Received: 26 August 2016 | Revised: 19 January 2017 | Accepted: 12 February 2017

DOI: 10.1002/humu.23197

SPECIAL ARTICLE

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

Predicting gene expression in massively parallel reporter assays: A comparative study

Anat Kreimer[1,2] | Haoyang Zeng[3] | Matthew D. Edwards[3] | Yuchun Guo[3] | Kevin Tian[3] | Sunyoung Shin[4] | Rene Welch[4] | Michael Wainberg[5] | Rahul Mohan[5] | Nicholas A. Sinnott-Armstrong[5] | Yue Li[6] | Gökcen Eraslan[7] | Talal Bin AMIN[8] | Ryan Tewhey[11,12] | Pardis C. Sabeti[11,12] | Jonathan Goke[8] | Nikola S. Mueller[7] | Manolis Kellis[6] | Anshul Kundaje[5] | Michael A Beer[9] | Sunduz Keles[4] | David K. Gifford[3] | Nir Yosef[1,10]

[1]Department of Electrical Engineering and Computer Science and Center for Computational Biology, University of California, Berkeley, Berkeley, California

[2]Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California, San Francisco, California

[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts

[4]Department of Statistics, Department of Biostatistics and Medical Informatics University of Wisconsin-Madison, Madison, Wisconsin

[5]Department of Genetics, Stanford University School of Medicine, Department of Computer Science, Stanford, California

[6]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts

[7]Computational Cell Maps, Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany

[8]Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore

[9]McKusick-Nathans Institute of Genetic Medicine, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland

[10]Ragon Institute of Massachusetts General Hospital, MIT and Harvard, Cambridge, Massachusetts

[11]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts

[12]Broad Institute, Cambridge, Massachusetts

Correspondence
Anat Kreimer and Nir Yosef, 304A Stanley Hall University of California, Berkeley, CA 94720.
Email: anat.kreimer@berkeley.edu; niryosef@berkeley.edu

Contract grant sponsors: NIH (U41 HG007346, R13 HG006650, U01HG007910).

For the CAGI Special Issue

Correction added on 10th April 2017, after first online publication: Two additional authors Ryan Tewhey and Pardis C. Sabeti are added to the authors list.

Abstract

In many human diseases, associated genetic changes tend to occur within noncoding regions, whose effect might be related to transcriptional control. A central goal in human genetics is to understand the function of such noncoding regions: given a region that is statistically associated with changes in gene expression (expression quantitative trait locus [eQTL]), does it in fact play a regulatory role? And if so, how is this role "coded" in its sequence? These questions were the subject of the Critical Assessment of Genome Interpretation eQTL challenge. Participants were given a set of sequences that flank eQTLs in humans and were asked to predict whether these are capable of regulating transcription (as evaluated by massively parallel reporter assays), and whether this capability changes between alternative alleles. Here, we report lessons learned from this community effort. By inspecting predictive properties in isolation, and conducting meta-analysis over the competing methods, we find that using chromatin accessibility and transcription factor binding as features in an ensemble of classifiers or regression models leads to the most accurate results. We then characterize the loci that are harder to predict, putting the spotlight on areas of weakness, which we expect to be the subject of future studies.

KEYWORDS

eQTLs, functional genomics, gene regulation, massively parallel reporter assays

1 | INTRODUCTION

Mapping genotype to phenotype has been the focus of many studies in the post genomic era, with an increasing focus on the noncoding genome (Farh et al., 2015; Hindorff et al., 2009; Maurano et al., 2012; Weingarten-Gabbay & Segal, 2014; Welter et al., 2014; Zhou et al., 2013). Gene expression has been and is still one of the most well-investigated phenotypes by such studies, starting with the pioneering work that modeled it as a function of sequence features of proximal promoter regions, focusing primarily on the occurrence, location, orientation, and cooperative interactions (Das, Banerjee, & Zhang, 2004; Segal, Raveh-Sadka, Schroeder, Unnerstall, & Gaul, 2008) of transcription factor (TF)-binding motifs, and K-mer frequencies (Beer & Tavazoie, 2004; Bussemaker, Li, & Siggia, 2001; Nguyen & D'Haeseleer, 2006). With the development of sequencing-based technologies for chromatin profiling (Thurman et al., 2012), methods for prediction of gene expression advanced accordingly, now adding experimentally measured chromatin properties as features (e.g., TF binding or histone modifications using ChIP-seq; chromatin accessibility using DNAse-seq or ATAC-seq) (Dong et al., 2012; Gonzalez, Setty, & Leslie, 2015; Marstrand & Storey, 2014; Natarajan, Yardimci, Sheffield, Crawford, & Ohler, 2012; Wilczynski, Liu, Yeo, & Furlong, 2012). With these new types of data, a number of related questions and challenges emerged. One set of studies aimed at annotating the state of the chromatin into broad categories (e.g., enhancers, insulators) based on observed chromatin features and/or DNA sequence (Ernst & Kellis, 2010; Hoffman et al., 2012) and then associate the resulting distal regulatory regions with the correct target gene (e.g., using chromatin conformation assays [Rao et al., 2014] or computational inference [Gonzalez et al., 2015]). Another set of studies aimed at predicting chromatin features (e.g., accessibility, TF binding) based on DNA sequences (Weingarten-Gabbay & Segal, 2014), and predict the dependence of epigenetic features on genetic variation (e.g., single-nucleotide variants [SNVs] [Ernst & Kellis, 2010; Erwin et al., 2014; Lee et al., 2015]). Recent methods, some of which are applied here, "closed the loop" and use DNA sequence alterations to predict changes in epigenetic features, which are then used as features for predicting the pertaining effects on the expression of the putative target genes.

One of the major hurdles in advancing this field and characterizing the regulatory "code" of the genome has been the lack of a well-controlled and scalable experimental system, which allows to investigate the direct effect of any sequence alteration of interest. A substantial progress to this end was the development of massively parallel reporter assays (MPRA)—a cost-effective, high-throughput activity screening of fully synthesized DNA regions (Fig. 1) (Smith et al., 2013; Weingarten-Gabbay & Segal, 2014). In MPRA, a library of thousands of putative regulatory DNA elements (each about 150 nt in length) with coupled unique tags is synthesized and used to generate a pool of plasmids; this pool is then transfected into cells, and the regulatory activity (as an enhancer or promoter) associated with the respective DNA element is assessed by sequencing the abundances of the expressed tags. Since oligoarrays can now be printed in a cost-effective fashion, MPRA provides a feasible means (albeit only in an episomal context (Inoue et al., 2017)) to systematically interrogate how regulatory activity is encoded in the DNA (Birnbaum et al., 2014; Kheradpour et al., 2013; Melnikov et al., 2012; Mogno, Kwasnieski, & Cohen, 2013; Patwardhan et al., 2012; Sharon et al., 2012; Smith et al., 2013), and estimate the effects of sequence variants.

Since MPRA is still a nascent technology, computational methods that make effective use of it are still emerging (Gertz, Siggia, & Cohen, 2009). Specifically, how to leverage MPRA to build better and more accurate models for predicting whether a DNA region of interest plays a regulatory role, and if so, how does this activity changes upon slight sequence variations (SNVs of short insertions or deletions [indels]), commonly observed in human cohorts (Zhou et al., 2013). The Critical Assessment of Genome Interpretation expression quantitative trait locus (eQTL) challenge is the first community effort aimed at advancing this type of studies. It is based on a comprehensive profiling of eQTLs observed in a subset of the Geuvadis database (Lappalainen et al., 2013) with MPRA, culminating in over 9,000 regulatory sequences, in their reference genome form and their alternative (SNV or indel) form (Tewhey et al., 2016). The goal of the challenge was twofold—first, participants were asked to predict the regulatory activity of each regulatory sequence (reference or alternative allele) in isolation. Then, the participants were asked to predict the differences between each pair of alleles.

In the following sections, we describe the results of this challenge and the lessons that can be learned via a meta-analysis of the competing methods. We start by summarizing the properties that were used by the participants as predictive features, divide these properties into several categories, and inspect the predictive ability of representative features from each category in isolation. We then move to inspect the predictive algorithms used by the participants and evaluate their overall performance using a range of metrics. As expected, we find that the predictions for the first part of the challenge (i.e., predicting the regulatory activity of each allele separately) were much more accurate than the second part (predict the differences between alleles), reflecting the difficulty in modeling the effects of nuanced sequence modifications. Furthermore, we find that overall the ranking of the participants is stable across specific subtasks and performance metrics, with the most promising methods belonging to the ones that "close the loop" as stated above. These methods use the DNA sequence as a primary feature for predicting epigenetic properties. These properties, in turn, are used to train an ensemble of models (e.g., using different learning algorithms) to provide a robust prediction of transcriptional activity and its dependence on sequence variation.

Focusing on loci that showed strong transcriptional activity in the MPRA, and taking a meta-analysis approach, we find that there are two distinct subsets of loci—one whose activity is predicted accurately by all (or most) competing methods, and another whose activity is poorly predicted by all competitors. Expectedly, we find that the "hard-to-predict" regions are associated with "paradigm-violating" properties, such as lack of accessibility or no apparent TF binding (as inferred by DNA-seq and ChIP-seq, or predicted by sequence-based models). While this is probably to some extent a result of inspecting regulatory activity in episomal setting, it may also point to knowledge gaps in defining the predictive features (e.g., unknown TF-binding preferences) or properly combining these features in a predictive model.

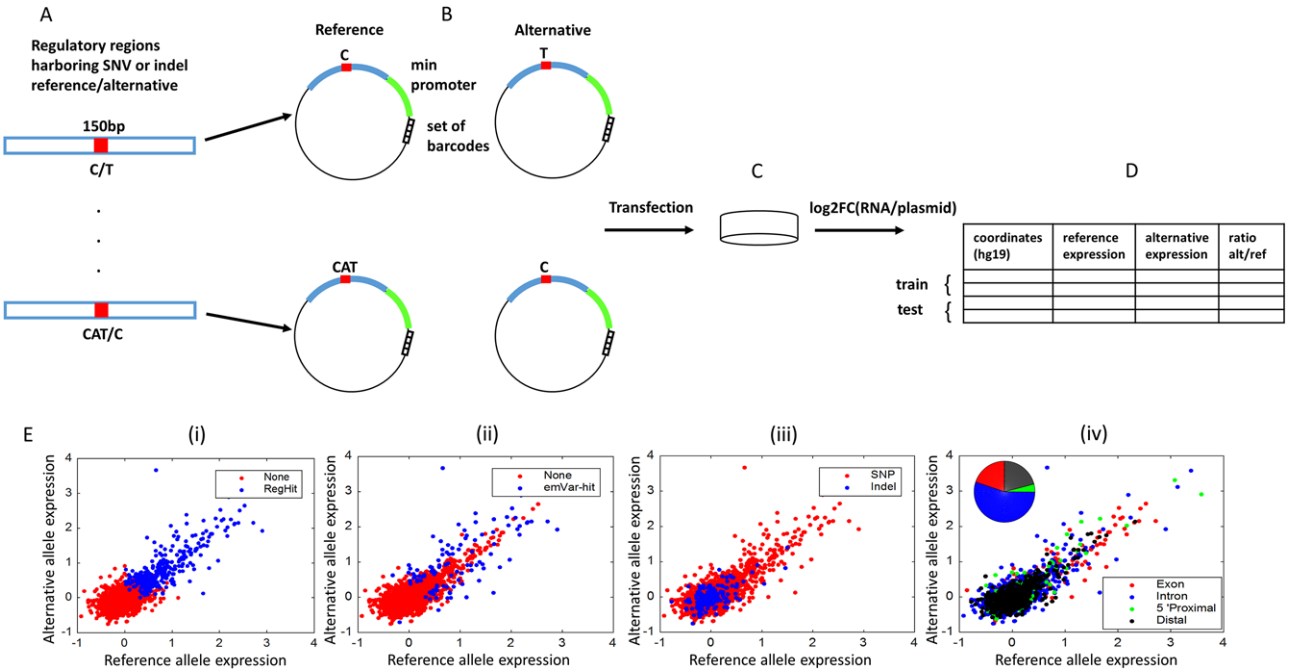

FIGURE 1 Experimental and challenge design. **A**: Selection of regulatory regions that harbor a short polymorphism (SNV or indel). **B**: Design of MPRA constructs for both reference and alternative alleles. **C**: Transfection into two different LCLs. **D**: Data provided for the challenge. **E**: Alternative versus reference allele expression for training set regions: (i) regulatory hit/non regulatory regions, (ii) emVar hit/non-emVar regions, and (iii) indel/SNP regions are marked in blue/red, respectively. (iv) Exon, intron, promoter, and distal regions are marked in red/blue/green/black, respectively

2 | THE EQTL-CAUSAL VARIANT CHALLENGE

Tewhey et al. (2016) used MPRA as a tool for investigating genetic variants that are statistically associated with changes to gene expression, considering both SNVs and indels. These eQTLs were inferred based on a collection of immortalized lymphoblastoid cell lines (LCLs) derived from a large set of individuals, where both genome sequences and transcription profiles are available (Tewhey et al., 2016). The eQTL-causal SNP Critical Assessment of Genome Interpretation challenge was based on a subset of these loci, and included 3,157 eQTLs inferred based on individuals of European ancestry (1000 Genomes Project Consortium et al., 2012; Lappalainen et al., 2013). Each eQTL in this collection is in turn associated with one or more variants (with an average of three variants per eQTL) whose individual effects are statistically indistinguishable due to linkage disequilibrium (LD; considering all loci that are in perfect LD with the top associated variant), leading to an overall set of 9,116 variants (8,570 SNVs and 546 indels; Fig. 1).

For each variant, a sequence of 150 bp was synthesized, which includes the surrounding genome sequence (using reference genome *hg19*) with the variant located at the central position (Fig. 1A). Each SNV was associated with two sequences that are identical except for the respective variation in position 76. For indels, the longer of the two alleles was designed as a 150-nt oligonucleotide; the shorter allele was then designed with the same flanking sequences as the longer allele (e.g., for a single-nucleotide indel TC/C: X[TC]Y and X[T]Y, where X and Y are 74-bp long DNA segments that flank the variant in the reference genome). To increase the accuracy and sensitivity of the assay, 20 nt barcodes were added to the oligos by emulsion PCR, such that each oligo is represented by an average of a thousand barcode tags within the plasmid pool (Fig. 1B). The plasmid library was electroporated into two LCLs (NA12878 and NA19239), using five and three technical replicates, respectively. Importantly, NA12878 is an ENCODE tier 1 cell line (ENCODE Project Consortium, 2012) and thus a large number of genomic assays performed on this cell line are publically available. Twenty-four hours after transfection, the GFP reporter mRNA was captured by hybridization and RNA sequencing of the 3'-UTR-adjacent barcodes was performed to quantify the influence of each 150-bp sequence on regulation of the reporter gene. RNA expression measured in barcode read counts was normalized relative to the input plasmid barcode counts determined by DNA sequencing, such that the reported MPRA output consisted of an estimate for the ratio between the number of transcripts (RNA-seq) and plasmids (DNA-seq) (Fig. 1C).

The resulting MPRA dataset was divided into a training set, which was made available to the participants, and a test set, which was held off and used for evaluation by independent assessors. The complete data set, including training and test subsets is provided in Supp. Table S1. The training set (Fig. 1D and E) consisted of MPRA results for 3,044 variants (2,874 SNVs and 170 indels) associated with 1,052 eQTLs. For each variant, the information available for training included: (1) the respective genomic coordinates (using the hg19 reference genome), and the position and type of the variant. (2) An estimated transcript to plasmid ratio for each allele (defined as log fold change [*log2FC*]; averaged across replicates). (3) An indication whether or not at least one of the two alleles exhibits a significantly high ratio of transcripts to DNA (regulatory hit). Significance of differential abundance of transcripts

versus plasmid input was evaluated using DE-seq2 (Love, Huber, & Anders, 2014) with a false discovery rate (FDR) cutoff of 1% to call hits. (4) Comparison of transcriptional activity between the two alleles. As before, this included a quantitative field indicating the fold change between the transcriptional activity of the two alleles (alternative/reference; *LogSkew*); and a binary field, indicating whether the difference is statistically significant (expression-modulating variants; emVar hits). Significance of allelic skew was evaluated using a *t*-test on the log-transformed RNA-seq/plasmid ratios across replicates with a FDR cutoff of 5% to call hits. (5) The name of the eQTL-associated gene and the association's coefficient (beta), *t*-statistic, and *P* value. As expected, hit regions have significantly higher expression (Fig. 1E (i and ii)).

The challenge consisted of two parts, each with its own test set. In the first part, participants were asked to predict the level of transcriptional activity (*log2FC*) for each allele, and determine for each variant whether at least one of the alleles is a regulatory hit. In addition, each prediction should have included a standard deviation, reflecting the confidence in the predicted values. The corresponding test set (Supp. Table S1) consisted of MPRA results for 3,006 variants (2,811 SNVs and 195 indels) associated with 1,050 eQTLs. In the second part, the participants were given variants that are confirmed regulatory hits and asked to predict the difference between the transcriptional activity of the two alleles, both quantitatively (*LogSkew*) and qualitatively (*emVar hits*). As before, each prediction should also include an estimate of statistical confidence. The corresponding test data consisted of MPRA results for 401 variants (370 SNVs and 31 indels) associated with 1,055 eQTLs (Supp. Table S1).

Seven groups participated in the challenge. Each group was allowed to submit multiple predictions, resulting in overall 20 submissions for the first part, and 13 submissions for the second part. The submissions spanned a wide array of predictive features and prediction algorithms, as summarized in Table 1.

3 | RESULTS

3.1 | Predictive features

The features that were used by the participating groups can be categorized into several classes. (1) Experimentally measured epigenetic properties, including TF-binding sites (TFBS), histone marks, chromatin accessibility (primarily by identifying DNase-hypersensitivity sites; henceforth abbreviated as DHS), and DNA methylation. To define these features, each reference allele is mapped to the reference human genome, and then queried against tracks of epigenetic properties (primarily from ENCODE [ENCODE Project Consortium, 2012] and the Epigenome Road map [Romanoski, Glass, Stunnenberg, Wilson, & Almouzni, 2015]), measured in LCLs and other cell lines. (2) Predicted epigenetic properties: this set of features covers similar properties as the experimentally derived ones (e.g., TFBS or DHS). However, instead of being directly measured, the properties are inferred based on the DNA sequence of the respective MPRA construct, using models trained on experimental data (e.g., protein-binding microarrays [Newburger & Bulyk, 2009] for TFBS, or DNase-seq [ENCODE Project Consortium, 2012] for DHS). A wide array of models for the prediction of epigenetic properties from sequence were used, from simple DNA-binding motif scoring (Grant, Bailey, & Noble, 2011) to more recent supervised learning algorithms such as *DeepBind* (Alipanahi, Delong, Weirauch, & Frey, 2015), gkm-support vector machine (SVM) (Ghandi, Lee, Mohammad-Noori, & Beer, 2014; Lee et al., 2015), and *Basset* (Kelley, Snoek, & Rinn, 2016). (3) Other locus-specific properties, including variant information (e.g., indication whether a variant is defined as a leading SNP [Tewhey et al., 2016]), and evolutionary conservation. (4) DNA k-mer frequencies.

Notably, the sequence-based features (feature classes 2 and 4) associate the reference and alternative alleles with different values (reflecting the differences in their respective DNA sequences). Conversely, in features that are based on direct characterization of the loci in the reference genome (feature classes 1 and 3), the two alleles are associated with the same value. Features of the former classes may therefore be more directly applicable for distinguishing between the two alleles at the second part of this challenge.

While the competing groups combined multiple feature sets using learning algorithms, we first wanted to explore the predictive capacity of each feature in isolation. To this end, we assembled a representative set from each feature class and measured its accuracy when applied on the test data sets (Supp. Table S1; Fig. 2). For class 1, we include 20 epigenetic properties, derived from experimental profiling of LCLs by the ENCODE consortium (ENCODE Project Consortium, 2012). These features include DHS (using DNase-seq), multiple histone modification, and TFBS (using ChIP-seq). These profiles were interpreted as binary, where a value of 1 indicates that the respective region overlaps with a peak of the respective signal (provided by the ENCODE unified pipeline [ENCODE Project Consortium, 2012]). For class 2, we included the number of predicted TFBS based on the presence of DNA-binding motifs form the ENCODE collection (Grant et al., 2011; Kheradpour & Kellis, 2014), or using a neural network model trained on protein-binding microarrays (Alipanahi et al., 2015). We also included the distance between the transcription start site of the MPRA construct and the nearest motif hit, and several sequence-based properties related to the DNA structure, including length of polyA/T subsequence representing nucleosome disfavoring sequences, GC content, and predicted DNA shape features including minor groove width, roll, propeller twist, and helix twist (Zhou et al., 2013). For class 3, we included evolutionary conservation scores, predicted by phastCons (Siepel et al., 2005). We did not include class 4 features (k-mers) in this analysis.

We used a number of tests to evaluate the accuracy of each feature (Supp. Note S1). For the regression tasks, that is, predicting the expression of the reference and alternate allele (*log2FC*) and their ratio (*LogSkew*), we applied several correlation measures (Person, Spearman, Kendall), considering either the entire test data, variants at the top 25% transcriptional activity (defined as the maximum *log2FC* of the two alleles) or absolute allelic skew, or a discretization of the data (predicted and observed) into quintiles. For the binary predictions (i.e., regulatory hit and *emVar* hit), we record the area under the receiver operating characteristic (ROC) and precision recall (PR) curves. To better

TABLE 1 Summary of submissions

Group	Features (feature classes 1–4)	Methods (part I)	Methods (part II)
1	Histone modifications in K562 cells (ENCODE Project Consortium, 2012) (class 1); evolutionary conservation (Siepel et al., 2005) (class 3); k-mer frequencies (class 4)	Regularized regression (e.g., elastic net [Zou & Hastie, 2005]; R, C), random forest (R, C), SVR (R), SVM (C)	Same as part I
2	Histone modifications, DHS, and TFBS in LCL (ENCODE Project Consortium, 2012) (class 1); predictions of DHS in 164 cell lines (Kelley et al., 2016) (class 2); predictions of TFBS (Cowper-Sal lari et al., 2012) and LCL-specific histone modifications and DHS based on ENCODE Project Consortium (2012) (class 2)	Ensemble of gradient boosting models (Pedregosa et al., 2011). Each model trained on a different feature subset (R, C).	Same as part I
3	k-mers (class 4)	Linear SVR (R) and SVM (C)	Same as part I
4	Part I: segmentation of genomic regions based on histone modifications in LCL (ENCODE Project Consortium, 2012; Ernst & Kellis, 2012) (class 1); predictions of TFBS, DHS, and histone marks (using Alipanahi et al. (2015) and Zhou & Troyanskaya, (2015)), with data from ENCODE Project Consortium (2012) and Romanoski et al. (2015) (class 2) Part II: allele-specific activity level predicted by the models in part I	Ensemble of models, using LASSO or random forest, and trained on different feature subsets (R). Ensemble of neural networks, trained different feature subsets (C).	Difference between predicted alleles' scores (R) Ensemble of classifiers (e.g., KNN; C)
5	Predictions of DHS (using Lee et al. (2015) with LCL data from ENCODE Project Consortium (2012)) (class 2)	Predicted alleles' DH scores are used directly (R,C)	Difference between DHS scores of the two alleles (Ghandi et al., 2014) (R, C)
6	Part I: histone modifications, DHS, DNA-methylation, and TFBS in LCL (ENCODE Project Consortium, 2012) (class 1); predictions of TFBS, and protein-binding sites in the transcribed RNA (using Alipanahi et al. (2015), Grant et al. (2011), and Hume, Barrera, Gisselbrecht, & Bulyk (2015)), with data from Alipanahi et al. (2015) and ENCODE Project Consortium (2012) (class 2) Part II: all of features form part I, plus allele-specific activity levels predicted by the models in part I	Random forest (R, C). The classifier used the results of the regression task as additional features.	Random forest (R, C)
7	Predictions of TFBS, DHS, and histone marks, using Zhou and Troyanskaya (2015), with data from ENCODE Project Consortium (2012); (class 2). 0/1 indicator of leading variant and eQTL *P* value (class 3)	Random forest	Same as part I

DHS, DNase-hypersensitivity sites; eQTL, expression quantitative trait locus; KNN, K nearest neighbor; TFBS, transcription factor-binding site.
Notes: Methods and features used by each group for the two parts of the challenge. Features are divided into four classes: (1) experimentally measured epigenetic properties, (2) predicted epigenetic properties, (3) other locus-specific properties, and (4) DNA k-mer frequencies. For the methods: R corresponds to the regression tasks (predicting Log2FC in part I or LogSkew in part II); C corresponds to the classification tasks (predicting regulatory hits in part I or emVar hits in part II).

account for the binary predicted values (i.e., class 1 features), we also applied a fold enrichment test by examining the overlap between the set of true positives (regulatory hits or *emVar* hits) and the predicted positives. For quantitative features (classes 1, 3), the predicted positives were defined as regions with value higher than the mean. The significance of each test was evaluated by the respective statistical test (correlation *P* values for the regression tasks; Kolmogorov–Smirnov test for ROC and PR; hypergeometric *P* value for the enrichment test). All *P* values were corrected using the Benjamini–Hochberg procedure, and only associations below a FDR of 5% are presented.

Consistent with the previous literature (Erwin et al., 2014; Kwasnieski, Fiore, Chaudhari, & Cohen, 2014; Smith et al., 2013), the most highly predictive features for the absolute expression levels (part I) are those related to TF binding (number of bound TF, inferred either computationally or experimentally) and chromatin accessibility (Fig. 2A). We note that there is an overall high correlation within features categories (Supp. Fig. S1) and negligible correlation between features and eQTL statistics (Supp. Fig. S2). Notably, the set of all MPRA regions is significantly enriched with majority of histone marks with respect to the entire genome (Supp. Table S2), as expected (Francois Aguet et al., 2017; Fromer et al., 2016). However, within those regions, we find the histone marks to have a low ability to predict eQTL strength (Supp. Fig. S2).

We also find a significant positive relationship in various histone modifications that are associated with active regulatory regions (e.g., H3K27ac). Considering the contribution of individual TFs, we find several regulators whose predicted binding sites are particularly predictive of regulatory activity of MPRA constructs (Fig. 2B). Interestingly, among the top TFs are Batf and Irf4 (supported by ChIP-seq data as well; Fig. 2C), which are highly expressed in LCL and are known to form a heterodimer that performs pioneer functions (i.e., recruitment of chromatin remodeling machinery) in T cell development (Ciofani et al., 2012). For part II, we do not observe any individual features that are significantly predictive, considering different ways to aggregate the

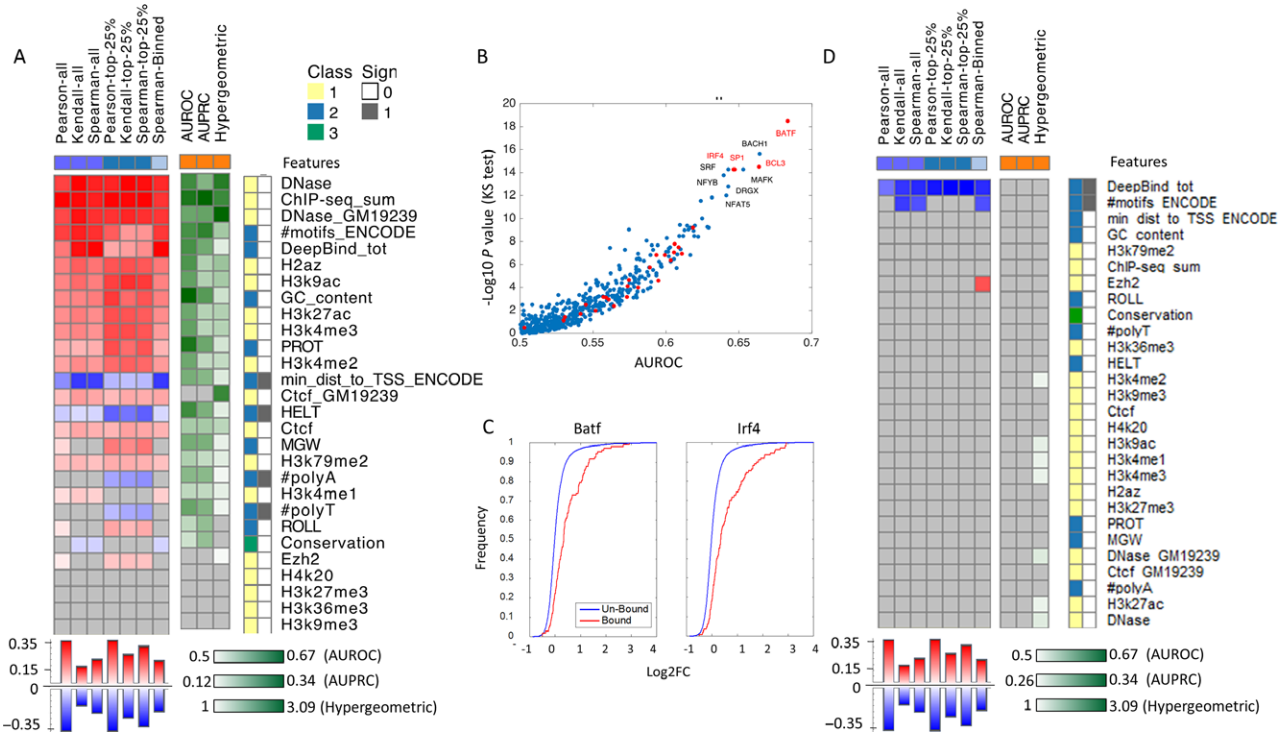

FIGURE 2 Individual feature accuracy using standard statistical tests. Features classes are divided in four categories. (1) Experimentally measured epigenetic properties (i.e., DHS, multiple histone modification and TFBS using ChIP-seq). (2) Predicted epigenetic properties (i.e., TFBS predictions: #motifs_ENCODE—number of predicted TFBS based on the presence of DNA-binding motifs from ENCODE), or DeepBind_tot—using a neural network model trained on protein-binding microarrays. min_dist_to_TSS_ENCODE—the distance between the transcription start site of the MPRA construct and the nearest motif hit, #PolyA, #PolyT—length of polyA/T subsequence, GC content, and DNA shape features: minor groove width (MGW), roll, propeller twist (PROT), and helix twist (HELT). DeepBind_tot feature was derived by marking the regions that score at the top 90% as hits for every TF, and for every region, count the number of TFs for which it has a hit. The aggregation method we use for these features is log fold between the alternate and reference allele and subtraction for DeepBind_tot. (3) Locus-specific properties (i.e., evolutionary conservation scores). (4) k-mer frequencies (not included in this analysis). For regression tasks, we applied several correlation measures (Person, Spearman, Kendall), considering either the entire test data (purple squares), variants at the top 25% of quantitative measurements (blue squares), or a binning of the data (light blue squares). For the binary predictions, we record the AUROC and AUPRC (orange squares). For both regression and classification tasks, we applied a hypergeometric test (orange square). The features are ranked based on the median performance across all tests and presented sorted from the most to the least predictive. Nonsignificant correlations are marked in gray; high positive/high negative/low correlation is marked in red/blue/white for regression and dark green/light green/white for classification, respectively. Features categories 1/2/3/4 (only the first three presented in this figure) are denoted by yellow/blue/green/pink and the sign of the correlation positive/negative is marked with white/gray. **A**: Part I: regression tasks include the expression of the reference and alternate allele; classification task includes regulatory hit prediction. **B**: Contribution of individual TFs for predicting regulatory activity of MPRA constructs measured by the minus log *P* value of the ks-test for AUROC per factor **C**: Cumulative distribution of regulatory activity for regions that are bound/un bound by two of the most predictive factors (BATF and IRF4). **D**: Part II: regression tasks include allelic skew and classification task includes emVar hit prediction

scores of the two alleles in feature class 2 (max, min, difference), and using a lenient cutoff of FDR<0.1 (Fig. 2D). The only exception was a weak positive signal from the differences in number of predicted TFBS, attesting to the complexity of predicting differences that hinge on a single nucleotide difference or few nucleotide differences. Since MPRA is subject to experimental error, as most other assays, we have tested the robustness of these results by resampling, and observed overall consistent results (Supp. Table S3).

3.2 | Combining predictive features with learning algorithms

The prediction algorithms for the first part of this challenge included a wide array of standard machine learning techniques, both for the classification task (regulatory hits; e.g., SVM, random forest, neural networks), and the regression tasks (*Log2FC*; e.g., support vector regression, regularized linear models, regression trees), with random forests being the most widely used method (Table 1). Some of the groups further used an ensemble of predictors, by varying either the type of prediction algorithm (e.g., combining linear models and random forest) or the set of predictive features (e.g., different classifiers, each using epigenetic features predicted by a different algorithm (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015)). For instance, group #2 applied gradient boosting for classification of regulatory hits, and reported an aggregate over the predicted scores of eight different models, trained on eight distinct feature sets, with each set including either experimentally derived epigenetic properties (e.g., ChIP-seq signal over the respective loci in LCL from ENCODE), or computationally

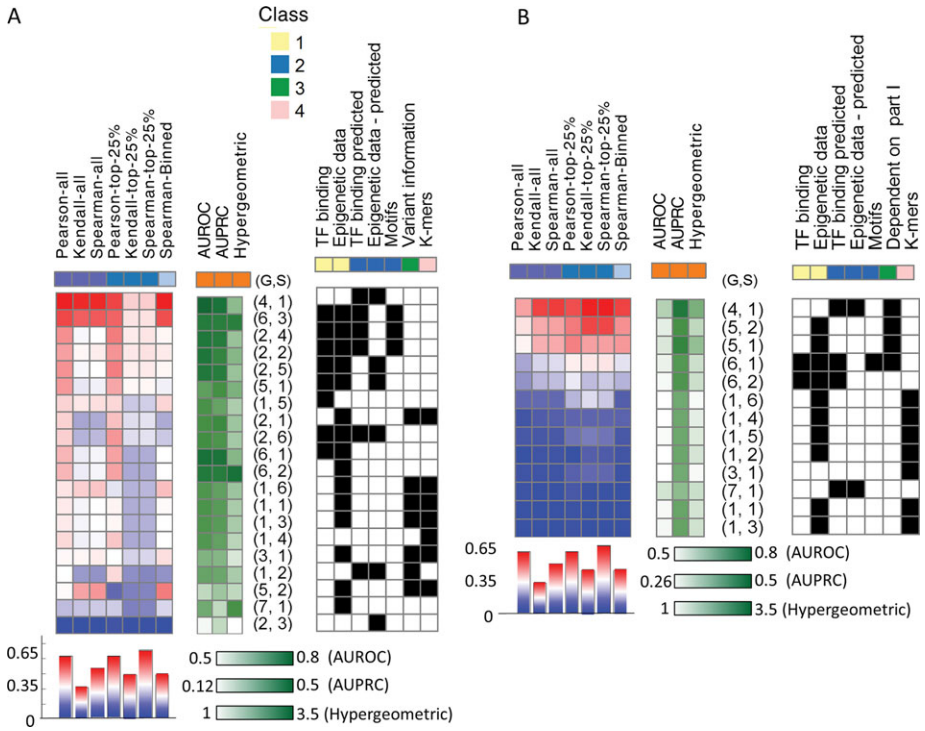

FIGURE 3 Summary of performance and features used per submission. The statistical tests used and features examined are similar to Figure 2. High/low performance is marked in red/blue and dark/light green, respectively. Group number and submission are denoted by (G,S); features use is indicated by black(1)/white(0) heat-maps. The submissions are ranked based on the median performance across all tests and presented sorted from high to low performance. **A**: Part I. **B**: Part II. The third class of features for part II indicates whether the model is dependent on predictions from part I (green square).

derived ones (e.g., DNase hypersensitivity predictions in 164 cell lines using Basset [Kelley et al., 2016]).

In part II, we observed two main strategies: the first was to treat this part independently from part I, and build predictive models using the feature categories summarized above. The alternative strategy was to use the individual activity of each allele, as predicted in part I, to infer their differential activity (quantitatively: *LogSkew*; and qualitatively: *emVar hits*). For instance, using a similar strategy as in the *deepSEA* framework (Zhou & Troyanskaya, 2015), group #4 (Zeng, Edwards, Guo, & Gifford, 2016) built a two-step classifier: in the first level, it applies the classifiers from part I to predict the transcriptional activity of each allele; in the second part, it uses the predicted activity of each allele and the difference between them as features for predicting emVar hits using an ensemble of classifiers, including regularized logistic regression, random forest, SVM, and K nearest neighbors.

3.3 | Performance evaluation

The groups submitted their predictions for the held-out test data sets, which we then used for evaluation. We used similar performance tests as above to rank the groups, and then derived an overall ranking by taking the median across tests (Fig. 3). Not surprisingly, we observe that the relative performance within each part is overall consistent across the different tasks, for instance, the accuracy of predicting transcriptional activities (*Log2FC*) is highly indicative of the performance in the related classification task (regulatory hit). The consistency between the two parts was less substantial (even though the top ranking group was similar), for instance, with group #5 performing well in part II, but less so in part I. Interestingly, this group based its predictions on feature class 2 only (sequence-based predictions of epigenetic properties). While their method was previously shown to be predictive of the effects of subtle sequence changes (mostly SNV) on chromatin accessibility (Ghandi et al., 2014), the lack of direct experimental measurements as features (i.e., class 1) might be the cause for the less favorable performance in part I compared to groups that used class 1 features. More generally, we observe a substantially worse performance in part II in comparison to part I, which is consistent with the single feature analysis above, and reflects the difficulty of predicting the effects of nuanced sequence modifications.

Next, we wanted to assess whether there are specific classes of models or combinations of features that are associated with better performance. To address this, we record for each submission the types of models and features that were used (Table 1; Fig. 3). We note that ensemble methods were generally better performing, highlighting the need for robust inference methodologies, and consistently with other applications of machine learning in biology (Marbach et al., 2012). Furthermore, it is clear that nonlinear methods perform better, an expected result given the plausibility of nonlinear (Das et al., 2004) and combinatorial (Spitz & Furlong, 2012) effects of the features. For part I, we observe that, generally, including TFBS as features (either predicted or experimentally derived) leads to better performance, which is consistent with the individual feature analysis (Fig. 3A). For part II, we find that relying on models trained in part I (i.e., using the predicted allele

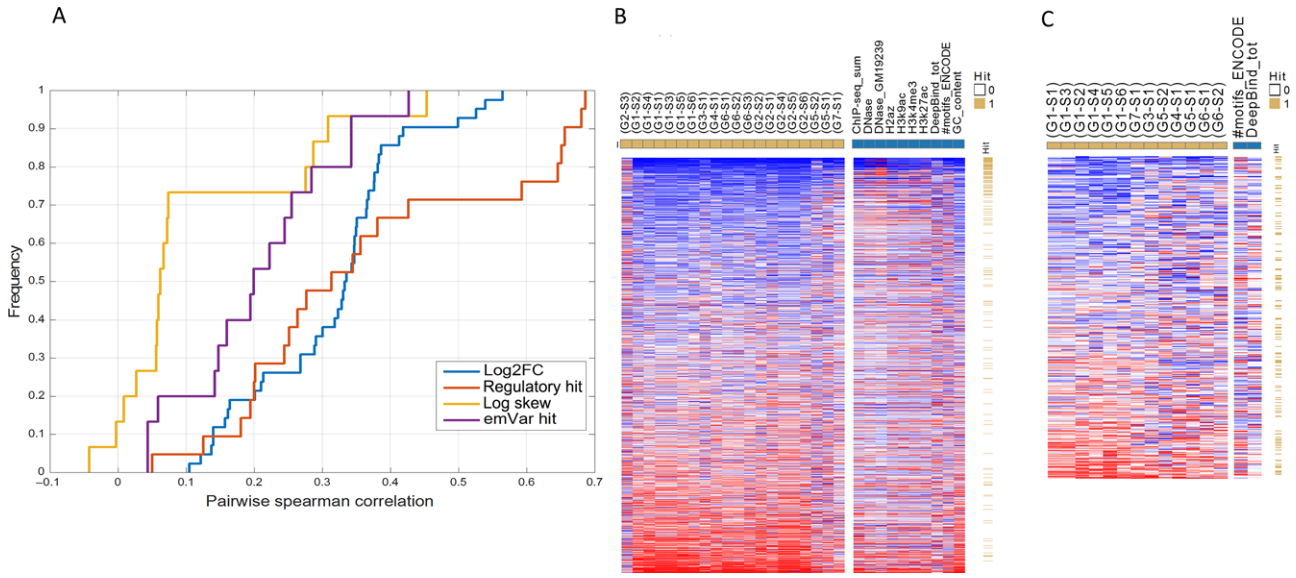

FIGURE 4 Regions hardness. Respective accuracy per region and submission is defined as the absolute difference between the observed and predicted rank, scaled by the expected difference (using random ranking). The "region hardness to predict" per part is defined as the mean rank across all tasks. Hard/easy to predict regions are denoted by red/blue, respectively, and sorted from easy to hard. **A**: Cumulative distribution of the Spearman correlation coefficient when comparing each pair of groups (taking the maximum over all possible pairs of submissions) for their regions accuracy per prediction (log2FC, regulatory hit, log skew, emVar hit). **B**: Left panel: heat-map of regions hardness for part I when using the predictions from all groups (yellow squares). The regions are sorted by their rank and denoted whether they are regulatory hits (yellow/white). Right panel: heat-map of regions hardness when using the top 10 features as predictors (blue squares). **C**: Left panel: heat-map of regions hardness for part II when using the predictions from all groups (yellow squares). The regions are sorted by their rank and denoted whether they are emVar hits (yellow/white). Right panel: heat-map of regions hardness when using the top two features as predictors (blue squares)

activity levels as features) leads to improved performance (Wilcoxon ranksum test *P* value 0.0028). (Fig. 3B).

3.4 | Where do we fail?

We next wanted to characterize the regions that were proven to be hard to predict. We start by ranking the MPRA constructs by their observed activity levels (Log2FC). For each competing submission and each MPRA construct, we then define the respective accuracy as the absolute difference between the observed and predicted rank, scaled by the difference expected by a random ranking (which becomes smaller the closer we are to the average). We note that this measure is more robust than taking the difference of the original (nonrank transformed) values, which due to the scaling of variance with the mean (as expected) leads to strong bias for highly active constructs (Supp. Fig. S3).

We then assess how consistent different groups are in their performance within a region. To this end, we record the Spearman correlation coefficient of region performance between every pair of groups (taking the maximal correlation among all pairs of submissions). The cumulative distribution of these coefficients (Fig. 4A) suggests that there is an overall agreement in regions performance for the predictions in part I across different submissions (Fig. 4B) and less coherent agreement for part II (Fig. 4C). Similarly, we observe an overall agreement in regions performance for the predictions in part I, based on the 10 most predictive features individually (Fig. 4B) but not for part II (Fig. 4C).

Given the consistency of submissions for part I, we focus our analysis on pinpointing which genomic and epigenetic features are associated with the ability to predict the activity level (Log2FC) of a region (i.e., region "hardness"). Considering all the variants in the test set of part I, we find that regions accurately predicted by all or most competing submissions are highly enriched with regulatory hits (Fig. 4B). This observation that the activity of truly active regions is generally easier to predict is expected since the activity level of clear nonhits is likely to fall within the regimen of noise. To gain a better understanding of what makes a region hard to predict, we focused our attention on regulatory hits. We used similar performance tests as above to evaluate the extent to which the different feature classes (Fig. 2) and other properties (Fig. 5A) is indicative for the difficulty in predicting the activity of a regulatory hit. First, we find that the measurement noise (evaluated based on reproducibility of the MPRA assay [Supp. Fig. S3] does not discriminate between hard-to-predict and easy-to-predict regions. Second, we find that hard-to-predict regions are associated with a lower transcriptional activity (as expected); however, this association (AUROC = 0.54) is not as strong as that observed for other features (Fig. 5A). Next, considering the contribution of individual TFs, we find several regulators whose binding sites are enriched in either hard or easy regions (Fig. 5B). Evidently, the TFs that are enriched in hard regions tend to have more binding sites across the genome (based on ChIP-seq in LCL; Fig. 5C), which—from the machine learning perspective—naturally makes them less powerful in discriminating active from inactive regions. More globally, we observe that hard-to-predict regions tend to have less TF-binding sites and reduced association with open chromatin and active chromatin marks in the genome, as well as lower GC content. These results reflect our overall conception of what characterizes an active region (Fig. 2A) and are in line with the

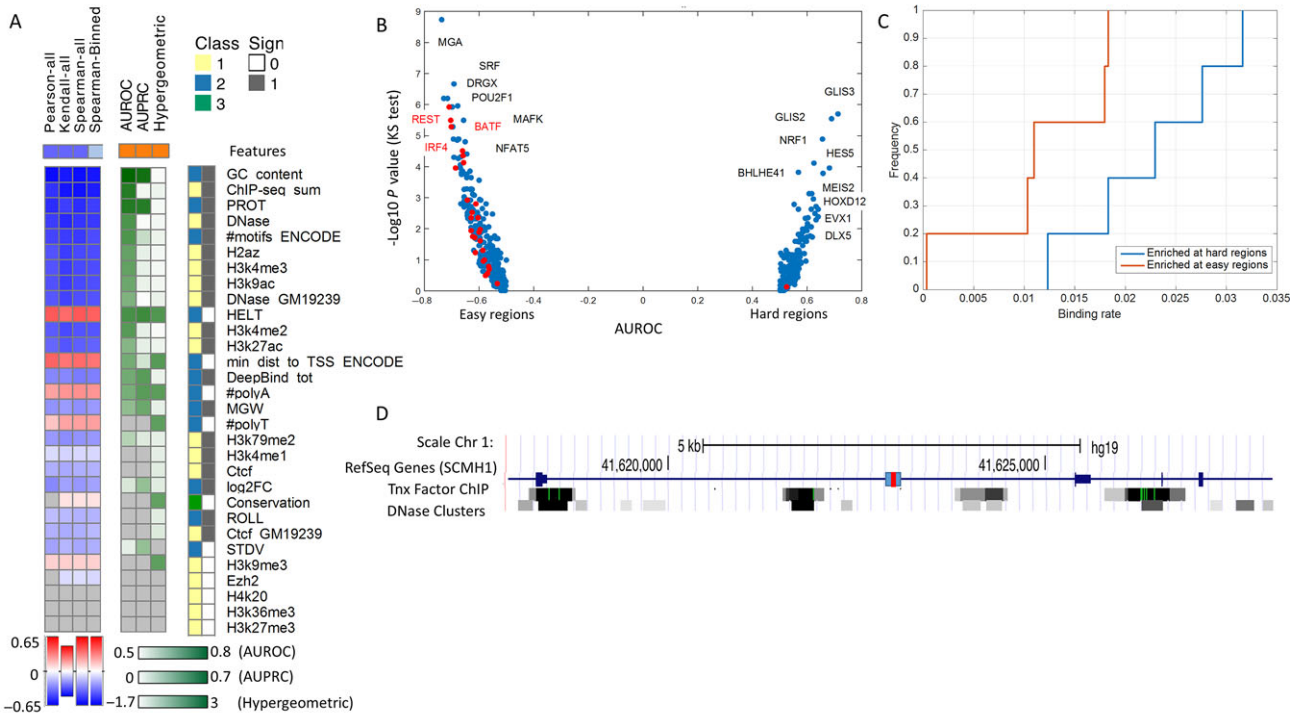

FIGURE 5 Features correlation with region hardness for regulatory hit regions in part I. **A**: Statistical tests used and features examined are similar to Figure 2. The features are sorted from the most predictive (blue/red) to the least predictive (white) to region's hardness, based on the mean rank across all methods across the three tests in part I. The positive class is defined as the top 100 hardest regions. Two additional features are included: noise (log(STD/Mean)) across replicates and log2FC. **B**: Contribution of individual TFs to the prediction of hard and easy regions, measured by the minus log *P* value of the ks-test for AUROC per factor. **C**: Frequency of binding sites across genome for TFs that are enriched in hard and easy regions. **D**: The genomic loci of a specific regulatory hit region with high expression levels that is hard to predict across all submissions. Gene annotations, DHS, and transcription factor ChIP tracks from ENCODE are shown. The 150-bp region is marked with a blue rectangular with a red mark in the middle indicating the eQTL.

previous literature, for instance, that the expression of genes whose promoters has a low GC content is more difficult to predict (Dong et al., 2012). Indeed, looking at individual cases, we find a number of regions that are highly active in the MPRA assay, but are not associated with any TFBS or accessible chromatin in LCL (Fig. 5D). While these apparent discrepancies may be related to the episomal nature of the MPRA assay, close investigation of such regions may be valuable for identifying genetic or epigenetic properties that are predictive of transcriptional activity, in addition to those employed in this challenge.

4 | DISCUSSION

The outcome of the Critical Assessment of Genome Interpretation eQTL challenge serves two main purposes. The first is providing a benchmark and encouraging the development of methods for predicting transcriptional activity of DNA regions, thus improving our understanding of the individual genetic and epigenetic properties that make up the regulatory code, and the appropriate way to model their interdependence in a predictive mathematical model. The second purpose takes a translational point of view—a given eQTL variant is usually associated with multiple loci that cannot be discriminated due to LD. The methods developed here thus join and enhance the published cohort of computational studies (e.g., Alipanahi et al., 2015; Ghandi et al., 2014; Kelley et al., 2016; Kircher et al., 2014; Ritchie, Dunham, Zeggini, & Flicek, 2014; Zeng, Hashimoto, Kang, & Gifford, 2016; Zhou & Troyanskaya, 2015) that prioritize likely causal variants in an LD block, based on the predicted allelic shift in chromatin state.

As opposed to existing body of computational studies, the task of identifying causal variants in high-throughput was tackled by Tewhey et al. (2016) experimentally—using a combined pipeline of eQTL analysis followed by MPRA of the identified loci. Evidently, the resulting MPRA data proved valuable for the identification of key loci in the original study (Tewhey et al., 2016) and for the development of predictive methods in this challenge. However, it is important to bear in mind that MPRA is conducted outside of the natural context of the chromatin and the cell's regulatory network, thus potentially leading to inaccuracies. Indeed, it is still not clear whether the MPRA constructs (Fig. 1) are capable of acquiring physiological chromatin in a manner comparable to endogenous loci (Inoue et al., 2017). While some studies successfully used MPRA to model interactions between TFs (Kwasnieski et al., 2014; Smith et al., 2013), other studies suggest that episomal assays may in some cases fail to reflect cooperative TF activity, due to differences in histone H1 stoichiometry and nucleosome positioning (Archer, Lefebvre, Wolford, & Hager, 1992; Hebbar & Archer, 2007; Hebbar & Archer, 2008; Smith & Hager, 1997). Future novel experimental approaches, including lentivirus-based MPRAs that can allow for integration into the genome (Inoue et al., 2017), will shed more light on the features that determine regions functionality. Future

challenges should focus on finer annotation of TFBS and epigenetic assays, which seem to encompass the majority of information regarding regions regulatory activity. Specifically, considering cellular context (e.g., pioneer factors as shown in Fig. 2, or RNA-seq data) when prioritizing features may improve prediction, as opposed to treating all TFs equally. Lastly, we note this challenge is based on MPRA in regions that harbor an eQTL; the results in Figure 1E indicate that distal regions (i.e., regions that do not intersect with promoters, introns, or exons) show lower transcriptional activity, which might be a result of an error in the eQTL association (which, naturally, tends to be more error prone for distal sites due to statistical burden).

In conclusion, while the task of predicting expression phenotype from genotype is immensely complex, this challenge has seen some promising methodologies. Development of such methods and pinpointing which genetic and epigenetic features contribute to regions functionality is essential to the study of human disease.

The answer key, predictions, and assessment are available on the CAGI Website: https://genomeinterpretation.org/content/4-eQTL-causal_SNPs

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65.

Aguet, F., Brown, A. A., Castel, S., Davis, J. R., Mohammadi, P., Segre, A. V., … Montgomery, S. B. (2017). Local genetic effects on gene expression across 44 human tissues. BioRxiv.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838.

Archer, T. K., Lefebvre, P., Wolford, R. G., & Hager, G. L. (1992). Transcription factor loading on the MMTV promoter: A bimodal mechanism for promoter activation. *Science*, *255*(5051), 1573–1576.

Beer, M. A., & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, *117*(2), 185–198.

Birnbaum, R. Y., Patwardhan, R. P., Kim, M. J., Findlay, G. M., Martin, B., Zhao, J., … Ahituv, N. (2014). Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genetics*, *10*(10), e1004592.

Bussemaker, H. J., Li, H., & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, *27*(2), 167–171.

Ciofani, M., Madar, A., Galan, C., Sellars, M., Mace, K., Pauli, F., … Littman, D. R. (2012). A validated regulatory network for Th17 cell specification. *Cell*, *151*(2), 289–303.

Cowper-Sal lari R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., … Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics*, *44*(11), 1191–1198.

Das, D., Banerjee, N., & Zhang, M. Q. (2004). Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences*, *101*(46), 16234–16239.

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., … Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, *13*(9), R53.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, *28*(8), 817–825.

Ernst, J., & Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods*, *9*(3), 215–216.

Erwin, G. D., Oksenberg, N., Truty, R. M., Kostka, D., Murphy, K. K., Ahituv, N., … Capra, J. A. (2014). Integrating diverse datasets improves developmental enhancer prediction. *PLoS Computational Biology*, *10*(6), e1003677.

Farh, K. K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343.

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., … Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442–1453.

Gertz, J., Siggia, E. D., & Cohen, B. A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, *457*(7226), 215–218.

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Computational Biology*, *10*(7), e1003711.

Gonzalez, A. J., Setty, M., & Leslie, C. S. (2015). Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nature Genetics*, *47*(11), 1249–1259.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018.

Hebbar, P. B., & Archer, T. K. (2007). Chromatin-dependent cooperativity between site-specific transcription factors in vivo. *The Journal of Biological Chemistry*, *282*(11), 8284–8291.

Hebbar, P. B., & Archer, T. K. (2008). Altered histone H1 stoichiometry and an absence of nucleosome positioning on transfected DNA. *The Journal of Biological Chemistry*, *283*(8), 4595–4601.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, *9*(5), 473–476.

Hume, M. A., Barrera, L. A., Gisselbrecht, S. S., & Bulyk, M. L. (2015). UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, *43*(Database issue), D117–D122.

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., … Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research*, *27*(1), 38–52.

Kelley, D. R., Snoek, J., & Rinn, J. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26*(7), 990–999.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., … Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, *23*(5), 800–811.

Kheradpour, P., & Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, *42*(5), 2976–2987.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315.

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., & Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, *24*(10), 1595–1602.

Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511.

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, *47*(8), 955–961.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., … Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804.

Marstrand, T. T., & Storey, J. D. (2014). Identifying and mapping cell-type-specific chromatin programming of gene expression. *Proceedings of the National Academy of Sciences*, *111*(6), E645–E654.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., … Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, *30*(3), 271–277.

Mogno, I., Kwasnieski, J. C., & Cohen, B. A. (2013). Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Research*, *23*(11), 1908–1915.

Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E., & Ohler, U. (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research*, *22*(9), 1711–1722.

Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, *37*(Database issue), D77–D82.

Nguyen, D. H., & D'Haeseleer, P. (2006). Deciphering principles of transcription regulation in eukaryotic genomes. *Molecular Systems Biology*, *2*, 2006.0012.

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., … Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, *30*(3), 265–270.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.

Ritchie, G. R., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature Methods*, *11*(3), 294–296.

Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L., & Almouzni, G. (2015). Epigenomics: Roadmap for regulation. *Nature*, *518*(7539), 314–316.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., & Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, *451*(7178), 535–540.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., … Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, *30*(6), 521–530.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050.

Smith, C. L., & Hager, G. L. (1997). Transcriptional regulation of mammalian genes in vivo. A tale of two templates. *The Journal of Biological Chemistry*, *272*(44), 27493–27496.

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., … Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, *45*(9), 1021–1028.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), 613–626.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., … Sabeti, P. C. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, *165*(6), 1519–1529.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82.

Weingarten-Gabbay, S., & Segal, E. (2014). The grammar of transcriptional regulation. *Human Genetics*, *133*(6), 701–711.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–D1006.

Wilczynski, B., Liu, Y. H., Yeo, Z. X., & Furlong, E. E. (2012). Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Computational Biology*, *8*(12), e1002798.

Zeng, H., Edwards, M. D.., Guo, Y., & Gifford, D. K. (2016). Accurate eQTL prioritization with an ensemble-based framework. BioRxiv.

Zeng, H., Hashimoto, T., Kang, D. D., & Gifford, D. K. (2016). GERV: A statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, *32*(4), 490–496.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, *12*(10), 931–934.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., … Rohs, R. (2013). DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research*, *41*(Web Server issue), W56–W62.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Kreimer A, Zeng H, Edwards MD, et al. Predicting gene expression in massively parallel reporter assays: A comparative study. *Human Mutation*. 2017;38:1240–1250. http://doi.org/10.1002/humu.23197