Published online 23 September 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 20* ***10597–10611***
doi: 10.1093/nar/gkz808

Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification

Gabriel E. Hoffman[1,2,3,*], Jaroslav Bendl[1,3], Kiran Girdhar[1,3], Eric E. Schadt[2,3,4] and Panos Roussos[1,2,3,5,6,7]

[1]Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [2]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [4]Sema4, Stamford, CT, USA, [5]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [6]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA and [7]Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, NY, USA

Received November 20, 2018; Revised August 28, 2019; Editorial Decision September 09, 2019; Accepted September 12, 2019

ABSTRACT

Identifying functional variants underlying disease risk and adoption of personalized medicine are currently limited by the challenge of interpreting the functional consequences of genetic variants. Predicting the functional effects of disease-associated protein-coding variants is increasingly routine. Yet, the vast majority of risk variants are non-coding, and predicting the functional consequence and prioritizing variants for functional validation remains a major challenge. Here, we develop a deep learning model to accurately predict locus-specific signals from four epigenetic assays using only DNA sequence as input. Given the predicted epigenetic signal from DNA sequence for the reference and alternative alleles at a given locus, we generate a score of the predicted epigenetic consequences for 438 million variants observed in previous sequencing projects. These impact scores are assay-specific, are predictive of allele-specific transcription factor binding and are enriched for variants associated with gene expression and disease risk. Nucleotide-level functional consequence scores for non-coding variants can refine the mechanism of known functional variants, identify novel risk variants and prioritize downstream experiments.

INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of loci associated with risk to human diseases (1). Yet progress in understanding the molecular etiology of disease and the development of novel therapies has been limited by the fact that these studies are often not able to identify a specific functional variant and mechanistically relevant gene due to linkage disequilibrium (LD) (1–4). Integrating independent biological knowledge has the potential to increase the resolution of the associated region and improve the interpretation of GWAS results (5–7). Most notably, risk variants are enriched in non-coding regulatory regions (8–10). While interpreting the functional consequences of protein coding variants has been remarkably successful and improved the understanding of the biology of human disease (11–14), the rules governing the functional effects of variants in non-coding regulatory DNA have been more challenging to decipher. Novel approaches are needed to interpret non-coding variants from ongoing whole genome sequencing projects, for example, of somatic variants in cancer (15) and *de novo* variants in autism (16).

Recent work has sought to better understand the regulatory genome by characterizing the epigenetic differences in transcription factor (TF) binding, chromatin accessibility and histone modifications between tissues and cell types (17–19). Yet, these epigenetic tracks can cover a substantial portion of the genome, even though polymorphisms at only a fraction of sites are presumed to have a functional consequence. Moreover, these efforts have generally not integrated genetic variation. Other efforts have focused on the effects of genetic variation on gene expression (20–22) as well as multiple epigenetic assays (23–26). Yet, these molecular trait QTL studies are subject to the same challenges with linkage disequilibrium as GWAS so they generally cannot pinpoint the functional variant, or predict the func-

*To whom correspondence should be addressed. Tel: +1 212 659 1635; Email: gabriel.hoffman@mssm.edu

tional consequence of a rare variant not observed in the dataset.

Recent progress in developing computational models able to predict TF binding, chromatin accessibility, and histone modifications from only the genome sequence in the surrounding region offers a novel paradigm to interpret the functional consequences of non-coding variants (27–31). These models leverage advances in deep learning (32) to use DNA sequence context in the predictive model of functional consequences. Yet with few exceptions (28), these computational approaches consider only the discrete absence versus presence of an epigenetic signal (27,29–31). Moreover, these methods rely on the sequence of the reference genome so they do not model the contribution of genetic variation driving the epigenetic signal. Based on the extensive contribution of genetic variation to molecular phenotypes (20–26), and the increasing availability of epigenetics datasets from multiple individuals paired with genetic data (23–26,33), integrating genetics into model training has the potential to improve prediction accuracy and increase power of variant impact predictions. Finally, although these methods are trained jointly across many datasets, they only consider a single experiment from a given cell type and assay.

Here we introduce a deep learning framework for functional interpretation of genetic variants (DeepFIGV). Our approach extends the method of Kelley *et al.* (28) by (i) performing model training on many epigenetic experiments for a particular assay and cell type instead just a single representative sample, (ii) integrating whole genome sequencing to create a personalized genome sequence for each individual, and (iii) modeling quantitative variation in the epigenetic readout rather than dividing the genome into two classes. We develop predictive models of quantitative epigenetic variation in chromatin accessibility from DNase-seq and histone modifications (H3K27ac, H3K4me3 and H3K4me1) from 75 lymphoblastoid cell lines (LCL) (23,26). By training the models on many experiments from the same cell type and assay, integrating whole genome sequencing, and modeling quantitative variation in the epigenetic signal, we identify genetic variants with functional effects on the epigenome.

MATERIALS AND METHODS

Epigenomic data from lymphoblastoid cell lines

The dataset comprises ChIP-seq experiments for 3 histone modifications (H3K27ac, H3K4me1 and H3K4me3) for 75 individuals (23) and DNase I hypersensitivity experiments for 69 individuals (26). All individuals are of Yoruban ancestry from the 1000 Genomes Project (34). Processed data was downloaded from the ChromoVar3D website (chromovar3d.stanford.edu). Peak coordinates and signal intensities for each sample and each DNase peak were extracted from DNase_removeBlacklist_Log10PvalueThreshold_5_DATA_MATRIX.gz, and corresponding files were used for the 3 histone modifications. VCF of variants from whole genome sequencing was obtained from the same website.

Deep learning with a convolutional neural network

An extended version of the basset software (30) was used to learn parameters in a predictive model mapping from genome sequence as input to epigenetic signal as output. The analysis was customized to take advantage of this particular dataset by (i) integrating genetic variation from whole genome sequencing, (ii) modeling the quantitative variation in the epigenetic signal and (iii) combining many experiments from the same cell type into a large single-task learning application. This customized analysis enables a focus on genetic variants with relatively small effects on the quantitative signal value, rather than the strong effect required to completely lose or gain a binding or histone modification event. Each of the four assays was analyzed separately using a single-task learning approach.

Constructing DNA sequence as input to neural network

Personalized genome sequences were constructed using the GRCh37 reference genome with sites modified according to biallelic SNPs in the whole genome sequence using the bcftools consensus command. At homozygous alternate sites the reference allele is simply replaced by the alternative allele. Heterozygous sites are represented using the IUPAC nucleotide codes (35), so that for example an A/C heterozygote is indicated with the characters 'M'. Only biallelic SNPs are considered, so there are 6 additional characters, one for each pair of nucleotides.

Homozygous sites are one-hot coded as a matrix of mostly 0's with 4 rows corresponding to 'A', 'T', 'C' and 'G'. Coding a 1 in the 'T' row indicates the presence of that nucleotide in the corresponding position in the genome sequence. Heterozygous sites are encoded with a value of 0.5 in the two corresponding rows. Thus the training data does not explicitly include any information about phasing of the SNPs or allele-specific signals.

For each peak interval called in the processed data, the genome sequence within a specified distance from the center of the peak was extracted and matched to the corresponding signal value. Peaks exceeding an assay-specific width cutoff were excluded from the analysis (Supplementary Table S1). An assay-specific window size (DNase: 300 bp, H3K27ac: 2000 bp, H3K4me3: 2000 bp, H3K4me1: 1400 bp) was used to extract regions from the personalized reference genomes (Supplementary Table S1). Larger window sizes have been shown to increase prediction performance (27), and we used the largest window size where encoding the DNA sequence from all peaks and all individuals to one-hot coded format could be computed on a machine with 256 Gb RAM. This high memory usage is a limitation of the current basset implementation (30).

Model training and testing

The default model architecture from Basset (30) was used to train a 3 layer deep neural network and 300 convolutional

filters each 19 bp wide. The model included rectified linear unit (ReLU) and max pooling in order to learning a non-linear function mapping from DNA sequence to epigenetic readout (Supplementary Table S2). A 30% dropout was applied to avoid overfitting. All training was performed on NVIDIA Tesla K20X GPU. Training on a single assay took between 14 and 45 GPU hours. We observed that changing the number of filters between 100 and 400 and changing the filter width between 10 and 20 bp did not produce a substantial change in prediction accuracy. Multiple restarts gave similar prediction accuracy.

The dataset was divided into training, validation and testing sets. In order to avoid overfitting, an early stopping approach was used where the parameter values in the model were learned from the training set, but the final values were selected to minimize the squared prediction error in the validation set. Training was stopped after 10 epochs with no decrease in error in the validation set. The prediction performance for each assay was reported based on the test set.

The training, validation and testing sets were specially constructed using a conservative approach in order to ensure independence of the three sets. Since the signal values at a given peak are relatively similar across individuals, including the same peak region, albeit from different individuals, in both the training and testing sets could overstate the prediction performance. Similarly, peaks from the same individual are generated under the same experimental conditions and are subject to technical batch effects. Thus, including peaks from the same individual in both the training and testing set could also overstate the prediction performance. In order to avoid this issue, the test set is composed of peaks on chr1–chr8 from 60% of individuals, the validation set is composed of peaks on chr9-chr15 from the next 20% of individuals, and the test set is composed of peaks on chr16-chr22 from last 20% of individuals. Thus, the three sets have no overlap in either peaks or individuals (Supplementary Table S1, Figure S1) to ensure a conservative estimate of prediction performance.

We describe the analysis workflow with numbers from DNase data; numbers for other assays are shown in Supplementary Table S1. There were 681 990 total DNase peak intervals from 69 individuals. In order to focus on peaks of approximately equal size, peaks exceeding 250 bp were excluded. This left 463 094 peaks (67.9% of total) with a mean width of 150.7 bp. Multiplying the number of remaining peaks by the number of individuals gives a dataset of 31 953 486 examples. Since DNase and histone modification ChIP-seq are not strand specific-assays, the reverse complement sequence gives the same epigenetic signal as the original sequence. Augmenting the dataset by including the reverse complement of each example doubles the number of sequence-signal pairs. Constructing the training set from peaks on chr1–chr8 from the first 60% of individuals gives 229 421 unique peak regions and 18 812 522 total examples.

Genomic correlates

Minor allele frequency across populations were obtained from gnomAD r2.0.2 (11). Transcription factor binding motifs were obtained from the JASPAR 2018 database (36) and genomic location of bindings sites were identified with FIMO (37). For genome-wide summaries, a TFBS score cutoff of 500 was used, correspond to a *P*-value cutoff of 1e–5. For specific lookups (i.e. Figures 4F and 6C), a more liberal cutoff of 400 (i.e. *P*-value of 1e–4) was used. Genomic locations from ChIP-seq experiments for transcription factors in LCL GM12878 (18) were downloaded from http://egg2.wustl.edu/roadmap/src/chromHMM/bin/COORDS/hg19/TFBS/gm12878/. List of genes expressed in LCLs were obtained from http://egg2.wustl.edu/roadmap/src/chromHMM/bin/COORDS/hg19/expr/gm12878/. ChromHMM tracks (17) for LCL GM12878 were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/E116_18_core_K27ac_dense.bed.gz. Genome annotation of sites were obtained from VEP v85 (38) provided by gnomAD (11). CpG islands were obtained from Annotatr (39).

Comparison to canonical motifs

Each convolutional filter in the first layer of the neural network reads in 19 bp at a time and uses a weight for each of the four nucleotides to transform the DNA sequence to the next layer of the neural network. Each filter is a 19×4 matrix of continuous values that can be treated as a position weight matrix (PWM) (30). Each PWM learned in the current dataset were then compared to PWM's of transcription factors from the JASPAR 2018 database (36). We applied the widely used software tomtom (40) to query a our new set of PWM's against a know database of PWM and generate p- and q-values. Since a given filter can show high similarity to multiple JASPAR motifs, only the best match is reported. Motifs were visualized using ggseqlogo (41).

Evaluating variant effects

Coordinates and alleles of SNPs were obtained from multiple public resources (Supplementary Table S3) and combined into a non-redundant list comprising 413 223 060 sites and 437 960 283 variants (due to multi-allelic sites). The delta between the predicted signal from the reference and alternative alleles was evaluated for each of the four epigenetic assays. The median and standard deviation of the delta values for each assay were obtained for 208 million biallelic SNVs from whole genome sequencing (WGS) from gnomAD r2.0.2 and were used to compute *z*-scores for the entire set of variants for the corresponding assay. This approach used sites distributed across the genome that were identified independently of their predicted functional consequence and avoids double counting multiallelic sites. The standard deviation was computed using a robust method (i.e. winsorized) where delta values below the 1st percentile or above the 99th percentile were set to the value at the corresponding cutoff. This approach reduced the effect of variants with extreme scores. Changing the cutoff values had a very minimal effect of the resulting *z*-scores.

Evaluating all variants for the four assays took 5929 GPU hours using 10 NVIDIA Tesla K20X GPUs.

Integration with molecular trait QTLs

We downloaded QTLs for gene expression, DNase and histone modifications on Yoruban individuals (23), QTLs for gene expression on LCLs from multiple European populations (21). Enrichment was evaluated by comparing the DeepFIGV absolute *z*-cores from the lead QTLs to the scores for to variants ranked between 5th and 10th. Statistical fine mapping results were obtained from multiple cell types (42). Rare variants associated with gene expression outliers from multiple tissues were obtained from GTEx (43). Enrichments are evaluated based on 2113 rare variants associated with outliers and 67 044 not associated with outliers.

Chromatin accessibility QTLs from brain homogenate

Raw ATAC-seq (44) data from 189 human post mortem brains (45) of European ancestry was aligned to GRCh38 with STAR aligner (46). To create a final peakset, we subsampled and merged BAM-files separately for schizophrenia-case and control samples. We subsequently called peaks separately on these two merged BAM files with MACS2 (47) at q-value < 0.01, and merged these two peaksets into a single consensus peakset. For each sample, the reads in each consensus peak were quantified by featureCounts (48). Only peaks with 1 counts per million in at least 10% samples were retained for QTL analysis on the TMM normalized $\log_2$ counts per million values (49). QTL analysis was performed with QTLtools (50) using 5 ancestry PC's and variants with MAF $> 5\%$. Covariates also included 10 PEER components (51) gender, and GC content of reads for each peak for each sample. QTL analysis was performed on variants within 2 kb of each peak boundary.

Cancer somatic variants driving gene expression

We downloaded somatic variants in tumors that were identified by whole genome sequencing and results from an eQTL analysis combining nearby variants and testing the association with proximal genes (15). We considered the 569 genes with cis-eQTLs at FDR $< 30\%$ and evaluated the DeepFIGV *z*-score for each of four epigenetic assays for the 2309 somatic variants in the proximal regions. The enrichment analysis compared these variants to somatic variants in this dataset that were not associated with gene expression changes and which were matched for distance to transcription start site.

Prediction of allele specific binding (ASB)

DeepFIGV scores were used to predict the presence and direction of ASB using sites identified from transcription factor ChIP-seq and DNase I hypersensitivity experiments in LCLs (52,53) and HeLa-S3 cells (53). For AlleleDB (52), the ASB status for 42 ChIP-seq targets across 14 individuals totaling 77 experiments were reported at a total of 276,589 sites with sufficient read coverage (accB.auto.v2.1.aug16.txt.gz). Shi *et al.* (53) reported ABS for 36 targets across 7 LCL and HeLa-S3 cells across 51 518 total sites (ASB_GM12878_HeLa_1based.txt, ASB_other_GMs_1based.txt). Sites from the two databases were combined to produce a non-redundant set, so sites identified for the same target and individual in both databases were not double counted. Sites were considered as ASB or non-ASB based on a Benjamini-Hochberg corrected *P*-value (beta-binomial for AlleleDB and binomial for Shi *et al.*) <0.05, or >0.99, respectively. Only assays with at least 20 ASB examples were considered.

Precision-recall (PR) curves and area under the PR curve (AUPR) were used to evaluate the classification performance since the ASB vs non-ASB class counts were very imbalanced. PR curves and AUPR of empirical and random classes classifiers were evaluated with the PRROC package (54).

No allele-specific signal was used in the training of DeepFIGV and no re-training was performed for the ASB analysis. The DeepFIGV *z*-scores for each of the 4 assays in the training set were extracted for each site, and the PR and AUPR were computed by the intersecting these scores with the combined ASB dataset. Classifying ASB sites from non-ASB sites used the absolute value of the *z*-scores, while classifying the direction of ASB used the *z*-score itself. ASB magnitude was encoded as the number of alternative reads at a site divided by the total number of reads at that site. Thus, a positive ASB magnitude corresponds to a positive DeepFIGV *z*-score indicating that the alternative allele is predicted to increase signal compared to the reference allele.

Computing disease enrichments: LD-score regression

Publicly available GWAS summary statistics were obtained for immune diseases as well as other representative diseases and traits. We performed a partitioned heritability analysis with LD-score regression (LDSC) (8) in order to quantify the contribution to the trait heritability of variants with high absolute DeepFIGV z-scores. The per-SNP heritability was computed after accounting for the other genomic annotations. Annotations included 28 provided with LDSC baseline model (i.e. TFBS, TSS, UTR, intron, promoter, enhancer, superenhancer, epigenetic assays multiple sources (H3K27ac, H3K4me1, H3K4me3, H3K9ac, DNase)) in addition to peak regions from the four assays in LCLs used by DeepFIGV. DeepFIGV was the only annotation with nucleotide level resolution; other annotations were 10s or 100s of bases wide.

This analysis was restricted to common variants outside of the MHC region. The per-SNP heritability was evaluated for site exceeding absolute z-score cutoffs for 1, 2, 3, 4 and 5. There was not a sufficient number of sites with larger scores, since only common variants were considered.

Computing disease enrichments: Candidate causal variants

Candidate causal SNPs identified from finemapping analysis of autoimmune diseases were obtained from http://www.broadinstitute.org/pubs/finemapping (10). The DeepFIGV z-scores were obtained for the 8741 candidate causal

SNPs for 39 traits. Enrichments for each trait were evaluated by comparing the number of candidate causal sites with a DeepFIGV absolute z-score exceeding a given cutoff to the expected value from a random set of sites from a null distribution. This null was constructed for each site by drawing 10 000 sites from across the genome matching the MAF, gene density, distance to nearest genes and number of sites within LD of 0.5 of the original site (55).

Comparison to other variant scoring methods

Variant-level scores were obtained from DeepSea (27) evaluated on 17 DNase datasets and deltaSVM (29) evaluated on 35 DNase datasets, and CAPE (56) evaluated on two datasets. In addition we included CADD (57) and LINSIGHT (58). Scores were obtained from: https://www.ncbi.nlm.nih.gov/research/snpdelscore/rawdata/

For DeepSea and deltaSVM the reported delta values for the predicted signal from the reference and alternative alleles were transformed to a z-score using the observed standard deviation. Analysis was performed on a shared set of 12 million variants.

Analysis of LCL MPRA results

Results were downloaded from Tewhey *et al.* (59) and variants were divided into three classes: (i) expression modulating variants that showed significant difference in expression between reference and alternative alleles, (ii) variants that drove expression but did not show allelic differences, and (iii) variants whose sequence did not drive expression in this assay. Enrichment of expression modulating variants (i.e. class 1) were compared to the other two classes as a function of high predicted epigenetic signal or DeepFIGV *z*-scores for each assay.

Predicting pathogenic from benign variants in ClinVar

ClinVar variants were downloaded on April 22, 2019. Variants labeled as 'Pathogenic' or 'Likely_pathogenic' was considered as the positive class and variants labeled 'Benign' or 'Likely_benign' were considered the negative class. Variants were then stratified based on variant consequence (i.e. missense, 3′ UTR, etc) and the performance of each prediction (DeepFIGV, CADD (57), GWAVA (60)) were evaluated in each strata. Due to the high degree of class imbalance, performance of each predictor was reported as the area under the precision recall curve (AUPR). We note that performance of a random prediction varied across variant consequence since it depends on the class ratios.

Evaluation of CAGI5 regulation saturation data

Since DeepFIGV predictions were learned on a different dataset, we evaluated concordance between DeepFIGV and CAGI5 relative expression values on the combined training + test set from CAGI5. Data was obtained from genomeinterpretation.org.

RESULTS

Deep learning maps from genome sequence to epigenetic signal

DeepFIGV combines the quantitative signal from epigenetic experiments across multiple individuals with whole genome sequencing into a single machine learning task (Figure 1). While standard molecular trait QTL analyses rely on the correlation between the epigenetic signal and a given genetic variant (Figure 1A, B), deep learning using a convolutional neural network explicitly models the DNA sequence context to train a predictive model (Figure 1C, D). Evaluating the predicted effect of each variant produces a large database of nucleotide-level scores (Figure 1E, F) that can be integrated with other analyses to refine the mechanism of known functional variants, identify novel risk variants and prioritize downstream experiments (Figure 1G).

Datasets from each of four epigenetic assays (DNase-seq and H3K27ac, H3K4me3, and H3K4me1 histone marks) were analyzed separately (Figure 2). Previous quantifications using liberal peak calling thresholds were used in order to capture a wide range of quantitative variation (23,26). The parameters of the convolutional neural network were chosen to minimize the least squares prediction error. Extensive steps were taken to avoid overfitting, and all prediction results are reported on a set of individuals and chromosomes that were excluded from the training set (see Materials and Methods, Supplementary Figure S1). Increasing the number of individuals in the training set and including genetic variation in the genome sequence of each individual decreased prediction error on withheld test data (Supplementary Figure S2). Although the model uses only DNA sequence in the predictions, the predicted DNase signal shows strong concordance in the test set with the observed signal in peak regions (Spearman rho = 0.485, Pearson R = 0.707) (Figure 2A). Focusing on the more robust (i.e. rank based) Spearman correlation metric shows that these predictive models give substantial accuracy for all four assays for the quantitative epigenetic signal (Figure 2B). Genomic intervals corresponding to ENCODE blacklisted regions were already excluded and removing additional genomic intervals based on low sequence uniqueness/mappability (61) did not change the results (Supplementary Table S4). Examining the predicted signal for DNase for a representative example in the test set along a segment of chromosome 22 shows notable concordance with the observed signal in peak regions (Figure 2C).

Functional impact scores from the predicted difference in epigenetic signal for the reference versus the alternate allele were evaluated for 438 million variants observed in previous sequencing projects (Supplementary Table S3). Subsequent analysis was performed on 208 million biallelic SNPs from the gnomAD database of 15 000 whole genome sequences (11). The delta value for each variant is defined as $\Delta = S_{ALT} - S_{REF}$ with terms representing the predicted epigenetic signal from the alternative and reference alleles, respectively. Thus, a positive delta value indicates that the alternative allele increases the epigenetic signal compared to the reference allele. As expected, the mean and median delta values for all assays were very close to zero (Figure

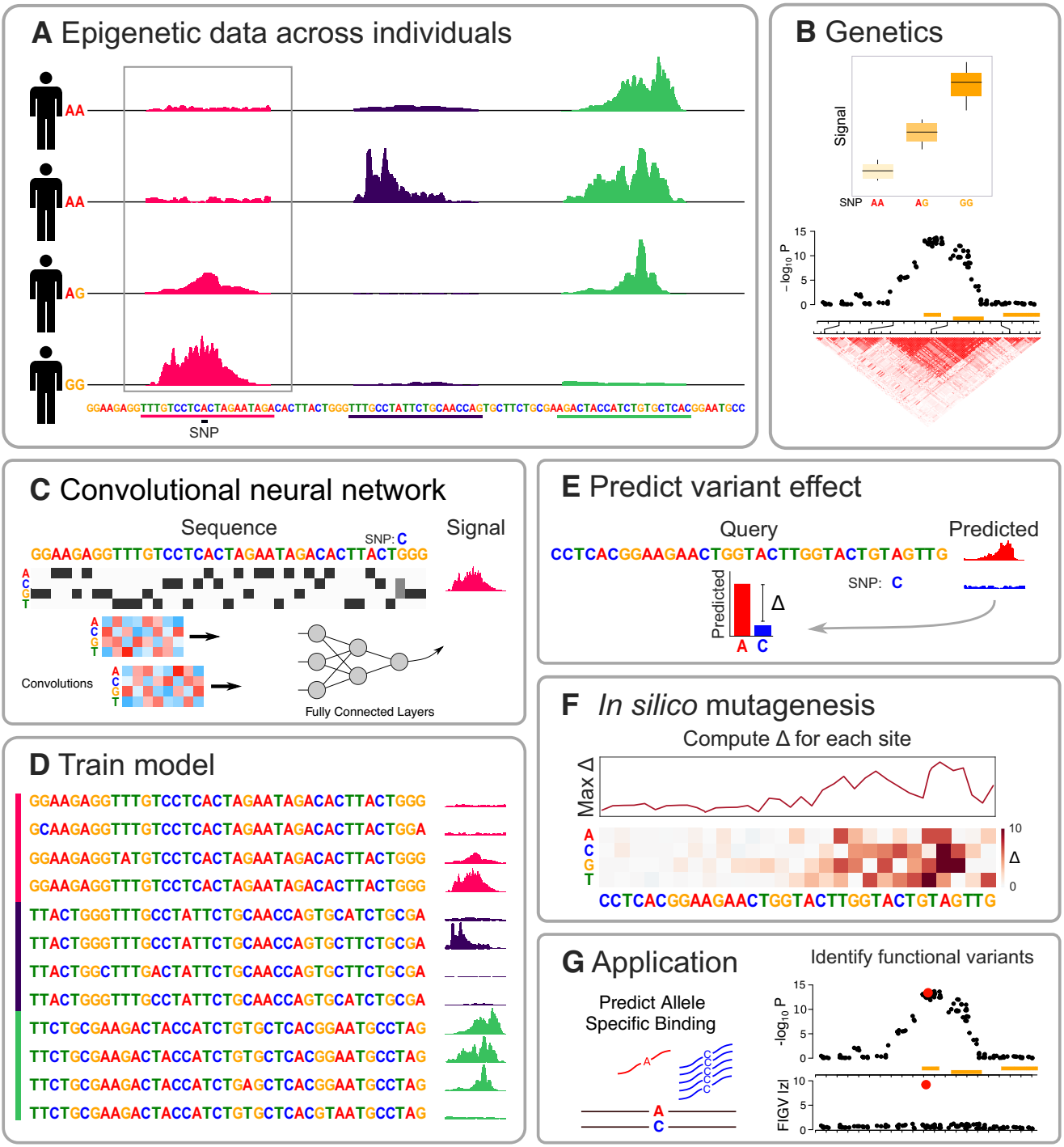

Figure 1. Computational workflow for Deep Functional Interpretation of Genetic Variants (DeepFIGV). (**A**) Quantitative signal from epigenetic assay (i.e. ChIP-seq, DNase-seq) across multiple individuals and genomic regions. (**B**) Standard genetic analysis stratifies quantitative signal by the allelic state at a given SNP, yet linkage disequilibrium complicates the interpretation of the functional variant. (**C**) DeepFIGV encodes a DNA sequences as an 'image' matrix of mostly zeros with a 1 (i.e. a dark box) indicating the presence of a particular nucleotide at that position. Heterozygous SNPs are encoded as a 0.5 for each allele. Convolutions are local matrix operations with parameter values learned from the data. A neural network uses the convolutions to predict the epigenetic signal from the DNA sequence. (**D**) Training the computational model links DNA sequences from many individuals to the epigenetic signal in each region. (**E**) The epigenetic signal is estimated for a query sequence with the reference and the alternate allele. The difference between the estimated signal values (i.e. delta) indicates the predicted effect of the variant. (**F**) *In silico* mutagenesis evaluates the delta value for every possible single nucleotide substitution. (**G**) DeepFIGV delta values are used to predict allele specific binding of transcription factors and identify candidate functional variants.

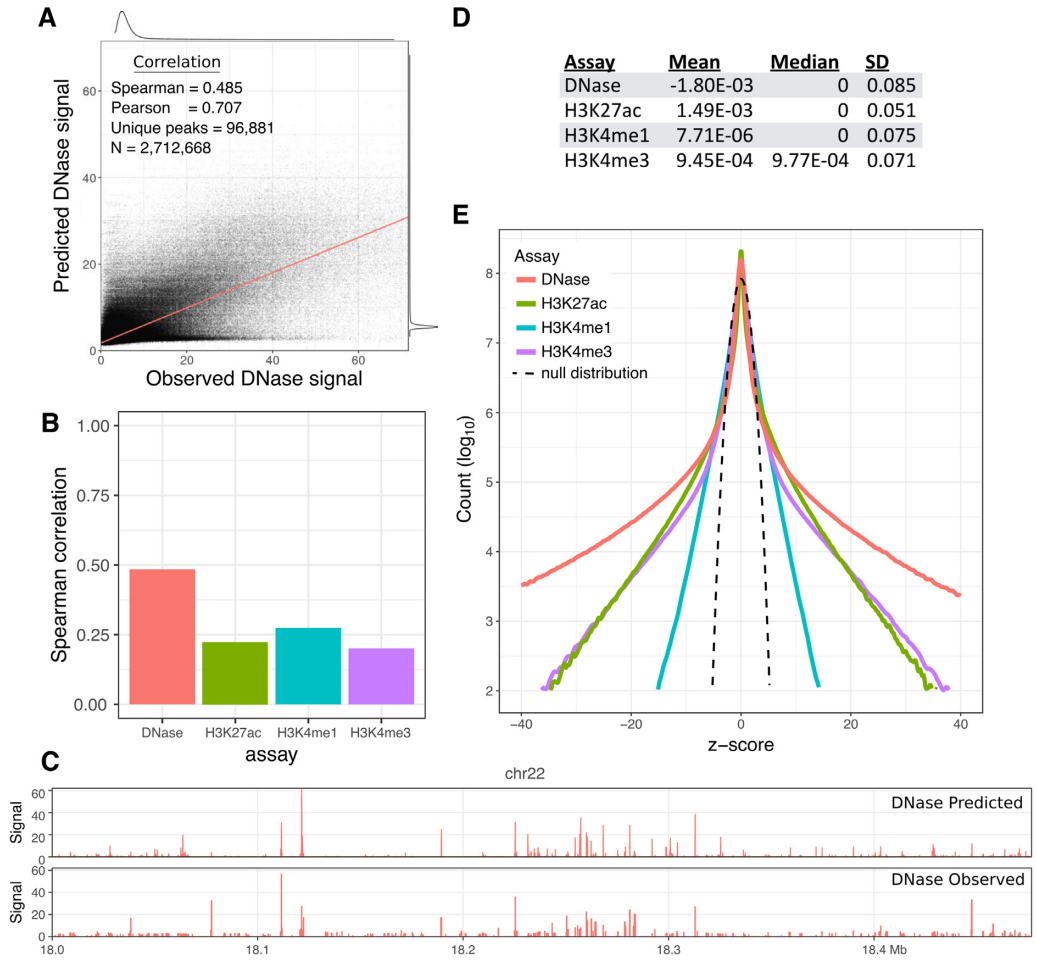

Figure 2. Evaluating DeepFIGV model and interpreting variant scores. (**A**) Predicted DNase signal compared to observed DNase signal evaluated on the test set. (**B**) Spearman correlation between predicted and observed epigenetic signal on the test set for four assays. (**C**) Predicted and observed DNase signal for peaks in a 300 kb segment on chr22 in the test set. (**D**) Mean, median and standard deviation of the delta scores for 208 million biallelic SNPs for each assay. (**E**) Density plot of *z*-scores for four assays. Dashed line indicates the null distribution of the z-scores, which is the standard normal distribution.

2D). Transforming these delta values to a standard scale (i.e. *z*-score) by dividing by the standard deviation for each assay shows an excess of variants with scores near zero compared to the standard normal distribution (Figure 2E). This is consistent with the vast majority of variants having no functional effect on the epigenome. Yet, there is an excess of variants with large effects on all four epigenetic assays, with DNase showing the highest excess followed by H3K27ac, H3K4me3 and finally H3K4me1.

Genomic correlates of predicted variant effects

Although no prior biological information is included in training the model, DeepFIGV recovers multiple aspects of known regulatory biology (Figure 3). The predictive model learned by the convolutional neural network is composed of a set of local sequence features called filters. Although learned *de novo*, the predictive sequences features extracted by these filters are often similar to known transcription factor bindings site (TFBS) motifs. Some filters have a direct correspondence to a known motif, but other filters model only a portion of a motif so that multiple filters combine to capture the signal encoded by the sequence (Figure 3A). Variants in TFBS motifs are enriched for the alternative allele decreasing the DNase signal (Figure 3B). TFBS nucleotides with high information content (i.e. high weight) in the position weight matrix have an even stronger enrichment for decreasing the DNase signal, consistent with variants being more likely to weaken rather than strengthen the affinity of a TFBS motif (Supplementary Figure S3). The TFBS enrichments are consistent with the biology of these assays: variants predicted to affect the open chromatin assay DNase are most enriched for TFBS motifs, followed by the H3K4me3 promoter mark and the H3K27ac active promoter and enhancer mark (Figure 3C). H3K4me1 is an enhancer mark that tags active or primed sequences and is not enriched for strong variants in TFBS.

The role of TFBS in regulating gene expression and epigenetics is well established, and consequences of variants in TFBS motifs are more interpretable than variants in other genome annotations (18,36,62). Yet despite notable enrichment in TFBS motifs, variants in these motifs account for a minority of variants with strong DeepFIGV scores. Only 30.1% of sites with an absolute z-score between 9 and 10,

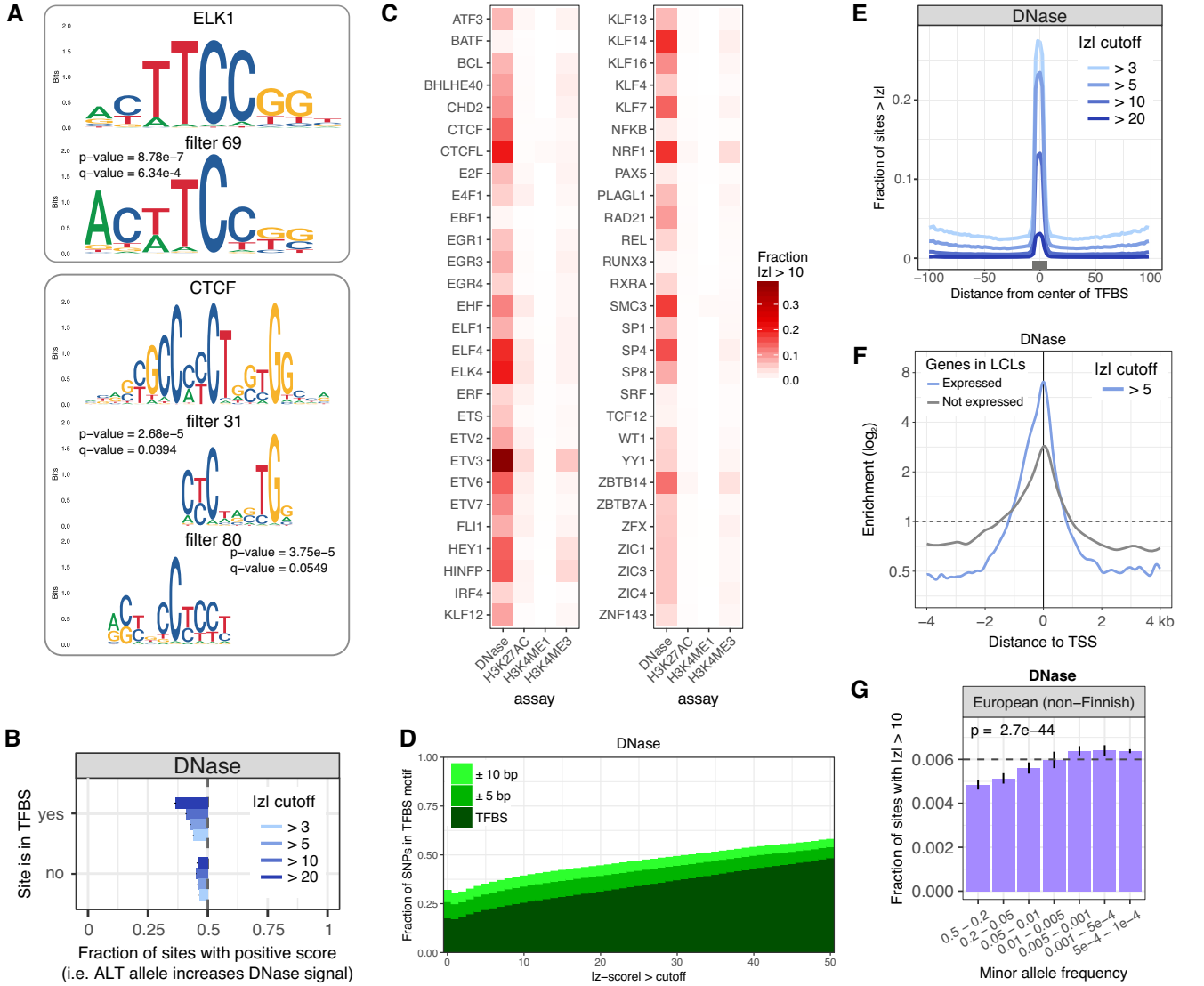

Figure 3. Genomic enrichments of predicted functional variants. (**A**) Canonical transcription factor binding motif along with the motif representation of convolutional filters learn from the DNase dataset. *P*-values indicate the probability of concordance this high between a canonical motif and convolutional filter occurring by chance given the motif database. Q-values correct for multiple testing since 300 convolutional filters were queried. (**B**) Ratio indicating the fraction of sites where the alternative allele increases the DNase signal (i.e has a positive DeepFIGV z-score) for DNase for four cutoffs. Ratios are shown for variants within or outside a TFBS motif. A value of 0.5 indicates an equal number of variants with positive and negative *z*-scores. A value <0.5 indicates an depletion of variants where the alternative allele increases the DNase signal, corresponding to an excess of variants where the alterative allele decreases DNase signal. (**C**) Fraction of variants in transcription factor binding sites that exceed a DeepFIGV absolute z-score of 10 for each of four assays. (**D**) Fraction of sites that are in a transcription factor binding site motif, or in the flanking 5 or 10 bp, for a range of DeepFIGV absolute z-score cutoffs for DNase. (**E**) Enrichment of variants near a TFBS motif exceeding four *z*-score cutoffs for DNase. Black box indicates median size of TFBS motif. (**F**) Enrichment of sites with absolute z-scores greater than 5 near the transcription start site of genes stratified by whether the genes are expressed in LCLs. Sites with absolute z-scores less than the genome-wide mean are used as the baseline for the enrichment. (**G**) Fraction of sites with absolute z-scores for DNase greater than 10 within 7 minor allele frequency bins based on non-Finnish Europeans from gnomAD. Dashed line indicates genome-wide fraction of sites. *P*-value is based a logistic regression where the response is a binary variable indicating if the absolute *z*-scores is >10 and the log minor allele frequency is the predictor. Error bars show 95% confidence intervals.

and 44.4% with an absolute *z*-score between 29 and 30 for DNase fall in a TFBS (Figure 3D, Supplementary Figure S4). Including flanking nucleotides within 5 or 10 bp increases these percentages, but for most z-score cutoffs, variants in or proximal to known TFBS motifs are a minority. While variants in TFBS motifs are enriched for variants predicted to affect DNase signal, the enrichment is not observed when expanding beyond these proximal nucleotides (Figure 3E for DNase, Supplementary Figure S3C for ChIP-seq). We note that these enrichments are not being driven by biases in peak location since variants located within peak intervals for each assay show a similar enrichment profile as variants not located within the peak intervals (Supplementary Figure S3D). Therefore, the majority of variants with strong predicted effects on all four assays do not fall in nor are they proximal to these known TFBS,

indicating that DeepFIGV models a more complicated relationship between genetic variants in epigenetic signal than is encoded by TFBS motifs alone.

Variant effects show a degree of cell type specificity as variants with strong predicted effects on DNase, H3K27ac and H3K4me3 are more enriched around the transcription start site (TSS) of genes expressed in LCLs, compared to genes not expressed in LCLs (Figure 3F, Supplementary Figure S5). Variants with strong predicted effects are also more enriched around the TSS of LCL-specific genes, compared to tissue-specific genes from each of 52 additional GTEx tissues (Supplementary Figure S6). Moreover, variants with strong predicted effects on DNase, H3K27ac and H3K4me3 are enriched in CpG islands, and ChromHMM tracks from LCLs (Supplementary Figures S7 and S8). Finally, variants with large predicted effects are depleted among common variants (minor allele frequency $> 1\%$) and are enriched in rare variants across multiple human populations, consistent with negative selection against variants that disrupt the epigenome (Figure 3G, Supplementary Figure S9).

Concordance with molecular trait QTLs

Lead cis-QTL variants (i.e. the local variant with the smallest *P*-value) for multiple assays are enriched for having strong predicted effect on the epigenome (Figure 4, Supplementary Figure S10). Variants that are lead cis-QTLs for DNase from the current dataset (23,26) are particularly enriched for having a strong predicted effect on DNase and H3K4me1, compared to variants ranked 5th–10th in the cis-QTL analysis (Figure 4A). Similarly, variants that are lead *cis*-QTLs for gene expression in an independent dataset of LCLs from European individuals (21) are most enriched for variants with a strong predicted effect on DNase (Figure 4B). In addition, variants that are lead cis-QTLs chromatin accessibility assayed by ATAC-seq in post mortem homogenate of human brain tissue from European individuals (45) are also enriched for variants with a strong predicted effect on DNase (Figure 4C). Rare variants associated with expression outliers in multiple tissue types (43) are enriched for variants with strong predicted effects on DNase (Figure 4D), but not ChIP-seq, compared to rare variants not associated with expression outliers. Moreover, somatic variants in cancer that drive changes in expression of nearby genes are enriched for variants with strong predicted effects on DNase, H3K4me3 and H3K27ac, compared to somatic variants matched for distance to transcription start site that were not associated with expression changes (15) (Figure 4E). Furthermore, candidate causal variants for expression QTLs identified by statistical fine mapping (42) are enriched for variants with strong predicted effects on DNase, compared to cis variants that have lower posterior probability (Figure 4F).

For example, rs11547207 is identified as an eQTL in LCLs from European individuals, but this SNP is in linkage disequilibrium with many nearby variants (Figure 4G). Statistical fine mapping indicates that this SNP has a high probably of being the causal variant in this region driving gene expression variation. Although analysis of the DNase signal from LCLs does not identify this SNP as a QTL, DeepFIGV directly models the sequence context of this variant and predicted a strong effect of the epigenetic signal in this region. *In silico* saturation mutagenesis in this region gives predictions at nucleotide-resolution and indicates that variants within $\sim$5 bp are also predicted to decrease the DNase signal while falling just upstream of a TFAP2A TFBS (36).

DeepFIGV variant scores predict allele specific binding

The predicted functional effect of genetic variants on each of the four epigenetic assays analyzed in DeepFIGV can identify allele-specific binding (ASB) of TFs in independent ChIP-seq experiments in LCLs (52,53) (Figure 5). Heterozygous variants can be divided into three categories based on ASB: (a) no allele specific effect, (b) ASB favoring the reference allele and (c) ASB favoring the alternative allele (Figure 5A). We evaluated the ability of DeepFIGV to distinguish between these categories even though no allele-specific information is included in model training. The predicted effect on the DNase signal can classify variants showing ASB versus no ASB for CCCTC-binding factor (CTCF) with an area under the precision recall (AUPR) curve of 0.202 while a random classifier gives an AUPR of 0.0493 (Figure 5B, C). This gives an AUPR increase of $0.202 - 0.0493 = 0.1527$ compared to random. Given a variant with an allele-specific effect, the predicted effect on DNase signal is able to classify the direction of the effect (i.e. favoring reference versus alternative) for CTCF with an AUPR of 0.704 compared to a random classifier of 0.36 (Figure 5D, E). This gives an AUPR increase of 0.344. Since the number of sites in each category varied substantially across TFs, we consider the increase in AUPR from the DeepFIGV score compared to a TF-specific baseline. DeepFIGV scores show an AUPR increase compared to random classifiers for predicting ASB events and predicting ASB direction for variants from independent assays of multiple transcription factors in LCLs (Figure 5F) and HeLa S3 cells (Supplementary Figure S11).

Concordance with large-scale functional experiments of variant impact

Scalable experimental approaches to measure the functional consequence of non-coding variants have recently been proposed (59,63–66). These massively parallel reporter assays (MPRA) couple thousands to millions of nucleotide sequences to a molecular readout that can be quantified by short read sequencing. Tewhey *et al.* (59) performed an MPRA of 32K variants in LCLs by inserting 150 bp sequences centered at the variant into an episomal vector. Based on experimental readout, sequences were divided into three classes: (1) expression modulating variants that showed significant difference in expression between reference and alternative alleles, (2) variants that drove expression but did not show allelic differences, and (3) variants whose sequence did not drive expression in this assay. Applying predictions from the DeepFIGV models is concordant with experimental readout for sequences and variants from these experiments (Figure 6). Evaluating each sequence based on predicted signal magnitude for the four

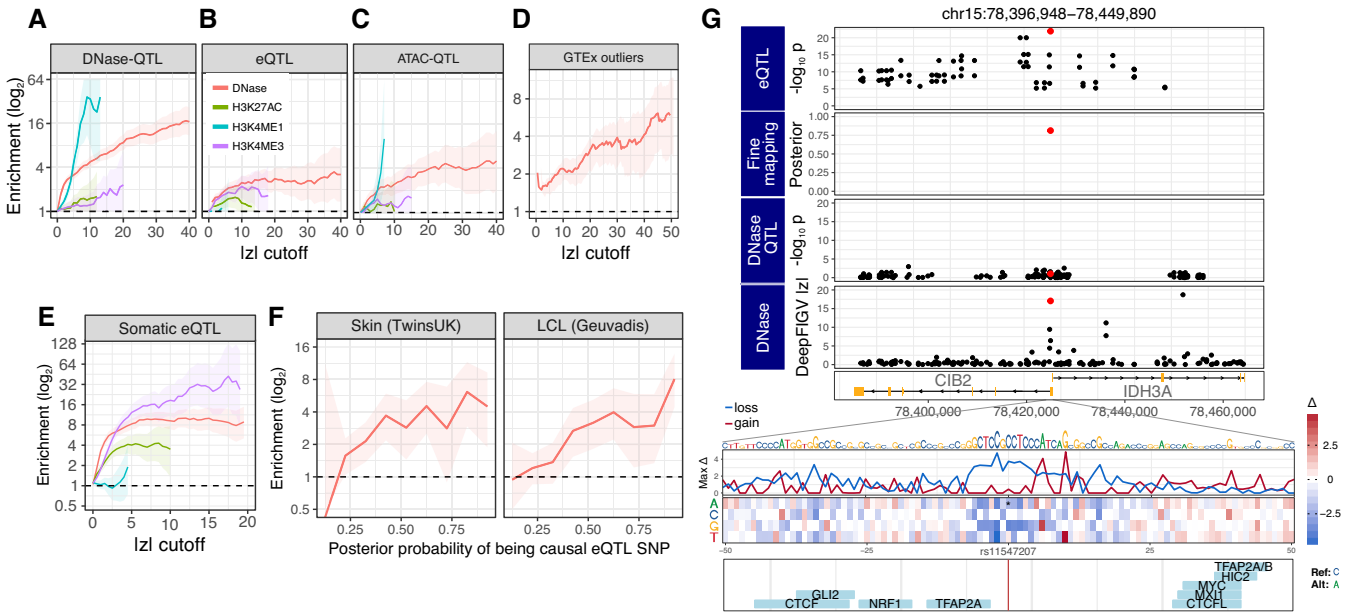

Figure 4. DeepFIGV scores predict results of molecular trait QTL analysis. (A–C) Lead variants from molecular trait QTL analysis from lymphoblastoid cell lines are enriched for SNPs with DeepFIGV absolute z-score exceeding a range of cutoffs. Enrichments are evaluated using (**A**) DeepFIGV scores for 4 assays for DNase-QTLs from LCL's of Yoruban ancestry, (**B**) expression QTLs from LCL's of European ancestry, and (**C**) chromatin accessibility QTL's assayed by ATAC-seq in human post mortem brain homogenate of European ancestry. Shaded regions indicated 95% confidence intervals. (**D**) Rare variants associated with gene expression outliers are enriched for DeepFIGV absolute z-score for DNase compared to rare variants not associated with outliers. Shaded regions indicated 95% confidence intervals. (**E**) Enrichment of somatic variants in cancer that drive gene expression changes (15) for strong DeepFIGV scores. (**F**) Candidate causal variants for expression QTLs with higher posterior probability are enriched for exceeding a DeepFIGV absolute z-score of 10 for DNase. Enrichments are shown for skin and LCL samples from TwinsUK, and LCL samples from GEAUVIDIS (21). Shaded regions indicated 95% confidence intervals. (**G**) DeepFIGV independently identifies candidate causal variant rs11547207 (shown in red) for QTL affecting expression of both CIB2 and IDH3A. eQTL analysis of GEUAVDIS identifies many correlated variants associated with these genes, but statistical fine-mapping identifies a signal candidate variant (42). Although this variant is not an DNase QTL in LCLs Yoruban individuals (23), DeepFIGV analysis on the same data identifies the same candidate causal variant identified by statistical fine mapping. *In silico* mutagenesis of 50bp around rs11547207 indicates that variants at nearby positions are predicted to decrease the DNase signal. Size of letters in DNA sequence is proportional to the maximum absolute delta at that position. Bottom panel shows TFBS motifs.

epigenetic assays shows that sequences with higher predicted signal are strongly enriched for driving expression in this experiment compared to sequences that do not drive expression (Supplementary Figure S12). As expected, variants with a high DeepFIGV z-score are enriched in variants with allelic effect (i.e. class 1) compared to variants that are in sequences that do not drive expression in this assay (i.e. class 3) (Figure 6A). Yet, despite the challenge of small experimental effect sizes between two alleles of a variant (59), variants found to drive changes in gene expression (i.e. class 1) are enriched for having strong predicted effects on DNase by DeepFIGV compared to variants that do not show an allelic effect (i.e. class 2) (Figure 6B).

Enrichment for disease risk variants and interpreting causal variants

Integrating DeepFIGV scores with large-scale genome-wide association studies shows that risk variants for common disease are enriched for variants predicted to impact the epigenome (Figure 7). We applied stratified LD-score regression (8) to evaluate the contribution of variants with different genomic annotations to disease risk. Analysis of 19 traits identified a contribution of variants passing multiple DeepFIGV *z*-score cutoffs to trait heritability, even after accounting for a baseline set of 32 genomic annotations (see Methods) (Figure 7A, Supplementary Figures S13,S14). Immune traits show the strongest contribution of DeepFIGV variants to trait heritability since the model was trained in LCLs (a B-cell lineage), yet there are also cell type autonomous effects and a contribution to non-immune traits. Further investigation of the impact of immune traits shows that candidate causal variants identified by statistical fine mapping (10) are enriched for variants with strong DeepFIGV effects (Figure 7B, Supplementary Figure S15).

DeepFIGV scores can elucidate the molecular mechanism of a causal variant and prioritize downstream experiments. Integrating DeepFIGV scores with candidate causal variants for inflammatory bowel disease (67) shows that for rs10748781, which has 99% posterior probability of being the causal variant in this region and is in a Stat4 TFBS, the alternative allele is predicted to decrease the DNase signal in this region in LCLs (Figure 7C). This result gives a specific cell type and biological assay to design a validation experiment. Moreover, this variant disrupts a CpG site, is a known DNA methylation QTL, and the methylation at nearby sites is predicted to mediate the effect of the variant on disease risk (68) (Supplementary Figure S16).

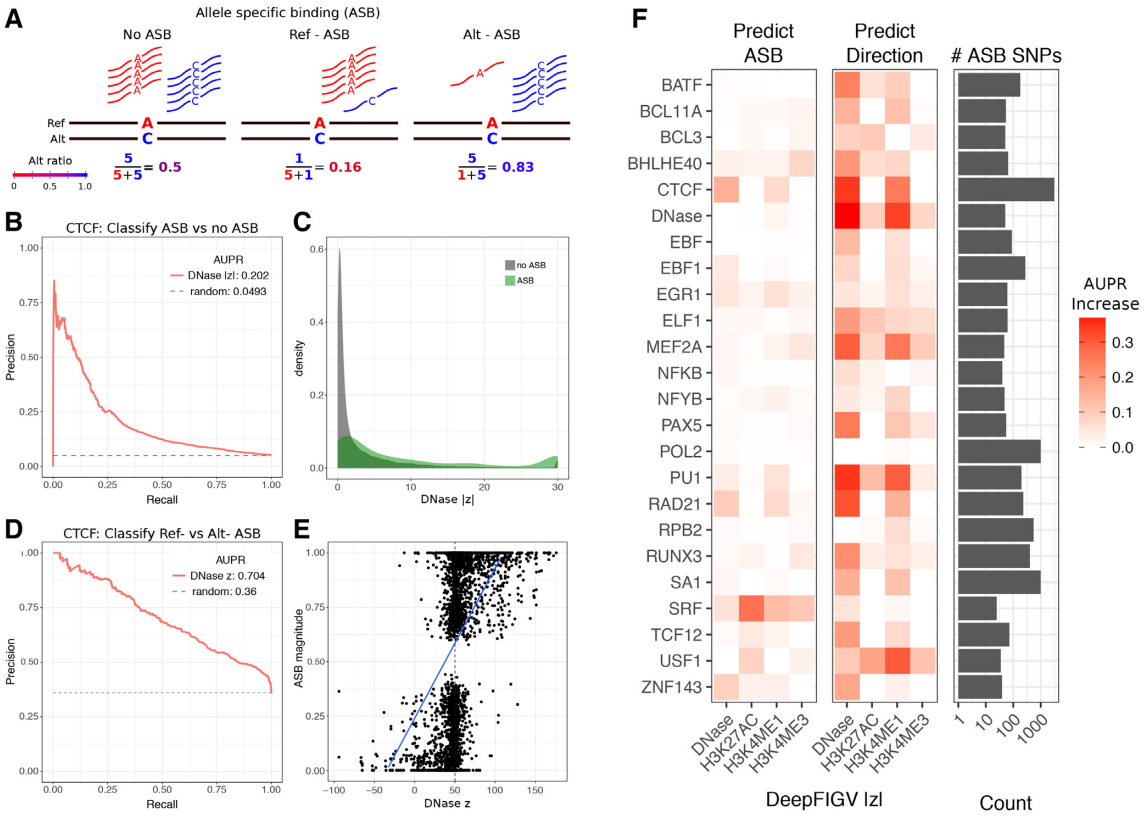

Figure 5. DeepFIGV scores predict allele specific transcription factor binding in LCLs. (**A**) Diagram illustrating 3 categories of allele specific binding (ASB): (i) no ASB, (ii) ASB favoring the reference allele, and 3) ASB favoring the alternative allele. (**B**) Precision-recall curve indicating performance of absolute DeepFIGV z-score for DNase in predicting ASB of CTCF. AUPR indicates the area under the precision-recall curve. Dashed line indicates the performance of a random predictor. (**C**) Density plot showing absolute DeepFIGV z-score for variants in (**B**) in the ABS or no ASB classes. (**D**) Precision-recall curve indicating performance of DeepFIGV *z*-score for DNase in predicting the directionality of ASB (reference versus alternative) for CTCF. AUPR indicates the area under the precision-recall curve. Dashed line indicates the performance of a random predictor. (**E**) Plot of ASB magnitude versus DeepFIGV DNase *z*-score from (**D**). (**F**) Increase in AUPR of predicting ASB status for DeepFIGV scores for four epigenetic assays compared to a TF-specific random predictor. Increase in AUPR is shown for predicting ASB versus no ASB (left) and predicting the directionality of ASB (reference versus alternative) (center). Right panel shows the number of ASB SNPs considered in each analysis.

Figure 6. Variants with large DeepFIGV *z*-scores are enriched for activity in experimental massively parallel reporter assay. (**A**) Variants that modulate gene expression in this assay are enriched for having large DeepFIGV z-scores compared to variants in sequences that do not drive expression. (**B**) Variants that modulate gene expression in this assay are enriched for having large DeepFIGV *z*-scores for DNase compared to variants that have no allelic effect in this experiment. Shaded regions indicate 95% confidence intervals.

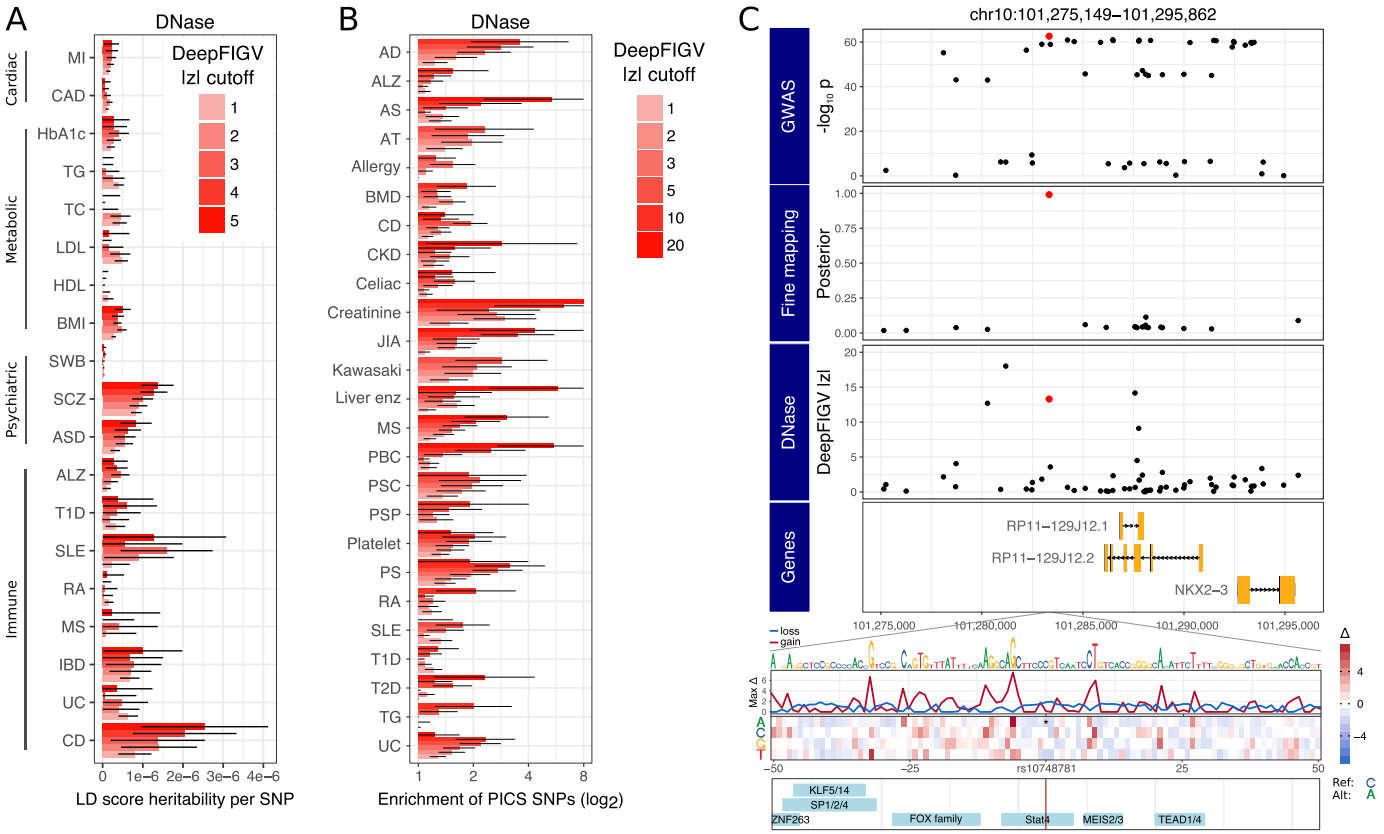

Figure 7. Disease risk variants are enriched for large DeepFIGV scores. (**A**) Linkage-disequilibrium score regression (LDSC) (8) partitioned heritability estimates for diseases in 4 categories. Heritability per SNP is computed for variants that exceed 5 cutoffs for DeepFIGV absolute *z*-score for DNase. Error bars indicate 2 standard deviations. (**B**) Enrichment of candidate causal variants for autoimmune disease (10) are variants exceeding six cutoffs DeepFIGV absolute z-score for DNase. Error bars indicate 2 standard deviations. (**C**) DeepFIGV elucidates molecular function of candidate causal variant for inflammatory bowel disease (67). GWAS identifies many correlated variants associated with disease risk, but statistical fine mapping identifies a single SNP (shown in red) as the candidate causal variant. This variant, rs10748781, disrupts a CpG site and is predicted to decrease the DNase signal in this region. *In silico* mutagenesis of 50bp around this SNP indicates that variants at nearby positions are predicted to decrease the DNase signal. Size of letters in DNA sequence is proportional to the maximum absolute delta at that position. Bottom panel shows TFBS motifs. Disease abbreviations: AD (Atopic dermatitis), ALZ (Alzheimer's), AS (Ankylosing spondylitis), ASD (Autism spectrum disorder), AT (Autoimmune thyroiditis), BMD (Bone mineral density), BMI (Body mass index), CAD (Coronary artery disease), CD (Crohn's disease), CKD (Chronic kidney disease) HbA1c (HbA1c protein level in blood), HDL (High-density lipoprotein), IBD (Inflammatory bowel disease), JIA (Juvenile idiopathic arthritis), LDL (Low-density lipoprotein), Liver enz (gamma glutamyl transferase), MI (myocardial infarction), MS (Multiple sclerosis), PBC (Primary biliary cirrhosis), PSC (Primary sclerosing cholangitis), PSP (Progressive supranuclear palsy), PS (Psoriasis), RA (Rheumatoid arthritis), SLE (Systemic lupus erythematosus), SWB (Subjective well-being), T1D (Type 1 diabetes), T2D (Type 2 diabetes), TC (total cholesterol), TG (Triglycerides), UC (Ulcerative colitis).

DISCUSSION

Translating findings of genetic studies to a molecular understanding of disease etiology and then to novel therapies has been hindered by the challenge of interpreting the functional consequence of genetic variants. There is a widely recognized need for accurate computational predictions of the functional impact of non-coding regulatory variants (69). Genomic annotations of the non-coding regions have generally taken one of four approaches. Evolutionary conservation or selection can identify functional regions of the genome, but consecutive nucleotides often have very similar scores and this approach does not give cell type- and assays-specific functional consequences (58,70,71). Epigenetic maps across multiple cell types, tissues and assays provide a functional interpretation, but peaks from these assays cover millions of nucleotides (17,18). Molecular trait QTL studies correlate genetic variants with gene expression or epigenetic signals, but interpretation of this correlation analysis is limited by linkage disequilibrium and is only applicable to variants observed in the dataset (20–22). Most recently, deep convolutional neural networks have been used to develop predictive models linking the genome sequence to splicing (72), protein binding (73), and epigenetic signals (27,29–31). Although these deep learning methods have been promising, their biological application has so far been limited.

Here, we present a deep learning framework that learns a predictive model linking DNA sequence to quantitative variation in epigenetic signal and evaluates the predicted functional impact of genetic variants on multiple assays. This framework models quantitative variation in the epigenome, integrates whole genome sequencing to create a personalized genome sequence for each individual, and trains on many experiments from the same cell type and assay. Because this framework fits a predictive model based on sequence context, it is less susceptible to issues of link-

age disequilibrium and can predict the functional impact of variants even if they are not observed in the training dataset.

Application to epigenetic assays of open chromatin (DNase-seq) and histone modifications (H3K27ac, H3K4me3 and H3K4me1) from 75 lymphoblastoid cell lines (LCL) (23,26) produces functional consequence scores that are concordant with other genomic annotations while capturing sequence context information beyond known TFBS motifs. We note that potential mechanisms of variants outside these motifs include affecting local DNA shape, DNA methylation or nucleosome positioning (74,75), but interpretation remains an open challenge (76). We demonstrate that these functional consequence scores inform molecular mechanism, are concordant with molecular trait QTL analysis, are predictive of allele-specific binding, and inform interpretation of risk variants for common disease. Moreover, these scores can prioritize variants for downstream experiments and indicate the appropriate cell type and functional assay. DeepFIGV scores are complementary to other non-coding variant scores, and compared to DeepSea (27) identifies more variants with extreme *z*-scores. (Supplementary Methods, Supplementary Figure S17).

Yet computational prediction of functional variants remains a challenging problem. DeepFIGV z-scores for DNase are correlated (spearman rho = 0.0802, $P = 5.32e–16$) with relative expression from an MPRA from the regulation saturation challenge from the Critical Assessment of Genome Interpretation (CAGI5, genomeinterpretation.org) (Supplementary Figure S18). But there is substantial room for improvement. Moreover, CADD (57) and GWAVA (60) show better performance in classifying pathogenic from benign variants in ClinVar (77) (Supplementary Figure S19). This result underscores that predicting the 'proximal' relationship between DNA and epigenetics is different that predicting the more complex relationship between DNA and higher-level disease phenotypes.

The differing performance of the prediction and biological enrichments across the four epigenetic assays is attributable to both biological and technical factors. These assays differ in the biological processes they measure. DNase measures open chromatin with high signal representing protein interacting with the DNA within a narrow region of ∼150 bp. Thus DNase signal is largely determined by the proximal DNA sequence and especially TF binding. Histone modification ChIP-seq is more complex readout of chromatin state with H3K4me3 at active promoters, H3K27ac at active promoters and enhancers, and H3K4me1 at active or primed enhancers. Due to spatial chromatin spreading, the role of trans-factors, and the increased width of these marks (300 bp to 1 kb), sequence-based prediction is known to be less accurate (27). Since genetic variants conferring disease risk or regulating gene expression can act through a number of mechanisms (9,20,23,26,67,69), the value of additional epigenetic assays depends on the accuracy of a predictive model as well as the regulatory mechanism of interest.

We note that since the current method is trained on the continuous epigenetic signal within peak regions learned from each assay, the method is dependent on the set of peaks. Moreover, there is likely meaningful epigenetic variation outside of strictly defined peak regions and, indeed, there is much interest in performing basepair-level predictions rather than summarizing the epigenetic signal for each peak (28,78). Thus although our predictions are peak-centric, and we observe that for DNase SNPs within DNase peaks are certainly associated with the observed basepair-level signal ($P = 6.5e–22$), SNPs outside peaks are still associated with basepair-level signal $P = 6.2e–5$) (Supplementary Figure S20).

Despite the remarkable experimental throughput of recent massively parallel report assays, these assays are limited to cell culture and they assay the function of the query sequence either in an episomal vector or through random insertion into the genome (59,63–66). Thus the degree to which results from MPRAs recapitulate function in the disease relevant cell type and natural genomic context remains unclear (65,79,80). In contrast, predictive models based on sequence context use natural genetic variation, are extensible to multiple biological assays, and evaluate sequences in their native chromosomal context. Moreover, they are applicable to cell culture, as well as cells from post mortem, biopsy or blood draws to more precisely target the relevant cell type.

The growth of large-scale resources pairing quantitative epigenetic assays with genetic data offers an opportunity to train rich predictive models on disease relevant cell types (25,33,81). Finally, we have developed a public resource of the DeepFIGV predicted functional scores for 438 million variants observed in previous sequencing projects available at deepfigv.mssm.edu.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Yungil Kim and Alexis Battle for providing data on gene expression outliers. We thank Anna Shcherbina and Anshul Kundaje for help accessing the ChromoVar3D data. We thank Judy Cho, Kristen Brennand, Andrew Chess and Schahram Akbarian for feedback throughout the project. We particularly thank Eugene Fluder and Hyung Min Cho for assistance with GPUs.

FUNDING

This work was supported by the National Institutes of Health (NIH) grants R01AG050986 (P.R.), R01MH109897 (P.R. and E.E.S.), U01MH116442 (P.R.), R01MH110921 (P.R.) and R01MH109677 (P.R.) and the Veterans Affairs Merit grants BX002395 and BX004189 (P.R.). Gabriel E. Hoffman is partially supported by a NARSAD Young Investigator Award 26313 from the Brain and Behavior Research Foundation, and a pilot grant from the Mount Sinai Alzheimer's Disease Research Center. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Funding for open access charge: NIMH [5R01MH109897-02].
Conflict of interest statement. None declared.

REFERENCES

1. Visscher,P.M., Wray,N.R., Zhang,Q., Sklar,P., McCarthy,M.I., Brown,M.A. and Yang,J. (2017) 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.
2. Spain,S.L. and Barrett,J.C. (2015) Strategies for fine-mapping complex traits. *Hum. Mol. Genet.*, **24**, R111–R119.
3. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS*, **106**, 9362–9367.
4. Pritchard,J.K. and Przeworski,M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.
5. Claussnitzer,M., Dankel,S.N., Kim,K.-H., Quon,G., Meuleman,W., Haugen,C., Glunk,V., Sousa,I.S., Beaudry,J.L., Puviindran,V. *et al.* (2015) FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.*, **373**, 895–907.
6. Kichaev,G., Yang,W.-Y., Lindstrom,S., Hormozdiari,F., Eskin,E., Price,A.L., Kraft,P. and Pasaniuc,B. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLos Genet.*, **10**, e1004722.
7. Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.
8. Finucane,H.K., Bulik-Sullivan,B., Gusev,A., Trynka,G., Reshef,Y., Loh,P.-R., Anttila,V., Xu,H., Zang,C., Farh,K. *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.
9. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
10. Farh,K.K., Marson,A., Zhu,J., Kleinewietfeld,M., Housley,W.J., Beik,S., Shoresh,N., Whitton,H., Ryan,R.J.H., Shishkin,A.A. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
11. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
12. Chong,J.X., Buckingham,K.J., Jhangiani,S.N., Boehm,C., Sobreira,N., Smith,J.D., Harrell,T.M., McMillin,M.J., Wiszniewski,W., Gambin,T. *et al.* (2015) The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.*, **97**, 199–215.
13. Peterson,T.A., Doughty,E. and Kann,M.G. (2013) Towards precision medicine: advances in computational approaches for the analysis of human variants. *J. Mol. Biol.*, **425**, 4047–4063.
14. Cline,M.S. and Karchin,R. (2011) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*, **27**, 441–448.
15. Zhang,W., Bojorquez-Gomez,A., Velez,D.O., Xu,G., Sanchez,K.S., Shen,J.P., Chen,K., Licon,K., Melton,C., Olson,K.M. *et al.* (2018) A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.*, **50**, 613–620.
16. Werling,D.M., Brand,H., An,J.-Y., Stone,M.R., Zhu,L., Glessner,J.T., Collins,R.L., Dong,S., Layer,R.M., Markenscoff-Papadimitriou,E. *et al.* (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.*, **50**, 727–736.
17. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
18. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
19. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
20. Aguet,F., Brown,A.A., Castel,S.E., Davis,J.R., He,Y., Jo,B., Mohammadi,P., Park,Y.S., Parsana,P., Segrè,A.V. *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
21. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
22. Fromer,M., Roussos,P., Sieberts,S.K., Johnson,J.S., Kavanagh,D.H., Perumal,T.M., Ruderfer,D.M., Oh,E.C., Topol,A., Shah,H.R. *et al.* (2016) Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.*, **19**, 1442–1453.
23. Grubert,F., Zaugg,J.B., Kasowski,M., Ursu,O., Spacek,D.V., Martin,A.R., Greenside,P., Srivas,R., Phanstiel,D.H., Pekowska,A. *et al.* (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, **162**, 1051–1065.
24. Waszak,S.M., Delaneau,O., Gschwind,A.R., Kilpinen,H., Raghav,S.K., Witwicki,R.M., Orioli,A., Wiederkehr,M., Panousis,N.I., Yurovsky,A. *et al.* (2015) Population variation and genetic control of modular chromatin architecture in humans. *Cell*, **162**, 1039–1050.
25. Chen, Ge,B., Casale,F.P., Vasquez,L., Kwan,T., Garrido-Martín,D., Watt,S., Yan,Y., Kundu,K., Ecker,S. *et al.* (2016) Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, **167**, 1398–1414.
26. Degner,J.F., Pai,A.A., Pique-Regi,R., Veyrieras,J.-B., Gaffney,D.J., Pickrell,J.K., De Leon,S., Michelini,K., Lewellen,N., Crawford,G.E. *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–394.
27. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931–934.
28. Kelley,D.R., Reshef,Y.A., Bileschi,M., Belanger,D., McLean,C.Y. and Snoek,J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, **28**, 739–750.
29. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.
30. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
31. Zhou,J., Theesfeld,C.L., Yao,K., Chen,K.M., Wong,A.K. and Troyanskaya,O.G. (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.*, **50**, 1171–1179.
32. LeCun,Y., Yoshua,B. and Geoffrey,H. (2015) Deep learning. *Nature*, **521**, 436–444.
33. Girdhar,K., Hoffman,G.E., Jiang,Y., Brown,L., Kundakovic,M., Hauberg,M.E., Francoeur,N.J., Wang,Y., Shah,H., Kavanagh,D.H. *et al.* (2018) Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nat. Neurosci.*, **21**, 1126–1136.
34. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
35. Comnish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Nucleic Acids Res.*, **13**, 3021–3030.
36. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., Van Der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.
37. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
38. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 1–14.
39. Cavalcante,R.G. and Sartor,M.A. (2017) Annotatr: genomic regions in context. *Bioinformatics*, **33**, 2381–2383.
40. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
41. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.
42. Brown,A.A., Viñuela,A., Delaneau,O., Spector,T.D., Small,K.S. and Dermitzakis,E.T. (2017) Predicting causal variants affecting

expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.*, **49**, 1747–1751.
43. Li, Kim,Y., Tsang,E.K., Davis,J.R., Damani,F.N., Chiang,C., Hess,G.T., Zappala,Z., Strober,B.J., Scott,A.J. *et al.* (2017) The impact of rare variation on gene expression across tissues. *Nature*, **550**, 239–243.
44. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
45. Bryois,J., Garrett,M.E., Song,L., Safi,A., Giusti-Rodriguez,P., Johnson,G.D., Shieh,A.W., Buil,A., Fullard,J.F., Roussos,P. *et al.* (2018) Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nature Communications*, **9**, 3121.
46. Dobin,A., Davis,Ca., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
47. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
48. Liao,Y., Smyth,G.K. and Shi,W. (2014) FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
49. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
50. Delaneau,O., Ongen,H., Brown,A.A., Fort,A., Panousis,N.I. and Dermitzakis,E.T. (2017) A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.*, **8**, 1–7.
51. Stegle,O., Parts,L., Durbin,R. and Winn,J. (2010) A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, **6**, e1000770.
52. Chen, Rozowsky,J., Galeev,T.R., Harmanci,A., Kitchen,R., Bedford,J., Abyzov,A., Kong,Y., Regan,L. and Gerstein,M. (2016) A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.*, **7**, 11101.
53. Shi,W., Fornes,O., Mathelier,A. and Wasserman,W.W. (2016) Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.*, **44**, 10106–10116.
54. Grau,J., Grosse,I. and Keilwagen,J. (2015) PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, **31**, 2595–2597.
55. Pers,T.H., Timshel,P. and Hirschhorn,J.N. (2015) SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics*, **31**, 418–420.
56. Alvarez,R.V., Li,S., Landsman,D. and Ovcharenko,I. (2018) SNPDelScore: combining multiple methods to score deleterious effects of noncoding mutations in the human genome. *Bioinformatics*, **34**, 289–291.
57. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
58. Huang, Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
59. Tewhey,R., Kotliar,D., Park,D.S., Liu,B., Winnicki,S., Reilly,S.K., Andersen,K.G., Mikkelsen,T.S., Lander,E.S., Schaffner,S.F. *et al.* (2016) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, **165**, 1519–1529.
60. Ritchie,G.R.S., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
61. Karimzadeh,M., Ernst,C., Kundaje,A. and Hoffman,M.M. (2018) Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.*, **46**, e120.
62. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
63. Starita,L.M., Ahituv,N., Dunham,M.J., Kitzman,J.O., Roth,F.P., Seelig,G., Shendure,J. and Fowler,D.M. (2017) Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.*, **101**, 315–325.
64. Ulirsch,J.C., Nandakumar,S.K., Wang,L., Giani,F.C., Zhang,X., Rogov,P., Melnikov,A., McDonel,P., Do,R., Mikkelsen,T.S. *et al.* (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, **165**, 1530–1545.
65. Ernst,J., Melnikov,A., Zhang,X., Wang,L., Rogov,P., Mikkelsen,T.S. and Kellis,M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.*, **34**, 1180–1190.
66. Arnold,C.D., Gerlach,D., Stelzer,C., Boryn,L.M., Rath,M., Stark,A., Boryn,Ł.M., Rath,M., Stark,A., Boryn,L.M. *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.
67. Huang, Fang,M., Jostins,L., Umićević Mirkov,M., Boucher,G., Anderson,C.A., Andersen,V., Cleynen,I., Cortes,A., Crins,F. *et al.* (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, **547**, 173–178.
68. Hannon,E., Weedon,M., Bray,N., O'Donovan,M. and Mill,J. (2017) Pleiotropic effects of trait-associated genetic variation on DNA methylation: utility for refining GWAS loci. *Am. J. Hum. Genet.*, **100**, 954–959.
69. Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
70. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
71. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
72. Xiong,H.Y., Alipanahi,B., Lee,L.J., Bretschneider,H., Merico,D., Yuen,R.K.C., Hua,Y., Gueroussov,S., Najafabadi,H.S., Hughes,T.R. *et al.* (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science*, **347**, 1254806–1254806.
73. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
74. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
75. Deplancke,B., Alpern,D. and Gardeux,V. (2016) The genetics of transcription factor DNA binding variation. *Cell*, **166**, 538–554.
76. Avsec,Ž., Kreuzhuber,R., Israeli,J., Xu,N., Cheng,J., Shrikumar,A., Banerjee,A., Kim,D.S., Beier,T., Urban,L. *et al.* (2019) The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.*, **37**, 592–600.
77. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
78. Avsec,Z., Weilert,M., Shrikumar,A., Alexandari,A., Krueger,S., Dalal,K., Fropf,R., McAnany,C., Gagneur,J., Kundaje,A. *et al.* (2019) Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. bioRxiv doi: https://doi.org/10.1101/737981, 21 August 2019, preprint: not peer reviewed.
79. Inoue,F., Kircher,M., Martin,B., Cooper,G.M., Witten,D.M., McManus,M.T., Ahituv,N. and Shendure,J. (2017) A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.*, **27**, 38–52.
80. Muerdter,F., Boryn,Ł.M., Woodfin,A.R., Neumayr,C., Rath,M., Zabidi,M.A., Pagani,M., Haberle,V., Kazmar,T., Catarino,R.R. *et al.* (2018) Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, **15**, 141–149.
81. PsychENCODE Consortium, Akbarian,S., Liu,C., Knowles,J.A., Vaccarino,F.M., Farnham,P.J., Crawford,G.E., Jaffe,A.E., Pinto,D., Dracheva,S. *et al.* (2015) The PsychENCODE project. *Nat. Neurosci.*, **18**, 1707–1712.