Ferlaino *et al. BMC Bioinformatics* (2017) 18:442
DOI 10.1186/s12859-017-1862-y

BMC Bioinformatics

RESEARCH ARTICLE Open Access

An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome

Michael Ferlaino[1,2*], Mark F. Rogers[3], Hashem A. Shihab[4], Matthew Mort[5], David N. Cooper[5], Tom R. Gaunt[4] and Colin Campbell[3]

Abstract

Background: Small insertions and deletions (indels) have a significant influence in human disease and, in terms of frequency, they are second only to single nucleotide variants as pathogenic mutations. As the majority of mutations associated with complex traits are located outside the exome, it is crucial to investigate the potential pathogenic impact of indels in non-coding regions of the human genome.

Results: We present FATHMM-indel, an integrative approach to predict the functional effect, pathogenic or neutral, of indels in non-coding regions of the human genome. Our method exploits various genomic annotations in addition to sequence data. When validated on benchmark data, FATHMM-indel significantly outperforms CADD and GAVIN, state of the art models in assessing the pathogenic impact of non-coding variants. FATHMM-indel is available via a web server at *indels.biocompute.org.uk*.

Conclusions: FATHMM-indel can accurately predict the functional impact and prioritise small indels throughout the whole non-coding genome.

Keywords: Indels, Non-coding genome, Variant prioritisation, Support vector machines

Background

The advent of next generation sequencing technologies has led to a rapid increase in identified genetic variation, including single nucleotide variants (SNVs), copy number variants, insertions and deletions (indels), in addition to larger scale DNA rearrangements. There are now a vast number of biomedical applications exploiting genomic sequence data and such data will play a crucial role in personalised medicine. As a consequence, interpretation of the functional impact of identified variants is of increasing importance. This has led to the development of accurate methods for assessing genomic tolerance and predictive techniques for discriminating between harmful (pathogenic) and neutral mutations [1–4].

In the past, there has been an emphasis on using sequencing technologies to study human exomes, rather than full genomes, owing to the reduced costs involved and a primary focus towards those regions of the genome deemed to be most functionally relevant. Accordingly, the vast majority of models for predicting the functional impact of indels have been restricted to their effect in the human exome – see e.g. [5–7].

However, the portion of the genome which codes for proteins accounts for only about 2% of the whole sequence, and it is becoming increasingly evident that non–coding portions of the genome play crucial functional roles in human development and disease [8]. For example, a germline deletion in the micro RNA MIR17HG leads to microcephaly [9], and a mutation in the promoter region of MIR146A is genetically associated with lupus [10]. Furthermore, most SNVs identified by genome wide association studies (GWASs) as correlated with increased risk of complex disease are located in non–coding regions [11].

*Correspondence: michael.ferlaino@bdi.ox.ac.uk
[1]Big Data Institute, University of Oxford, Oxford OX3 7LF, UK
[2]Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford OX3 9DU, UK
Full list of author information is available at the end of the article

Given examples like these, in this paper we focus on the association between non–coding variants and disease by developing a model for predicting the functional impact of indels in non–coding regions of the human genome. Our method can be seen as a generalisation of FATHMM [1] for prediction beyond point mutations. A web–based implementation of FATHMM–indel is available at *indels.biocompute.org.uk*.

Methods

Data collection

We developed a machine learning approach to classify the functional effects of small indels, that is, variants where the sequence change involves up to 20 base pairs. The term indel refers to micro insertions/deletions, i.e. mutations that either insert or delete a DNA string to the wildtype sequence.

Pathogenic non–coding indels were collected from the CinVar database [12]. From data downloaded on 8th January 2017, we extracted pathogenic mutations (clinical significance 5) *not* annotated as somatic. Neutral (likely benign) non–coding indels were collected from the exome variant server (EVS) data release ESP6500SIV2 [13]. We considered variants recorded in individuals of African ancestry since European and Asian populations have been subject to bottlenecks which might have resulted in pathogenic indels with relatively high minor allele frequencies (MAFs) – see e.g. [7]. Thus, to increase the probability that EVS mutations were truly benign polymorphisms, we only selected variants with MAF $\geq$ 1% in individuals with African ancestry. In addition to using database annotations to collect micro insertions/deletions in non–coding regions, we further exploited Ensembl GRCh37 (release 85) annotations. By using annotated coding sequence regions, we were able to verify that all examples in our data sets did *not* fall within genomic regions annotated as protein coding.

Repeats are extremely challenging genomic elements to sequence as they are characterised by high sequencing error rates. For example, repeats are strongly affected by polymerase slippage which can potentially alter the length of the repetitive sequence mutation [14]. For these reasons, we conservatively filtered all repeats from our data sets. These steps combined yielded 2 523 pathogenic and 9 783 neutral examples.

FATHMM-indel's features

We used a variety of data sources which potentially carry information about an indel's pathogenic status. Previous work on SNVs has shown that the best predictive models exploit information about sequence conservation in the vicinity of a mutation [2, 15]. Intuitively this makes sense as we expect that mutations occurring in highly conserved regions of a genome are more likely to have deleterious impact compared to those that occur in evolutionary variable regions. However, conservation metrics used to evaluate SNVs are based on distinct nucleotide positions within the human genome [1, 16, 17]. Hence, to study small indels, we must either revise these methods to produce conservation scores for longer ranges, or devise a method that uses existing single–nucleotide scores. Here we adopted the latter approach: to obtain conservation features for small regions, we treated each insertion or deletion as a series of mutations in the reference sequence. All features are described in details in the Additional file 1 (Supplementary Materials).

FATHMM-indel's model

We used a support vector machine (SVM) [18, 19] as our binary classifier, as SVMs have produced highly accurate classifiers for a variety of bioinformatics domains – see, e.g. [2, 15, 20, 21]. Kernel methods such as SVMs can easily handle structured data, such as strings and graphs, which are abundant in bioinformatic applications. Furthermore, support vector machines allow straightforward integration of heterogeneous biological data.

SVMs use kernel matrices to encode the similarity of data objects. Kernels have been derived for a number of different object types, from continuous and discrete variables, through to graph and sequence data (see e.g. [18] for an overview). In this work, we used a Gaussian kernel with precision γ and a "cost" parameter C to lessen the influence of noise in the data.

SVMs can be used to prioritise variants using Platt scaling [22]. Given a test instance $\mathbf{z}$, SVMs compute an "uncalibrated" score

$$f(\mathbf{z}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{z}) + b \tag{1}$$

$\mathbf{K}$ represents the kernel matrix encoding the similarity between data points. The dual parameters α_i (Lagrange multipliers) and b (bias) are learned from training data. The sum in (1) runs over all training examples $\mathbf{x}_i$ with class labels $y_i = \pm 1$. The score $f(\mathbf{z})$ can be interpreted as a confidence measure since, the larger the modulus $|f(\mathbf{z})|$, the greater the confidence of the prediction. $f(\mathbf{z})$ can be converted into a standardised score $\sigma(\mathbf{z}) \in [0, 1]$ by fitting of a logistic function

$$\sigma(\mathbf{z}) = \frac{1}{1 + \exp(Af(\mathbf{z}) + B)} \tag{2}$$

The parameters A and B are learned using maximum likelihood estimation on training data. Exploiting this approach, FATHMM–indel can prioritise variants by returning a score σ for each test mutation. A data point $\mathbf{z}$ is predicted as pathogenic (positive class) if $\sigma(\mathbf{z}) \geq 0.5$

whilst it is predicted as neutral (negative class) otherwise. Indels with largest scores σ are the most likely to be pathogenic.

The kernel machine we used is characterised by two hyperparameters (C, γ) that need to be optimised in order to select the best model to validate against currently published methods (see Results). One of the most popular protocols used for model selection is cross validation. However, it has empirically been shown that cross validation is susceptible of overfitting the *model selection* criterion and, consequently, provide an optimistic estimate of generalisation performance [23]. To control again this potential bias, we performed model selection using a rigorous nested cross validation (NCV) protocol. NCV is comprised of two (nested) loops of cross validation where the inner loop is used for hyperparameter tuning whilst the outer loop is used for performance assessment (Fig. 1). The data set is randomly split into ten *stratified* folds to ensure that each fold (approximately) contains the same number of examples for both classes. In each iteration of the outer loop, nine folds are used to create a tuning set whilst the remaining fold is used for testing. In the inner loop, a grid search is performed on the tuning set in order to select the optimal hyperparameters. A parameter space is created by setting up possible ranges for the hyperparameter values and an SVM is trained at each grid position in such space. The optimal model is selected by implementing ten–fold cross validation and accuracy is used as performance metric. Lastly, the best model is deployed on the testing set to assess the performance of the classifier. This procedure is repeated ten times (the number of stratified folds) and performance is evaluated using sensitivity, specificity, balanced accuracy, and area under the ROC curve (AUC).

Results

FATHMM-indel's performance evaluation

The data collected are substantially imbalanced with many more neutral than pathogenic instances. Therefore, in order to annotate a balanced training set, it is necessary to subsample the majority (EVS) class. A data set can be created by selecting 2 523 pathogenic indels and randomly drawing 2 523 data points from EVS mutations. Using such a set, FATHMM–indel's performance could be evaluated under nested cross validation.

However, it is crucial to establish whether our model is robust against subsampling of EVS mutations. Accordingly, we created 50 data sets comprised of 2 523 pathogenic and 2 523 randomly selected neutral indels. Our model was trained and tested on each set under nested cross validation. Performance was assessed by calculating averaged statistics and standard errors (SEs) across all 50 data sets. FATHMM–indel achieved an average performance of 89% sensitivity, 89% specificity, 89% balanced accuracy, and 0.95 AUC (Table 1). The small standard errors recorded show that our method is robust against subsampling of EVS indels.

In the next section, we compare our model with published methods on benchmark data. The results from this section indicate FATHMM–indel's performance is insensitive to subsampling of the neutral class. Therefore, to validate our method against state of the art models, we trained FATHMM–indel using a data set of 2 523 pathogenic and 2 523 randomly sampled neutral indels. The hyperparameters were set to the values which recorded highest balanced accuracy under nested cross validation experiments ($C = 10, \gamma = 0.01$).

Validation against published methods

In this section we compare our method with CADD [2] and GAVIN [24] – two state of the art models for predicting the impact of non–coding indels. These methods allow comprehensive validation of FATHMM–indel as they are capable of assessing mutation tolerance throughout the *whole* non–coding genome (i.e. they are not restricted to specific units, e.g. splice sites).

CADD is a prioritisation tool capable of measuring deleteriousness by computing "C scores" for genetic variants.

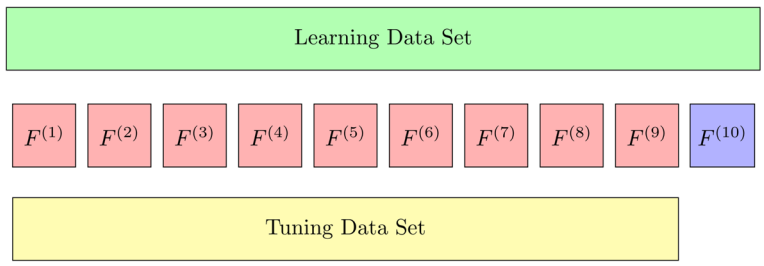

Fig. 1 Nested cross validation. To implement nested cross validation, we split the data set into ten stratified folds. The figure shows *one* out of ten NCV loops. For each NCV iteration, an independent testing set ($F^{(10)}$ in the figure) is left out to assess FATHMM-indel's performance. The remaining folds (red sets in the figure) are merged to create the tuning set used to learn, under cross validation, the optimal values of the hyperparameters. Crucially, a different fold is used as testing set in each iteration, fully exploiting all data to evaluate FATHMM-indel's performance

Table 1 NCV experiment results. FATHMM-indel's performance across 50 data sets created by randomly subsampling the neutral (EVS) class

Sensitivity (SE)	Specificity (SE)	Balanced accuracy (SE)	AUC (SE)
0.886 (0.005)	0.891 (0.005)	0.889 (0.004)	0.950 (0.003)

The small standard errors (SEs) indicate it is consistent to use one random EVS sample to train the final model

Table 2 Validation, on benchmark data, against published methods

	Sensitivity	Specificity	Balanced accuracy	MCC
FATHMM-indel	0.905	0.887	0.896	0.793
CADD	0.669	0.934	0.802	0.626
GAVIN	0.611	0.934	0.773	0.576

CADD's ability to assess the functional impact of mutations was achieved by training an SVM to discriminate between fixed derived alleles in humans (depleted of deleterious variants) and simulated mutations (enriched with deleterious variants). CADD can also be used to classify the impact of mutations by selecting an optimal threshold for *C* scores. As suggested by CADD's authors (through their model web server), all indels with scaled *C* scores of *at least* 15 were predicted as pathogenic.

In addition to predicting the functional class of mutations, FATHMM–indel can also prioritise each variant by computing a score σ (see Methods). For both CADD and FATHMM–indel, the higher the score, the higher the confidence the mutation is functional in disease.

GAVIN is a computational framework that, amongst others, exploits minor allele frequency data to calibrate its predictions. GAVIN does not rank mutations but only classifies the functional impact of a test indel as either pathogenic or neutral.

To perform an unbiased validation against CADD and GAVIN, we annotated a *balanced* benchmark data set comprised of mutations *not* used during the training of any model. Pathogenic indels were obtained from the human gene mutation database (HGMD) release 2014.v4 [25] whilst neutral instances comprised EVS indels with MAF $\geq$ 1%. We restricted our validation examples to mutations that can be scored by all methods and, according to our data collection protocol, we did not consider variants located in repeat regions. Furthermore, we exploited database and Ensembl annotations to ensure all validation indels were *not* located in coding regions. This procedure yielded a benchmark data set with 853 pathogenic (HGMD) indels and 853 neutral (EVS) indels.

Performance was measured using sensitivity, specificity, balanced accuracy, and Matthews correlation coefficient (MCC). The results of our empirical validation on benchmark data are detailed in Table 2. FATHMM–indel recorded the best performance, achieving a balanced accuracy of 90% compared to 80% for CADD and 77% for GAVIN. The substantial improvement in performance attained by our model is also highlighted by the high MCC value, showing how FATHMM–indel's predictions have the strongest correlation with the true class labels. Furthermore, the high sensitivity achieved by our model demonstrates FATHMM–indel's ability to identify truly pathogenic variants. This underlines the potential practical usefulness of our model in, for example, clinical settings where it is crucial not to erroneously categorise pathogenic mutations. Both CADD and GAVIN manifest a bias towards assessing the impact of validation indels as neutral. This has allowed CADD and GAVIN to reach high specificities but very low sensitivities due to the high number of false negatives (FNs). GAVIN recorded the highest value of false negatives (FN = 332, 39% of benchmark pathogenic indels), followed by CADD (FN = 282), whereas FATHMM–indel is characterised by the lowest number FN = 81 (9% of benchmark pathogenic indels). The somewhat lower specificity of our model is a consequence of a slightly higher false positive rate as 11% of benchmark neutral indels were erroneously predicted as pathogenic by FATHMM–indel, whilst 7% of validation neutral indels were miscategorised by CADD and GAVIN.

Since both CADD and FATHMM–indel score variants for prioritisation, it is possible to further compare these models' performance by means of ROC curves and corresponding AUC statistics. For binary classification, a ROC curve displays the true positive rate (sensitivity) as a function of the true negative rate (1 − specificity). The points of the curve are computed by varying the decision threshold from the most positive (pathogenic) data point to the most negative (neutral) one. This allows us to comprehensively validate these models and analyse their performance over the range of possible classification thresholds. The area under the ROC curve, known as AUC, measures the ranking quality of a classification hypothesis [26]. A perfect classifier would have unit AUC whereas random guessing would achieve an AUC of 0.5. The ROC curves of FATHMM–indel and CADD, obtained using the benchmark data set, are visualised in Fig. 2. FATHMM–indel was the best performing method achieving an AUC of 0.956 compared to 0.921 of CADD.

In our validation experiments on benchmark data, FATHMM–indel has shown significant performance improvements over published models. This also validates the ability of FATHMM–indel to generalise to other data sets and establish FATHMM–indel scores as informative metrics for variant prioritisation.

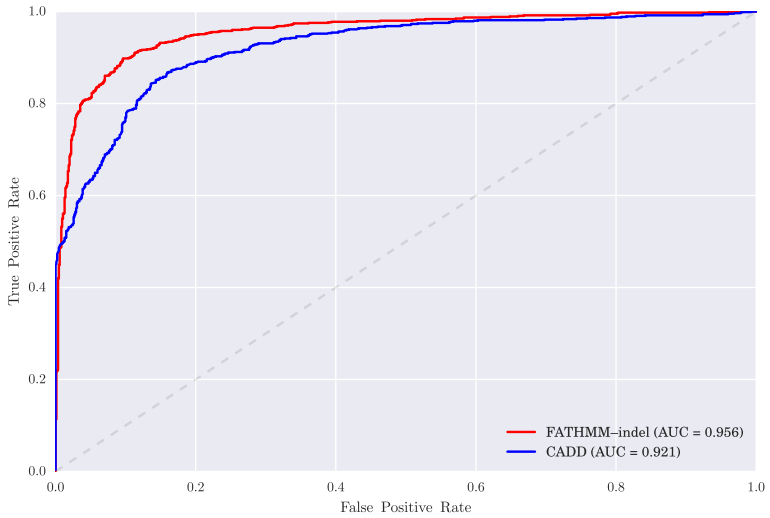

Fig. 2 Empirical ROC curves for FATHMM-indel and CADD. Performance comparison, on benchmark data, between FATHMM-indel and CADD. ROC curves display sensitivities and false positive rates at all possible cutoff levels. Therefore, they can be used to assess the performance of a model independently of the decision threshold

FATHMM-indel for population genetics

To further assess the validity of our approach, we collected non–coding indels from the latest data release (phase 3) of the 1 000 genomes (1KG) project [27]. Amongst its principal goals, the 1KG project aims at analysing the distribution of common and rare mutations in order to provide a broad representation of human genetic variation. In the project's final phase (phase 3), 2 504 genomes were reconstructed from apparently healthy individuals which are stratified into the 5 "continental" populations of East Asia (EAS), South Asia (SAS), Europe (EUR), Africa (AFR), and America (AMR). The 1KG data set also annotates the allele frequency (AF) for each continental population as well as the allele frequency for the global (GLB) sample. This allows to comprehensively analyse private (population specific) alleles and shared variants.

By collecting small variants not located in repetitive regions, we were able to score 1,466,000 non–coding indels from 1KG data. FATHMM–indel classified the vast majority of mutations as neutral, achieving an accuracy of 96%. This represents additional evidence supporting the informativeness of FATHMM–indel's scores for assessing genomic tolerance of non–coding variants.

Exploiting AF data, it is possible to analyse how evolutionary pressures are acting outside the exome by considering the frequency spectrum of indels predicted as pathogenic. We examined the distribution of 1KG indels by binning variants into three categories (Fig. 3). Rare indels have AF < 0.01, low frequency indels have AF $\in [0.01, 0.05]$, whereas common indels have AF > 0.05. Purifying selection removes disadvantageous alleles by reducing their frequency in a population. Therefore, common indels are less likely to be pathogenic than rare indels. We observed this phenomenon across all continental and global populations where the highest percentages of pathogenic indels are rare. Within the continental populations, AMR recorded the highest ratio (55%), followed by EUR (48%), AFR (47%), SAS (47%), and EAS (45%). This trend is even more prominent in the global population where the vast majority (70%) of pathogenic indels are rare. Non–rare variants shared across populations are typically older than non–rare private mutations and, therefore, less likely to be pathogenic.

Furthermore, by looking at common indels, we can analyse how bottlenecks have differentially effected populations. A drastic reduction in population size followed by a rapid growth enables deleterious variants to accumulate at high frequency [28]. European and Asian have been subjects of severe bottlenecks [27, 28] and, as can be seen in Fig. 3, these populations harbour higher ratios of pathogenic indels which are common. EAS has the highest percentage (41%) of disadvantageous common indels, followed by SAS (37%) and EUR (37%). Conversely, the African population is characterised by a much smaller ratio of pathogenic and common indels (29%). Interestingly, at least for the indels that we were able to score, the distribution of AMR indels is much more similar to the AFR frequency spectrum as, for instance, only 28% of pathogenic indels are common in the American population.

Discussion

We presented FATHMM–indel, an integrative method to assess mutation tolerance throughout the *whole*

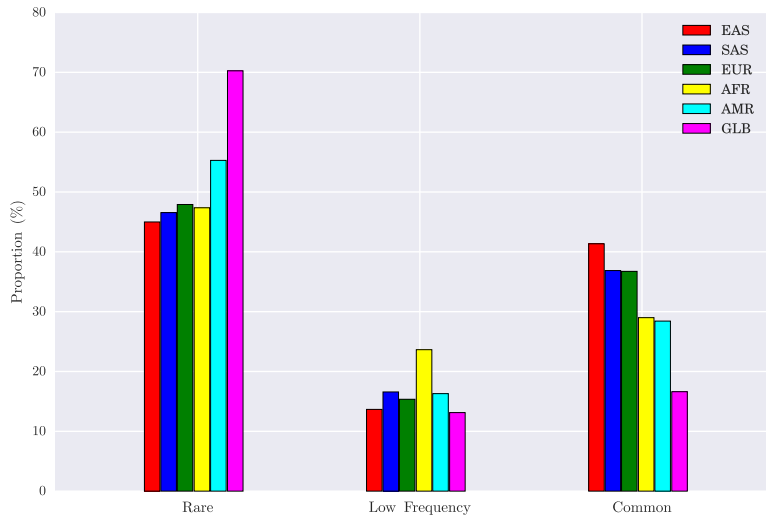

Fig. 3 Frequency spectrum for 1 KG indels predicted as pathogenic. Comparison between non-coding variants across populations and stratified according to allele frequency (AF<1% for rare indels and AF>5% for common indels)

non–coding genome. When validated on benchmark data, FATHMM–indel outperformed CADD and GAVIN, state of the art models for predicting the functional impact of non–coding variants. In addition to predicting the functional class (pathogenic or neutral) of an indel, our method is capable of prioritising variants by computing a standardised score (σ) for each test mutation. This introduces an additional level of flexibility by enabling the implementation of cautious classification to only consider predictions with highest confidence. Given the distribution of FATHMM–indel scores over validation indels, it is possible to cautiously classify our benchmark data set. For example, one can predict an indel with a score bigger than the 80th percentile (0.967) as pathogenic, whilst a mutation with a σ smaller than the 20th percentile (0.034) as neutral. This restricts the number of variants classified to 40% of all benchmark indels but, crucially, allows FATHMM–indel to achieve almost perfect performance with a balanced accuracy of 98%. The interplay between balanced accuracy and the proportion of benchmark indels cautiously classified is comprehensively visualised in Fig. 4. Cautious classification could be extremely useful in, for instance, medical genetics research where, from a "pool" of putative variants, one is interested in selecting only a small subset of *candidate* mutations for experimental validation.

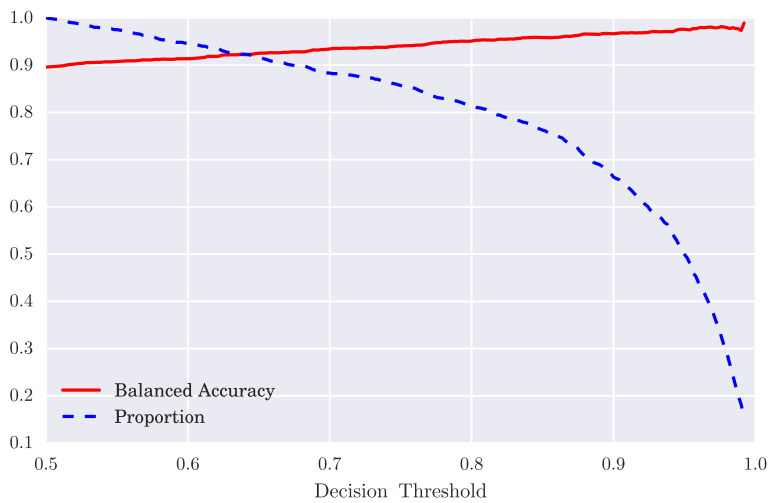

Fig. 4 Cautious classification of benchmark indels. Balanced accuracy, over validation data, as a function of the decision threshold. By selecting only predictions with highest confidence, FATHMM-indel is capable of achieving almost perfect classification

Given current estimates quantifying that *at least* 5% of the human genome is evolutionary constrained [29], it is crucial to deepen our understanding of how selective pressures are acting on non–coding elements. The distribution and evolution of deleterious alleles are fundamental in elucidating the genetic architecture of human disease. In this work, we have also shown how FATHMM–indel can be valuable to discover and analyse differences in non–coding mutation loads across populations.

Our model is available through a web server at *indels.biocompute.org.uk*. By uploading a file in (simplified) VCF, users can submit batches of indels. For a large submission of 10,000 variants, the web server returns FATHMM–indel scores within 30 min (on average).

FATHMM–indel was developed by harvesting knowledge from multiple genomic sources and performing integration at the level of data, where all features are annotated in one data set and similarities between examples are encoded in a unique kernel. As an avenue for future research, it would be interesting to investigate whether it is possible to further boost FATHMM–indel's performance by implementing multiple kernel learning (MKL). Within an MKL approach, multiple data sources are arranged into several feature *groups*, each with its own kernel matrix – see, e.g [30]. Further data sources are available thanks to the efforts of projects like the encyclopedia of DNA elements (ENCODE) consortium [31], which also aims at mapping functional and regulatory elements located *outside* protein coding regions. For example, it would be possible to annotate an additional feature group from transcription factor binding sites data, which have recorded excellent predictive power in assessing genomic tolerance of non–coding SNVs [15].

Currently, as a consequence of our data collection protocol, FATHMM–indel is unable to accurately prioritise non–coding variants located in repetitive regions. Before all repeats were filtered from our training data, 1% of pathogenic indels were repeats whilst 21% of neutral indels were located in repetitive elements. Annotating a training set by random sampling of repetitive sequences would lead to over representation of repeats within the neutral class and, consequently, result in the introduction of a potential confounding factor. Hence, extending FATHMM–indel's capabilities to prioritise repeats warrants further and careful analyses that we leave to future work.

Conclusions

We developed FATHMM–indel, an integrative computational model for predicting indel pathogenicity. Although the vast majority of genetic alterations lie outside the exome, there is a lack of methods *specifically* designed to predict the impact of indels throughout the *whole* non–coding genome. We developed our model to fill in this gap, to aid in predicting the biological consequences of non–coding variants. We envisage FATHMM–indel as a useful annotation tool that could be used, for example, to prioritise causative variants, like those identified in GWASs, for downstream studies to analyse the phenotypic impact of non–coding indels.

Additional file

Additional file 1: Supplementary Materials. In this PDF file, we report a detailed description of all the features used during the development of FATHMM-indel. (PDF 150 kb)

Abbreviations

AF: Allele frequency; AFR: Africa; AMR: America; AUC: Area under the ROC curve; CADD: Combined annotation dependent depletion; EAS: East Asia; EUR: Europe; EVS: Exome variant server; FATHMM: Functional analysis through hidden Markov models; FN: False negative; GAVIN: Gene aware variant interpretation; GLB: Global; GWAS: Genome wide association study; HGMD: Human gene mutation database; Indel: Insertion or deletion; MAF: Minor allele frequency; MCC: Matthews correlation coefficient; NCV: Nested cross validation; ROC: Receiver operating characteristic; SAS: South Asia; SNV: Single nucleotide variant; SVM: Support vector machine; VCF: Variant call format; 1KG: 1000 genomes

Acknowledgements

Not applicable.

Funding

MF is supported by MRC grant MR/M01326X/1. MFR is supported by EPSRC grant EP/K008250/1. MM and DNC gratefully acknowledge the financial support of Qiagen Inc. through a licence agreement with Cardiff University. TRG is supported by MRC IEU grant MC UU 12013/8.

Availability of data and materials

Most data sets used to develop FATHMM-indel are freely available for download, as VCF files, through our web server at *indels.biocompute.org.uk* (section Downloads). The only data set not freely available was annotated from the HGMD database. The most up to date HGMD release (HGMD professional) is available to academic, clinical and commercial users under license via QIAGEN Inc. Lastly, FATHMM-indel is available at the accompanying website *indels.biocompute.org.uk*.

Authors' contributions

MF performed all experiments, analyses, wrote the manuscript, and developed the web server. MFR helped with the design of covariates, the writing of the manuscript, and the testing of the web server. HAS, MM, and DNC resourced data and provided feedback about the manuscript. TRG and CC conceived the study, helped with the writing of the manuscript and the testing of the web server. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Data used are all from secondary sources, where primary ethics approval had been obtained for data acquisition. The details of the project were passed by Dr Birgit Whitman (Head of Research Governance, University of Bristol), who has confirmed that, as a secondary usage, no passage through the university ethics committee is required.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Big Data Institute, University of Oxford, Oxford OX3 7LF, UK. [2]Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford OX3 9DU, UK. [3]Intelligent Systems Laboratory, University of Bristol, Bristol BS8 1UB, UK. [4]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK. [5]Institute of Medical Genetics, Cardiff University, Cardiff CF14 4XN, UK.

Received: 23 May 2017 Accepted: 2 October 2017
Published online: 06 October 2017

References

1. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. Hum Mutat. 2013;34:57–65.
2. 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;46:310–5.
3. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11:294–6.
4. Quang D, Chen Y, Xie X. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31:761–763.
5. Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, Ryan M, Karchin R. Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (vest-indel). Hum Mutat. 2016;37:28–35.
6. Folkman L, Yang Y, Li Z, Stantic B, Sattar A, Mort M, Cooper DN, Liu Y, Zhou Y. Ddig-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics. 2015;31:1599–1606.
7. Hu J, Ng PC. Predicting the effects of frameshifting indels. Genome Biol. 2012;13:R9. doi:10.1186/gb-2012-13-2-r9.
8. Esteller M. Non-coding rnas in human disease. Nat Rev Genet. 2011;12:861–74.
9. de Pontual L, Yao E, Callier P, Faivre L, Drouin V, Cariou S, Van Haeringen A, Geneviève D, Goldenberg A, Oufadem M, Manouvrier S, Munnich A, Vidigal JA, Vekemans M, Lyonnet S, Henrion-Caude A, Ventura A, Amiel J. Germline deletion of the mir-17-92 cluster causes skeletal and growth defects in humans. Nat Genet. 2011;43:1026–30.
10. Luo X, Yang W, Ye DQ, Cui H, Zhang Y, Hirankarn N, Qian X, Tang Y, Lau YL, de Vries N, Tak PP, Tsao BP, Shen N. A functional variant in microrna-146a promoter modulates its expression and confers disease risk for systemic lupus erythematosus. PLoS Genet. 2011;7(6):e1002128. doi:10.1371/journal.pgen.1002128.
11. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24:102–10.
12. 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;42:980–5.
13. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, NHLBI Exome Sequencing Project, Akey JM. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493:216–20.
14. Narzisi G, Schatz MC. The challenge of small-scale repeats for indel discovery. Front Bioeng Biotechnol. 2015;3:8. doi:103389/fbioe.2015.00008.
15. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536–43.
16. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genom Res. 2005;15:1034–50.
17. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.
18. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press; 2004.
19. Campbell C, Ying Y. Learning with Support Vector Machines. USA: Morgan and Claypool; 2011.
20. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4(10):e1000173. doi:10.1371/journal.pcbi.1000173.
21. Afsar Minhas F, Ross ED, Ben-Hur A. Amino acid composition predicts prion activity. PLoS Comput Biol. 2017;13(4):e1005465. https://doi.org/10.1371/journal.pcbi.1005465.
22. Platt J. Probabilities for sv machines In: Smola J, Bartlett PL, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers. Massachusetts: MIT Press; 1999. p. 61–74.
23. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079–107.
24. van der Velde KJ, de Boer EN, van Diemen CC, Sikkema-Raddatz B, A bbott KM, Knopperts A, Franke L, Sijmons RH, de Koning TJ, Wijmenga C, Sinke RJ, Swertz MA. Gavin: Gene-aware variant interpretation for medical sequencing. Genome Biol. 2017;18:6. doi:10.1186/s13059-016-1141-7.
25. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136:665–77.
26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology. 1982;143:29–36.
27. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
28. Lohmuelle KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD. Proportionally more deleterious genetic variation in european than in African populations. Nature. 2008;21:994–7.
29. Pheasant M, Mattick JS. Raising the estimate of functional human sequences. Genome Res. 2007;17:1245–53.
30. Gönen M, Alpaydin E. Multiple kernel learning algorithms. J Mach Learn Res. 2011;12:2211–68.
31. The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. Nature. 2012;489:57–74.