CERENKOV: Computational Elucidation of the Regulatory Noncoding Variome

Yao Yao*
Oregon State University
EECS
yaoyao@oregonstate.edu

Zheng Liu*
Oregon State University
EECS
liuzhen@oregonstate.edu

Satpreet Singh
Oregon State University
EECS
singhsa@oregonstate.edu

Qi Wei
Oregon State University
EECS
weiq@oregonstate.edu

Stephen A. Ramsey
Oregon State University
Department of Biomedical Sciences
Corvallis, Oregon 97330
stephen.ramsey@oregonstate.edu

ABSTRACT

We describe a novel computational approach, CERENKOV (Computational Elucidation of the REgulatory NonKOding Variome), for discriminating regulatory single nucleotide polymorphisms (rSNPs) from non-regulatory SNPs within noncoding genetic loci. CERENKOV is specifically designed for recognizing rSNPs in the context of a post-analysis of a genome-wide association study (GWAS); it includes a novel accuracy scoring metric (which we call average rank, or AVGRANK) and a novel cross-validation strategy (locus-based sampling) that both correctly account for the "sparse positive bag" nature of the GWAS post-analysis rSNP recognition problem. We trained and validated CERENKOV using a reference set of 15,331 SNPs (the `OSU17` SNP set) whose composition is based on selection criteria (linkage disequilibrium and minor allele frequency) that we designed to ensure relevance to GWAS post-analysis. CERENKOV is based on a machine-learning algorithm (gradient boosted decision trees) incorporating 246 SNP annotation features that we extracted from genomic, epigenomic, phylogenetic, and chromatin datasets. CERENKOV includes novel features based on replication timing and DNA shape. We found that tuning a classifier for AUPVR performance does not guarantee optimality for AVGRANK. We compared the validation performance of CERENKOV to nine other methods for rSNP recognition (including GWAVA, RSVP, DeltaSVM, DeepSEA, Eigen, and DANQ), and found that CERENKOV's validation performance is the strongest out of all of the classifiers that we tested, by both traditional global rank-based measures ($\langle$AUPVR$\rangle$ = 0.506; $\langle$AUROC$\rangle$ = 0.855) and AVGRANK ($\langle$AVGRANK$\rangle$ = 3.877). The source code for CERENKOV is available on GitHub and the SNP feature data files are available for download via the CERENKOV website.

ACM-BCB '17, August 20-23, 2017, Boston, MA, USA

DOI: http://dx.doi.org/10.1145/3107411.3107414

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Computational biology**; **Bioinformatics**;

KEYWORDS

SNP, GWAS, noncoding, rSNP, SNV, machine learning

1 INTRODUCTION

1.1 The rSNP detection problem

Human genome-wide association studies (GWAS) have led to the discovery of 21,751 unique SNP-trait associations in more than 2,400 studies (based on millions of individuals) that are linked to human traits [55]. However, GWAS functional interpretation has been largely limited to *coding regions* in which single nucleotide polymorphisms (SNPs) can be mapped to functional consequence predictions using well-established methods [45]. This limits the yield of knowledge from GWAS because 90% of human GWAS-identified SNPs are located in noncoding regions [33]. Within noncoding regions, a key roadblock to identifying the molecular mechanisms underlying trait variation is the difficulty of pinpointing the probable causal noncoding SNPs [48]—the so-called *regulatory SNPs* (rSNPs)—that are associated with the trait. Various correlates of rSNPs are known [31], with expression quantitative trait locus (expression QTL, or eQTL) associations being one validated example [36]. But in general, how to computationally integrate various types of genomic, phylogenetic, epigenomic, transcription factor binding site (TFBS), and chromatin-structural rSNP correlates in order to discriminate regulatory SNPs from non-regulatory noncoding SNPs is a fundamental problem in computational biology.

* These authors contributed equally to this work.

Machine learning (ML) classification approaches based on sets of experimentally verified regulatory single nucleotide variants (rSNVs) from databases such as the Human Gene Mutation Database (HGMD), ORegAnno [35] or ClinVar (as well as neighboring "control SNVs" (cSNVs) that are presumed non-functional) have shown promise for advancing the field of computational rSNP recognition.

1.2 Previous approaches

Previous computational approaches to the problem of discriminating functional noncoding SNPs (i.e., regulatory SNPs or rSNPs) from nonfunctional SNPs can be divided into two categories, "rSNV-supervised" approaches that in which ground-truth sets of regulatory single nucleotide variants (rSNVs, such as from HGMD) were used in model training, and "rSNV-unsupervised" approaches that were not trained using ground-truth sets of rSNVs (and were instead trained using other sets of variants, typically much larger than the training sets used in the rSNV-supervised methods).

1.2.1 rSNV-unsupervised approaches. Among the rSNV-unsupervised approaches, various classification models have been proposed including probabilistic graphical models (such as in the fitCons variant scoring method [16]), support vector machine (SVM; such as in the CADD variant scoring method [24] and in the DeltaSVM variant scoring method [29]), deep neural networks (such as in the DANN variant scoring method [41]), and spectral methods (such as in the Eigen-PC method [21]). Somewhat counter-intuitively, some previous studies [21, 29] of rSNV-unsupervised methods have reported greater accuracy than rSNV-supervised methods [43] that were used for comparison, with Eigen-PC appearing to have the strongest such comparative performance results. These previous results suggest that state-of-the-art rSNV-unsupervised scoring methods, such as Eigen-PC, may be useful as *features* within an rSNV-supervised method.

1.2.2 rSNV-supervised approaches. A variety of classification algorithm types and feature-set types have been used for rSNV-supervised methods to identify regulatory variants. Montgomery *et al.* [34] integrated 23 variant annotation features (including transcription factor binding sites, TFBS) within a support vector machine (SVM) model, using a small reference set (104 rSNVs). Torkamani & Schork [51] integrated 28 TFBS, epigenomic, and chromatin features within a naïve Bayes classifier, also using a small reference set (102 rSNVs). Zhao et al. [58] derived 158 variant annotation features from population haplotype, gene annotation, and phylogenetic datasets and employed the Random Forest [4] algorithm, using a larger reference set (445 rSNVs). Ritchie et al. [43] engineered 175 features from several hundred epigenomic, population genetic, phylogenetic, and TFBS datasets; they used Random Forest and a moderate-sized reference set (1,614 rSNVs) to train their classifier, which they called GWAVA. In a method called DeepSEA, Zhou & Troyanskaya [59] employed deep convolutional neural networks (CNNs) in order to reduce the dimension of 1,000 bp of flanking sequence for a SNV to a single allele-dependent score. They trained 919 such CNNs using published epigenomic and chromatin measurements and integrated the allele-dependent scores within a gradient boosted decision tree algorithm, using a moderate-sized reference set (2,997 rSNVs). In a method called DANQ, Quang & Xie [42] used the same 919 chromatin measurements as Zhou & Troyanskaya used for training DeepSEA, but they used a hybrid network architecture combining CNN and deep recurrent neural network (RNN) components. In the RSVP method, Peterson *et al.* [39] derived a set of 2,237 SNP annotation features that was substantially expanded (versus previous approaches) by including a SNP's nearest-gene functional annotations and its nearest-gene multi-tissue expression profile; their feature-set also included SNP annotations based on local genomic replication timing measurements (Repli-chip [19] and Repli-seq [17] assay data from cell lines). Using a moderate-sized ground truth set (1,999 rSNVs), Peterson *et al.* trained an ensemble decision tree classifier. Thus, in terms of supervised methods for rSNP classification, there has been a steady progression in the comprehensiveness of the feature-sets and in the variety of algorithmic approaches that have been used.

1.2.3 DNA shape as a potential feature. Previous computational and experimental studies have demonstrated that local pentameric DNA sequence is predictive of three-dimensional DNA shape distortions (such as roll, propeller twist, and helical twist) [44, 60] and that specific features of the local DNA shape can improve discrimination of regulatory from non-regulatory DNA [57] in general and for transcription factor-specific prediction of TFBS [32]. Montgomery *et al.* have reported [34] that a specific DNA physical parameter (bendability) is beneficial for discriminating regulatory from non-regulatory variants. We have previously developed a computational model, `regshape,` that incorporates four sequence-predicted DNA shape parameters to predict transcription factor binding sites [57], for which allele-dependent scores may be useful within an rSNP detection model.

1.3 Limitations of previous approaches

An important limitation of all of the previous rSNV-supervised approaches to the rSNP recognition problem (of which we are aware) is that none of the models were trained and validated (i) using locus-by-locus assignment of variants to cross-validation (CV) folds and (ii) using a reference set of SNPs that was *selected for relevance to the application of GWAS post-analysis* and assigned to CV folds on a locus-by-locus basis. In the GWAS post-analysis application context, trait-linked SNPs are analyzed within *linkage disequilibrium blocks,* and thus, the SNPs are partitioned into loci and the goal is to identify the functional noncoding variant *within each locus.*[1] Thus, it is the *rank of the classifier's rSNP prediction score within a locus matters more than the global (locus-agnostic) ranking of rSNP prediction scores.* In every previous rSNV-supervised approach to this problem (of

[1] Thus, the rSNP recognition problem for GWAS post-analysis could be described as having a "sparse positive bag" structure [5].

which we are aware), SNVs are assigned to CV folds without regard to the SNVs' genomic positions; such "variant-level sampling", while standard practice, makes it impossible to use CV to measure the accuracy of the method on a per-locus basis and it makes the results not representative of how the method would perform in identifying a candidate causal rSNP within a "new" region identified by a GWAS.

A second limitation of previous approaches concerns selection of reference variants. Intrinsic to the GWAS approach is that the local linkage disequilibrium (LD) block needs to have a sufficiently high minor allele frequency (MAF; greater than 0.05) in order for GWAS to be sufficiently powered to detect an association of the block with the trait at genome-wide significance ($p \leq 5 \times 10^{-8}$). In the previous rSNP recognition methods development efforts of which we are aware, a significant fraction of the variants that were used for training are unrealistic in the context of a GWAS post-analysis application domain, either due to the variant having a low (i.e., less than 0.05) MAF,[2] or due to the variant not being in linkage disequilibrium with any known functional non-coding SNP. Further, no previous approach to this problem has combined a comprehensive SNP annotation feature-set (including replication timing data, chromatin segmentation information, expression quantitative trait locus annotations, and transcription factor binding site (TFBS) information for all available TFBS models) within a high-performance gradient boosted decision trees classification algorithm.

1.4 Our approach

We have developed a method, CERENKOV (Computational Elucidation of the REgulatory NonKOding Variome) for identifying rSNPs within GWAS regions. The method is in the "rSNV-supervised" category of approaches (but we note that it includes SNP scores from a state-of-the-art rSNV-unsupervised method, Eigen-PC). CERENKOV is based on the integration of 246 SNP annotation features (including a novel feature based on allele-dependent difference in a DNA shape score that we have previously shown is predictive of TFBS [57]) within a gradient boosted decision tree classification algorithm. Our approach for training and testing CERENKOV was focused on maximizing the relevance and generalizability of our model and its performance assessments for the specific application of GWAS post-analysis.

1.4.1 New performance measure: AVGRANK. In addition to measuring the performance of our classifiers using traditional locus-independent metrics (area under the receiver operating characteristic (AUROC) curve and area under the precision versus recall (AUPVR) curve), we propose and demonstrate a novel performance metric based on averaging, over all rSNPs, the ranks of the rSNP prediction scores of all ground-truth rSNPs within their loci (defined by LD as described above). Specifically, we define, for the sets $L \subset Z_+$ of numbered loci, $S \subset Z_+$ of numbered SNPs, and $R \subset S$ of numbered rSNPs,

$$\text{AVGRANK} = \langle \text{rank}(y_r, \vec{y}_{S_{l(r)}}) \rangle_{r \in R \text{ sth. } |S_{l(r)}|>1}, \quad (1)$$

where $l(r)$ is the locus $l \in L$ for rSNP $r \in R$, $S_l \subset S$ is the set of all SNPs in locus l, y_s is the classifier prediction score for SNP s to be in the rSNP class, $\vec{y}_Q$ is the vector of prediction scores of SNPs $Q \subset S$, $\text{rank}(q, \vec{w})$ is the rank of the score $q \in [0,1]$ in the (length W) vector of scores $\vec{w} \in [0,1]^W$ (with lowest score having rank 1, and with tied scores given identical rank assignments), and $\langle\rangle$ denotes arithmetic mean. We propose that the classifier with the best performance for GWAS post-analysis should minimize the AVGRANK.

1.4.2 The `OSU17` reference SNP set. In order to maximize the relevance of our method to GWAS post-analysis, it is critical to train and test using *common variants*, i.e., SNPs within their local linkage disequilibrium blocks in the genome. Thus, we trained and validated CERENKOV using only SNPs with MAF ≥ 0.05 (as opposed to MAF ≥ 0.01 for the two key previous studies [43, 59]). Further, for the set of control SNPs (cSNPs) that we used as negative examples for training and validation, we used only SNPs that were in strong linkage disequilibrium ($r^2 \geq 0.8$) with, and located no more than 50 kbp distance from, a "positive example" rSNP.

1.4.3 Locus-based sampling. In order to maximize the relevance of our method to GWAS post-analysis as well as enable the assessment of CERENKOV's performance using both global rank-based measures (e.g., AUPVR and AUROC) and the new AVGRANK measure, we trained and tested CERENKOV using a novel (so far as we are aware) method for assigning SNPs to CV folds, which we call *locus-based sampling.* In brief, for each replication of a k-fold CV assessment of CERENKOV's performance, we assign *loci* (i.e., all SNPs together within a given locus) to CV folds, in such a manner that the SNPs overall and the rSNPs are both equipartitioned across the folds. We used the same fold assignments for CERENKOV that we used for the models to which we compared CERENKOV.

2 METHODS

2.1 The `OSU17` reference SNP set

We obtained minor allele frequencies (MAFs) for all SNPs from the dbSNP-based [46] `snp146` SQL table hosted at the UCSC Genome Browser [23] site; for SNPs not in `snp146`, we obtained MAFs from the 1,000 Genomes (1KG) Project Phase 3 [1] Variant Call Format (VCF) file. For the representative set of rSNPs for training/evaluation, we obtained 1,659 SNPs from HGMD (Rel. 2016) that satisfied all of the following criteria: (i) the SNP was marked as `regulatory` in HGMD; (ii) the `disease` field did not not contain `cancer`; (iii) MAF ≥ 0.05; (iv) the SNP was not an indel and not contained within a CDS (based on the complete set of transcripts from the Ensembl 75 gene annotation build); and (v) the SNP was not exclusively mapped to the Y chromosome (due to the lack of phased haplotype data available for proxy SNP searching). For each of these rSNPs, we used the SNP

[2] Databases of rSNVs such as HGMD have an ascertainment bias for low-MAF variants [27], and this can potentially bias the classifier in a way that does not generalize to the GWAS post-analysis application.

Annotation and Proxy Search (SNAP) tool [22] to identify SNPs that are in LD ($r^2 \geq 0.8$ in 1KG Phase 1, with HapMap used instead of 1KG for chromosome X), and we filtered to include only SNPs within 1 kbp of an rSNP, that were not contained within a CDS, that have MAF ≥ 0.05, and that are not themselves on the list of rSNPs. Overall, this filtering procedure produced a list of 13,672 cSNPs. The combined set of 15,331 SNPs (which we call the `OSU17` reference SNP set) was thus designed as an appropriate ground-truth set for the application of GWAS post-analysis. Overall, the class imbalance of `OSU17` is ∼8.24 (cSNP/rSNP).

2.2 Extracting the CERENKOV features

The CERENKOV feature extraction software is based on Python and SQL. We extracted 246 SNP features for each of the `OSU17` SNPs, using data from SNP annotation databases, epigenomic and chromatin datasets, phylogenetic conservation scores, and DNA shape-based scores (Table 1).

2.2.1 Features extracted from UCSC. We used the `snp146` UCSC SQL table as the initial source for SNP annotations (GRCh37 assembly coordinate system). We extracted additional SNP annotation information by (i) coordinate-based joins to other genome annotation tracks in the UCSC database; and (ii) by joining with non-UCSC data sources using the SNP coordinate. For triallelic and quadrallelic SNPs, we used the two most frequent alleles, for the purpose of obtaining features that depend on allele-dependent scores. We derived DNase I hypersensitive site (DHS) features from data tracks from published genome-wide assays with high-throughput sequencing-based detection (DNase-seq) from the ENCODE project [50] (the `master` peaks are summary peaks combining data from DHS experiments in 125 cell types; the `uniform` DHS peaks are from DHS experiments in individual cells, processed using the ENCODE uniform peaks analysis pipeline [28]). The `ENCODE_TFBS` feature is presented in the table as a single feature for conciseness, but in fact it is 158 separate binary features, one for each TF for which genome-wide TF binding site data (from chromatin immunoprecipitation with high-throughput sequencing readout, or ChIP-seq) peak data from the ENCODE Uniform Peaks analysis are available [28]. For replication timing features, we processed track-specific BigWig files from UCSC to obtain the timing scores at individual SNP positions. For ChromHMM, Segway and lamina-associated domains (LAD) annotations, we used the SQL tables from UCSC.

2.2.2 Features extracted from Ensembl. We used the BioMart tool to download (i) TFBS motif occurrences (based on the 2014 release of the Jaspar database [40]) and ChromHMM chromatin segmentation labels from Ensembl Regulation 75 and (ii) GENCODE transcription start sites (from Ensembl Genes 75) with which we computed signed TSS distances.

2.2.3 GTEx feature. We obtained SNP-to-gene associations for 13 tissues (adipose, artery/aorta, artery/tibial, esophagus/mucosa, esophagus/muscularis, heart left ventricle, lung, skeletal muscle, tibial nerve, sun-exposed skin, stomach, thyroid, and whole blood) from GTEx Analysis Version 4 from the GTEx project data portal. For each SNP, we selected the minimum association p-value across genes and tissues.

2.2.4 DNA shape feature. We used the `regshape` R package [57] and computed the difference in the regulatory potential scores for the local 11 bp sequence centered on the SNP, for both the major and next-to-major SNP alleles.

2.3 Features for the other classifiers that we compared to CERENKOV

2.3.1 GWAVA. For the 15,331 `OSU17` SNPs, we extracted the 175 GWAVA [43] SNP annotation features using the GWAVA software (version 1.0). For the GWAVA features that overlapped with CERENKOV features, we compared feature vectors directly to verify consistency.

2.3.2 DeltaSVM. For DeltaSVM, we obtained 19 bp sequences centered on each of the `OSU17` SNPs, using the GRCh37 genome assembly and inserting the two possible SNP alleles into the central base position. We used the DeltaSVM software (3/30/2015 version), using the 19mer sequence files as input and using 226 pre-built 19mer-based chromatin models from the DeltaSVM website as input, to produce a 15,331×226 feature matrix.

2.3.3 RSVP. For RSVP, we used the RSVP software [39] version 1.0.1. We mapped the `OSU17` SNPs to their nearest gene using `Annovar` [53] release 2016Feb01. In addition, the flanking five nucleotides on either side of the SNPs were also included in the input to the perl script, `script_RSVP.pl`, which output a 15,331×2,238 feature matrix for RSVP classifier. For analyzing RSVP features with Random Forest, we imputed missing values using the column-wise mean.

2.3.4 DeepSEA. We extracted 1,000 bp haplotype sequences flanking the `OSU17` SNPs (for the two most common alleles for each of the SNPs) using the 1KG Phase 3 VCF data (see Sec. 2.1). For each of the 15,331 pairs of 1 kbp sequences, we obtained scores from the 919 DeepSEA CNN models using the web tool. As in the original publication [59], for each SNP and each of the 919 CNNs, we obtained two features consisting of (i) the absolute difference of scores between the two alleles and (ii) the absolute difference of the logit-transformed scores for each of the two alleles. This overall procedure produced a 15,331×1,838 feature matrix.

2.3.5 DANQ. For DanQ [42], we extracted 919 hybrid CNN-RNN scores for each of the two alleles for each of the 15,331 `OSU17` SNPs, using the DANQ software and bundled model files in HDF5 format (Jan. 13, 2016). For each SNP and each of the 919 CNNs, we obtained two features consisting of the absolute difference of scores between the two alleles and the absolute difference of the logit-transformed scores between the two alleles. This produced a 15,331×1,838 feature matrix which we cross-checked against DeepSEA for consistency.

feature(s)	*feature type*	*raw data src.*	*feature description*
normChromCoord	continuous	UCSC	the SNP coordinate (normalized to chrom. length)
majorAlleleFreq	continuous	UCSC/1KG	the major allele frequency (1KG)
minorAlleleFreq	continuous	UCSC/1KG	the next-to-major allele frequency (1KG)
phastCons	continuous	UCSC	46-way placental mammal phastCons score [47]
GERP++	continuous	UCSC	bp-level GERP++ [10] score
avg_GERP	continuous	UCSC	avg. GERP score [8] in ±100 bp window
DNAShapeScore	continuous	UCSC	diff. of 11 bp `regshape` [57] score between alleles
avg_daf	continuous	1KG	average derived allele frequency in ±1 kbp region
avg_het	continuous	1KG	average heterozygosity rate in ±1 kbp region
maf1kb	continuous	UCSC/1KG	average of the MAF values for all SNPs in ±1 kbp window
eqtlPvalue	continuous	GTEx	$-\log_{10}\min(p)$ for GTEx eQTL for the SNP, across 13 tissues [14]
GC5Content	integer (0-5)	UCSC	GC content in a 5 bp window
GC7Content	integer (0-7)	UCSC	GC content in a 7 bp window
GC11Content	integer (0-11)	UCSC	GC content in a 11 bp window
local_purine	integer (0-11)	UCSC	number of purine bases in local 11 bp window
local_CpG	integer (0-10)	UCSC	number of CpG dinucleotides in 11 bp window
ss_dist	integer	UCSC	signed distance to nearest exon boundary
tssDistance	integer	Ensembl75	signed distance to nearest Ensembl TSS
gencode_tss	integer	GENCODE	signed distance to nearest GENCODE TSS
tfCount	integer	UCSC	sqrt(count) of ENCODE ChIP-seq TFBS overlap. SNP
uniformDhsScore	integer	UCSC	sum scores of ENCODE uniform DHS peaks overlap. SNP
uniformDhsCount	integer	UCSC	count of ENCODE uniform DHS peaks overlap. SNP
masterDhsScore	integer	UCSC	sum scores of ENCODE master DHS peaks overlap. SNP
masterDhsCount	integer	UCSC	count of ENCODE master DHS peaks overlap. SNP
chrom	categorical (23)	UCSC	the chromosome to which the SNP maps
nestedrepeat	categorical (2)	UCSC	SNP is in a RepeatMasker [6] DNA repeat
simplerepeat	categorical (2)	UCSC	SNP is in a Tandem Repeats Finder [2] repeat
cpg_island	categorical (2)	UCSC	SNP is in an epigenome-predicted CpG island [3]
geneannot	categorical (4)	UCSC	classifies SNP location as CDS, intergenic, UTR, or intron
majorAllele	categorical (4)	UCSC/1KG	the major allele for the SNP
minorAllele	categorical (4)	UCSC/1KG	the next-to-major allele for the SNP
pwm	categorical (22)	Ensembl75	ID of the Jaspar 2014 [40] motif in which SNP is a match
chromhmm	6×categ. (26)	UCSC	ChromHMM label in Gm12878, H1hesc, HeLaS3, HepG2, HUVEC and K562 cells
segway	6×categ. (26)	UCSC	Segway label in Gm12878, H1hesc, HeLaS3, HepG2, HUVEC and K562 cells
ch_comb_WEAKENH	categorical (4)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ch_comb_ENH	categorical (6)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ch_comb_REP	categorical (7)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ch_comb_TSSFLANK	categorical (5)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ch_comb_TRAN	categorical (7)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ch_comb_TSS	categorical (7)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ch_comb_CTCFREG	categorical (7)	Ensembl75	ChromHMM label in Ensembl Reg. Seg. build
ENCODE_TFBS	158×categ. (2)	UCSC	158 features for SNP being in an ENCODE TFBS [13] peak
FsuRepliSeq	16×continuous	UCSC	Replication Timing by Repli-chip [19] from ENCODE/FSU
UwRepliSeq	16×continuous	UCSC	Replication Timing by Repli-seq [17] from ENCODE/UW
SangerTfbsSummary50kb	continuous	Ensembl75	Summary of Ensembl TFBS peaks from 18 human cell types
NkiLad	categorical (2)	UCSC	SNP is in a Lamina Associated Domain (NKI study [15], Tig-3 cells)
vistaEnhancerCnt	categorical (2)	UCSC	count of VISTA [52] HMR-Conserved Non-coding Human Enhancers [38] overlap. SNP
vistaEnhancerTotalScore	categorical (2)	UCSC	sum scores of VISTA [52] HMR-Conserved Non-coding Human Enhancers [38]
eigen	continuous	Eigen	Eigen-PC v1.1 raw score [21]

Table 1: The 246 SNP features that are used in CERENKOV. Abbreviations are as follows: UCSC, UC Santa Cruz Genome Browser portal; 1KG, 1,000 Genomes Project; Ensembl75, Ensembl Release 75 [9]; GENCODE, the GENCODE project release 19 [18]; ENCODE, Encyclopedia of DNA Elements [49]; FSU, Florida State University; UW, University of Washington; NKI, Netherlands Cancer Institute; GTEx, the genotype tissue-expression project; GERP, the Genomic Evolutionary Rate Profiling score; CDS, coding DNA sequence; UTR, untranslated region; MAF, minor allele frequency; HMR, human-mouse-rat; TSS, transcription start site.

2.3.6 Eigen, CADD, DANN, fitCons. For Eigen [21], we downloaded genome-wide nucleotide-level scores (version 1.1) and filtered them to obtain the raw Eigen-PC scores for the `OSU17` SNPs. For CADD [24], we downloaded genome-wide nucleotide-level scores (version 1.3). For DANN [41], we downloaded genome-wide nucleotide-level scores (released Nov. 13, 2014) and filtered them to obtain scores for the `OSU17` SNPs. For fitCons [16], we downloaded genome-wide feature files (version 1.01) for highly significant scores (`fc-hu-0.bw`) and extracted the scores at the locations of the `OSU17` SNPs. For all four of these score types, we used the published per-SNP scores directly to rank validation-set SNPs for computing AVGRANK, AUROC, and AUPVR.

2.4 Machine learning

For the machine learning framework, we used the R statistical computing environment (version 3.2) [20] under Ubuntu 16.04.1 LTS running on Amazon EC2 within an m4.16xlarge instance (or across multiple such instances in parallel).

2.4.1 Random Forest. We used the R package `ranger` [56] version 0.6.0. We used `ranger` in decision-tree mode (i.e.,

not in probabilistic forest mode). For GWAVA [43] and RSVP [39], we used the published hyperparameters. To validate the equivalency of this classification algorithm vs. the original GWAVA python implementation, we cross-checked the AUROC performance of `ranger` (with the GWAVA features and the published GWAVA `region` SNP set) against the `RandomForest` implementation in `scikit-learn` [37] (version 0.14.1) and found nearly identical performance between `ranger` and `scikit-learn`, $\langle \text{AUROC} \rangle = 0.71$.

2.4.2 Gradient Boosted Decision Trees. For the gradient boosted decision trees (GBDT) classifier, we used the R API for `xgboost` [7] version 0.6-4 (hereafter, `xgboost`-GBDT). We used gradient boosted trees (`booster=gbtree`) and binary "logistic" classification as the objective, with the default loss function (`objective=binary:logistic`). We used ten-fold CV [25] with *locus-level sampling* as described in Sec. 3, in which we assigned rSNPs to folds (stratifying on the number of cSNPs per rSNP), and then assigned cSNPs to the *same fold to which it's LD-linked rSNP was assigned.* Thus, in the case of locus-level sampling, an rSNP and its linked cSNPs are always assigned to the same CV fold. For every prediction performance metric we report, the fold composition was exactly the same across all of the rSNP recognition models studied. We used the `xgboost`-GBDT classification algorithm for the studies in Fig. 1b, Fig. 1c, Fig. 2, and for the classifiers CERENKOV, RSVP_XGB, Deepsea_XGB, DANQ_XGB, and deltaSVM_XGB in Fig. 3. We used `base_score` = 0.1082121 (the rSNP/cSNP class imbalance). We obtained the feature importance scores using the `xgb.importance` method.

2.4.3 Tuning CERENKOV. We tuned the `xgboost`-GBDT classifier with a hyperparameter septuple grid size of 3,888, with locus-based sampling. The tuning hyperparameter tuple that maximized the validation $\langle \text{AUPVR} \rangle$ was: $\eta = 0.1$, $\gamma = 10$, `nrounds` = 30, `max_depth` = 6, `subsample` = 1.0, `colsample_bytree` = 0.85, and `scale_pos_weight` = 1; we used these hyperparameter values for all subsequent analyses using `xgboost`-GBDT. [In contrast, the hyperparameter tuple that minimized the validation $\langle \text{AVGRANK} \rangle$ was: $\eta = 0.1$, $\gamma = 100$, `nrounds` = 30, `max_depth` = 6, `subsample` = 0.85, `colsample_bytree` = 0.85, and `scale_pos_weight` = 8].

2.5 *t*-SNE and Statistical testing

For the unsupervised analysis, we used the *t*-distributed stochastic neighbor embedding (*t*-SNE) implementation in the R package `Rtsne` on the standardized, numeric-encoded feature data; results were consistent across multiple runs of *t*-SNE. For averaging and statistical testing of performance data, we used logit-transformed AUROC and AUPVR values [12]. We compared different combinations of classifiers and feature-sets using 95% confidence intervals (CI) on the average measures, that we estimated using bootstrap with 1,000 iterations.

3 RESULTS

3.1 Comparing AUPVR and AVGRANK

We first analyzed the distributions of two SNP classes (rSNPs and cSNPs) within an unsupervised two-dimensional embedding of their 246-dimensional feature vectors, using *t*-SNE (see Sec. 2.5). This analysis revealed that the majority of rSNPs are located in dense clusters in the embedding space, suggesting that effective machine learning-based discrimination of rSNPs from cSNPs would be possible using this feature-set (**Fig. 1a**). Next, we used a supervised approach to investigate the degree of concordance between the traditional AUPVR accuracy measure and the AVGRANK accuracy measure that we have proposed for the GWAS rSNP detection problem. For this analysis we used the gradient boosted decision trees classification algorithm implementation in `xgboost` [7] and our 15,331×246 matrix of SNP feature data. We trained and tested `xgboost`-GBDT (using 24 independent replications of 10-fold [25] CV with locus-based sampling) for each of 3,888 different tuples of `xgboost` hyperparameters (see Sec. 2.4.3) and for each hyperparameter tuple, we computed $\langle \text{AUPVR} \rangle$ and $\langle \text{AVGRANK} \rangle$ on the validation set SNPs, across the 240 independent samples. We found that while $\langle \text{AVGRANK} \rangle$ and $\langle \text{AUPVR} \rangle$ are strongly correlated on the large scale (Fig. 1b), the relationship between $\langle \text{AVGRANK} \rangle$ and $\langle \text{AUPVR} \rangle$ for the stronger-performing hyperparameter tuples (for which $\langle \text{AUPVR} \rangle > 0.42$) deviates from a linear relationship (Fig. 1c). Specifically, we found that the hyperparameter tuple that maximizes the $\langle \text{AUPVR} \rangle$ has an $\langle \text{AUPVR} \rangle$ of 0.506 (95% CI, 0.500–0.512), whereas the hyperparameter tuple that minimizes the $\langle \text{AVGRANK} \rangle$ has an $\langle \text{AUPVR} \rangle$ of 0.491 (95% CI, 0.485–0.496). Thus, in this example using `xgboost`-GBDT and the CERENKOV features on the `OSU17` SNP set, we found that tuning a classifier to maximize validation AUPVR does not guarantee optimality of the classifier in terms of AVGRANK accuracy (we argue that the latter, AVGRANK, is more relevant to the GWAS application domain).

3.2 CERENKOV feature importance

In order to better understand the contributions of different categories of features to classification accuracy for rSNP recognition, we analyzed the frequencies with which each of the 246 features in CERENKOV was used in a tree split by the `xgboost`-GBDT algorithm, using the hyperparameters that we had selected to maximize validation-set AUPVR (Sec. 3.1). We additively aggregated the feature scores into 14 feature categories such as "SNP-to-gene distances," "chromatin segmentation," and "TFBS annotations." As shown in Fig. 2, and consistent with previous SNP annotation-based studies (all of which used different filtering criteria for cSNP selection than we used in this study), the distance between a SNP and the transcription start site of the nearest genes has the strongest feature importance overall [34, 39, 43, 51], and that SNP annotation based on the gene context (3′ UTR, 5′ UTR, intron, or intergenic) also has high importance [43]. Furthermore, consistent with the Peterson *et al.* study [39],

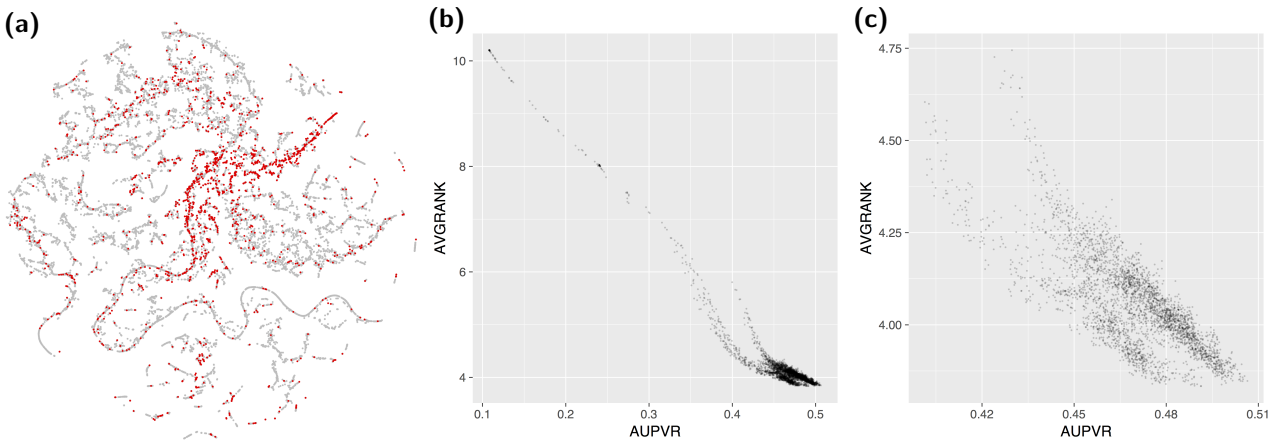

Figure 1: The AUPVR and AVGRANK performance measures are functionally distinct. (a) t-SNE embedding of the 246-dimensional featureset for the `OSU17` set of SNPs; rSNPs are colored in red, showing local concentrations of rSNPs. (b) Scatter plot of validation ⟨AVGRANK⟩ and ⟨AUPVR⟩ values for CERENKOV, for 3,888 `xgboost`-GBDT hyperparameter tuples. (c) Zoom of the lower-right of (b), showing that optimizing for average AVGRANK and optimizing for average AUPVR are not equivalent, for the `OSU17` SNPs.

Measure	Mean	95% CI
AUPVR	0.505	0.503–0.508
AUROC	0.855	0.854–0.856
AVGRANK	3.877	3.843–3.909

Table 2: Validation-set performance measures for CERENKOV on the `OSU17` reference SNP set.

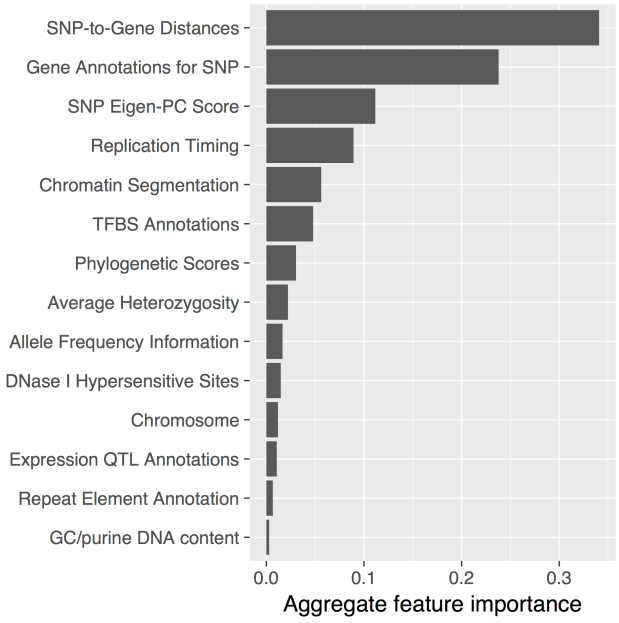

Figure 2: Importance scores for 14 categories of features used in CERENKOV. Bar length represents the aggregate frequency with which all features in the indicated category are used for a tree split.

we found that SNP annotations based on replication timing experimental measurements ("Replication Timing") had high feature importance. The feature importance analysis also yielded a novel finding, that the Eigen-PC score [21] strongly contributed to CERENKOV's accuracy; ours if the first supervised method (of which we are aware) in which Eigen-PC scores are incorporated with SNP annotations for rSNP recognition. Out of the 246 features, we found that the DNA shape-based score ranked 105 in terms of its feature importance, and nearly as highly as the uniform DNase I hypersensitive site count (`uniformDhsCount`) feature.

3.3 Comparing CERENKOV to other methods for rSNP recognition

Having identified an `xgboost`-GBDT hyperparameter set that maximizes validation-set ⟨AUPVR⟩ for rSNP/cSNP discrimination, we precisely measured the validation-set performance of CERENKOV by AUPVR, AUROC, and AVGRANK, using 200 replications with 10-fold CV, and using bootstrap resampling of the results to obtain 95% confidence intervals (see Sec. 2.5). We found that on the `OSU17` set of SNPs, CERENKOV has a validation-set ⟨AUPVR⟩ of 0.505, an ⟨AUROC⟩ of 0.855, and an ⟨AVGRANK⟩ of 3.877 (Table 2).

Next, we compared CERENKOV to nine other methods for prioritizing functional noncoding SNPs: DeltaSVM [29], RSVP [39], DANN [41], fitCons [16], CADD [24], DeepSEA [59], DANQ [42], Eigen [21], and GWAVA [43]. For the single-score-per-SNP methods (CADD, Eigen, DANN, fitCons) that were not trained using ground-truth rSNPs from HGMD in the original reference studies in which the methods were reported, we used the published per-SNP scores directly to

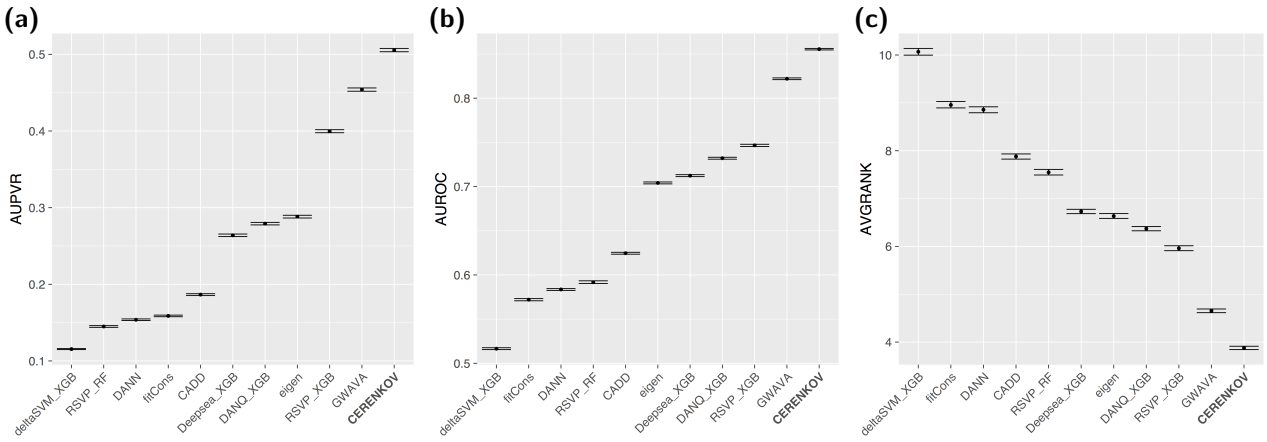

Figure 3: Validation performance of CERENKOV improves upon nine methods for rSNP detection to which we compared it, by both global rank-based measures (AUPVR (a) and AUROC (b)) and our AVGRANK method (c). Marks, average performance in 200 replications of 10-fold CV (2,000 samples); bars, bootstrap 95% confidence interval. XGB means `xgboost`-GBDT was used; RF means that Random Forest was used.

rank validation-set SNPs for computing AVGRANK and AUPVR. For the multi-feature methods, where possible, we used the classification algorithms as described in the original reference publications for the methods (i.e., for GWAVA, Random Forest; for DeepSEA, `xgboost`-GBDT) using the hyperparameters from the original publications.[3] For RSVP, since we did not have access to the MATLAB `treefit` implementation, we used both Random Forest (with imputed feature values; see Sec. 2.3.3) with hyperparameters matching those in the RSVP article, as well as `xgboost`-GBDT using our optimal set of hyperparameters. For DANQ, based on the high degree of similarity of the DANQ approach to the DeepSEA approach (identical chromatin datasets used for training the deep neural networks), we used the `xgboost`-GBDT method and the DeepSEA hyperparameter values; we post-processed the 919×2 DANQ scores for each SNP exactly as for DeepSEA. On identical assignments of genomic loci to CV folds (based on our locus-sampling approach), and with a 200-fold outer replication loop (thus yielding 2,000 performance samples per classifier), we compared the validation-set performance of the ten classifiers with CERENKOV, based on AUPVR, AUROC, and AVGRANK. We found that the average performance of CERENKOV was superior to the other ten classifiers, by all three performance measures (Fig. 3; bars indicate 95% confidence intervals). Consistent with the observation (Fig. 1b,c) that AUPVR and AVGRANK are closely related but that the relationship is not a simple monotonic function, the rankings of classifiers by AUPVR and AVGRANK in Fig. 3a and Fig. 3c are not identical; for example, Eigen outperforms DANQ_XGB by AUPVR, but DANQ_XGB outperforms Eigen by AVGRANK.

[3]The DeepSEA method produces one probability score for each of 919 convolutional neural network (CNN) models for each of two SNP alleles. We computed absolute score differences and absolute logit-transformed score differences (for the two alleles) for each CNN model and for each SNP, exactly as described in the DeepSEA method publication.

4 CONCLUSION AND DISCUSSION

We report a new framework and classifier, CERENKOV, for scoring noncoding SNPs based on their regulatory potential. CERENKOV—by virtue of its training-set construction criteria (locus-based, MAF ≥ 0.05), its novel performance measure (AVGRANK), and its novel CV approach (locus-based sampling)—is specifically designed for the problem of identifying candidate causal noncoding SNPs in GWAS post-analysis. We have demonstrated, using side-by-side comparisons on identical assignments of SNPs to CV folds, that CERENKOV's performance exceeds that of the nine other functional noncoding SNP prioritization methods to which we compared it, by both classical global rank-based measures (AUPVR and AUROC) and by the new GWAS-oriented performance measure (AVGRANK) that we proposed. CERENKOV's validation-set AUROC performance, 0.855 (95% CI of 0.854–0.856), compares favorably with the AUROC[4] (0.84) of the recently published PRVCS rSNP predictor [30].

The source code for CERENKOV is available on GitHub at `github.com/ramseylab/cerenkov` under the Apache 2.0 open-source software license, and the feature files that were used in the comparative analysis of CERENKOV with the other published methods are available on the CERENKOV website at `lab.saramsey.org/cerenkov`. By making the software, the data files, and in particular the `OSU17` SNP set (with benchmark results) available, we hope to accelerate development of methods for noncoding SNP functional analysis.

We anticipate that CERENKOV's performance may be improved through several possible enhancements that we are investigating, including new features and the use of a custom `xgboost`-GBDT loss function that is specifically designed to minimize AVGRANK. An appealing extension of

[4]The published [30] AUROC for the PRVCS classifier was based on an HGMD-based set of ground-truth SNPs with similar class imbalance (10.7) to that of the `OSU17` set of SNPs used here. Efforts to directly compare these classifiers on on the `OSU17` set of SNPs are ongoing.

CERENKOV would be to combine deep neural network-based approaches based on the local 1 kbp sequence haplotype (recognizing that the local haplotype is important to contextualizing functional SNP alleles [54]), with CERENKOV's current matrix of 246 features; such a hybrid "neural network plus decision trees" approach has shown promise in image classification [26]. It is presently unclear what the minimum attainable validation AVGRANK score would be expected to be, for the `OSU17` SNP set; undoubtedly, precision values are dampened by "latent positives" in the training dataset, i.e., high-scoring cSNPs that are simply undiscovered rSNPs. Using machine learning techniques that are specifically designed to address "positives-plus-unlabeled" problems [11] (such as the rSNP detection problem studied here) is another appealing avenue for future investigation.

ACKNOWLEDGMENTS

This work was supported by the NIH (award HL098807 to S.A.R.), the Medical Research Foundation of Oregon (New Investigator Award to S.A.R.), Oregon State University (Health Sciences award to S.A.R.), the Oregon State University Center for Genome Research and Biocomputing (in-kind contribution of compute time to S.A.R.), the PhRMA Foundation (Research Starter Grant in Informatics to S.A.R.) and the NSF (awards 1557605-DMS and 1553728-DBI to S.A.R.).

REFERENCES

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. 2015. A global reference for human genetic variation. *Nature* 526, 7571 (2015), 68–74.

[2] G Benson. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27, 2 (1999), 573–580.

[3] Christoph Bock, Jörn Walter, Martina Paulsen, and Thomas Lengauer. 2007. CpG Island Mapping by Epigenome Prediction. *PLOS Computat Biol* 3, 6 (2007), e110.

[4] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[5] Razvan C. Bunescu and Raymond J. Mooney. 2007. Multiple Instance Learning for Sparse Positive Bags. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML-2007)*. ACM, Corvallis, OR, 105–112.

[6] Juan Caballero, Arian F A Smit, Leroy Hood, and Gustavo Glusman. 2014. Realistic artificial DNA sequences as negative controls for computational genomics. *Nucleic Acids Res* 42, 12 (2014), e99.

[7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *arXiv.org* 1603.02754 (2016), 1–13.

[8] Gregory M Cooper, Eric A Stone, George Asimenos, NISC Comparative Sequencing Program, Eric D Green, Serafim Batzoglou, and Arend Sidow. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15, 7 (2005), 901–913.

[9] Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E Hunt, Sophie H Janacek, Nathan Johnson, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Fergal J Martin, Thomas Maurel, William McLaren, Daniel N Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P Wilder, Amonida Zadissa, Bronwen L Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M J Searle, Giulietta Spudich, Stephen J Trevanion, Andy Yates, Daniel R Zerbino, and Paul Flicek. 2015. Ensembl 2015. *Nucleic Acids Research* 43, Database issue (2015), D662–9.

[10] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol* 6, 12 (2010), e1001025.

[11] C Elkan and K Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Las Vegas, NV, 213–220.

[12] Vadim Farztdinov and Fionnuala McDyer. 2012. Distributional fold change test - a statistical approach for detecting differential expression in microarray experiments. *Algorithms Mol Biol* 7, 1 (2012), 29.

[13] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P Boyle, Philip Cayting, Alexandra Charos, David Z Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Frietze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O'Geen, Zhengqing Ouyang, E Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E Reddy, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 7414 (2012), 91–100.

[14] GTEx Consortium. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 6235 (2015), 648–660.

[15] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, and Bas van Steensel. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 7197 (2008), 948–951.

[16] Brad Gulko, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics* 47, 3 (2015), 276–283.

[17] R Scott Hansen, Sean Thomas, Richard Sandstrom, Theresa K Canfield, Robert E Thurman, Molly Weaver, Michael O Dorschner, Stanley M Gartler, and John A Stamatoyannopoulos. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* 107, 1 (2010), 139–144.

[18] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James G R Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E Antonarakis, and Roderic Guigo. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7 Suppl 1 (2006), S4.1–9.

[19] Ichiro Hiratani, Tyrone Ryba, Mari Itoh, Tomoki Yokochi, Michaela Schwaiger, Chia-Wei Chang, Yung Lyou, Tim M Townes, Dirk Schübeler, and David M Gilbert. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biology* 6, 10 (2008), e245.

[20] Ross Ihaka and Robert Gentleman. 1995. R: A Language for Data Analysis and Graphics. *J Comp Graph Stat* 5, 3 (1995), 299–314.

[21] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics* 48, 2 (2016), 214–220.

[22] Andrew D Johnson, R E Handsaker, S L Pulit, M M Nizzari, C J O'Donnell, and P I W de Bakker. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 24 (2008), 2938–2939.

[23] Donna Karolchik, Angie S Hinrichs, and W James Kent. 2009. The UCSC Genome Browser. In *Cur Protoc Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[24] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genet* 46, 3 (2014), 310–315.

[25] R Kohavi. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Selection and Model Estimation. In *Proc Int Joint Conf Artif Intel*. ACM, San Francisco, CA, 1137–1143.

[26] P Kontschieder, M Fiterau, and A Criminisi. 2015. Deep neural decision forests. In *Proc Int Conf Comput Vision*. IEEE, Santiago, Chile, 1467–1475.

[27] Gregory V Kryukov, Len A Pennacchio, and Shamil R Sunyaev. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80, 4 (2007), 727–739.

[28] S G Landt, G K Marinov, A Kundaje, P Kheradpour, F Pauli, S Batzoglou, B E Bernstein, P Bickel, J B Brown, P Cayting, Y Chen, G DeSalvo, C Epstein, K I Fisher-Aylor, G Euskirchen, M Gerstein, J Gertz, A J Hartemink, M M Hoffman, V R Iyer, Y L Jung, S Karmakar, M Kellis, P V Kharchenko, Q Li, T Liu, X S Liu, L Ma, A Milosavljevic, R M Myers, P J Park, M J Pazin, M D Perry, D Raha, T E Reddy, J Rozowsky, N Shoresh, A Sidow, M Slattery, J A Stamatoyannopoulos, M Y Tolstorukov, K P White, S Xi, P J Farnham, J D Lieb, B J Wold, and M Snyder. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* 22, 9 (2012), 1813–1831.

[29] Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nature Genet* 47, 8 (2015), 955–961. gkm-SVM.

[30] Mulin Jun Li, Zhicheng Pan, Zipeng Liu, Jiexing Wu, Panwen Wang, Yun Zhu, Feng Xu, Zhengyuan Xia, Pak Chung Sham, Jean-Pierre A Kocher, Miaoxin Li, Jun S Liu, and Junwen Wang. 2016. Predicting regulatory variants with composite statistic. *Bioinformatics* 32, 18 (2016), 2729–2736.

[31] M J Li, B Yan, P C Sham, and J Wang. 2015. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief Bioinformatics* 16, 3 (2015), 393–412.

[32] Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W Wasserman. 2016. DNA Shape Features Improve Transcription Factor Binding Site Predictions *In Vivo*. *Cell systems* 3, 3 (2016), 278–286.e4.

[33] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 6099 (2012), 1190–1195.

[34] Stephen B Montgomery, Obi L Griffith, Johanna M Schuetz, Angela Brooks-Wilson, and Steven J M Jones. 2007. A survey of genomic properties for the detection of regulatory polymorphisms. *PLOS Comput Biol* 3, 6 (2007), e106.

[35] Stephen B Montgomery, O L Griffith, M C Sleumer, C M Bergman, M Bilenky, E D Pleasance, Y Prychyna, X Zhang, and S J M Jones. 2006. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* 22, 5 (2006), 637–640.

[36] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. 2010. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genet* 6, 4 (2010), e1000888.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12 (2011), 2825–2830.

[38] Len A Pennacchio, Nadav Ahituv, Alan M Moses, Shyam Prabhakar, Marcelo A Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D Lewis, Ingrid Plajzer-Frick, Jennifer Akiyama, Sarah De Val, Veena Afzal, Brian L Black, Olivier Couronne, Michael B Eisen, Axel Visel, and Edward M Rubin. 2006. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444, 7118 (2006), 499–502.

[39] Thomas A Peterson, Matthew Mort, David N Cooper, Predrag Radivojac, Maricel G Kann, and Sean D Mooney. 2016. Regulatory Single-Nucleotide Variant Predictor Increases Predictive Performance of Functional Regulatory Variants. *Hum Mutat* 37, 11 (2016), 1137–1143.

[40] E Portales-Casamar, S Thongjuea, A T Kwon, D Arenillas, X Zhao, E Valen, D Yusuf, B Lenhard, W W Wasserman, and A Sandelin. 2009. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 38, Database (2009), D105–D110.

[41] Daniel Quang, Yifei Chen, and Xiaohui Xie. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 5 (2015), 761–763.

[42] Daniel Quang and Xiaohui Xie. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 44, 11 (2016), e107.

[43] Graham R S Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. 2014. Functional annotation of noncoding sequence variants. *Nature Methods* 11, 3 (2014), 294–296.

[44] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. 2009. The role of DNA shape in protein–DNA recognition. *Nature* 461, 7268 (2009), 1248–1253.

[45] Marc A Schaub, Alan P Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* 22, 9 (2012), 1748–1759.

[46] S T Sherry, M H Ward, M Kholodov, J Baker, L Phan, E M Smigielski, and K Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29, 1 (2001), 308–311.

[47] A Siepel. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 8 (2005), 1034–1050.

[48] Barbara E Stranger, Eli A Stahl, and Towfique Raj. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187, 2 (2011), 367–383.

[49] The ENCODE Project Consortium. 2011. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology* 9, 4 (2011), e1001046.

[50] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489, 7414 (2012), 75–82.

[51] Ali Torkamani and Nicholas J Schork. 2008. Predicting functional regulatory polymorphisms. *Bioinformatics* 24, 16 (2008), 1787–1792.

[52] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. 2007. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Research* 35, Database issue (2007), D88–92.

[53] Kai Wang, Mingyao Li, and Hakon Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38, 16 (2010), e164.

[54] Lucas D Ward and Manolis Kellis. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnol* 30, 11 (2012), 1095–1106.

[55] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42, Database issue (2014), D1001–6. accessed in 2016.

[56] Marvin N. Wright and Andreas Ziegler. 2015. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv.org* 1508.04409 (2015), 1–17.

[57] Jichen Yang and Stephen A Ramsey. 2015. A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites. *Bioinformatics* 31, 21 (2015), 3445–3450.

[58] Yiqiang Zhao, Wyatt T Clark, Matthew Mort, David N Cooper, Predrag Radivojac, and Sean D Mooney. 2011. Prediction of functional regulatory SNPs in monogenic and complex disease. *Hum Mutat* 32, 10 (2011), 1183–1190.

[59] Jian Zhou and Olga G Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* 12, 10 (2015), 931–934.

[60] T Zhou, L Yang, Y Lu, I Dror, A C Dantas Machado, T Ghane, R Di Felice, and R Rohs. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 41, W1 (2013), W56–W62.