenic or likely pathogenic SNPs were downloaded from the NCBI Variation Viewer (http://www.ncbi.nlm.nih.gov/variation/view/). Interestingly, 14 of these pathogenic SNPs, including 4 in the 3'UTR and 6 in the second intron, lie in this 600 bp functional segment. In the upstream half of the second intron, prediction scores were substantially lower. No pathogenic variants could be found in that region. Overall, using 0.5 as the cutoff, 89.6% (69 out of 77) of the pathogenic SNPs located at functional locations. Within the HBB promoter, 75% (6 out of 8) of the pathogenic variants located at functional locations (Fig. 2E). Moreover, bumps of high scores could be observed at the known protein binding sites in the HBB promoter[26]. When comparing the entire HBB promoter, known protein binding sites, and the pathogenic

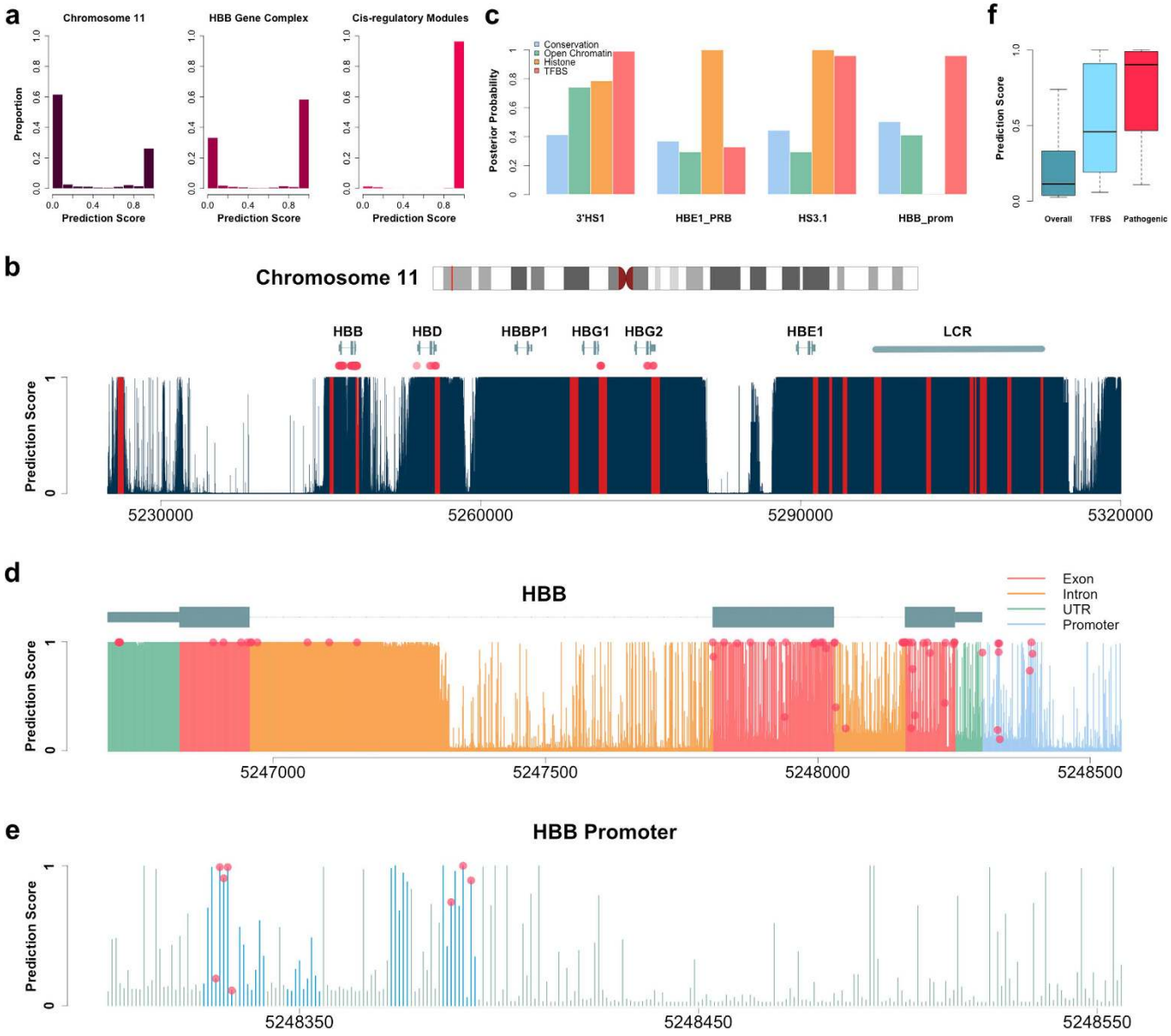

Figure 2. Functional prediction for the HBB gene complex. (**a**) Histogram of the prediction scores in chromosome 11, HBB gene complex, and the 23 CRMs. 32.2%, 62.2% and 97.0% are predicted as functional, respectively. (**b**) Prediction results for the HBB complex. Dark blue bars show the prediction score at each location. All the 23 CRMs are marked in red. There appears to be fewer than 23 red bars because some of the CRMs are very close to each other. Red dots indicate the locations of known pathogenic SNPs downloaded from the NCBI Variation Viewer. (**c**) The posterior probabilities given a single group of annotations could be used to measure the relative contribution of different sources of information (See **Methods**). Four CRMs are plotted to illustrate that prediction scores are driven by different annotations in different CRMs. (**d**) Prediction results for the HBB gene and its promoter. The promoter, UTRs, introns and exons are marked with different colors. Red dots show the prediction scores of the pathogenic variants. (**e**) Prediction results for the HBB promoter. Known protein binding sites in the HBB promoter are marked in blue. Red dots show the prediction scores of the pathogenic variants. (**f**) Boxplot of the prediction scores of HBB promoter, known protein binding sites, and pathogenic variants.

variants within the promoter, there was a substantial increase in prediction score (Fig. 2F). All of this evidence suggests that important functional segments could still be detected locally even in a generally lower-scored region.

Prediction for ZRS, an enhancer of the SHH gene. Zone of polarizing activity regulatory sequence (ZRS) is one of the most studied developmental enhancers. It is located in the fifth intron of

Name	Start	Stop	Mean Score
HS5	5312534	5312694	1.000
HS4	5309419	5309707	1.000
HS3.2	5306814	5307392	1.000
HS3.1	5306356	5306418	1.000
HS3	5305882	5306169	1.000
HS2_neg	5302090	5302174	1.000
HS2_pos	5301795	5302089	1.000
HS1	5296894	5297517	1.000
HBE1_NRA	5294082	5294308	1.000
HBE1_PRA	5293982	5294081	1.000
HBE1_NRB	5292886	5292928	1.000
HBE1_PRB	5292690	5292886	0.999
HBE1_up	5291344	5291610	1.000
HBE1_prom	5291175	5291343	1.000
HBG2_up	5276215	5276745	1.000
HBG2_prom	5276011	5276214	1.000
HBG1_up	5271291	5271813	0.999
HBG1_prom	5271086	5271290	1.000
HBG1_3'enh	5268365	5269114	1.000
HBD_prom	5255713	5256160	1.000
HBB_prom	5248301	5248556	0.253
HBB_3'enh	5245876	5246140	1.000
3'HS1	5226013	5226493	0.987

Table 1. Mean prediction scores of the known CRMs in the HBB gene complex. *Coordinates are based on hg19.

the protein-coding gene LMBR1, approximately 1 Mb upstream of SHH's transcriptional start site[27,28]. Through linkage mapping of several large families with preaxial polydactyly (PPD) and triphalangeal thumb, an associated locus of approximately 500 kb was identified on chromosome 7q36. In later studies, the region was further narrowed down to the fifth intron of LMBR1[29–31]. As a highly conserved 774 bp region in this intron, ZRS has been intensively studied. It has been shown to be crucial for limb development not only in humans, but also in mice, dogs, cats, and even chickens[27].

We investigated the prediction results in gene LMBR1. A highly scored plateau could be observed in its fifth intron (Fig. 3A). The mean predicted score for this intron was 0.595. This was higher than the mean score of the entire LMBR1 transcript (0.385), of all the introns in LMBR1 (0.384), and even of all the exons in LMBR1 (0.448). These results showed strong signs of function in the fifth intron. The ZRS region got an even higher mean predicted score 0.871, which confirmed its importance (Fig. 3B). When observing its surrounding region, ZRS could be easily identified as a dense region with high prediction scores (Fig. 3C). Moreover, the ZRS region serves as one of the most well studied examples for pathogenic variants in an enhancer. A total of 13 single nucleotide variants in ZRS have been identified to cause human limb malformations[27]. All these 13 SNVs were predicted to be highly functional, with the mean prediction score 0.987 (Fig. 3D).

In conclusion, our method successfully identified the fifth intron of LMBR1 as a functional region. It also further confirmed the importance of ZRS. It is notable that the large number of identified pathogenic variants in ZRS is possibly subject to the ascertainment bias. In fact, mutations in ZRS did not account for the limb malformation in all the studied families[32]. Our prediction in the surrounding regions has the potential to guide future studies.

Prediction for Functional Elements in the Human X-inactivation Center. X-chromosome inactivation, originally described 50 years ago[33], is the mechanism for X-chromosome dosage compensation in mammals. The long non-coding RNA Xist has been shown to be both necessary and sufficient to induce X-chromosome inactivation in mouse ES cells[34]. The surrounding genomic region, often referred to as the X-inactivation center (Xic for mouse and XIC for human), contains several crucial regulatory elements for mouse X-inactivation[35]. However, recent studies have suggested the existence of substantial variations in the mechanism of achieving X-inactivation among species[36–39]. We applied GenoCanyon on

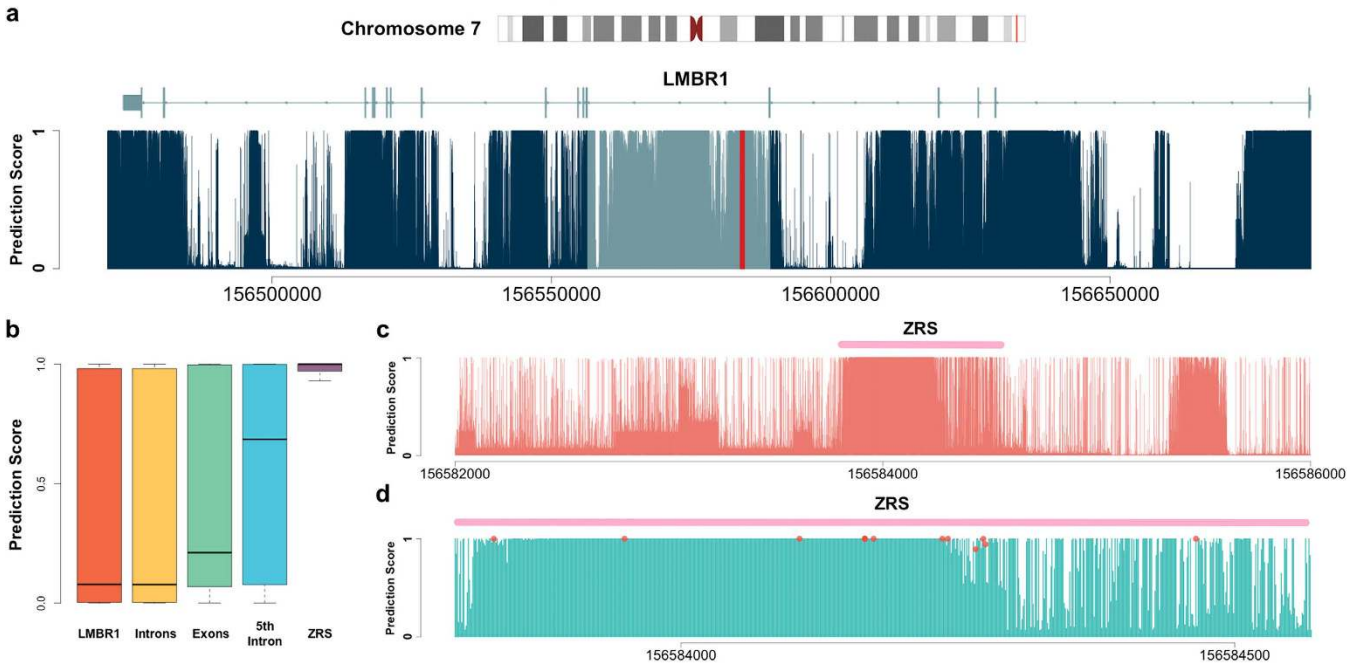

Figure 3. Prediction results for the SHH enhancer in LMBR1. (**a**) Prediction scores in the LMBR1 gene. The fifth intron and ZRS are highlighted in light blue and red, respectively. (**b**) Boxplot of the prediction scores in LMBR1, 16 introns, 17 exons, the 5th intron, and ZRS. The results highlighted the function in the 5th intron of LMBR1 and confirmed the importance of ZRS. (**c**) Prediction results for the surrounding region of ZRS, which is highlighted in pink. An obvious highly scored plateau can be observed at ZRS. (**d**) The prediction results within the ZRS. 13 pathogenic variants are discovered in ZRS. The predicted scores at their locations are marked with red dots. There appears to be only 11 dots because three variants all reside at location 156584166 (hg19).

human XIC to predict the functional potential of the orthologs of known regulatory elements in mouse models (Fig. 4A; Table 2).

Xist and its antisense ncRNA Tsix, as well as two upstream ncRNAs Ftx and Jpx have all been shown to haven cis-regulatory roles in mouse X-inactivation[40–43]. Our prediction confirmed the function of the master ncRNA XIST in human. Both the XIST gene and its transcribed regions got nearly perfect prediction scores. Moreover, a XIST-specific peak of high score could be observed on Fig. 4A, showing satisfying resolution of prediction. Studies suggesting a truncated form of TSIX in human have led to some debate in its function. Compared to its mouse ortholog, the human TSIX gene has lost the CpG island as well as the enhancer elements Dxpas34 and Xite[44]. In our prediction, TSIX got mean score 0.383, which is low for such an active genomic region. When considering only the region that does not overlap with XIST, the number even dropped to 0.197. A recently discovered lncRNA, Linx, has been hypothesized to take part in Tsix expression in mice[45]. In the mouse genome, the Linx gene lies between two protein-coding genes Nap1l2 and Cdx4. However, its human ortholog has not yet been discovered. The intergenic region between human NAP1L2 and CDX4 has a low mean prediction score 0.016, which argues against not only the existence of LINX in human, but also TSIX function. Jpx and Ftx both showed the potential to activate X-inactivation in mice[42,43]. But the functions of their human orthologs have not been studied[37]. The mean prediction scores of the transcribed regions in JPX and FTX are 0.256 and 0.304, respectively. This suggested only moderate functional potential of these two human lncRNAs. However, both scores received a substantial boost when the entire gene was considered. On Fig. 4A, several functional peaks could also be clearly observed in the untranscribed regions in JPX and FTX. These results might guide the detection of novel regulatory elements in human XIC.

Besides the mentioned lncRNA genes, the human XIC also contains 6 protein-coding genes, NAP1L2, CDX4, CHIC1, ZCCHC13, SLC16A2, and RLIM. It is notable that most of their exons clearly reside in the functional peaks in Fig. 4A, showing the ability of GenoCanyon to capture the functional landscape of this genomic region. We calculated the mean prediction scores for all the RefSeq transcripts of these genes. In CDX4, CHIC1, and SLC16A2, all the transcript scores were substantially larger than the scores of untranscribed regions. Among the 6 protein-coding genes, Rlim (also referred to as Rnf12) produces the U3 ubiquitin ligase that acts in a dose-dependent manner on the initiation of X-inactivation[46]. The human RLIM gene has a high mean predicted score 0.973. Two of its RefSeq transcripts both got 0.930 as

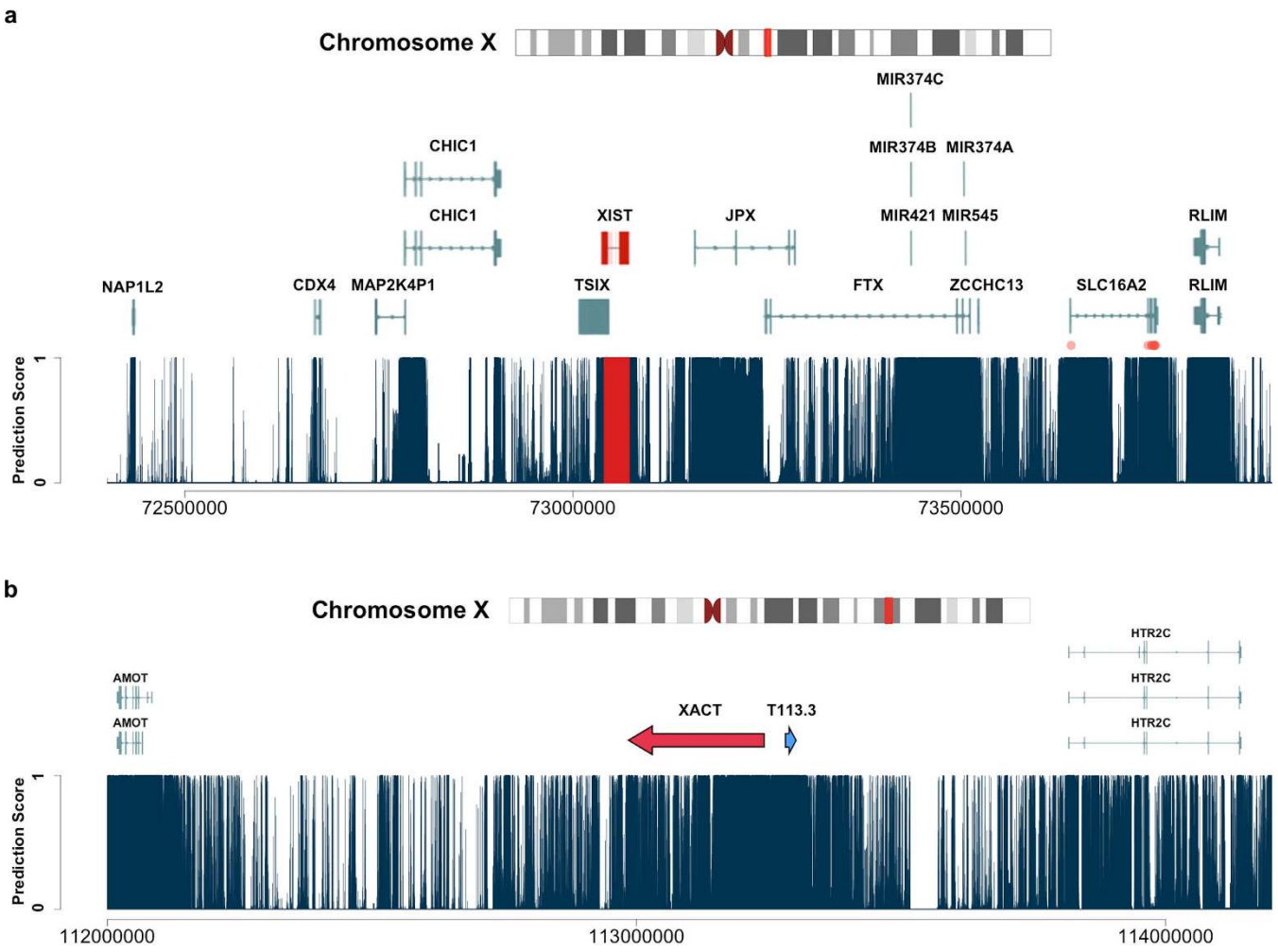

Figure 4. Prediction results for regions involved in human X-inactivation. Each dark blue line shows the prediction score at a single base. (**a**) Functional prediction for the human XIC. All the RefSeq transcripts in this region are plotted. The master lncRNA XIST is highlighted in red. Red dots show the locations of known pathogenic variants downloaded from the NCBI variation viewer. (**b**) Functional prediction for the intergenic region between AMOT and HTR2C on chromosome Xq23. A red and a blue arrow represent the recently discovered transcripts XACT and T113.3, respectively.

the mean score, which is also very high. In Fig. 4A, the RLIM gene perfectly lies in an isolated functional plateau, which suggests its strong functional potential in human. It has been observed that the homologous pairing of two regions (Tsix/Xite and Xpr/Slc16a2) might have impacts on Xist upregulation[47,48]. In human XIC, the TSIX/XITE region has been truncated, but the region surrounding the SLC16A2 gene showed its functional potential in our prediction. The exons of SLC16A2 lie in two large separate functional peaks, suggesting the importance of the transcribed region as well as a large bulk of untranscribed region in SLC16A2. Whether these regions serve as the human XPR remains to be investigated. More interestingly, 8 pathogenic SNPs in SLC16A2 have been submitted to ClinVar[49]. These variants were believed to be involved in Allan-Herndon-Dudley syndrome, showing that SLC16A2 has its crucial function in other processes as well. The other genes in Xic have not been related to X-inactivation yet. Our prediction suggested that the exons of NAP1L2, CDX4, and CHIC1 all showed different levels of functional potential, which is not surprising because of their protein-coding nature. The human XIC also contains several microRNA genes and one pseudogene MAP2K4P1. MAP2K4P1 did not get a high score, which was in agreement with its pseudogene status. The microRNA transcript might partially explain the large functional plateau near the 5' end of FTX.

XACT, a recently discovered lncRNA coating the active X chromosome in human pluripotent cells, has been shown to take part in X-inactivation initiation uniquely in human[37,38]. It lies in a 1.7 Mb large intergenic region between protein-coding genes AMOT and HTR2C. A shorter transcript T113.3 upstream of XACT was also identified. But its function has not been studied. We investigated this region using GenoCanyon. The AMOT gene and the HTR2C exons both showed substantial functional

Gene Name	Gene Mean Score	Transcript ID	Transcript Mean Score
NAP1L2	0.988	NM_021963.3	0.988
CDX4	0.234	NM_005193.1	0.554
CHIC1	0.229	NM_001039840.2	0.931
		NM_001300884.1	0.955
ZCCHC13	0.184	NM_203303.2	0.184
SLC16A2/XPCT	0.575	NM_006517.4	0.951
RLIM/RNF12	0.973	NM_016120.3	0.930
		NM_183353.2	0.930
XIST	0.999	NR_001564.2	0.998
TSIX	0.383	NR_003255.2	0.383
JPX	0.501	NR_024582.1	0.256
FTX	0.438	NR_028379.1	0.304
MAP2K4P1	0.161	NR_029423.1	0.095

Table 2. Prediction results for the 6 protein-coding genes, 4 lncRNA genes, and 1 pseudogene in the human XIC, as well as all their transcripts in RefSeq.

potential. A clear plateau of high scores could also be observed in the intergenic domain (Fig. 4B). The mean prediction score for the entire intergenic region, XACT, and T113.3 were 0.148, 0.383, and 1.000, respectively. Although the mean predicted score for XACT was only moderate, it still confirmed the functional signal in such a lowered-scored intergenic domain. Also, our prediction suggested the importance of T113.3 and its surrounding region.

Investigating the Ability of Classifying Variants. GenoCanyon was not designed as a variant classifier. However, enrichment in prediction score is still expected for the known pathogenic variants. We downloaded all the annotated variants from ClinVar in June 2014[49]. The subset of single nucleotide variants annotated as "Pathogenic", "Likely Pathogenic", or "Pathogenic/Likely Pathogenic" was treated as the positive set. Similarly, the subset of SNVs annotated as "Benign", "Likely Benign", or "Benign/Likely Benign" was treated as the negative set. The positive set contained 19,242 variants, and the negative set contained 8,874 variants. The mean prediction score in the positive set and the negative set were 0.912 and 0.735, respectively. When using 0.5 as the natural cut-off, the sensitivity was as high as 0.915, with a low specificity of 0.263. The AUC was 0.727.

It is worth noting that GenoCanyon measures the functional potential of genomic locations, not the tolerability of specific variants. The transcribed regions in a crucial protein-coding gene should be expected to have a high functional score. However, it would still be natural to observe many tolerable synonymous SNPs in that gene. All these tolerable SNPs become "false-positives" in the analysis above, leading to a low specificity. Moreover, many of the known "benign" variants are by-products of association studies. Their properties were investigated because they lie in candidate regions in the disease pathway, which explains why the mean prediction score of benign variants was also high. On the other hand, if a variant were shown to be pathogenic in experiments, the underlying region would surely have some functions related to the disease. In this sense, the high sensitivity of GenoCanyon suggests that it may be a good indicator of its prediction ability. Finally, the performance of supervised-learning-based methods is highly sensitive to the choice of training data. For example, when using common variants with matched regions as the negative training set, the performance of GWAVA on its own training data dropped substantially (AUC = 0.71)[14].

Discussion

The HBB gene cluster, ZRS, and the X-inactivation center all have been paradigms for studying the complex genomic regulatory network. The prediction results in these regions showed that GenoCanyon is capable of detecting functional regions in the human genome, which is a unique feature most existing whole-genome annotation tools do not have. With the wide adoption of next-generation sequencing, GenoCanyon may help researchers focus on candidate regions that are likely to be functional and reduce the spurious signals among the overwhelming genomic information.

Throughout this article, we have discussed the differences between GenoCanyon and variant classifiers in that GenoCanyon measures the functional potential of genomic locations instead of the pathogenicity of a specific variant and a high score does not necessarily imply deleteriousness. However, in some scenarios that variants distribute across the entire genome, GenoCanyon may still serve well as a conservative tool for noise reduction. For example, sequencing technology is rapidly becoming a focus

of efforts in genomic epidemiology. However, the overwhelming number of rare variants in the human genome brings the issue of extreme multiple testing. It has been discussed recently that the sample size required for a well-powered RVAS (rare variants association study) using sequencing is similar to that of a traditional GWAS (genome-wide association study)[50]. Without a huge cohort, true signals could be easily overshadowed by extreme yet spurious observations. In this case, GenoCanyon could be used to filter the SNPs and reduce 2/3 of the tests as more than 2/3 of the human genome is less likely to be functional. Moreover, the high sensitivity of GenoCanyon ensures that the true signal is still kept in the dataset. The ability of predicting functional potential at each nucleotide is another useful feature of GenoCanyon. In association studies, genetic variants are used as markers capturing signals for nearby regions. Therefore, for each SNP, the mean prediction score for its surrounding region may serve well as a prior in post-GWAS prioritization. Existing variant classifiers cannot achieve this task because they only predict the deleteriousness of genotyped variants. It is worth noting that most of the annotation data have a resolution ranging from tens to hundreds of nucleotides due to the limitation of current experimental techniques. However, data input of these annotations is at nucleotide level, which makes it possible to measure the functional potential for each base pair.

Based on unsupervised learning, GenoCanyon does not suffer from the highly biased knowledge of the non-coding DNA. More importantly, the model can be generalized in many directions. Firstly, the ENCODE annotations used in GenoCanyon were clustered across several or even nearly a hundred different cell lines. Therefore, the current functional regions predicted by GenoCanyon are in fact the union of functional elements in different cell types. Using the annotations for one single cell type, a cell type-specific functional prediction tool could be built under the same framework. In studies where several candidate cell types are of interest, prediction based on the cell-type-specific models would have higher specificity. Secondly, the model can be extended to other species. The functional elements in model organisms are generally better studied. Such tools for different species could potentially benefit the multi-species comparison and help detecting functional orthologs in human. Thirdly, in order to simplify the model, we transformed the biochemical annotations into binary variables (See Methods). Therefore, the information of signal strength has not been used. When these information as well as more annotations are incorporated using more complex modeling techniques, the specificity may be improved. Finally, the current model assumes the leading role of genetic function, and treats conservation measures and the biochemical signals as consequences. Among different annotations, conditional independence was also assumed (Fig. 1). However, it would be interesting to investigate the correlations among variables in either the functional or the non-functional group. In that case, statistical graphical models could be implemented to make the model more flexible. These are all very interesting directions to generalize GenoCanyon. However, complex models lead to higher variance, intensive computation, and less interpretability. Dealing with these trade-offs has never been trivial. The good prediction results show that GenoCanyon has reached a nice balance. The current powerful features as well as its generalizable potential make GenoCanyon a unique and useful tool for whole-genome annotation.

Methods

Statistical Model. For each location in the human genome, define Z to be the latent indicator of function, where Z = 1 indicates that location is functional and 0 otherwise. We selected 22 different annotations corresponding to either conservation score or biochemical activity, including 2 genomic conservation measures, 2 indicators of open chromatin, 8 histone modifications, and 10 TFBS peaks (Supplementary Table 1). These annotations are selected because their functional impacts are relatively well studied and easier to model. DNA methylation is not included in the model because the gene silencing mechanism requires modeling the functional impact of methylation to other nucleotides that are possibly far away, which is a challenging task. Genomic data for all the 22 annotations were downloaded from the UCSC Genome Browser except GERP (Supplementary Table 3). We denote the vector of all the annotations as $\boldsymbol{A}$.

$$\boldsymbol{A} = (A_1,\ A_2,\ \ldots,\ A_{22}). \tag{1}$$

When a genomic location is functional (Z = 1), we assume that the annotations have a joint probability density $f(\boldsymbol{A}|Z = 1)$; similarly, when a genomic location is non-functional (Z = 0), we assume that the annotations have another joint density $f(\boldsymbol{A}|Z = 0)$. Since Z is unknown, the distribution of the observed data would be a mixture of $f(\boldsymbol{A}|Z = 1)$ and $f(\boldsymbol{A}|Z = 0)$. Instead of modeling direct causal relationships among these 22 annotations, we assume that they are connected only through Z. In other words, the 22 annotations are all modeled to be consequences of Z. Under these assumptions, the 22 different annotations are conditionally independent when Z is given[51]. Therefore, the conditional joint density of $\boldsymbol{A}$ given Z can be factorized as

$$f(\boldsymbol{A}|Z = c) = \prod_{i=1}^{22} f_i(A_i|Z = c)\ ,\ c = 0, 1 \tag{2}$$

Finally, for a genomic location, assume π to be the prior probability of being functional, i.e.

$$\pi = P(Z = 1) \tag{3}$$

Then, given the annotations, the posterior probability of Z = 1 can be used as a reasonable functional measure when the parameter estimates are plugged in.

$$\begin{aligned} P(Z = 1|\boldsymbol{A}) &= \frac{\pi f(\boldsymbol{A}|Z = 1)}{\pi f(\boldsymbol{A}|Z = 1) + (1 - \pi) f(\boldsymbol{A}|Z = 0)} \\ &= \frac{\pi \prod_{i=1}^{22} f_i(A_i|Z = 1)}{\pi \prod_{i=1}^{22} f_i(A_i|Z = 1) + (1 - \pi) \prod_{i=1}^{22} f_i(A_i|Z = 0)} \end{aligned} \tag{4}$$

We chose GERP[52] and PhyloP[53] as the conservation measures because both of them are approximately normally distributed and therefore easier to model. PhyloP46way was chosen instead of PhyloP100way because a large phylogenetic distance would bring too little conserved signal as well as many incomplete data. All the other annotations were cell-type-specific, so we coded them into binary variables to cluster the signal across cell lines. If signal was detected in at least one cell line, we coded the corresponding $A_i = 1$. Otherwise, $A_i = 0$. For DNase I, FAIRE, and TFBS, there were downloadable cluster files on the UCSC Genome Browser. A total of 125, 25, and 91 cell lines were clustered, respectively. We made our own histone peak cluster files across 16 cell lines from the Broad histone track on ENCODE (Supplementary Table 4). The 8 histone modifications were chosen because they are relatively well-studied[54]. We chose the top 10 Transcription Factors with the highest binding site coverage after being transformed into binary variables.

Finally, normal distribution and Bernoulli distribution were used to model the continuous and binary annotations, respectively.

$$f_i(A_i|Z = c) = \frac{1}{\sqrt{2\pi}\,\sigma_{ic}} \exp\left(-\frac{(A_i - \mu_{ic})^2}{2\sigma_{ic}^2}\right),\ i = 1, 2;\ c = 0, 1 \tag{5}$$

$$f_i(A_i|Z = c) = p_{ic}^{A_i}(1 - p_{ic})^{1-A_i},\ i = 3, \ \ldots, \ 22;\ c = 0, 1 \tag{6}$$

Estimation

In total, our model has 49 parameters.

$$\boldsymbol{\Theta} = (\pi, \boldsymbol{Q_1}, \boldsymbol{Q_0}, \boldsymbol{P_1}, \boldsymbol{P_0}) \tag{7}$$

where

$$\boldsymbol{Q_c} = (\mu_{1c}, \mu_{2c}, \sigma_{1c}, \sigma_{2c}),\ c = 0, 1 \tag{8}$$

$$\boldsymbol{P_c} = (p_{3c}, p_{4c}, \ \ldots, p_{22,c}),\ c = 0, 1 \tag{9}$$

The GWAS Catalog[55] was downloaded from the NHGRI GWAS Catalog website (http://www.genome.gov/gwastudies/) in July 2014. It contained 13,070 unique SNPs that were significant in GWAS studies. For each SNP, we marked the interval between its 500 bp upstream and 499 bp downstream. In this way, 13,070 intervals were collected. Each interval spanned 1 kbp. After deleting the overlapping coordinates, the entire region spanned 12,801,840 bp. Each significant SNP in the GWAS Catalog hints the existence of functional elements nearby. These functional elements differ in their sizes and in the distance to the probed SNP. Since each interval was 1,000 bp in length and a large number of intervals were collected, the whole collection was a large enough and reasonably chosen set on which we could learn the distributions of annotations in both functional and non-functional groups. All the 22 annotations were then collected at each location in this set. The PhyloP scores and GERP scores were not available at 221,643 and 28,741 locations, respectively. After removing these locations, the final dataset contained 12,580,197 genomic locations. None of the other annotations have the issue of incomplete data. Finally, the Expectation-Maximization (EM) algorithm was used to estimate the parameters. As expected, the estimates showed solid differences between the functional and non-functional groups (Supplementary Table 5). We also tried replacing the missing conservation measures with the neutral score 0. Then the entire 12,801,840 locations were used to estimate the parameters. Little differences in parameter estimates were observed between the two approaches (Supplementary Table 6). Moreover, in order to test if the estimates are stable under different choices of datasets, we randomly sampled two subsets on chromosome 1, containing 2,000,000 and 6,000,000 bp, respectively. After adding these locations into the original 12,801,840 bp dataset, the parameters were estimated using the EM algorithm again. No substantial differences were observed in the estimates (Supplementary Tables 7 and 8). Based on these

results, the GWAS-loci-based dataset containing 12,801,840 bp seems to contain enough functional elements for accurate parameter estimation, and is general enough so that genome heterogeneity does not have a strong impact on estimation. Finally, in order to check the sensitivity of our model to the perturbation in annotation data, we re-fitted the model multiple times after removing several annotations (Supplementary Table 9). The parameter estimates remained consistently stable in all these cases, suggesting that the framework we propose is robust to the choice of annotations. The stable estimates of marginal parameters also show that the potential correlations among annotations do not have a strong impact on model fitting.

Marginal Effect of Different Annotations. For each binary annotation A_j (j = 3, …, 22), its effect on the final prediction can be measured using the odds ratio.

$$\begin{aligned}\mathrm{OR}_j &= \frac{P\left(Z=1\middle|A_1,\ldots,A_j=1,\ldots,A_{22}\right)}{P\left(Z=0\middle|A_1,\ldots,A_j=1,\ldots,A_{22}\right)} \times \frac{P\left(Z=0\middle|A_1,\ldots,A_j=0,\ldots,A_{22}\right)}{P\left(Z=1\middle|A_1,\ldots,A_j=0,\ldots,A_{22}\right)} \\ &= \frac{f_j\left(A_j=1\middle|Z=1\right)}{f_j\left(A_j=1\middle|Z=0\right)} \times \frac{f_j\left(A_j=0\middle|Z=0\right)}{f_j\left(A_j=0\middle|Z=1\right)} \\ &= \frac{p_{j1}\left(1-p_{j0}\right)}{p_{j0}\left(1-p_{j1}\right)}, j=3,\ldots,22\end{aligned} \tag{10}$$

We calculated the odds ratios for all 20 binary annotations (Supplementary Table 5). The annotation with the least effect was the histone modification H3K27me3. According to our estimation, the probabilities to detect the H3K27me3 signal in functional and non-functional classes are almost the same (0.80 and 0.72). In fact, H3K27me3 has been discovered to be associated with Polycomb-repressed regions[56,57], which could partially explain the phenomenon. All the other binary annotations showed variable yet substantial signals of function. The marginal effect of a continuous annotation depends on its value. The interpretation is also less straightforward. More importantly, although these statistics could help us gain some intuition of how each annotation works marginally, the final prediction relies on all of them. The effectiveness of the method needs to be tested as a whole.

In order to visualize the relative contribution of different sources of information (Fig. 2C), posterior probabilities given a particular group of annotations were calculated for each location.

$$\mathrm{P}(\text{Functional} \mid \text{Conservation}) = P(Z=1|A_1, A_2) \tag{11}$$

$$\mathrm{P}(\text{Functional} \mid \text{Open Chromatin}) = P(Z=1|A_3, A_4) \tag{12}$$

$$\mathrm{P}(\text{Functional} \mid \text{Histone}) = P(Z=1|A_5, \ldots, A_{12}) \tag{13}$$

$$\mathrm{P}(\text{Functional} \mid \text{TFBS}) = P(Z=1|A_{13}, \ldots, A_{22}) \tag{14}$$

Then, for each CRM, the mean posterior probabilities were plotted.

Estimating the Functional Proportion. After plugging in the parameter estimates, the prediction score could be calculated using formula (4). If the PhyloP or the GERP score was not available, the neutral value 0 was used. Using the cutoff 0.5, 33.3% of the human genome was predicted to be functional. However, it is notable that the EM algorithm also gave an estimate for the functional proportion, 42.7% in our case (Supplementary Table 5). This estimation was based on the 12,580,197 locations we chose, which might not represent the entire genome. 42.7% could be treated as the prior knowledge, but the final prediction will be driven by the actual annotations at each location. Therefore, 33.3% would still be a better estimation. To see if the prior had a strong effect, we estimated the functional proportion of chromosome 22 using different values for π while keeping other parameters unchanged. When using 0.3 and 0.5 as the π values, the estimated functional proportions were 0.376 and 0.389, respectively. Compared to the original estimate 0.383, there was not a substantial change.

Figures and Web Application. All figures were plotted using R. The "ggbio" package was used to plot the chromosomes and transcripts[58]. The GenoCanyon web application was developed using the "shiny" package in R. The "bigmemory" package was implemented to access and manipulate massive datasets[59]. The GenoCanyon web application is available at http://genocanyon.med.yale.edu. The web server is implemented using Apache running on CentOS version 6.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921, doi:10.1038/35057062 (2001).
2. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome research* **11,** 863–874, doi:10.1101/gr.176601 (2001).
3. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7,** 248–249, doi:10.1038/nmeth0410-248 (2010).
4. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7,** 575–576, doi:10.1038/nmeth0810-575 (2010).
5. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics : TIG* **24,** 344–352, doi:10.1016/j.tig.2008.04.005 (2008).
6. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 9362–9367, doi:10.1073/pnas.0903103106 (2009).
7. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74, doi:10.1038/nature11247 (2012).
8. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1318948111 (2014).
9. Doolittle, W. F., Brunet, T. D., Linquist, S. & Gregory, T. R. Distinguishing between "function" and "effect" in genome biology. *Genome biology and evolution* **6,** 1234–1237, doi:10.1093/gbe/evu098 (2014).
10. King, D. C. *et al.* Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research* **15,** 1051–1060, doi:10.1101/gr.3642605 (2005).
11. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40,** D930–934, doi:10.1093/nar/gkr917 (2012).
12. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22,** 1790–1797, doi:10.1101/gr.137323.112 (2012).
13. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587, doi:10.1126/science.1235587 (2013).
14. Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nature methods* **11,** 294–296 (2014).
15. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46,** 310–315 (2014).
16. Eberwine, J., Sul, J. Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat Methods* **11,** 25–27 (2014).
17. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640, doi:10.1126/science.1105136 (2004).
18. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology* **30,** 1095–1106, doi:10.1038/nbt.2422 (2012).
19. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482, doi:10.1038/nature10530 (2011).
20. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324,** 389–392, doi:10.1126/science.1169050 (2009).
21. Meader, S., Ponting, C. P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome research* **20,** 1335–1343, doi:10.1101/gr.108795.110 (2010).
22. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337,** 1675–1678, doi:10.1126/science.1225057 (2012).
23. Sankaran, V. G. *et al.* A functional element necessary for fetal hemoglobin silencing. *The New England journal of medicine* **365,** 807–814, doi:10.1056/NEJMoa1103070 (2011).
24. Xu, J. *et al.* Transcriptional silencing of {gamma}-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes & development* **24,** 783–798, doi:10.1101/gad.1897310 (2010).
25. Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100,** 3077–3086, doi:10.1182/blood-2002-04-1104 (2002).
26. Giardine, B. *et al.* Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature genetics* **43,** 295–301, doi:10.1038/ng.785 (2011).
27. VanderMeer, J. E. & Ahituv, N. cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental dynamics : an official publication of the American Association of Anatomists* **240,** 920–930, doi:10.1002/dvdy.22535 (2011).
28. Makrythanasis, P. & Antonarakis, S. E. Pathogenic variants in non-protein-coding sequences. *Clinical genetics* **84,** 422–428 (2013).
29. Heutink, P. *et al.* The gene for triphalangeal thumb maps to the subtelomeric region of chromosome 7q. *Nature genetics* **6,** 287–292, doi:10.1038/ng0394-287 (1994).
30. Heus, H. C. *et al.* A physical and transcriptional map of the preaxial polydactyly locus on chromosome 7q36. *Genomics* **57,** 342–351, doi:10.1006/geno.1999.5796 (1999).
31. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7548–7553, doi:10.1073/pnas.112212199 (2002).
32. Gurnett, C. A. *et al.* Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly. *American journal of medical genetics. Part A* **143,** 27–32, doi:10.1002/ajmg.a.31563 (2007).
33. Lyon, M. F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190,** 372–373 (1961).
34. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137, doi:10.1038/379131a0 (1996).
35. Augui, S., Nora, E. P. & Heard, E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nature reviews. Genetics* **12,** 429–442, doi:10.1038/nrg2987 (2011).
36. Yang, C. *et al.* X-chromosome inactivation: molecular mechanisms from the human perspective. *Human genetics* **130,** 175–185, doi:10.1007/s00439-011-0994-9 (2011).
37. Vallot, C. & Rougeulle, C. Long non-coding RNAs and human X-chromosome regulation: a coat for the active X chromosome. *RNA biology* **10,** 1262–1265, doi:10.4161/rna.25802 (2013).
38. Vallot, C. *et al.* XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. *Nature genetics* **45,** 239–241, doi:10.1038/ng.2530 (2013).
39. Migeon, B. R., Chowdhury, A. K., Dunston, J. A. & McIntosh, I. Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. *American journal of human genetics* **69,** 951–960, doi:10.1086/324022 (2001).
40. Chow, J. & Heard, E. X inactivation and the complexities of silencing a sex chromosome. *Current opinion in cell biology* **21,** 359–366, doi:10.1016/j.ceb.2009.04.012 (2009).

41. Lee, J. T., Davidow, L. S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature genetics* **21,** 400–404, doi:10.1038/7734 (1999).
42. Tian, D., Sun, S. & Lee, J. T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143,** 390–403, doi:10.1016/j.cell.2010.09.049 (2010).
43. Chureau, C. *et al.* Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region. *Human molecular genetics* **20,** 705–718, doi:10.1093/hmg/ddq516 (2011).
44. Chureau, C. *et al.* Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome research* **12,** 894–908, doi:10.1101/gr.152902 (2002).
45. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385, doi:10.1038/nature11049 (2012).
46. Barakat, T. S. *et al.* The trans-activator RNF12 and cis-acting elements effectuate X chromosome inactivation independent of X-pairing. *Molecular cell* **53,** 965–978, doi:10.1016/j.molcel.2014.02.006 (2014).
47. Bacher, C. P. *et al.* Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation. *Nature cell biology* **8,** 293–299, doi:10.1038/ncb1365 (2006).
48. Xu, N., Tsai, C. L. & Lee, J. T. Transient homologous chromosome pairing marks the onset of X inactivation. *Science* **311,** 1149–1152, doi:10.1126/science.1122984 (2006).
49. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42,** D980–985, doi:10.1093/nar/gkt1113 (2014).
50. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* **111,** E455–464, doi:10.1073/pnas.1322563111 (2014).
51. Pearl, J. *Causality: models, reasoning and inference*. Vol. 29 (Cambridge Univ Press, 2000).
52. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15,** 901–913, doi:10.1101/gr.3577405 (2005).
53. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20,** 110–121, doi:10.1101/gr.097857.109 (2010).
54. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
55. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42,** D1001–1006, doi:10.1093/nar/gkt1229 (2014).
56. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837, doi:10.1016/j.cell.2007.05.009 (2007).
57. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560, doi:10.1038/nature06008 (2007).
58. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome biology* **13,** R77, doi:10.1186/gb-2012-13-8-r77 (2012).
59. Kane, M. J., Emerson, J. W. & Weston, S. Scalable Strategies for Computing with Massive Data. *Journal of Statistical Software* **55,** 1–19 (2013).

Acknowledgments

This study was supported by the National Institutes of Health grants R01 GM59507 and U01 HG005718, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Yale World Scholars Program sponsored by the China Scholarship Council.

Author Contributions

Q.L. and H.Z. designed the project. Q.L. wrote the initial draft and performed the analyses. Y.H. collected the annotation datasets. Q.L., J.S. and Y.C. developed the web server. K.C. advised on web server development. H.Z. advised on statistical and genetic issues.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Lu, Q. *et al.* A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. *Sci. Rep.* **5**, 10576; doi: 10.1038/srep10576 (2015).