Yao *et al. BMC Bioinformatics* (2019) 20:63
https://doi.org/10.1186/s12859-019-2637-4

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

CERENKOV2: improved detection of functional noncoding SNPs using data-space geometric features

Yao Yao[1,2], Zheng Liu[1,2], Qi Wei[1,2] and Stephen A. Ramsey[1,2*]

Abstract

Background: We previously reported on CERENKOV, an approach for identifying regulatory single nucleotide polymorphisms (rSNPs) that is based on 246 annotation features. CERENKOV uses the xgboost classifier and is designed to be used to find causal noncoding SNPs in loci identified by genome-wide association studies (GWAS). We reported that CERENKOV has state-of-the-art performance (by two traditional measures and a novel GWAS-oriented measure, AVGRANK) in a comparison to nine other tools for identifying functional noncoding SNPs, using a comprehensive reference SNP set (`OSU17`, 15,331 SNPs). Given that SNPs are grouped within loci in the reference SNP set and given the importance of the data-space manifold geometry for machine-learning model selection, we hypothesized that within-locus inter-SNP distances would have class-based distributional biases that could be exploited to improve rSNP recognition accuracy. We thus defined an intralocus SNP "radius" as the average data-space distance from a SNP to the other intralocus neighbors, and explored radius likelihoods for five distance measures.

Results: We expanded the set of reference SNPs to 39,083 (the `OSU18` set) and extracted CERENKOV SNP feature data. We computed radius empirical likelihoods and likelihood densities for rSNPs and control SNPs, and found significant likelihood differences between rSNPs and control SNPs. We fit parametric models of likelihood distributions for five different distance measures to obtain ten log-likelihood features that we combined with the 248-dimensional CERENKOV feature matrix. On the `OSU18` SNP set, we measured the classification accuracy of CERENKOV with and without the new distance-based features, and found that the addition of distance-based features significantly improves rSNP recognition performance as measured by AUPVR, AUROC, and AVGRANK. Along with feature data for the `OSU18` set, the software code for extracting the base feature matrix, estimating ten distance-based likelihood ratio features, and scoring candidate causal SNPs, are released as open-source software CERENKOV2.

Conclusions: Accounting for the locus-specific geometry of SNPs in data-space significantly improved the accuracy with which noncoding rSNPs can be computationally identified.

Keywords: SNP, GWAS, noncoding, rSNP, Data space, Machine learning

Background

The rSNP detection problem

Human genome-wide association studies (GWAS) have led to the discovery of genetic variant-to-trait associations in thousands of studies collectively involving millions of individuals [1]. Functional interpretation of genetic loci identified through GWAS has primarily focused on *coding regions* in which single nucleotide polymorphisms (SNPs) can be mapped to consequence predictions based on amino acid changes [2]; however, 90% of human GWAS-identified SNPs are located in *noncoding* regions [3]. Within a noncoding trait-associated region, it is difficult to pinpoint the regulatory SNP (or rSNP) that is causal for trait variation [4]. Various types of SNP annotations that correlate with functional rSNPs are known [5], for example, phylogenetic sequence conservation [6] and expression quantitative trait locus (expression QTL,

*Correspondence: stephen.ramsey@oregonstate.edu
[1]School of Electrical Engineering and Computer Science, Oregon State University, 97330 Corvallis, OR, USA
[2]Department of Biomedical Sciences, Oregon State University, 106 Dryden Hall, 97330 Corvallis, OR, USA

or eQTL) association [7]. But the general problem of how to integrate various types of genomic, phylogenetic, epigenomic, transcription factor binding site (TFBS), and chromatin-structural rSNP correlates in order to identify rSNPs is a fundamental challenge in computational biology. Progress on this problem has been spurred by the growth of literature-curated databases of experimentally validated rSNPs such as the Human Gene Mutation Database [8] (HGMD), ORegAnno [9] or ClinVar [10]. While various approaches to the rSNP recognition problem have been proposed that do not involve training based on an example set of experimentally validated rSNPs (we call such methods "unsupervised" approaches) [11–21], converging lines of evidence from our work [22] and others' [23–26] suggest (but are not *entirely* consistent on this point [21]) that approaches that are supervised by example sets of experimentally validated rSNPs significantly improves accuracy with which rSNPs can be discriminated from nonfunctional noncoding SNPs.

Many types of genomic data have been used to derive SNP annotation features that have proved useful in supervised models for rSNP recognition [22]. The picture emerging from dozens of studies over the past ten years is that increasing the breadth and diversity of such SNP annotation features improves rSNP detection, and thus there has been a steady increase in the number of features that are used in machine-learning approaches for this problem, from 23 features [23], to 28 features [27], to 158 features [28], to 175 features [24], to 246 features in our previous work [22]. The dimensionality of feature-spaces has rapidly increased in the last few years, with reports of rSNP recognition models that incorporate 919 features [16, 26, 29] derived from epigenomic data from the Encyclopedia of DNA Elements (ENCODE) project [30] and 2132 features [25] derived from the Gene Ontology [31]. However, in our previous work [22] we found that a model with a 246-dimensional feature space clearly outperformed models [25, 26, 29] with significantly higher-dimensional feature spaces. This suggests that feature-feature correlation within, and sparsity of, high-dimensional feature-sets may lead to diminishing returns in terms of improving rSNP detection accuracy.

A variety of supervised classification algorithms have been proposed for identifying functional noncoding SNPs, including the support vector machine (SVM) [17, 19, 23, 32], naïve Bayes [27], ensemble decision tree algorithms [24, 25, 28], probabilistic graphical models [18, 33], deep neural networks [20, 26, 29], weighted sum of feature ranks [34], and our work using regularized gradient boosted decision trees [22] and deep residual networks [35]. Recently, there have been several proposals of hybrid methods such as combining recurrent and convolutional neural networks [26] and integrating deep neural networks with regularized gradient boosted decision trees [29]. Beyond binary classification, regression-based approaches have been proposed for detecting rSNPs, including linear regression [36] and a mixture-of-regressions model [37]. Overall, there has been a shift toward models with higher parametric complexity as the sizes of example sets of experimentally validated rSNPs has increased [22].

Novelty and performance of our previous CERENKOV method

In our previous work [22], we described CERENKOV (Computational Elucidation of the REgulatory NonKOding Variome), a machine-learning approach for rSNP recognition that incorporated four key innovations. First, CERENKOV incorporated a within-group-rank-based measure of classification accuracy, which we called AVGRANK. AVGRANK more realistically models the costs associated with incorrect predictions in post-GWAS SNP analysis than typical measures of accuracy like area under the receiver operating characteristic (AUROC) curve or area under the precision-vs-recall (AUPVR) curve. We found that optimizing a model to maximize AUPVR does not guarantee optimality for AVGRANK, and thus, that both measures should be taken into account in evaluating the performance of a computational model for rSNP recognition. Second, in CERENKOV we used a state-of-the-art regularized gradient boosted decision tree (xgboost) classification algorithm [38], which improved upon the rSNP recognition performance that could be achieved (on an identical feature-set) using the previously-proposed classification algorithms Random Forest and Kernel Support Vector Machine [22]. Third, for CERENKOV we engineered 246 SNP-level features from phylogenetic, genomic, epigenomic, chromatin structural, cistromic, population genetic, replication-timing, and functional genomic datasets. Fourth, we trained, validated, and performance-benchmarked CERENKOV using a reference set of 15,331 SNPs (the `OSU17` SNP set) comprising 1659 experimentally validated human rSNPs and 13,672 neighboring "control" SNPs (cSNPs) that are each in strong linkage disequilibrium with at least one rSNP. We selected the `OSU17` SNPs to represent noncoding loci that would be expected to be encountered in a post-GWAS analysis, based on population minor allele frequency [22]. We compared the accuracy of CERENKOV to nine other published rSNP recognition models (DeltaSVM [19], RSVP [25], DANN [20], fitCons [18], CADD [17], DeepSEA [29], DANQ [26], Eigen [21], and GWAVA [24]) and found that CERENKOV's performance significantly improved upon the current state-of-the-art, by AUPVR, AUROC, and AVGRANK.

Introducing CERENKOV2

In this work we report on CERENKOV2, a next-generation machine-learning approach for rSNP recognition that improves upon our previous approach, CERENKOV [22] in terms of accuracy and insights into the data-space geometry of the problem. In addition to using a significantly expanded reference set of SNPs [the OSU18 SNP set (see "The OSU18 reference SNP set" section), which has 39,083 SNPs for model benchmarking], we have incorporated new engineered features into CERENKOV2 that are based on likelihood ratios of average SNP-to-neighboring-SNPs distances for various types of distance measures, as described below. By taking account geometric properties of the distribution of SNPs in data space (as described in detail in the next section), CERENKOV2 achieves significantly better rSNP recognition performance than CERENKOV.

The importance of data-space geometry

It is a well-established principle in machine-learning that understanding the manifold structure of cases in data-space can help guide appropriate selection of a classification model and/or geometric features that enable more accurate classification [39, 40]. Data-space inter-sample distance measures are fundamental to many machine-learning algorithms such as k-Nearest-Neighbors [41] (k-NN), and in the case of k-NN, the choice of distance measure can be a key determinant of the accuracy of the classifier [42]. Given that (1) rSNPs and cSNPs are grouped into genetic loci in which the within-locus SNPs are in linkage disequilibrium with one another (making rSNP recognition a *grouped* machine-learning problem), and (2) in the reference SNP set, each associated locus has at least one rSNP in it and usually many cSNPs (such that the problem has a "sparse positive bag" structure [43, 44]), we hypothesized that within-locus SNP-SNP distances in data space may be informative for discriminating rSNPs from cSNPs. But despite the importance of the choice of data-space metric in many machine-learning applications and in clustering [45], the potential utility of data-space metric-based features for improving accuracy of computational recognition of rSNPs has not to our knowledge been systematically explored. Here we report on the first effort (of which we are aware) to improve rSNP detection performance by systematically incorporating data-space geometric features, specifically, intralocus SNP-SNP distances in feature space.

Data-space geometric features for rSNP recognition

Based on our initial observation that SNPs within the same locus tend to be clustered in data space, we investigated whether there are class label-specific biases in the locus-based average SNP-to-neighboring-SNPs distances that could be exploited to improve accuracy for discriminating rSNPs from cSNPs. In mathematical terms, for a SNP s, we denote by $L(s)$ the set of SNPs within the same locus as s (for details on the selection of cSNPs that are within the same locus as an rSNP, see "Methods" section). Then, for a given locus s and a given distance metric $d(\cdot,\cdot)$, we define an intralocus average SNP-to-neighboring-SNP distance or "intralocus radius" $\lambda_{s|d}$ by

$$\lambda_{s|d} = \frac{1}{|L(s)|-1} \sum_{s' \in L(s), s' \neq s} d(s,s'). \quad (1)$$

One such metric would be the Pearson distance defined as $d(s,s') = 1 - r(s,s')$, where $r(s,s')$ is the Pearson correlation coefficient [46] between the feature vectors of SNPs s and s'. With Pearson distance being applied, we found that the distribution of intralocus radii for rSNPs were markedly different from cSNPs', with rSNPs often having higher intralocus radii than cSNPs, i.e., $\lambda_{r|\text{Pearson}} > \lambda_{c|\text{Pearson}}$. Given the sparsity of rSNPs in the genome (cSNPs outnumber rSNPs 14.5 to one in the OSU18 SNP set) and the typically large linkage disequilibrium-defined locus sizes in the human genome [47], the locus neighborhood for any given s in general mostly contains cSNPs. Together, these observations suggest that in feature-space, the SNPs of a given locus have an "atom"-like structure with respect to Pearson distance–a core rSNP and a "cloud" of cSNPs with higher average distance from the it (Fig. 1).

Based on this initial observation, we systematically calculated intralocus radii for each SNP in the OSU18 reference SNP set, using five different distance measures (Canberra [48], Euclidean [49], Manhattan [50], cosine [51], and Pearson) applied to both scaled and unscaled feature data (for a total of ten combinations). We found significant differences between the distributions of the ten intralocus radius values conditioned on the two classes (rSNPs and cSNPs). Based on this, we parametrically modeled the intralocus radius distributions (see "Analysis of intralocus radius distributions for rSNPs and cSNPs" section) and thereby obtained log-likelihood ratios that we incorporated into the feature set for CERENKOV2 (see "Using data-space geometric features in CERENKOV2" section). We quantified the relative importance of the distance based features in the context of the CERENKOV2 base feature-set (see "CERENKOV2 feature importance" section). Finally, we compared the classification performance of CERENKOV2—including the new distance-based features—with that of CERENKOV on the OSU18 reference SNP set (see "Comparison of CERENKOV2 vs. CERENKOV performance" section) and found that CERENKOV2 had significantly better performance than CERENKOV, by AUROC, AUPVR, and AVGRANK. The complete feature data for the OSU18 training and validation SNP set are available online and

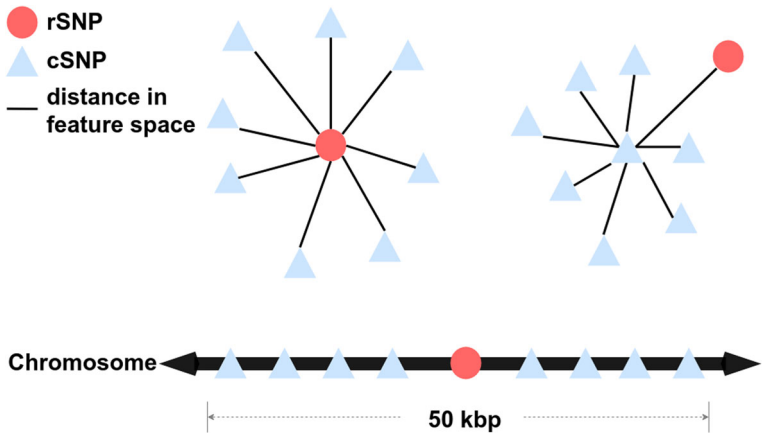

Fig. 1 The geometric idea behind the intralocus distance features that are used in CERENKOV2. Top panel, SNPs from the same locus form a data-space "cloud." Triangles and circles, SNPs; black lines, distances between a central SNP and the other SNPs within the locus. Bottom panel, SNPs shown in their chromosomal context

the software code for CERENKOV2 is freely distributed to the scientific community online under an open-source license (see "Availability of data and materials" section).

Results

Analysis of intralocus radius distributions for rSNPs and cSNPs

We computed intralocus radii for each of the `OSU18` SNPs (see "Computing the geometric features" section) using ten combinations of distance measures and data matrices: Canberra distance, Euclidean distance (L^2 norm), Manhattan distance, cosine distance (defined as 1.0 minus the cosine similarity) and Pearson distance, each on unscaled data and min-max scaled data (the latter set of distance measures will be designated with the suffix "(scaled)" in each case). We first analyzed the intralocus SNP-SNP radius distributions for the two SNP classes (rSNPs and cSNPs) within 248-dimensional feature-space using kernel density estimation for radius values conditioned on the class label (rSNP or cSNP) of the reference SNP. As seen in Fig. 2 (see also Additional file 1: Table S2), there are class label-dependent differences in the skewness and kurtosis, indicting that geometric biases exist between rSNPs and cSNPs in data-space.

For cosine and Pearson distances, the intralocus radius distributions for rSNPs are slightly more skewed to the left and more platykurtic than the distributions for cSNPs; in terms of Euclidean and Manhattan distances, the intralocus radius distributions for rSNPs are left-skewed and more leptokurtic, while the cSNPs' are right-skewed and less leptokurtic; for the rest distances, the intralocus radius distributions for cSNPs are slightly more skewed to the right and more leptokurtic than the distributions for rSNPs (see also Additional file 1: Table S2).

Analysis of intralocus radius likelihood ratios (rSNP vs. cSNP)

The intralocus radius distribution analysis suggested that taking account of the intralocus radius likelihood for the SNP conditioned on a possible class label (rSNP or cSNP) would be useful for discriminating rSNPs from cSNPs. To visualize the potential class-label discriminating power of each of the ten methods for computing intralocus radii, we empirically estimated the rSNP/cSNP log-likelihood ratios (LLRs) for the ten different methods for computing intralocus radii using binned counts of SNPs for posterior probability estimation (Fig. 3). Consistent with the differences seen in the density distributions (Fig. 2), we found that log-likelihood ratios were significantly different from zero for the majority of bins for intralocus radii computed, for each of the ten distance measures except for cosine (unscaled) and Pearson (unscaled).

Next, we extracted features from intralocus radii for use in the CERENKOV classifier, using sets of SNPs that were reserved for training within a cross-validation framework (see "Gradient boosted decision trees" section). In order to avoid issues with zero-count bins associated with the limited number of SNP loci within a single cross-validation fold, we used a parametric approach: instead of empirically estimating likelihood ratios, for each of the ten methods for computing intralocus radii we fit parametric distributions to the radius values (conditioned on the class label of the reference training SNP). We then applied the fitted parametric models to compute log-likelihood ratios for both the training and validation sets of SNPs and integrated those ten log-likelihood ratios as feature vectors, yielding a 258-column feature matrix input for classification which we compared to performance (using the same classification algorithm) of the original 248-column feature matrix.

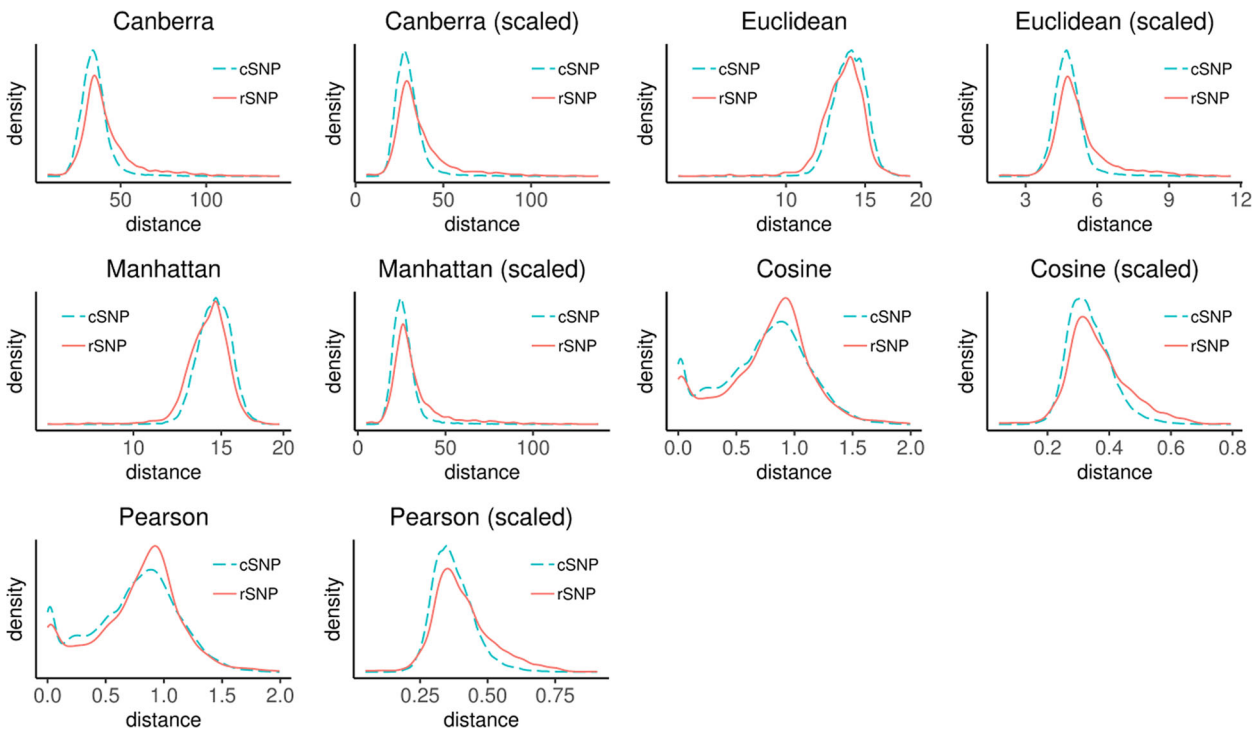

Fig. 2 Distributions of intralocus radii computed using five different distance measures (Canberra, Euclidean, Manhattan, cosine, and Pearson) applied to scaled and unscaled feature data, conditioned on the type of reference SNP (rSNP or cSNP) for the intralocus radius calculation. Results shown are for all `OSU18` SNPs (see "The `OSU18` reference SNP set" section). Significant differences in the rSNP likelihoods vs. cSNP likelihoods are evident for Canberra, Canberra (scaled), Euclidean (scaled), Manhattan (scaled), cosine, and Pearson methods for computing intralocus radii. Modest differences in rSNP vs. cSNP likelihoods were evident for the cases of Euclidean and Manhattan methods for computing intralocus radii

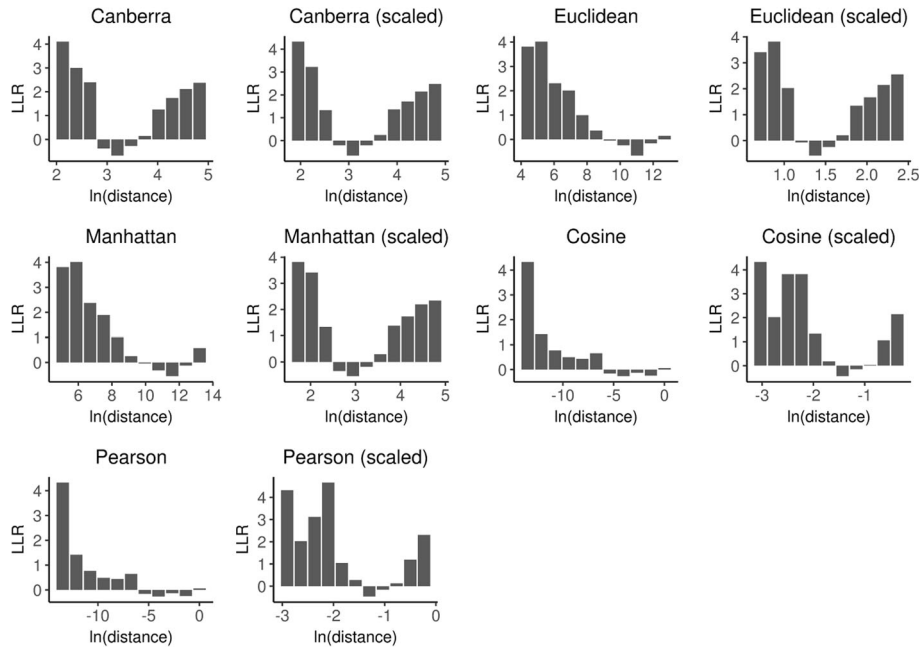

Fig. 3 Empirically estimated log-likelihood ratios (rSNP/cSNP) based on intralocus radii computed using ten methods. Results shown are for all `OSU18` SNPs (see "The `OSU18` reference SNP set" section). LLR, log-likelihood ratio (natural logarithm); ln, natural logarithm

Using data-space geometric features in CERENKOV2

On an identical starting set of reference SNPs (`OSU18`, see "The `OSU18` reference SNP set" section) and identical assignments of SNPs to cross-validation folds, we compared the performance of the CERENKOV classification algorithm incorporating the 248-column feature matrix (without intralocus radii-based features) with the performance of the CERENKOV algorithm incorporating a 258-column feature matrix (including intralocus radii-based features). Using ten independent replications of five-fold cross-validation with grouped sampling based on locus ("locus-based sampling", see "Gradient boosted decision trees" section) and using three metrics (AUPVR, AUROC, and AVGRANK [22]), we measured performance separately for classification using the two feature matrices and using xgboost hyperparameters selected to maximize training-set AUPVR (see "Hyperparameter tuning" section). For the classification algorithm we used a high-performance implementation of regularized gradient boosted decision trees (`xgboost` [38], hereafter, `xgboost`-GBDT). For the two models, the inputs to `xgboost` were thus a 39,083×248 feature matrix and a 39,083×258 feature matrix, respectively. We trained and tested `xgboost`-GBDT (using ten independent replications of five-fold [52] cross-validation with locus-based sampling [22]) with the optimal `xgboost` hyperparameters (see "Hyperparameter tuning" section).

Comparison of CERENKOV2 vs. CERENKOV performance

Within the above-described cross-validation framework, we found that the inclusion of the ten geometric features improved validation-set AUPVR from 0.358 to 0.402 ($p < 10^{-25}$), AUROC from 0.830 to 0.839 ($p < 10^{-18}$), and AVGRANK from 11.172 to 10.994 (lower is better for AVGRANK [22]; $p < 0.004$) (Fig. 4 and Additional file 1: Table S1). From these results, we concluded that the addition of the ten geometric features based on the intralocus radius of SNPs in data-space significantly improved performance for rSNP recognition.

CERENKOV2 feature importance

In order to better understand the contributions of different categories of features—particularly geometric features—to rSNP recognition accuracy, we separately trained a Random Forest algorithm on the 258-column feature matrix for the `OSU18` reference SNP set (see "The `OSU18` reference SNP set" section) and then obtained permutation [53] and Gini impurity [54]-based estimates of the importance of each of the 258 features (Fig. 5). Consistent with findings from the Peterson et al. study [25], SNP annotations based on replication timing experimental measurements ("repliseq") had highest overall feature importance; however, the ten log-likelihood-ratio features that were based on data-space geometry strongly contributed to accuracy for rSNP recognition.

Application of CERENKOV2 to identify trait-associated noncoding SNPs

To illustrate the biological utility of CERENKOV2, we used CERENKOV2 to compute rSNP prediction scores for noncoding SNPs in the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) database. We identified two noncoding SNPs that are trait-associated in GRASP and that have CERENKOV2 rSNP prediction scores greater than 0.7: rs2239633

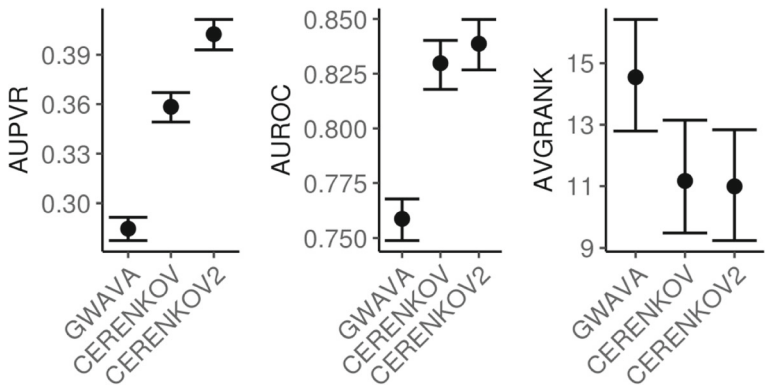

Fig. 4 Performance of GWAVA, CERENKOV and CERENKOV2 on the `OSU18` reference SNP set, by three performance measures. Marks, sample arithmetic mean of validation-set performance; bars, estimated 95% confidence intervals (see "Gradient boosted decision trees" section); `GWAVA`, based on the GWAVA's Random Forest model with 174 features [24]; `CERENKOV`, our previous model with the base 248-column feature matrix; `CERENKOV2`, our current model consisting of the base feature matrix plus ten log-likelihood features derived from intralocus radii and fitted using training data only; AUPVR, area under the precision-vs-recall curve (higher is better); AUROC, area under the receiver operating characteristic curve (higher is better); AVGRANK, intralocus average score rank (lower is better [22])

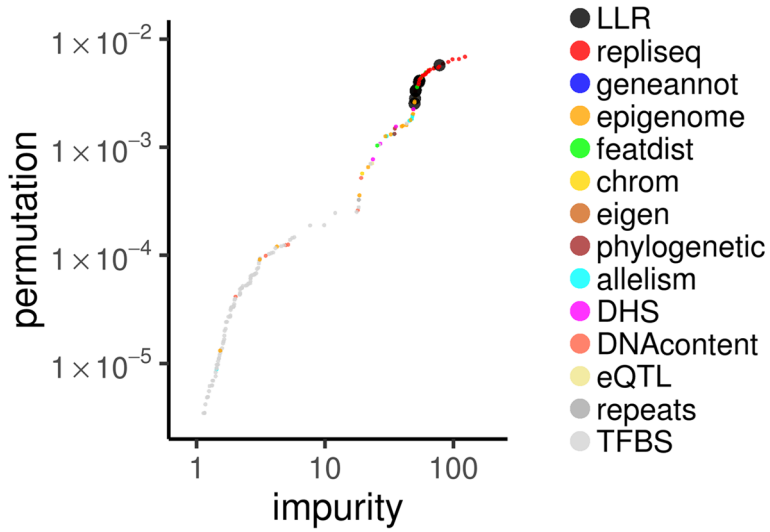

Fig. 5 Gini and permutation importance values of 258 features in 14 categories (colored marks). Feature category labels as follows: "LLR", log-likelihood ratio (the new data-space geometric features); "repliseq", replication timing; "geneannot", gene-model annotation-based; "epigenome", epigenomic segmentation [67, 68]; "featdist", SNP location-related; "chrom", the chromosome; "eigen", based on the Eigen [21] score; "phylogenetic", phylogenetic interspecies local sequence conservation [6, 80, 81]; "allelism", allele and MAF-related; "DHS", DNase I hypersensitive site; "DNAContent", local nucleotide frequences; "eQTL", expression quantitative trait locus [75]; "repeats", genomic repeat annotation; "TFBS", transcription factor binding site (see "Extracting the nongeometric features" section and Ref. [22] for details)

(associated with acute lymphoblastic leukemia), and rs11071720 (associated with mean platelet volume in circulation, and with gene expression of *TPM1* in blood. This illustrates how CERENKOV can be used to filter GWAS summary results to identify noncoding SNPs that have high potential to have a mechanistic (gene regulatory) interpretation.

Discussion

We anticipate that CERENKOV2's performance may be improved through several possible enhancements. An appealing extension would be to combine deep neural network-based approaches based on the local 1 kbp sequence haplotype (recognizing that the local haplotype provides important correlates of functional SNP alleles [55]), with CERENKOV2's current set of 258 SNP features. Our previous work [35] has demonstrated that a classifier (Res2s2am) based on a deep residual network architecture has state-of-the-art performance on the related problem of discriminating trait-associated noncoding SNPs from control noncoding SNPs. While the present work significantly improves rSNP recognition accuracy, the validation-set AVGRANK performance values (averaging nearly 11) clearly demonstrate that further gains in accuracy are needed in order to fully realize the potential of integrative, data-driven computational approaches to substantially accelerate the search for causal noncoding GWAS variants. 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 [56] (such as the rSNP detection problem studied here) or semi-supervised learning algorithms [57] would seem to offer a principled approach to handling the issue of latent positives among the cSNPs. Given the extent to which common features (e.g., replication timing, local GC content, phylogenetic sequence conservation, chromatin accessibility, and transcription factor binding sites [22]) are used by many supervised tools for rSNP recognition, the results from our analysis of the performance of CERENKOV2 suggest that accounting for the intralocus data-space geometry of SNPs may be broadly useful for advancing bioinformatics for post-GWAS SNP analysis.

Conclusion

CERENKOV2 significantly improves upon our previous framework and classifier, CERENKOV, in its ablity to score noncoding SNPs based on their regulatory potential. CERENKOV2—by virtue of its training-set construction criteria (locus-based, MAF $\geq$ 0.05) and its novel feature set including geometric ones—is specifically designed for the problem of identifying candidate causal noncoding SNPs in GWAS summary regions. We have demonstrated, using side-by-side comparisons on identical assignments of SNPs to cross-validation folds, that CERENKOV2's performance exceeds that of our previous CERENKOV, by both classical global rank-based measures (AUPVR and AUROC) and by the GWAS-oriented performance measure (AVGRANK) that we previously

proposed. In particular, CERENKOV2's validation-set AUPVR performance, 0.404, is a significant improvement over CERENKOV's AUPVR of 0.359 on the same reference SNP set (OSU18). The results reported in this work are based on a significantly expanded reference SNP set (OSU18, which has more than double the number of SNPs in the OSU17 reference set), which should increase the generalizability and robustness of the performance results reported herein.

The source code, feature data files, and instructions for installing and running CERENKOV2 are freely available online (see "Availability of data and materials" section). By making the software, the data files, and in particular the OSU18 SNP set (with benchmark results) available, we hope to accelerate development of methods for functional analysis of noncoding SNPs and ultimately increase the yield of molecular insights from GWAS.

Methods

The OSU18 reference SNP set

We obtained minor allele frequencies (MAFs) for all SNPs from the dbSNP-based [58] snp146 SQL table hosted at the UCSC Genome Browser [59] site. For the representative set of rSNPs for training/evaluation, we obtained 2,529 rSNPs in total from HGMD (Rel. 2017.2), ORegAnno (Rel. 2015.12.22) and ClinVar that satisfied all of the following criteria: (i) for all SNPs from HGMD, they were marked as regulatory in HGMD and the disease field did not contain cancer; the other SNPs from ORegAnno or ClinVar were of GRCh37 (hg19) assembly; (ii) MAF $\geq$ 0.05; (iii) the SNP was not an indel and not contained within a coding DNA sequence (CDS; based on the complete set of transcripts from the Ensembl 75 gene annotation build); and (iv) 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 Annotation and Proxy Search (SNAP) tool [60] to identify SNPs that are in LD ($r^2 \geq 0.8$ in 1,000 Genomes (1KG) Phase 1 [61], with data from the International HapMap Project [62] used instead of 1KG for chromosome X), and we filtered to include only SNPs within 50 kbp of an rSNP, that were not contained within a CDS, that have MAF $\geq$ 0.05, and that are not themselves on the list of rSNPs. Overall, this filtering procedure produced a list of 36,554 cSNPs. The combined set of 39,083 SNPs (which we call the OSU18 reference SNP set) was thus designed as an appropriate reference set for the application of post-GWAS SNP analysis. Overall, the class imbalance of OSU18 is ~14.454 (cSNP/rSNP).

Extracting the nongeometric features

The CERENKOV feature extraction software is based on Python and SQL. We extracted 248 SNP features for each of the OSU18 SNPs, using information and measurements from SNP annotation databases, epigenomic and chromatin datasets and phylogenetic conservation scores (Table 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 [63] (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 [64]). The ENCODE_TFBS feature is presented in Table 1 as a single feature for conciseness, but in fact it is 160 separate binary features, one for each transcription factor (TF) for which genome-wide TFBS data (from chromatin immunoprecipitation with high-throughput sequencing readout, or ChIP-seq) and peak data (from the ENCODE Uniform Peaks analysis) are available [64]. For replication timing features, we processed track-specific BigWig files for Repli-seq [65] and Repli-chip [66] experiments from UCSC to obtain the timing scores at individual SNP positions. For ChromHMM [67], Segway [68] and lamina-associated domains (LAD) [69] annotations, we used the SQL tables from UCSC. We used BED file downloads to obtain annotations for DNA repeat elements predicted by RepeatMasker [70], DNA repeat elements predicted by Tandem Repeats Finder [71], epigenome-based CpG island predictions produced by the Bock et al. software pipeline [72], and VISTA enhancer predictions [73].

Features extracted from Ensembl

We used the BioMart tool to download (i) TFBS motif occurrences (based on the 2014 release of the Jaspar database [74]) and ChromHMM chromatin segmentation labels from Ensembl Regulation 75 and (ii) GENCODE transcription start sites (TSS; from Ensembl Genes 75) with which we computed signed TSS distances.

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

Table 1 The 248 SNP features used in CERENKOV

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 [6]
`GERP++`	continuous	UCSC	bp-level GERP++ [80] score
`avg_GERP`	continuous	UCSC	avg. GERP score [81] in ±100 bp window
`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 [75]
`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 [70] DNA repeat
`simplerepeat`	categorical (2)	UCSC	SNP is in a Tandem Repeats Finder [71] repeat
`cpg_island`	categorical (2)	UCSC	SNP is in an epigenome-predicted CpG island [72]
`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 [74] 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`	160×categ. (2)	UCSC	160 features for SNP being in an ENCODE TFBS [84] peak
`FsuRepliSeq`	16×continuous	UCSC	Replication Timing by Repli-chip [66] from ENCODE/FSU
`UwRepliSeq`	16×continuous	UCSC	Replication Timing by Repli-seq [65] from ENCODE/UW

Table 1 The 248 SNP features used in CERENKOV (*Continued*)

Feature(s)	*Feature type*	*Raw data src.*	*Feature description*
`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 [85], Tig-3 cells)
`vistaEnhancerCnt`	categorical (2)	UCSC	count of VISTA [73] HMR-Conserved Non-coding Human Enhancers [86] overlap. SNP
`vistaEnhancerTotalScore`	categorical (2)	UCSC	sum scores of VISTA [73] HMR-Conserved Non-coding Human Enhancers [86]
`eigen`	continuous (2)	Eigen	Eigen & Eigen-PC v1.1 raw scorea [21]

Abbreviations are as follows: UCSC, UC Santa Cruz Genome Browser portal; 1KG, 1,000 Genomes Project; Ensembl75, Ensembl Release 75 [82]; GENCODE, the GENCODE project release 19 [83]; ENCODE, Encyclopedia of DNA Elements [30]; 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

muscle, tibial nerve, sun-exposed skin, stomach, thyroid, and whole blood) from the Genotype Tissue Expression (GTEx) Project [75] Analysis Version 4 from the GTEx project data portal. For each SNP, we selected the minimum association p-value across genes and tissues.

Computing the geometric features

For each distance metric $d(\cdot,\cdot)$, we first computed the intralocus radius $\lambda_{s|d}$ for each SNP s in our `OSU18` dataset, in the data-space of all the 248 features (categorical data were binary-encoded which expanded the dimension of the data space to 587); then we separated those intralocus radii according to SNP classes, making two sets $\Lambda_{r|d} = \{\lambda_{r|d} | r \text{ is an rSNP}\}$ and $\Lambda_{c|d} = \{\lambda_{c|d} | c \text{ is a cSNP}\}$. For empirical estimation of likelihoods, we used the R `hist` function with 11 bins on $\Lambda_{r|d}$ and $\Lambda_{c|d}$ and then gathered the bin counts, $\left\{N_{r|d}^{(1)}, \ldots, N_{r|d}^{(11)}\right\}$ and $\left\{N_{c|d}^{(1)}, \ldots, N_{c|d}^{(11)}\right\}$ respectively. The empirical likelihood ratio for bin i can be computed with formula $LR_d(i) = \frac{N_{r|d}^{(i)}}{N_{c|d}^{(i)}}$. For fitting parametric density distributions for intralocus radii, we used the `fitdistrplus` package (version 1.0.9) [76] in R and we used the normal distribution for cosine and Pearson distances and the log-normal distribution for the other eight combinations of distance function and data scaling/non-scaling. Akaike information criterion was leveraged (AIC) [77] for distribution selection. Then for each distance metric $d(\cdot,\cdot)$, 2 probability density functions, $p_{r|d}(\cdot)$ and $p_{c|d}(\cdot)$, can be estimated from $\Lambda_{r|d}$ and $\Lambda_{c|d}$. And for any given SNP s, its likelihood ratio is defined as the ratio of its probability densities, i.e. $LR_d(s) = \frac{p_{r|d}(\lambda_{s|d})}{p_{c|d}(\lambda_{s|d})}$. This likelihood can be interpreted as the extent to which SNP s inclines to be an rSNP, observing its intralocus radius.

For loci where only one SNP (rSNP in all cases) was located, we set its likelihood ratio to 1. For each of the `OSU18` SNPs, and using the parametric distributions fitted as described above, we computed log-likelihood-ratio scores for each of the ten combinations of distance metric and scaled/unscaled data listed in "Data-space geometric features for rSNP recognition" section. [The rationale for using min-max scaling for the data matrix for Canberra, Euclidean, and Manhattan distances was to reduce the impact of high-variance continuous features]. The ten columns of log-likelihood-ratio data were then appended to `OSU18` dataset as ten new features during our machine learning processes.

Machine learning

The feature extraction and distance computation were done in Python 3 under Ubuntu 16.04 and would take about two hours with a single core of an Intel Core i7-4790 CPU. Peak RAM usage was approximately 12 GB.

For the machine learning framework, we used the R statistical computing environment (version 3.4.4) [78], also under Ubuntu 16.04. The complete machine-learning process required 25 min in total for the three models (GWAVA, CERENKOV and CERENKOV2).

Random forest

In order to compare CERENKOV2 with GWAVA [24], we annotated `OSU18` dataset with the GWAVA program and then applied Random Forest algorithm to the gained GWAVA feature matrix. Specifically, we used the R package `ranger` [79] version 0.6.0 with the published hyperparameters. To make a fair comparison, we adapted the same cross-valiation settings and performance measurements to CERENKOV2's (see "Gradient boosted decision trees" section below). In addition, Random Forest is also applied to illustrate CERENKOV2 feature importances (see "CERENKOV2 feature importance" section).

Gradient boosted decision trees

For the gradient boosted decision trees (GBDT) classifier, we used the R API for `xgboost` [38]

version 0.6.4.1. 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 cross-validation [52] with *locus-based sampling*, 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-based sampling, an rSNP and its linked cSNPs are always assigned to the same cross-validation fold. Especially for those 10 geometric features, distribution parameters were estimated only on training data to prevent data leakage. For every prediction performance metric we report, the fold composition was exactly the same across all of the rSNP recognition models studied. We set `base_score` = 0.06918531 (the rSNP/cSNP class imbalance). We estimated 95% confidence intervals on the sample mean using 1,000 iterations of bootstrap resampling [52].

Hyperparameter tuning

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 AUPVR was: $\eta = 0.1$, $\gamma = 10$, `nrounds` = 30, `max_depth` = 7, `subsample` = 1.0 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 AVGRANK was: $\eta = 0.1$, $\gamma = 100$, `nrounds` = 30, `max_depth` = 6, `subsample` = 0.85, `colsample_bytree` = 0.85, and `scale_pos_weight` = 8).

GRASP database

We downloaded the full GRASP 2.0.0.0 catalog in tab-delimited value (TSV) format and joined the GRASP data with the CERENKOV2 prediction matrix using the dbSNP refSNP ID as the join key. We then filtered the resulting data matrix to include only SNPs whose GRASP trait-association *P*-values were less than the accepted human genome-wide significance level (5×10^{-8}) and whose CERENKOV rSNP prediction score was at least 0.7.

Additional file

Additional file 1: Supplementary Tables. This PDF file contains 2 supplementary tables. The first one provides a view of comparison of validation-set performance measures between GWAVA, CERENKOV and CERENKOV2 on the `OSU18` reference SNP set. The second one lists the skewnesses and kurtoses of intralocus radii computed using Canberra, Euclidean, Manhattan, cosine, and Pearson distances, applied to scaled and unscaled feature data, and conditioned on the type of reference SNP (rSNP or cSNP). (PDF 90 kb)

Abbreviations

AIC: Akaike information criterion; AUPVR: Area under precision-vs-recall curve; AUROC: Area under receiver operating characteristic curve; AVGRANK: Average ranks of the prediction scores for all ground-truth rSNPs within their loci; CERENKOV: Computational Elucidation of the REgulatory NonKOding Variome; CDS: Coding sequence; cSNP: Control SNP; DHS: DNase I hypersensitive site; ENCODE: Encyclopedia of DNA elements; GBDT: Gradient boosted decision tree; GERP: Genomic evolutionary rate profiling; GTEx: Genotype tissue-expression project; GWAS: Genome-wide association study; HGMD: Human gene mutation database; HMR: Human-mouse-rat; *k*-NN: *k*-nearest-neighbors; LAD: Lamina-associated domain; LLR: Log-likelihood ratio; MAF: Minor allele frequency; NKI: Netherlands cancer institute; QTL: Quantitative trait locus; rSNP: Regulatory SNP; SNAP: SNP annotation and proxy search; SNP: Single nucleotide polymorphisms; SVM: Support vector machine; TF: Transcription factor; TFBS: Transcription factor binding site

Acknowledgements

Not applicable.

Funding

This work was supported by the Medical Research Foundation of Oregon (New Investigator Award to SAR), Oregon State University (Health Sciences award to SAR), the PhRMA Foundation (Research Starter Grant in Informatics to SAR) and the National Science Foundation (awards 1557605-DMS and 1553728-DBI to SAR).

Availability of data and materials

The source code and instructinos for installing and running CERENKOV2 are available on GitHub at `https://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 are freely available online (and hyperlinked from the CERENKOV2 project README on GitHub).

Authors' contributions

Designed the study: SAR, YY, ZL; wrote the software: YY, ZL, SAR, QW; carried out the computational analyses: YY; wrote the manuscript: YY, SAR; edited the manuscript: YY, SAR, ZL, QW. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 5 July 2018 Accepted: 18 January 2019
Published online: 06 February 2019

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(Database issue):1001–6. accessed in 2016.
2. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22(9):1748–1759.
3. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–1195.
4. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367–83.
5. Li MJ, Yan B, Sham PC, Wang J. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. Brief Bioinforma. 2015;16(3):393–412.

6. Siepel A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
7. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLOS Genet. 2010;6(4):1000888.
8. Krawczak M, Cooper DN. The human gene mutation database. Trends Genet. 1997;13(3):121–2.
9. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics. 2006;22(5):637–40.
10. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):980–5.
11. Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Züchner S, Hauser MA. SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics. 2005;21(22):4181–186.
12. Macintyre G, Bailey J, Haviv I, Kowalczyk A. is-rSNP: a novel technique for *in silico* regulatory SNP detection. Bioinformatics. 2010;26(18):524–30.
13. Xiao R, Scott LJ. Detection of cis-acting regulatory SNPs using allelic expression data. Genet Epidemiol. 2011;35(6):515–25.
14. Riva A. Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. BMC Genomics. 2012;13 Suppl 4:7.
15. Li MJ, Wang LY, Xia Z, Sham PC, Wang J. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Res. 2013;41(Web Server issue):150–8.
16. Bryzgalov LO, Antontseva EV, Matveeva MY, Shilov AG, Kashina EV, Mordvinov VA, Merkulova TI. Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. PLoS ONE. 2013;8(10):78833.
17. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
18. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47(3):276–83.
19. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet. 2015;47(8):955–61. gkm-SVM.
20. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761–3.
21. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214–20.
22. Yao Y, Liu Z, Singh S, Wei Q, Ramsey SA. Cerenkov: Computational elucidation of the regulatory noncoding variome. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. Boston: ACM; 2017. p. 79–88.
23. Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJM. A survey of genomic properties for the detection of regulatory polymorphisms. PLoS Comput Biol. 2007;3(6):106.
24. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294–6.
25. Peterson TA, Mort M, Cooper DN, Radivojac P, Kann MG, Mooney SD. Regulatory Single-Nucleotide Variant Predictor Increases Predictive Performance of Functional Regulatory Variants. Hum Mutat. 2016;37(11): 1137–43.
26. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44(11):107.
27. Torkamani A, Schork NJ. Predicting functional regulatory polymorphisms. Bioinformatics. 2008;24(16):1787–92.
28. Zhao Y, Clark WT, Mort M, Cooper DN, Radivojac P, Mooney SD. Prediction of functional regulatory SNPs in monogenic and complex disease. Hum Mutat. 2011;32(10):1183–90.
29. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4.
30. The ENCODE Project Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. 2011;9(4):1001046.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
32. Andersen MC, Engström PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol. 2008;4(1):5.
33. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24(1):14–24.
34. Ryan NM, Morris SW, Porteous DJ, Taylor MS, Evans KL. SuRFing the genomics wave: an R package for prioritising SNPs by functionality. Genome Med. 2014;6(10):79.
35. Liu Z, Yao Y, Benjamin W, Wei Q, Ramsey SA. Res2s2am: Deep residual network-based model for identifying functional noncoding snps in trait-associated regions. In: Pacific Symposium on Biocomputing. Hawaii: World Scientific; 2019.
36. Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D. Learning a Prior on Regulatory Potential from eQTL Data. PLOS Genet. 2009;5(1):1000358.
37. Shin S, Keleş S. Annotation Regression for Genome-Wide Association Studies with an Application to Psychiatric Genomic Consortium Data. Stat Biosci. 2017;9(1):50–72.
38. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. arXiv.org. 2016;1603.02754:1–13.
39. Zhang J, Wang J, Huang H. Manifold learning for visualizing and analyzing high-dimensional data. IEEE Intell Syst. 2010;25:54–61. https://doi.org/10.1109/MIS.2010.8.
40. Francois D, Wertz V, Verleysen M. Choosing the Metric: A Simple Model Approach. In: Jankowski N, Duch W, Grabczewski K, editors. Meta-Learning in Computational Intelligence. Heidelberg: Springer; 2011. p. 97–115.
41. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
42. Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus. 2016;5(1):1304.
43. Bunescu RC, Mooney RJ. Multiple instance learning for sparse positive bags. In: Proceedings of the 24th Annual International Conference on Machine Learning (ICML-2007). Corvallis: ACM; 2007. p. 105–12.
44. Natarajan N. Learning with positive and unlabeled examples. 2015. PhD thesis, University of Texas at Austin.
45. Anderberg MR. Cluster analysis for applications. 1973. Technical report, Office of the Assistant for Study Support Kirtland AFB N MEX.
46. Pearson K. Note on regression and inheritance in the case of two parents. Proc R Soc Lond. 1895;58:240–2.
47. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19:491–504.
48. Lance GN, Williams WT. Mixed-data classificatory programs i - agglomerative systems. Aust Comput J. 1967;1(1):15–20.
49. Dunford N, Schwartz JT. Linear Operators Part I: General Theory. New York: Interscience; 1958.
50. Krause EF. Taxicab Geometry: An Adventure in non-Euclidean Geometry. Chelmsford: Courier Corporation; 1975.
51. Singhal A, et al. Modern information retrieval: A brief overview. IEEE Data Eng Bull. 2001;24(4):35–43.
52. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Selection and Model Estimation. In: Proc Int Joint Conf Artif Intel. San Francisco: ACM; 1995. p. 1137–43.
53. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10): 1340–7.
54. Breiman L, Friedman J, Olshen R, Stone CJ. Classification and Regression Trees. Abingdon: Taylor & Francis; 1984.
55. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. 2012;30(11):1095–106.
56. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas: ACM; 2008. p. 213–20.

57. Zhu X, Goldberg AB. Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn. 2009;3(1):1–130.
58. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
59. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. In: Cur Protoc Bioinformatics. Hoboken: Wiley; 2009.
60. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24(24):2938–939.
61. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
62. International HapMap Consortium. The International HapMap Project. Nature. 2003;426(6968):789–96.
63. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414): 75–82.
64. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22(9):1813.
65. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci. 2010;107(1):139–44.
66. Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schübeler D, Gilbert DM. Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. 2008;6(10):245.
67. Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215.
68. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9(5):473.
69. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453(7197):948.
70. Caballero J, Smit AFA, Hood L, Glusman G. Realistic artificial DNA sequences as negative controls for computational genomics. Nucleic Acids Res. 2014;42(12):99.
71. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
72. Bock C, Walter J, Paulsen M, Lengauer T. CpG Island Mapping by Epigenome Prediction. PLoS Computat Biol. 2007;3(6):110.
73. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35(Database issue):88–92.
74. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2009;38(Database):105–10.
75. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60.
76. Delignette-Muller ML, Dutang C. fitdistrplus: An R package for fitting distributions. J Stat Softw. 2015;64(4):1–34.
77. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. New York: Springer; 1998. p. 199–213.
78. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comp Graph Stat. 1995;5(3):299–314.
79. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. arXiv.org. 2015;1508.04409:1–17.
80. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol. 2010;6(12):1001025.
81. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7): 901–13.
82. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):662–9.
83. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:4–19.
84. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91–100.
85. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453(7197):948–51.
86. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. In vivo enhancer analysis of human conserved non-coding sequences. Nature. 2006;444(7118):499–502.