

IW-Scoring: an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome

Jun Wang^{1,*}, Abu Z. Dayem Ullah¹ and Claude Chelala^{1,2}

¹Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK and ²Centre for Computational Biology, Life Sciences Initiative, Queen Mary University of London, London, UK

Received September 01, 2017; Revised January 08, 2018; Editorial Decision January 18, 2018; Accepted January 24, 2018

ABSTRACT

The vast majority of germline and somatic variations occur in the noncoding part of the genome, only a small fraction of which are believed to be functional. From the tens of thousands of noncoding variations detectable in each genome, identifying and prioritizing driver candidates with putative functional significance is challenging. To address this, we implemented IW-Scoring, a new Integrative Weighted Scoring model to annotate and prioritise functionally relevant noncoding variations. We evaluate 11 scoring methods, and apply an unsupervised spectral approach for subsequent selective integration into two linear weighted functional scoring schemas for known and novel variations. IW-Scoring produces stable high-quality performance as the best predictors for three independent data sets. We demonstrate the robustness of IW-Scoring in identifying recurrent functional mutations in the *TERT* promoter, as well as disease SNPs in proximity to consensus motifs and with gene regulatory effects. Using follicular lymphoma as a paradigmatic cancer model, we apply IW-Scoring to locate 11 recurrently mutated noncoding regions in 14 follicular lymphoma genomes, and validate 9 of these regions in an extension cohort, including the promoter and enhancer regions of *PAX5*. Overall, IW-Scoring demonstrates greater versatility in identifying trait- and disease-associated noncoding variants. Scores from IW-Scoring as well as other methods are freely available from <http://www.snp-nexus.org/IW-Scoring/>.

INTRODUCTION

Over 98% of the human genome does not encode proteins. Despite its past reference as ‘junk DNA’ when first discov-

ered, noncoding sequences are now recognized as functionally important, possessing millions of regulatory elements and noncoding RNA genes. The annotation of potential regulatory sequences, through the Encyclopedia of DNA Elements (ENCODE) (1), Roadmap Epigenomics Consortium (2) and the FANTOM5 project (3,4), is revolutionizing our understanding of noncoding sequences and revealed that large stretches of the human genome (~80%) is evidently associated with DNA transcription to RNA, chromatin marks and other epigenomic elements. It has also established that many of the genetic variants associated with disease and diverse traits uncovered by genome-wide association studies (GWAS) are located within these noncoding regions, residing in or near ENCODE and Epigenome defined locations. However, compared to protein-coding regions, our understanding of noncoding regulatory elements remains poor.

Overall the significance of non-coding mutations remains an underexplored area of cancer genomics. With the rapid emergence of whole-genome and regulatory region targeted survey, many recurrent noncoding mutations have been identified in various cancer types. For example the prevalence of *TERT* promoter mutations has been established in melanoma (5,6), gliomas and a subset of tumours in tissues with low rates of self-renewal (7). Moreover, *TERT* promoter mutations significantly correlate with survival and disease recurrence in bladder cancer, demonstrating the clinical significance for noncoding mutations (8). In chronic lymphocytic leukaemia (CLL), whole genome sequencing (WGS) of 150 tumour/normal pairs alongside with DNase-seq and chromatin immunoprecipitation sequencing (ChIP-seq) identified recurrent mutations in the 3' UTR region of *NOTCH1* gene and in the active enhancer of *PAX5* (9). In a subset of T cell acute lymphoblastic leukaemia (T-ALL), somatic mutations were found in the intergenic region to create MYB-binding motifs, which resulted in a super-enhancer upstream of the *TALI* oncogene (10). Most recently, three significantly mutated promoters have been identified based on deep sequencing in 360 primary breast

*To whom correspondence should be addressed. Tel: +44 20 7882 8688; Fax: +44 20 7882 3884; Email: j.a.wang@qmul.ac.uk

cancers, and more such regions remain to be discovered (11). Whole-genome sequencing (WGS) data is increasingly available, especially through TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium), prompting more pan-cancer style studies to look for significantly mutated regulatory elements across cancer types (12–15).

To systematically study these noncoding variants and assess their functional/pathogenic potential, they first require careful annotation, by determining the regulatory regions they map to, e.g. ENCODE, Epigenome Roadmap and FANTOM5 defined elements, and their potential target genes. However, identifying potential noncoding pathogenic/driver/functional variants and distinguishing them from benign/passenger/non-functional variants, remains challenging, due to their abundance, cell/tissue type specificity and complex modes of action (16). In the last few years, many studies have taken an integrative approach of combining available noncoding annotation features to provide scores for the likely functional impact of noncoding variants. Most of these methods, including CADD (17), GWAVA (18), FATHMM-MKL (19) and Genomiser (20), used machine-learning algorithms to develop classifiers integrating a range of annotations such as regulatory features, conservation metrics, genic context and genome-wide properties to differentiate disease-associated/deleterious variants from benign/neutral variants in the model training. Some other methods, such as DeepSEA (21) and DeltaSVM (22), chose to directly learn regulatory sequence codes from large-scale chromatin-profiling data generated from ENCODE, while FitCons (23) opted to estimate the selective pressure for those functionally important genomic regions on the basis of patterns of polymorphism and divergence. Additional methods, like FunSeq2 (24) and Eigen (25), developed a weighted scoring system to combine the relative importance of various annotation features to separate functional and non-functional variants. The advantage of these weighted scoring methods is that they do not rely on any labelled gold-standard training data of disease-associated and putatively benign variants, which are often inaccurate and impractical for the extent of variants linked with GWAS and cancer studies.

Most of the methods discussed above simply provide a continuous functional score for each variant without further information on the prediction accuracy and functional consequence in affected regulatory elements. A level of confidence or significance is also needed for each estimated score as these methods often score the same variants differently with varied performances across different data types. Thus, a class ensemble approach that combines the predictions of all these functional methods with a weighted scheme would offer a powerful approach to summarise multiple predictive evidences, increase specificity and rank variants.

Here, we explore eleven most commonly used functional annotation methods and scoring systems for genetic variations in the noncoding genome. Considering the improved robustness and generalisability of a class ensemble approach compared to a single scoring method (26), we then propose a new integrative weighted scoring method, named IW-Scoring. Using an unsupervised spectral ap-

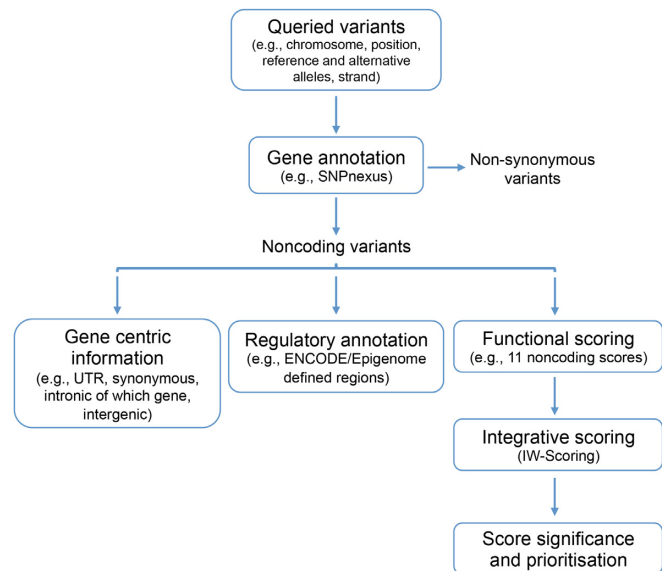


Figure 1. The IW-Scoring framework.

proach based on feature covariance, IW-Scoring selectively amalgamates some of those eleven methods (i.e. full set or subsets) into a single ‘ensemble-like’ algorithm and delivers two separate linear weighted functional scoring schemas for known and novel variations, respectively. Across independent test data sets, IW-scoring produces consistently high-quality stable performance in differentiating functionally significant variations from others, while the constituent methods show mixed performances at best. Finally, we demonstrate the utility of IW-Scoring in identifying and prioritizing functional variants in GWAS, expression quantitative trait locus (eQTL) and cancer studies.

MATERIALS AND METHODS

Overview of IW-Scoring framework and data resources

The workflow in IW-Scoring consists of four top-level modules: (i) gene annotation, (ii) regulatory annotation, (iii) functional scoring and (iv) score integration and significance inference (Figure 1). With queried variants as input, IW-Scoring framework uses the four modules to annotate and score variants, and rank noncoding ones based on their predicted functional significance. Within the gene annotation module, the queried variants are first annotated against Ensembl (version 75, GRCh37/hg19 genome assembly) gene annotation to identify and filter out any variants that can potentially lead to non-synonymous changes for any transcript of a gene using SNPnexus (27). A gene centric view/summary is provided for filtered noncoding variants within genes (e.g. synonymous, UTR and intronic), within 1 kb up- or downstream of genes, or in intergenic regions. For intergenic variants, we also collect their nearest up- and downstream genes and distances to them.

The next step is to further annotate noncoding variants against ENCODE/Epigenome/FANTOM5 defined regions along with Ensembl Regulatory Build annotation to identify overlapping regulatory elements of var-

ious chromatin (e.g. DNase I), polymerase and histone (e.g. H3K4me1/2/3, H3K27ac, H3K36 and H3K9) marks, transcription factor (TF) binding sites (e.g. CTCF, FOXA1, NFKB, c-Myc, c-Jun, p300, etc.) and predicted promoters/enhancers/TSS, with the supporting cell and tissue types also reported.

Next, noncoding functional scores of 11 different scoring systems derived from eight studies, including CADD v1.3 (17), DeepSEA (21), Eigen (Eigen and Eigen-PC) (25), fit-Cons (23), FunSeq2 (24), FATHMM-MKL (19), GWAFA (region, TSS and unmatched scores) (18) and ReMM (20) (Supplemental Table S1), are extracted for all noncoding variants allowing missing values for some scores. The pre-computed genome-wide scores are used for all systems, except DeepSEA for which the standalone version of the tool is executed locally.

Finally, an integrative score, i.e. IW-Scoring, is calculated for each variant as the linear weighted sum of the functional scores obtained from different systems (see below for details), with associated statistical significance also derived.

Assembly of training variant set

For the training set, we extracted all variants from the 1000 Genomes Project data set (28) that were not present in dbNSFP (29) v3.0, and were either (i) within 1 kb upstream of the gene start site and 1 kb downstream of the gene end site, or (ii) within 5' and 3' UTR of a gene or (iii) synonymous variants in coding regions, leading to a total of 712 259 noncoding variants.

Construction of integrative noncoding scores

We used an unsupervised spectral approach similar to Eigen/Eigen-PC (25) to derive a weighted linear combination of individual scores. To achieve this, we first need to learn the weights for different functional scoring systems to be incorporated in our scoring system, and then develop a strategy to calculate integrative scores and associated significance for query variants considering the missing values and data rescaling.

Estimation of the weight for constituent scores. The weight estimation procedure was implemented as following:

- (1) *Training data set of functional scores:* for the training data of 712 259 noncoding variants, we obtained functional scores from the 11 scoring systems, allowing missing values for certain scores. DeepSEA functional significance scores (P -values) were $-\log_2$ transformed to enable data integration.
- (2) *Data rescaling:* noncoding scores (continuous variables) were rescaled to have a mean of zero and a variance of one, individually. The minimum and maximum for the original and rescaled values of each scoring system were retained for the data transformation of queried data sets (Supplemental Table S1).
- (3) *Covariance estimation:* a covariance matrix was calculated as pairwise correlations between any two of the 11 scoring systems. This allows certain scores to be used for variants with missing values when scores from

other methods are available. This ensures we estimate the weights based on all observed variants accurately.

- (4) *Weight estimation via eigendecomposition:* the estimated weights for the individual scoring systems were calculated using the eigendecomposition of the covariance matrix. The lead eigenvector (i.e. the one with the greatest eigenvalue) was used to assign the weights to the corresponding scoring systems.

In order to identify whether the calculated weight values are sensitive to the training data, we conducted the same procedure with subsets of the original training dataset, comprising of random selection of 20%, 40%, 60% and 80% of the variants respectively. In addition, we conducted the procedure multiple times with different random selection of 20% of the variants. In all cases, the changes in derived weight values for different training datasets were insignificant, suggesting that the weights are not sensitive to the training data set (Supplementary Figure S1).

We then calculated IW-scores, a weighted sum of rescaled values of different noncoding scores, for all training variants, in order to determine a distribution of integrative scores. We imputed missing values in the rescaled scoring matrix using the Amelia R package (30). Amelia performs multiple imputation of missing data taking into account potential interdependencies between variables, combining the expectation-maximization with bootstrapping algorithm. The final imputed values were calculated as a mean of 10 imputations.

Calculating integrative scores and associated significance for queried variant sets. The IW-Scoring framework to calculate integrative scores for any set of queried variants is outlined below:

- (1) *Score extraction and missing value imputation:* similar to the training data processing, we first extract scores from different methods. For variants with missing values from certain scoring systems, imputations are enforced. Considering the usual small sample size for queried variant sets, we employ a strategy to merge them with a set of 100 000 variants randomly selected from the training set where all values are available, in order to increase the imputation accuracy. We then impute the missing values using Amelia based on the average of 10 imputations.
- (2) *Data rescaling:* we use the training data set as a reference for all queried variants to be rescaled to. For each scoring system, we rescale the values of queried variants to fit into the rescaled distribution of the training set based on the parameters (e.g. the minimum and maximum values) derived from the original and rescaled scores. The rescaled values need to satisfy the following equation,

$$\frac{Max_{Ai} - m_i}{m_i - Min_{Ai}} = \frac{Max_{Ri} - x_i}{x_i - Min_{Ri}} \quad (1)$$

where Max_{Ai} and Min_{Ai} , Max_{Ri} and Min_{Ri} are the maximum and minimum values of the original and rescaled scores for scoring system i in the training set, respectively. Variables m_i and x_i are the original and rescaled values of scoring system i for the queried variant. Thus, the solu-

tion to this equation of calculating x_i is

$$x_i = \frac{Max_{Ri} \times (m_i - Min_{Ai}) + Min_{Ri} \times (Max_{Ai} - m_i)}{Max_{Ai} - Min_{Ai}}. \quad (2)$$

- (3) *Integrative scoring*: after rescaling, the integrative scores are calculated as the weighted sum of rescaled values of all scoring systems for queried variants, as

$$IW_score = \sum_i W_i x_i \quad (3)$$

where W_i is the weight of scoring system i derived from the training set.

- (4) *Scoring significance and variant ranking*: the integrative scores for queried variants are compared to the lognormal distribution of all 712 259 scores in the training set to determine the significance levels. Queried variants are further ranked based on IW-scores and P -values derived.

Performance comparison between IW-scores and other scores

We used three independent data sets to evaluate the variant differentiating (functional/deleterious from non-functional/benign variants) performances between IW-scoring and other scores. These data sets included non-coding variants from the ClinVar database (31) version 2016/11/01 and National Human Genome Research Institute (NHGRI)-EBI GWAS Catalog (<http://www.ebi.ac.uk/gwas/home>) version 2016/11/21. For cancer data set, we selected noncoding variants curated in the COSMIC database (32) Version 79. The three test data sets chosen are 'gold-standard' for benchmarking variant functional scores and annotation. They have been widely used in many studies, including those featuring CADD, Eigen, GWAVA, DeepSea and FATHMM methods. The performance of each scoring method differentiating functional from non-functional variants was assessed using the average receiver operating characteristic (ROC) curves and the area under curve (AUC) using pROC R package (33).

Software availability

The workflow for IW-Scoring described above has been implemented as a web-based tool that allows users to annotate any list of known and novel variants against various non-coding annotation and scoring databases, as well as to calculate IW-scores for them. This integrated single web tool is freely available for use at <http://snp-nexus.org/IW-Scoring>.

RESULTS

IW-Scoring for noncoding variants in training data set

We first assessed individual functional scores and their correlation/covariance structure using 712 259 training noncoding variants. Correlations among different functional scores for the training variants showed that scores from Eigen, DeepSEA, FATHMM noncoding, ReMM and CADD were more closely correlated, whereas GWAVA scores (i.e. unmatched and TSS) were more similar to each other (Figure 2A). Under the assumption of conditional independence among individual functional scores given the true state of a variant (either functional or non-functional),

this correlation/covariance structure could be used to determine the weight for each scoring system when combined to differentiate variants. We are aware of the limitation of our assumption, as many functional scores tend to use similar information for prediction, thus likely to be correlated when scoring variants. However, our inclusion of a wide range of functional scores based on different subsets of regulatory annotations and different algorithms (either machine-learning classifier, or sequence pattern recognition, or sequence conservation or weighted summarizing score) minimises the inter-dependency effect and ensure the functional scoring based on the most diverse evidences. It is worth noting that functional scores were available for >95% of training variants for all methods, except for FunSeq2 (84%), Eigen-PC (87%) and Eigen (91%), partially due to the missing values for chromosome X variants for Eigen (Supplementary Table S1).

Using the eigendecomposition of the covariant matrix derived from the training set, we calculated the weights for different scores (Supplementary Table S1). Eigen, DeepSEA, FATHMM noncoding, ReMM and CADD had the highest weights, while fitCons, GWAVA TSS and Eigen-PC had the lowest weights, half of those for Eigen and DeepSEA. Our impression is that evolution and conservation based measures, like fitCons, may not be appropriate choice to score functional noncoding variants, as the regulatory elements are often lineage/species-specific (34). We then computed the integrative scores for all training variants as the weighted sum of different functional scores (rescaled values) where the scores appeared to follow a lognormal distribution (mean of 0.028 and standard deviation of 2.211) with a long tail in the direction of positive values (Figure 2B). This distribution could then be used to infer the statistical significance for all queried variants. We also compared the integrative scores amongst different feature types using the Ensembl Regulatory Build annotation, which revealed that variants in promoter regions had the highest scores, followed by those in promoter flanking regions and enhancers (Supplementary Figure S2).

Benchmark of IW-Scoring and comparison with other methods

Using three independent validation sets, we assessed the performance of IW-Scoring against other methods in differentiating pathogenic/functional noncoding variants from benign/non-functional ones.

ClinVar pathogenic and benign noncoding variants. We selected noncoding single nucleotide variants (SNV) from the ClinVar database, including 3'/5' UTR, intergenic upstream / downstream, intronic, and synonymous coding variants, resulting in in total 769 pathogenic SNVs (true positives) and 11 173 benign SNVs as control (true negatives). As shown in the average receiver operating characteristic (ROC) curves when comparing all 11 individual functional scores (Figure 3A), ReMM, Eigen and FATHMM noncoding gave the best performances in differentiating pathogenic from benign variants with the area under the curve (AUC) values ranging 0.76–0.78 (Supplementary Table S2). The performances of DeepSEA and

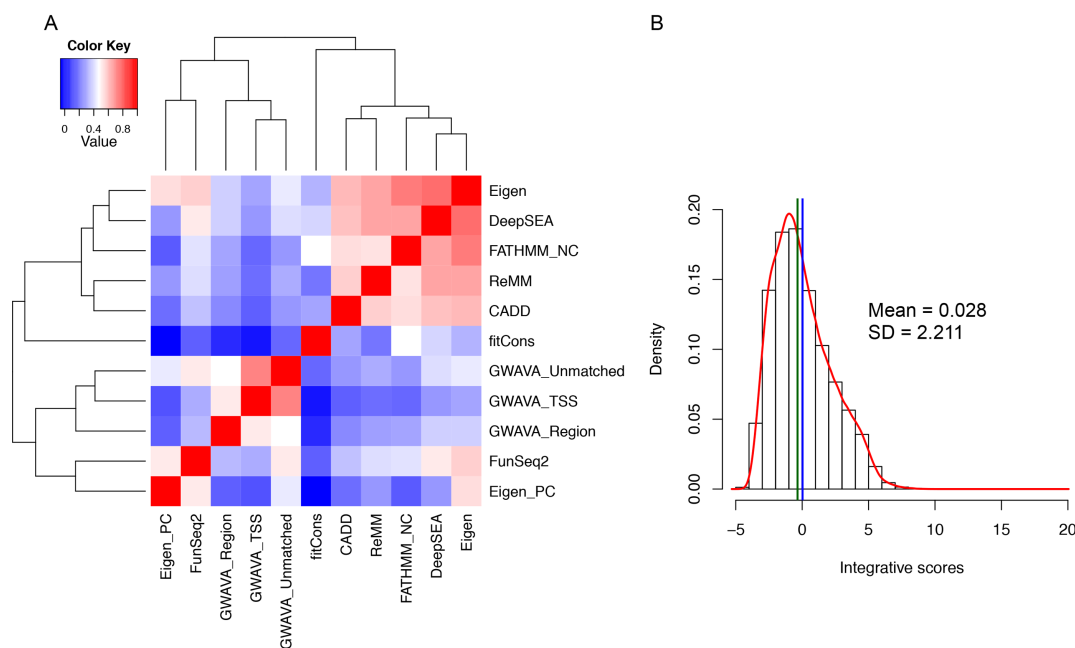


Figure 2. Correlations between 11 functional scores and the distribution of integrative scores of 712 259 noncoding variants in the training set. (A) Correlation matrix of 11 functional scores based on variants in the training set. (B) Distribution of integrative scores based on 11 scoring systems for all variants in the training set. The mean and standard deviation were shown. Mean and median values were indicated by the blue and green lines, respectively.

CADD were the next best with AUC of 0.72, whereas fitCons and GWAVA unmatched classifiers performed poorly (AUC = 0.58 and 0.62, respectively). We also performed the Wilcoxon rank-sum tests to compare functional scores between pathogenic and benign noncoding variants for each scoring system. Although the overall results were highly significant for all methods, P values from the tests were the most significant for ReMM ($P = 1.5 \times 10^{-149}$), FATHMM noncoding ($P = 2.9 \times 10^{-121}$) and Eigen ($P = 2.5 \times 10^{-114}$), while P values for fitCons ($P = 6.0 \times 10^{-11}$) and GWAVA unmatched ($P = 7.3 \times 10^{-13}$) were the least significant, consistent with the AUC results.

We next calculated IW-Scoring values for all ClinVar pathogenic and benign noncoding variants. The AUC for the integrative scores was 0.78 (95% CI: 0.77–0.80), marginally higher than that of the best performing score above, ReMM (Figure 3B). P value from the Wilcoxon rank-sum test was also the lowest for the IW-scores ($P = 1.1 \times 10^{-151}$). We also simply added up all 11 individual scores as the ‘sum-up’ scores, and the AUC for this was 0.78 (95% CI: 0.76–0.79), almost on a par with the weighted scores. As fitCons performed the worst out of all methods, we excluded fitCons from the training set and recalculated weights for all the remaining scores, followed by the calculation of IW-scores for all selected ClinVar noncoding variants. The AUC for this set of integrative scores to distinguish pathogenic from benign variants was 0.80 (95% CI: 0.78–0.81), the highest of all scores (Figure 3B). This best performance was also supported by the most significant P value derived from the Wilcoxon rank-sum test ($P = 1.5 \times 10^{-165}$) (Figure 3C; Supplementary Table S2). Furthermore, we also computed the significance P -values by comparing IW-scores with the lognormal distribution of those derived from the training set for selected Clin-

Var pathogenic and benign noncoding variants. Overall, P -values for pathogenic variants were much more significant than those for benign variants (combined P -value using Wilcoxon’s method (15), $P = 0.0003$ for pathogenic variants and $P = 0.15$ for benign variants) (Figure 3D). Within a queried variant set, IW-scores and associated P -values could be further used to rank and prioritise variants. It is worth noting that the performance of IW-Scoring did not improve when we excluded other methods with poor performances (e.g. GWAVA unmatched) or just used the best performing methods only (e.g. ReMM, Eigen and FATHMM noncoding). As shown in the distribution of scores across all tools for pathogenic and benign variants (Supplementary Figure S3), IW-score (excluding fitCons) is close to taking the maximum of individual scores, increasing the specificity greatly. It is also important to note that IW-Scoring, along with ReMM and DeepSEA, resolved 100% of the selected ClinVar variants, with CADD and FATHMM noncoding scoring 98% variants, while three methods had scores for <90% selected variants, GWAVA (89%), Eigen-PC (86%) and FunSeq2 (only 55%) (Supplementary Table S2).

We also benchmarked variant distinguishing performances between IW-Scoring and LINSIGHT (35), a very recent noncoding variant scoring method based on the existing INSIGHT-fitCons evolutionary conservation framework but with vastly improved prediction power. Our results show that IW-Scoring excluding fitCons produced marginally better results than LINSIGHT (AUC = 0.79) (Figure 3B), but LINSIGHT only resolved 55% of all tested ClinVar noncoding variants (Supplementary Table S2).

GWAS significant noncoding SNPs. The second validation test data set we used was the noncoding trait-associated SNPs from the National Human Genome Research Insti-

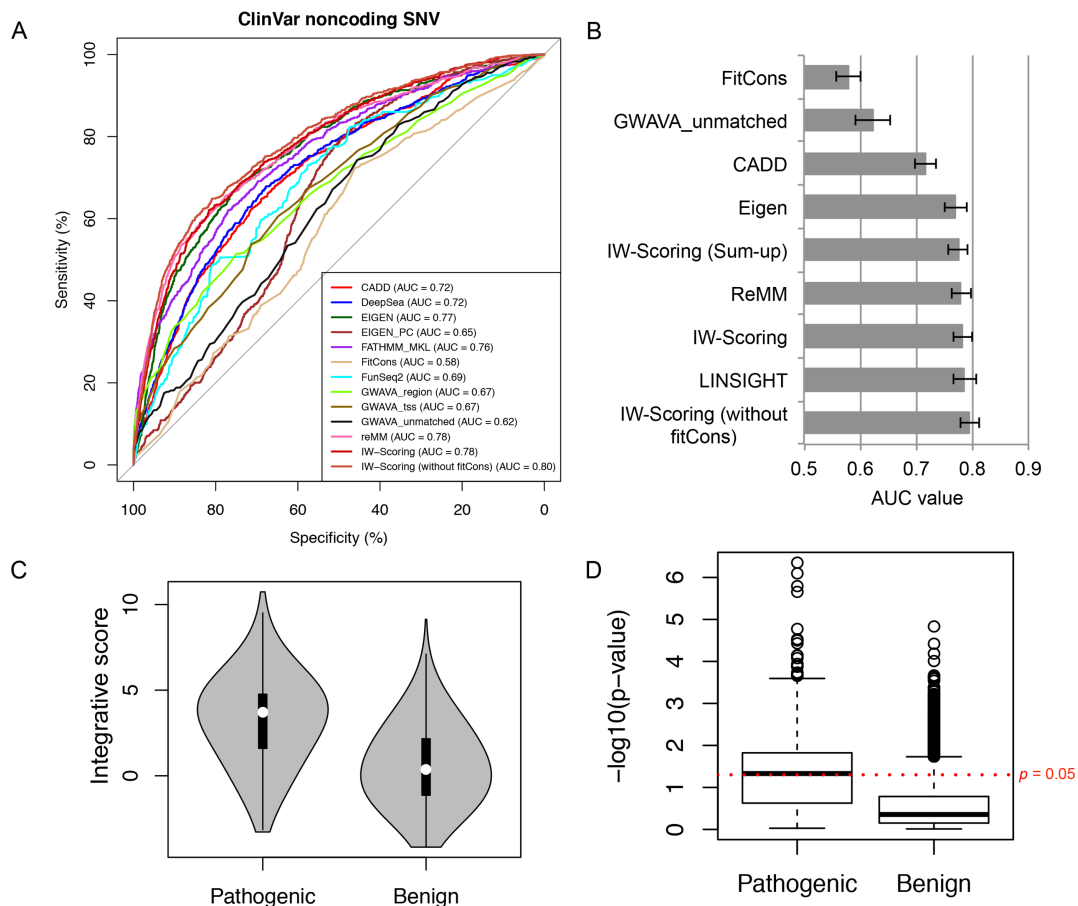


Figure 3. Comparison of the performances of different functional scores with IW-Scoring using the ClinVar pathogenic and benign noncoding variants. (A) ROC curves of IW-Scoring and 11 individual functional scores with AUC values noted, measuring the accuracy to differentiate pathogenic and benign variants. (B) AUC values for IW-Scoring and selected individual scores to differentiate pathogenic and benign variants. 95% CI bars were also shown. (C) Violin plots of integrative scores (without fitCons) for ClinVar pathogenic and benign variants. (D) Boxplot of *P*-values associated with integrative scores (without fitCons) for pathogenic and benign variants.

tute (NHGRI)-EBI GWAS Catalog. We identified 19,797 GWAS significant noncoding SNPs, consisting of intergenic, intronic, synonymous and regulatory region variants. For negative ‘non-functional’ variants, we randomly selected more than twice the number of 1000 Genomes intergenic and intronic SNPs with minor allele frequency (MAF) > 5% in the population ($n = 45\,808$). In general, the performances of all methods were much poorer for this data set when compared to ClinVar data set. This is probably because most of these GWAS significant SNPs are not truly causal, but in linkage disequilibrium (LD) with the genuine causal ones, which consisted of only 5% of all GWAS catalogued SNPs (36). The AUCs for IW-Scoring and GWAVA unmatched classifiers were very comparable (0.591 and 0.595, respectively), producing the best results of all (Figure 4A). Eigen-PC, DeepSEA, FunSeq2, GWAVA TSS and Eigen were the next best (AUC range 0.580–0.589), while ReMM and CADD performed relatively poorly (AUC, 0.533 and 0.536, respectively). FitCons appeared to be inadequate in handling this set of SNP data with AUC below 0.500 (Figure 4A; Supplementary Table S3). Thus, we excluded fitCons from the integrative scores. In contrast with the ClinVar analysis, this exclu-

sion did not improve the discriminating power of the integrative scores (AUC = 0.590) (Figure 4A). When comparing the *P* values from the Wilcoxon rank-sum test between GWAS significant and non-functional variants, the integrative scores led to the highest significance ($P = 3.7 \times 10^{-301}$) along with GWAVA unmatched, followed by Eigen-PC, DeepSEA, FunSeq2, GWAVA TSS and Eigen (Supplementary Table S3). Again, IW-Scoring was informative for all selected 65 605 variants, with ReMM, DeepSEA, FATHMM noncoding, FunSeq2 and CADD all scoring >95% variants. Eigen-PC was the only score system that had scores for <90% variants (89%) (Supplementary Table S3). Note that AUC for LINSIGHT (0.575) was significantly lower than that for IW-Scoring, and higher than that for FATHMM noncoding (Figure 4A). Different from the results for ClinVar variants, LINSIGHT managed to resolve over 99% of all tested variants in this set.

Noncoding cancer mutations from COSMIC. Next, we compared the different scoring systems to determine their ability to differentiate potentially functional from non-functional noncoding cancer mutations in the COSMIC database (32). As GWAVA only scores known SNPs, we

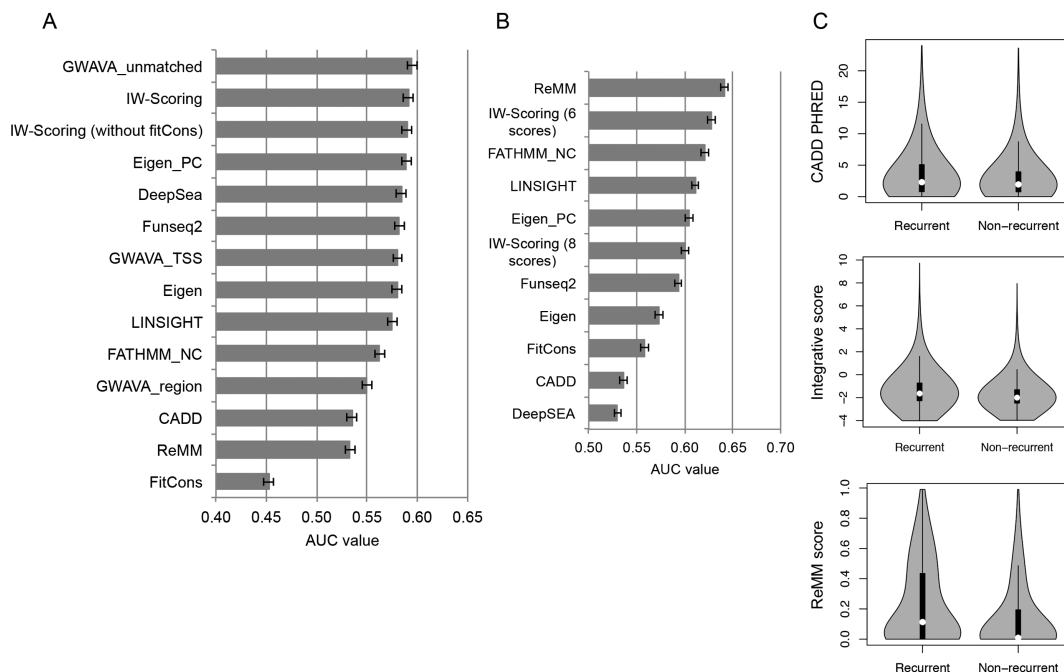


Figure 4. Comparison of the performances of different functional scores using the GWAS and COSMIC noncoding data sets. (A) AUC values for IW-Scoring and individual scores in the ability to differentiate between GWAS and randomly selected 1000G noncoding SNPs. 95% CI bars were also shown. (B) AUC values for IW-Scoring and individual scores in the ability to differentiate COSMIC recurrent from non-recurrent noncoding mutations. (C) Violin plots of functional scores between recurrent and non-recurrent noncoding mutations for CADD, integrative scores (six scores) and ReMM.

omitted all three GWAVA scores from this test. Since the large-scale discovery of true functional noncoding mutations is still very much lacking, for potential ‘functional’ ones, we identified all COSMIC noncoding mutations that had been identified in more than two samples and were also absent or present in <1% of samples in the 1000 Genome Project and NHLBI GO Exome Sequencing Project (ESP) data sets, restricting our analysis to 68 969 noncoding mutations. We further limited this set of mutations to those within the annotated regulatory regions only from ENCODE, Epigenome Roadmap and Ensembl Regulatory Build, leading to 34,813 potentially functional mutations as true positives. For the control set, we randomly selected ~4 million COSMIC non-recurrent noncoding mutations, and identified those with MAF larger than 1% in the 1000 Genome data set. We further excluded mutations within the annotated regulatory regions, leading to 57 866 noncoding mutations as the non-functional control. The AUC was the highest for ReMM (0.64), followed by FATHMM noncoding (0.62), Eigen-PC (0.60) and IW-Scoring of eight scoring systems (0.60) (Figure 4B; Supplementary Table S4). CADD and DeepSEA appeared to perform the poorest in differentiating the selected recurrent from non-recurrent noncoding mutations, with AUC < 0.55. Thus, we excluded these two functional scores from IW-Scoring, and the new integrative scores based on the six remaining scoring systems significantly improved the discriminating power with an AUC of 0.63, which was only marginally lower than that for the best performing method, ReMM (Figure 4B; Supplementary Table S4). In general, recurrent noncoding mutations within the regulatory regions had significantly higher functional scores on average than their non-

recurrent counterpart for all scoring systems. The extent of this difference was the highest for ReMM and IW-Scoring compared to others (Figure 4C; Supplementary Table S4). Again, Eigen and Eigen-PC resolved the lowest number of selected COSMIC variants. Note that IW-Scoring combining six functional scores also outperformed evolutionary conservation LINSIGHT framework (AUC = 0.61) in differentiating recurrent from non-recurrent noncoding mutations (Figure 4B; Supplementary Table S4).

In summary of the benchmark results across three data sets, the performance of IW-Scoring was stable and ranked consistently among the best performing methods. However, the performances of other methods appeared to vary greatly among data types (Supplementary Table S5).

IW-Scoring usage and workflows

To further assist end-users to apply IW-Scoring to annotate and score their queried variants, here we discuss the standard usage of our framework. We have also developed a useful single web portal where scores from IW-Scoring and all other methods were available for query. IW-Scoring contains separate workflows for scoring known and novel variants. Furthermore, based on the detailed evaluation of IW-Scoring performance on the test data sets (Figures 3 and 4), each workflow is designed to provide two sets of scores. For known variants, we provide two IW-scores with the associated *P*-values: IW-score (K11) that aggregates scores from 11 functional scoring systems including fitCons, and IW-score (K10) that excludes fitCons scores. This approach is useful for ranking and narrowing down variants indicated in GWAS and QTL studies, as well as rare known vari-

ants in hereditary diseases. As GWAVA only scores known variants, the workflow for novel variants excludes GWAVA scores from the aggregate calculation and provides two integrative scores: IW-score (N8) that aggregates scores from the other eight scoring systems, and IW-score (N6) that further excludes CADD and DeepSEA scores. This workflow is preferred for variants identified in cancer and other somatic diseases. In situations where users do not know which options to choose, they can use both workflows to generate IW-scores. The percentage of queried variants that are scored by GWAVA can further guide users in interpreting which IW-scores would be more reliable.

Application of IW-Scoring to study noncoding variants from human data sets

Having assessed the performance of IW-Scoring and other available functional scores in a range of test data sets, next we aimed to use our scoring system to evaluate key noncoding variants derived from association mapping, eQTL and cancer studies.

Disease SNPs near consensus motifs. We first assessed the functional significance for known disease-associated variants that fall near consensus motifs of transcription factor binding sites. We used candidate causal variants generated from Farh *et al.* (36), where transcription and *cis*-regulatory elements annotations for primary immune cells generated from Epigenome Roadmap Project were used and integrated. In this study, Farh *et al.* investigated the effects of disease SNPs in altering TF binding, and identified a notable AP-1 binding motif-disrupting SNP rs17293632 associated with Crohn's disease. We extracted all nearby known variants, within 5 kb up- and downstream of rs17293632 ($n = 187$), and calculated the IW-scores for them. Our results show that IW-score (K10) for the causal SNP rs17293632 is the highest (score = 9.48, $P = 7.32 \times 10^{-13}$) among all the SNPs investigated (Figure 5A). rs17293632 is located within an intron of *SMAD3*, a region enriched for H3K4me1, H3K27ac, DNase I and TF ChIP-seq peaks, also with a higher degree of sequence conservation (Figure 5A). The assessment of IW-scores further highlighted three additional SNPs with functional potential in this region, rs28514342 (5.47, $P = 0.006$), rs12324077 (4.98, $P = 0.012$) and rs193193326 (4.50, $P = 0.020$), which were all within 5 kb from the disease-associated rs17293632, thus likely to be in the same linkage disequilibrium (LD) block.

We also compared the performances of IW-Scoring and other methods in differentiating rs17293632 from nearby non-associated SNPs (Supplementary Figure S4). Although the majority of the methods prioritized rs17293632 as the highest causal SNP, the average difference in functional impact scores between rs17293632 and all nearby background SNPs was maximized in IW-scores (K11 and K10), along with DeepSEA and FunSeq2 (Figure 5A). The differentiating performances and scoring differences of GWAVA (all three scores) along with ReMM and fitCons were poor for this case. FitCons and GWAVA unmatched scores failed to produce significant differences between rs17293632 and the nearby known SNPs (Figure 5A).

Disease SNPs with gene regulatory effects. We also assessed the functional significance for disease SNPs with gene regulatory effects. Combining epigenome data and a data set mapping variants in peripheral blood gene expression, Farh *et al.* identified two eQTL SNPs in the *IKZF3* locus with independent effects on *IKZF3* expression. Based on the GWAS and eQTL significance, rs12946510 is associated with decreased *IKZF3* expression and increased multiple sclerosis (MS) risk, while rs907091 is associated with increased expression but with no association with MS risk. Interestingly, IW-score (K10) for the causal SNP rs12946510 is high (5.35, $P = 0.007$), whereas the IW-score for the eQTL SNP rs907091 with no disease effect is not significant (-0.105 , $P = 0.525$), in line with the scores for all nearby non-functional SNPs (Figure 5B). These results indicated that IW-scores are probably more sensitive to disease causative effects linked with phenotypic differences, rather than gene expression differences of nearby targeted genes. Therefore, by combining IW-scores and gene expression differences, we could further pinpoint those SNPs with both disease association and gene regulatory effects. *IKZF3* 3' downstream region is enriched for H3K27ac, DNase I peaks and TF binding sites; therefore, many variants in this region appeared to have higher IW-scores compared to those in other regions (Figure 5B).

When comparing the differentiating results between IW-Scoring and all other methods for these two SNPs, IW-scores (K11 and K10) were the best performing ones, producing the largest difference in functional scores between causal and non-causal variants, and clearly prioritizing the disease associated SNP rs12946510 (Figure 5B). Although other methods (e.g. Eigen-PC and Eigen) also prioritized the causal SNP correctly, the score difference between causal and non-causal SNPs was much smaller compared to that of IW-scores. In particular, fitCons, GWAVA TSS, FunSeq2, GWAVA unmatched and region scores all failed to differentiate and prioritise rs12946510 from the non-associated counterpart rs907091 (Figure 5B).

Recurrent functional noncoding mutations in cancer. We assessed the functional significance of recurrent noncoding mutations identified in cancer, using *TERT* promoter mutations as the representative example. First, we identified somatic mutations curated in COSMIC and Huang *et al.* (37), as well as the known SNPs in this region, including ~15 kb up- and downstream flanking regions. In order to generate comparable scores between somatic mutations and known SNPs, we used the IW-Scoring novel variant workflow for both. Intriguingly, higher IW-scores were observed around the promoter region marked by strong DNase I signal and conserved TF binding sites, and gradually decreased to non-functional levels moving away from the promoter (Figure 5C). For the two most frequently recurrent mutations, chr5:1 295 228 C>T and chr5:1 295 250 C>T (reverse strand, hg19), the integrative scores IW-score (N8) were 4.25 ($P = 0.021$) and 4.09 ($P = 0.025$), respectively. The scores for other less frequent mutations, chr5:1 295 228 C>A and chr5:1 295 161 A>C (reverse strand), were 3.12 ($P = 0.068$) and 3.47 ($P = 0.048$). Encouragingly, this provided further confirmation of the accuracy and validity of our integrative approach for somatic variants. Interest-

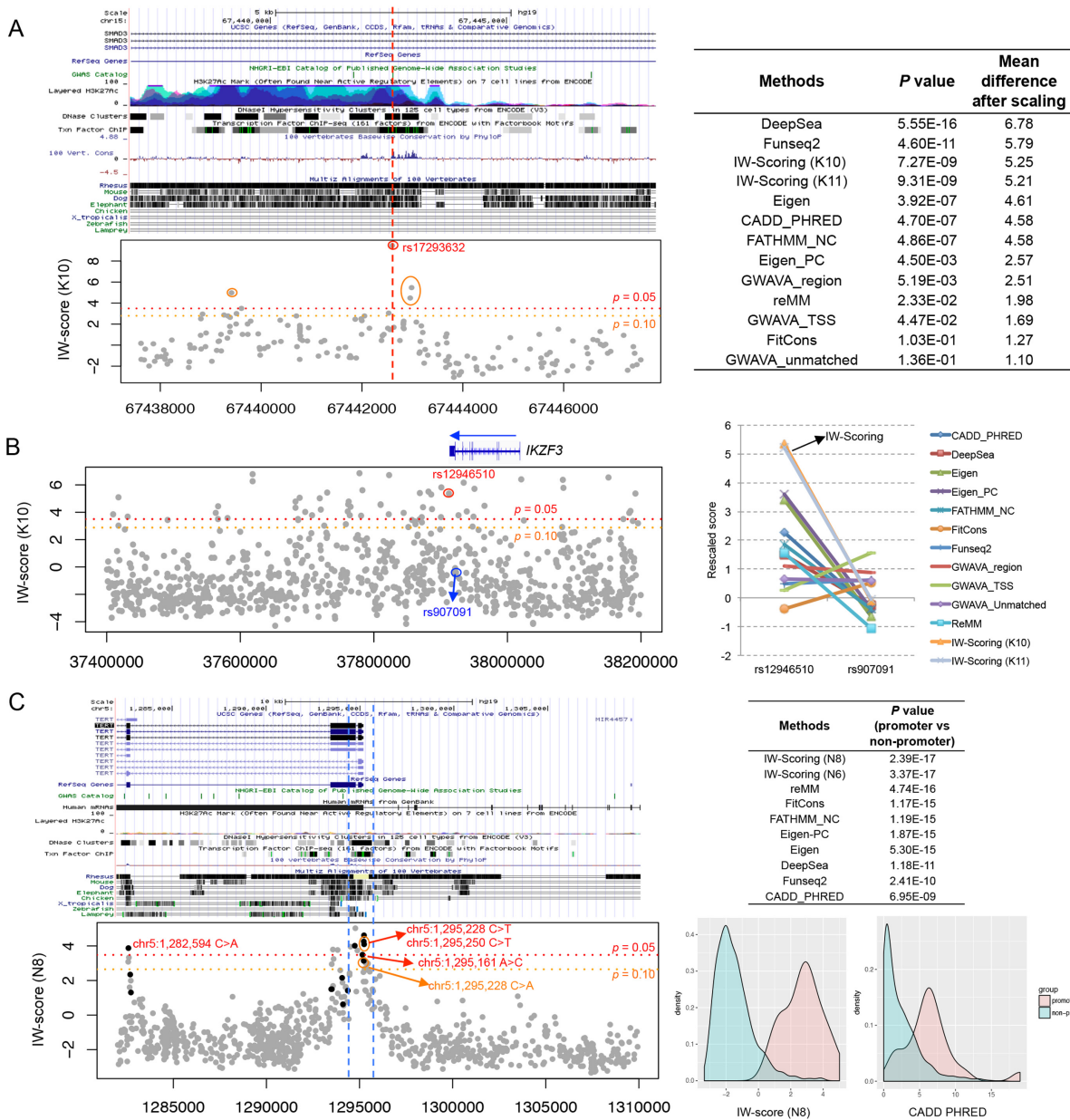


Figure 5. Application of IW-Scoring in scoring variants identified in GWAS, eQTL and cancer studies. (A) Disease SNPs near consensus motifs in *SMAD3* intronic region. Integrative scores (IW-score K10) were shown for all nearby known SNPs along with the disease-associated candidates. The score of rs17293632 was compared to the mean value and standard deviation of all nearby known variants, and P-value was derived and shown for each method. The mean difference in scores between rs17293632 and all other nearby SNPs after rescaling (mean of zero and SD of one, for the convenience of cross-comparison) for each method was also shown at the right panel. (B) Disease SNPs with gene regulatory effects nearly 3' UTR of *IKZF3*. Integrative scores (IW-score K10) were shown for 1,000 randomly selected known SNPs nearby, and two GWAS and eQTL candidates, circled with red and blue circles, respectively. Functional scores of rs12946510 and rs907091 for all methods were also extracted and compared, rescaled to fit into the rescaled distribution of the training set. The plots were shown at the right panel. (C) *TERT* promoter mutations and known SNPs in the region, with their integrative scores calculated. Black and grey dots denote somatic mutations and SNPs, respectively. Functional scores between *TERT* promoter (defined by the DNase I mark of the GM12878 cell line) and non-promoter regions were compared and P-values were derived at the right panel.

ingly, we also identified a silent mutation within a DNase I hypersensitive site, chr5:1 282 594 C>A (reverse strand), c.1719C>A, p.L573L, with a score of 3.88 ($P = 0.032$) with strong pathogenic potential (Figure 5C).

We also assessed the correlation between predicted functional scores and positions in relation to *TERT* promoter for all noncoding variants and mutations in and near *TERT* for all other methods (Supplementary Figure S5). Although

most of the methods demonstrated reduced scores with the distance away from the promoter, IW-scores (N8 and N6) provided the most informative and reliable correlation and the largest difference of scores for noncoding variants between *TERT* promoter and non-promoter regions based on the Wilcoxon rank-sum test (Figure 5C; Supplementary Figure S6). This further consolidates our conviction that the

performance of IW-Scoring is effective and stable regardless of data types, unlike other methods.

Landscape of noncoding mutations in follicular lymphoma. We next used IW-Scoring to investigate the landscape of noncoding mutations in cancer. We chose to examine WGS datasets in a haematological cancer, follicular lymphoma (FL) where the landscape of the coding mutations has been extensively studied recently (11,38–40). More than 90% of the tumours have mutations in genes encoding epigenetic regulators suggests that these tumours rely on epigenetic deregulation (41). However, little is known about the repertoire of non-coding mutations in FL.

First we used a cohort of 14 WGS cancer samples from six FL patients from our previous study (11). We calculated IW-scores for all 93 078 unique noncoding somatic mutations. As expected, mutations in defined regulatory regions generally had significantly higher scores than those in non-regulatory regions (Wilcoxon test, $P = 3.53 \times 10^{-49}$, for Ensembl Regulatory Build; $P = 2.71 \times 10^{-89}$ for ENCODE annotation). For example, within Ensembl Regulatory Build defined regions, mutations in promoters had the highest IW-scores on average, followed by those in open chromatin and promoter flanking regions, while mutations in TF binding sites had relatively low scores (Figure 6A). This is largely reflected by the occurrence of mutations in various ENCODE annotated regions (Supplementary Figure S7). Among different ENCODE TF binding regions, mutations within *POU2F2*, *BCLAF1*, *FOXA2*, In1 and Brg1 binding sites seemed to have the highest pathogenic potential on average, further emphasizing the importance of B-cell specific activator (*POU2F2*, also known as Oct-2) (mean IW-score, -0.47 , compared to the mean of -1.15 for all mutations in TF binding regions), Bcl-2 associated protein (*BCLAF1*) (mean IW-score, -0.63), and chromatin modification and remodelling (*FOXA2*, In1/*SMARCB1* and Brg1/*SMARCA4*, mean IW-score of -0.70 , -0.83 and -0.85 , respectively) in FL development (Figure 6B; Supplementary Table S6).

By using a $P < 0.05$ significance threshold for the IW-scores, this allowed the identification and refinement of the list of non-coding mutations to 1375 ($\sim 1.5\%$) variants (Supplementary Table S7). To highlight recurrent functional mutations in regulatory regions, we next examined the density of these candidate functional mutations and searched for regions where functional mutations were clustered within short inter-distance of each other (< 10 kb). We identified 11 such clusters with enriched functional mutations ($n \geq 3$) (Supplementary Table S7). Many of these are known targets of aberrant somatic hypermutations (aSHM) in B-cell lymphomas, such as *BCL2*, *BCL6*, *BCL7A*, *CXCR4* and *PAX5* (42,43). However, many recurrent functional mutations were also found outside the typical targeted regions of aSHM for these genes, i.e. within ~ 2 kb downstream of transcription start sites (TSS). To what extent these mutations were associated with aSHM remains unclear.

We were able to validate these mutation clusters in an extended cohort of 36 FL patients derived from the ICGC Malignant Lymphoma Project where whole-genome simple somatic mutations (SSM) were available (ICGC Data Portal Release 23). Nine out of 11 clusters (82%) were also

significantly enriched for functional mutations in this larger cohort (Supplementary Table S8), although many more mutational clusters enriched for functional mutations were discovered (data not shown). For example, we found in total 64 functional mutations in *PAX5* 5' upstream/UTR and first intron, and only seven of these were within the usual aSHM target regions (Supplementary Table S9). The majority of these mutations were within *PAX5* first intron that contains an active promoter characterized by a DNase I hypersensitive site with strong H3K4me3 and H3K27ac signals in a lymphoblastoid B cell line GM12878 (Figure 6C). Interestingly, we further identified two clusters of mutations ~ 300 – 330 kb upstream of the *PAX* gene. One cluster was within *ZCCHC7* intronic region with five potential functional mutations, and the other cluster of 24 mutations was located at the 3' downstream region of *ZCCHC7* with the most significant mutation as chr9:3 7371 916 G>A (3.29 , $P = 0.058$) (Figure 6C and D). Both regions seemed to contain active enhancers marked by DNase I, H3K4me1 and H3K27ac peaks profiled in GM12878 (Figure 6C). The latter cluster has been previously described in several B-cell related malignancies, indicating this region served as a *PAX5* enhancer (9). By comparing FL mutations and their IW-scores among these three regions, our results further suggest that mutations in *PAX5* promoter were commonly more deleterious than those in enhancers (Wilcoxon test, $P = 4.64 \times 10^{-4}$, for promoter versus enhancer region 2) (Figure 6C and D). Future studies are required to further understand the gene expression regulatory effect and mechanism of these mutations in those regions.

In light of other methods likely offering similar observations, we evaluated the functional scores against coordinates for all identified noncoding mutations in chromosome 9, for ReMM and FATHMM noncoding (selected for their decent performances in differentiating potential functional from non-functional noncoding cancer mutations), as well as for CADD (selected for its relatively poor performance) (Supplementary Figure S8). The results demonstrated that clusters enriched for mutations with high functional scores were not clearly evident based on those methods, compared to the result from IW-Scoring (Figure 6C). We further quantified the differences in scores of noncoding mutations between *PAX5* promoter and first intron, and the rest of chromosome 9, and IW-Scoring produced the largest difference in scores between the two categories (Wilcoxon test, $P = 1.32 \times 10^{-61}$), followed by FATHMM noncoding, ReMM and CADD (Supplementary Figure S8), demonstrating the higher detection power of IW-Scoring identifying potential functional mutation clusters in the genome.

DISCUSSION

The noncoding regions of the genome harbour a substantial fraction of total DNA sequence variations, and the functional contribution of these variants to complex traits, genetic diseases and tumourigenesis is still very poorly defined. Only a small fraction is believed to be truly functional and pathogenic. How to identify and prioritise these key functional driver events has become critical in the era of routine whole genome surveys and studies. We describe here an integrative approach, IW-Scoring, for noncoding

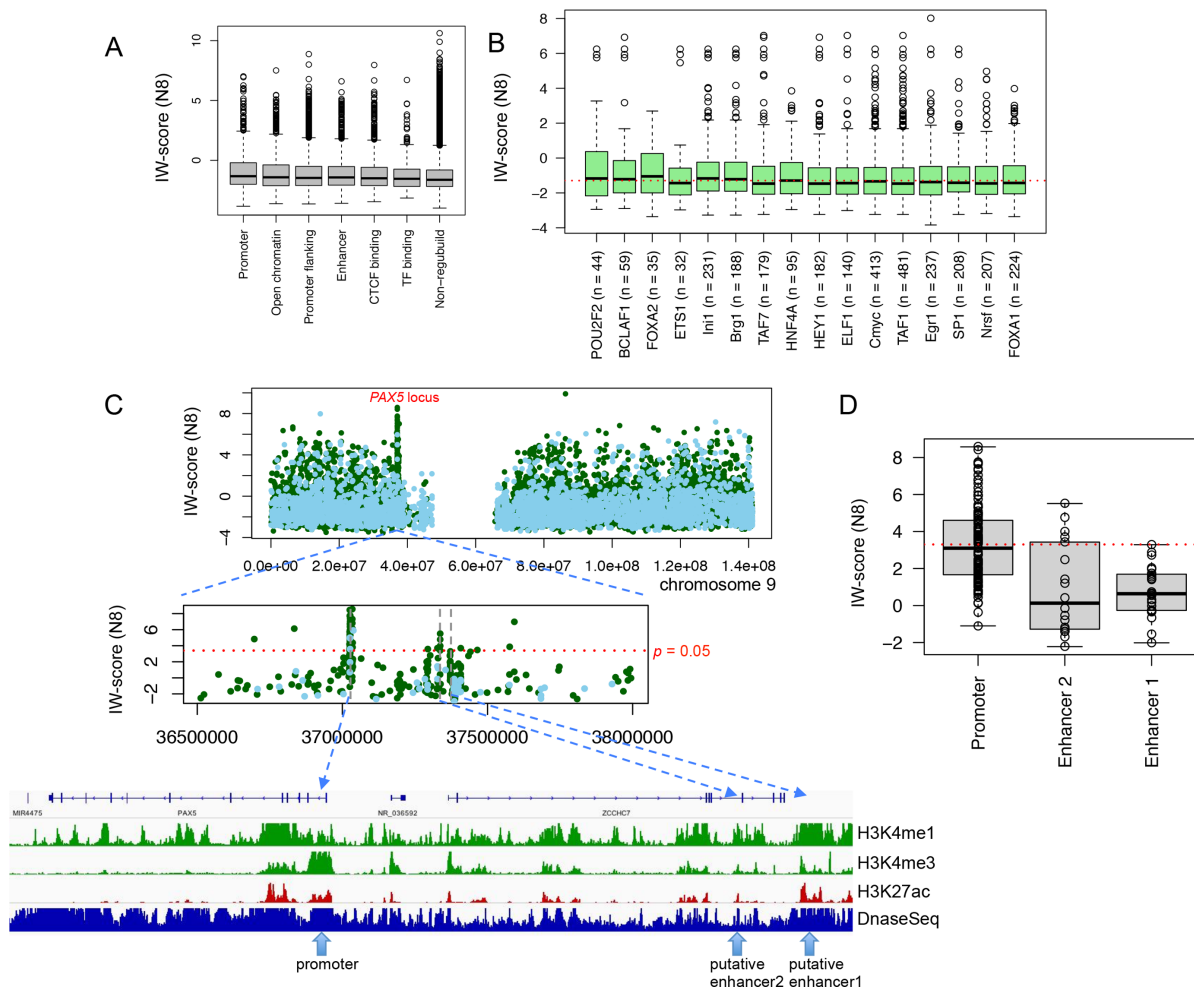


Figure 6. Integrative scores IW-score (N8) of somatic mutations in follicular lymphoma. (A) Integrative scores for mutations within Ensembl Regulatory Build annotated regions. (B) Integrative scores for mutations within top mutated TF binding sites. The mean value was shown by the red dotted line. (C) Recurrent mutations in *PAX5* locus. Integrative scores were shown for all noncoding mutations in whole chromosome 9 first. *PAX5* locus was further zoomed in with three mutations clusters further identified, *PAX5* promoter, putative enhance regions 1 and 2. Mutations from ICGC FL samples were shown by solid green dots, while mutations from Okosun *et al.* were shown by solid light blue dots. Signals of epigenetic histone marks were for GM12878 cell line. (D) Integrative scores for mutations across three *PAX5* associated regions in boxplots. Red dotted line indicated the level of $P = 0.05$.

variant annotation and functional scoring. We showed that our approach outperforms most other functional scoring methods in differentiating functional/pathogenic from non-functional/benign variants in a variety of independent data sets: (i) validated clinically relevant variants, (ii) GWAS significant variants and (iii) recurrent COSMIC cancer mutations. We also demonstrated its powerful application in identifying functional mutations in FL noncoding genome.

Our integrative approach has several advantages when compared to other available methods. First, it starts with embedded gene and regulatory annotation modules, allowing for easy access to the gene centric information and overlapping regulatory elements from a wide range of annotation resources, as well as available functional scores for queried variants. Second, it uses an unsupervised spectral approach to assign weights to available functional scores, and integrates these into a weighted sum. This approach yields the final scores and functional calls based on multiple evidences with a level of significance also derived to

further improve the prioritisation. Thirdly, the nature of this integration meant that the performance of IW-scores was stable and ranked consistently among the best performing methods across all test variant data types. The performances of other methods, however, appeared to vary among data types, therefore, caution is needed to interpret these scores when different data types are studied and compared. In contrast with other tools, IW-Scoring benefitted by scoring the highest proportion of variants in all data sets via imputations based on interdependencies between variables. These advantages ensure that our method is versatile and can be applied to various studies from known SNPs, to rare germline variants and to somatic novel mutations.

A common feature of IW-Scoring and many other methods (e.g. CADD, Eigen and GWAVA) is that disease/phenotype and tissue/cell specificities with their related annotations were all combined into a single functional score. This technique significantly reduced the dimensionality of scoring output without having to produce

scores for each tissue and phenotype, as well as for different chromatin/histone marks, making the data post-processing much easier and more straightforward especially for a large number of variants. However, IW-Scoring still allows for the functional variants associated with specific tissues, cells and features to be identified through the regulatory annotation module. This is currently lacking in many other methods, although some algorithms have chosen to focus on the identification of disease/tissue specific risk variants recently (22,44). Compared to most available methods, we believe our approach is optimally balanced between summarized and detailed evidences for the diverse range of users.

The approach we have adopted here is similar to Eigen and Eigen-PC, however our framework is specifically designed for noncoding variants, incorporating most recent and a wider range of noncoding annotation and functional features/scores. Via a vigorous weight learning process, strong weights were assigned to the block of closely correlated scores (Eigen, DeepSEA, FATHMM noncoding, ReMM and CADD). We have demonstrated that the derived IW-scores provide consistently high-quality performance across various data sets where the individual constituent scores fail to do so, demonstrating the accuracy, validity and stability of our approach. Such ensemble based approach with different estimated weights has been shown to perform better than any single component classifier (26), and has been widely used in various bioinformatics problems (44,45). The weighted integration technique based on the eigendecomposition of the covariance matrix also offers the flexibility to incorporate any other correlated genome-wide functional scores/features into the integrative scores. Many other integration approaches for continuous output can also be explored. For example, decision tree and random forest have been shown to perform well to predict the functional consequences of non-synonymous variants combining various scores (46). Their performances on noncoding variants in comparison to our IW-Scoring approach are yet to be seen. However, as suggested by our results, incorporating more annotation scoring systems does not necessarily improve the performance of integrative scores, especially if these additional scoring systems are very much uncorrelated with the existing ones. Integrative scores can be mostly useful when combined with additional layers of genetic data, such as gene expression profiling by RNA-seq, to look for differential expression of target genes. This combinatorial approach has recently been applied to identify recurrent noncoding regulatory mutations in pancreatic cancer (47). We are aware that IW-Scoring and all other algorithms are useful and powerful tools to prioritise mutations and associated genes from hundreds of thousands of noncoding mutations. However, these findings still need to be validated in the wet-lab to determine their true oncogenic potential.

In this age of WGS and diagnostics, we believe IW-Scoring offers great versatility in discovering noncoding disease causal variants. With the collection of a cohort of samples with the same disease or phenotype, one can identify recurrently mutated noncoding regions enriched for functional mutations predicted by IW-Scoring, which could drive the disease initiation and development, with or with-

out the presence of coding drivers. This will certainly trigger the new wave of noncoding biomarker discovery. We propose that IW-Scoring can play an integral part in the search for significant regulatory variations in complex traits and diseases.

AVAILABILITY

IW-Scoring is freely available for use at <http://snp-nexus.org/IW-Scoring>.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jude Fitzgibbon, Jessica Okosun and Trevor Graham from Barts Cancer Institute for their valuable comments.

Authors' contributions: J.W. conceived the study. J.W., D.U. and C.C. designed and performed the experiments. J.W. wrote the manuscript. J.W., D.U. and C.C. edited and reviewed the manuscript. All authors read and approved the final manuscript.

FUNDING

Cancer Research UK Centre of Excellence Award to Barts Cancer Centre [C16420/A18066]; Barts and The London Charity Grant [MGU0344]; D.U. is funded by Pancreatic Cancer Research Fund Tissue Bank Grant. Funding for open access charge: Charity Open Access Fund (COAF) award.

Conflict of interest statement. None declared.

REFERENCES

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
3. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
4. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
5. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K. *et al.* (2013) TERT promoter mutations in familial and sporadic melanoma. *Science*, **339**, 959–961.
6. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L. and Garraway, L.A. (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science*, **339**, 957–959.
7. Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A. Jr, Friedman, A.H., Friedman, H., Gallia, G.L., Giovannella, B.C. *et al.* (2013) TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 6021–6026.

8. Rachakonda,P.S., Hosen,I., de Verdier,P.J., Fallah,M., Heidenreich,B., Ryk,C., Wiklund,N.P., Steineck,G., Schadendorf,D., Hemminki,K. *et al.* (2013) TERT promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17426–17431.
9. Puente,X.S., Bea,S., Valdes-Mas,R., Villamor,N., Gutierrez-Abril,J., Martin-Subero,J.I., Munar,M., Rubio-Perez,C., Jares,P., Aymerich,M. *et al.* (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **526**, 519–524.
10. Mansour,M.R., Abraham,B.J., Anders,L., Berezovskaya,A., Gutierrez,A., Durbin,A.D., Etchin,J., Lawton,L., Sallan,S.E., Silverman,L.B. *et al.* (2014) Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, **346**, 1373–1377.
11. Rheinbay,E., Parasuraman,P., Grimsby,J., Tiao,G., Engreitt,J.M., Kim,J., Lawrence,M.S., Taylor-Weiner,A., Rodriguez-Cuevas,S., Rosenberg,M. *et al.* (2017) Recurrent and functional regulatory mutations in breast cancer. *Nature*, **547**, 55–60.
12. Araya,C.L., Cenik,C., Reuter,J.A., Kiss,G., Pande,V.S., Snyder,M.P. and Greenleaf,W.J. (2016) Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat. Genet.*, **48**, 117–125.
13. Melton,C., Reuter,J.A., Spacek,D.V. and Snyder,M. (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.*, **47**, 710–716.
14. Weinhold,N., Jacobsen,A., Schultz,N., Sander,C. and Lee,W. (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.
15. Wilkinson,B. (1951) A statistical consideration in psychological research. *Psychol. Bull.*, **48**, 156–158.
16. Khurana,E., Fu,Y., Chakravarty,D., Demichelis,F., Rubin,M.A. and Gerstein,M. (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, **17**, 93–108.
17. Kircher,M., Witten,D.M., Jain,P., O’Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
18. Ritchie,G.R., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
19. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
20. Smedley,D., Schubach,M., Jacobsen,J.O., Kohler,S., Zemojtel,T., Spielmann,M., Jager,M., Hochheiser,H., Washington,N.L., McMurry,J.A. *et al.* (2016) A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.*, **99**, 595–606.
21. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
22. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.
23. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
24. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
25. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
26. Dietterich,T.G. (2000) Ensemble methods in machine learning. *Lect. Notes Comput. Sci.*, **1857**, 1–15.
27. Dayem Ullah,A.Z., Lemoine,N.R. and Chelala,C. (2012) SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.*, **40**, W65–W70.
28. Genomes Project,C., Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
29. Liu,X., Wu,C., Li,C. and Boerwinkle,E. (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.
30. Honaker,J., King,G. and Blackwell,M. (2011) Amelia II: a program for missing data. *J. Stat. Softw.*, **45**, 1–47.
31. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
32. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
33. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
34. Ward,L.D. and Kellis,M. (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, **337**, 1675–1678.
35. Huang,Y.F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
36. Farh,K.K., Marson,A., Zhu,J., Kleinewietfeld,M., Housley,W.J., Beik,S., Shores,N., Whitton,H., Ryan,R.J., Shishkin,A.A. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
37. Huang,F.W., Bielski,C.M., Rinne,M.L., Hahn,W.C., Sellers,W.R., Stegmeier,F., Garraway,L.A. and Kryukov,G.V. (2015) TERT promoter mutations and monoallelic activation of TERT in cancer. *Oncogenesis*, **4**, e176.
38. Krysiak,K., Gomez,F., White,B.S., Matlock,M., Miller,C.A., Trani,L., Fronick,C.C., Fulton,R.S., Kreisel,F., Cashen,A.F. *et al.* (2017) Recurrent somatic mutations affecting B-cell receptor signaling pathway genes in follicular lymphoma. *Blood*, **129**, 473–483.
39. Green,M.R., Kihira,S., Liu,C.L., Nair,R.V., Salari,R., Gentles,A.J., Irish,J., Stehr,H., Vicente-Duenas,C., Romero-Camarero,I. *et al.* (2015) Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E1116–E1125.
40. Pasqualucci,L., Khiabanian,H., Fangazio,M., Vasishta,M., Messina,M., Holmes,A.B., Ouillette,P., Trifonov,V., Rossi,D., Tabbò,F. *et al.* (2014) Genetics of follicular lymphoma transformation. *Cell Rep.*, **6**, 130–140.
41. Araf,S., Okosun,J., Koniali,L., Fitzgibbon,J. and Heward,J. (2016) Epigenetic dysregulation in follicular lymphoma. *Epigenomics*, **8**, 77–84.
42. Khodabakhshi,A.H., Morin,R.D., Fejes,A.P., Mungall,A.J., Mungall,K.L., Bolger-Munro,M., Johnson,N.A., Connors,J.M., Gascoyne,R.D., Marra,M.A. *et al.* (2012) Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget*, **3**, 1308–1319.
43. Odegard,V.H. and Schatz,D.G. (2006) Targeting of somatic hypermutation. *Nat. Rev. Immunol.*, **6**, 573–583.
44. Chen,L., Jin,P. and Qin,Z.S. (2016) DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol.*, **17**, 252.
45. Yang,P.Y., Yang,Y.H., Zhou,B.B. and Zomaya,A.Y. (2010) A review of ensemble methods in bioinformatics. *Curr. Bioinform.*, **5**, 296–308.
46. Knecht,C., Mort,M., Junge,O., Cooper,D.N., Krawczak,M. and Caliebe,A. (2016) IMHOTEP—a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants. *Nucleic Acids Res.*, **45**, e13.
47. Feigin,M.E., Garvin,T., Bailey,P., Waddell,N., Chang,D.K., Kelley,D.R., Shuai,S., Gallinger,S., McPherson,J.D., Grimmond,S.M. *et al.* (2017) Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma. *Nat. Genet.*, **49**, 825–833.