eXtasy: variant prioritization by genomic data fusion

Alejandro Sifrim[1,2,4], Dusan Popovic[1,2,4], Leon-Charles Tranchevent[1,2], Amin Ardeshirdavani[1,2], Ryo Sakai[1,2], Peter Konings[1,2], Joris R Vermeesch[3], Jan Aerts[1,2], Bart De Moor[1,2] & Yves Moreau[1,2]

Massively parallel sequencing greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce eXtasy, an approach to prioritize nonsynonymous single-nucleotide variants (nSNVs) that substantially improves prediction of disease-causing variants in exome sequencing data by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization.

Rare variation identified by exome sequencing is particularly useful for discovering the cause of rare monogenic disorders. By selecting only nSNVs and loss-of-function mutations that are not present in healthy populations or unaffected samples, we can discard a large proportion of exomic variation as likely neutral. However, despite such aggressive filtering, several thousand candidate causal mutations remain, and we need predictive methods to prioritize variants for further validation. Several computational methods take into account biochemical, evolutionary and structural properties of mutations to assess their potential deleteriousness[1–5], but most of these methods suffer from high false positive rates when predicting the impact of rare nSNVs[4]. A plausible explanation for this poor performance is that many of the scrutinized variants are mildly deleterious and subject to weak purifying selection[6,7] but are not specific to the disease of interest.

To assess this hypothesis and to improve variant prioritization, we propose a genomic data fusion methodology[8] that integrates multiple strategies to detect the deleteriousness of mutations and prioritizes them in a phenotype-specific manner. A key innovation included in our strategy is a method for gene prioritization[9], which scores mutated genes according to their similarity to known disease genes by fusing heterogeneous genomic information. We also integrate haploinsufficiency prediction scores[10] that predict the probability that the function of a gene is affected if it is present in a functionally haploid state. To integrate these data sources, we use Random Forests learning[11] and train our model on the Human Gene Mutation Database (HGMD) of human disease-causing mutations[12] compared against three control sets: common polymorphisms and two independent sets of rare variation (**Supplementary Note 1**).

After generating and annotating the different data sets, we inspected the distribution of deleteriousness prediction scores across the positive and control sets (**Supplementary Fig. 1**). All predictions seemed to score the positive set high and thus showed a high sensitivity. In the control sets, most methods correctly classified common polymorphisms as benign, yet they classified a substantial proportion of rare variation as deleterious, leading to low precision. Control variants seemed to occur more often in genes predicted to maintain functionality in a haploid state (haplosufficient), whereas disease-causing variants showed no clear pattern. Disease-causing variants were primarily found in top-ranked genes after gene prioritization. By contrast, control variants showed a homogeneous distribution of gene prioritization ranks, which is to be expected under the assumption that they are prioritized for randomly selected phenotypes.

By integrating these different scores, we aimed to boost our ability to discriminate between putatively mildly deleterious rare variants and actual disease-causing variants. We evaluated several commonly used classification approaches and chose a Random Forests algorithm because it outperformed all other classification algorithms on this task (**Supplementary Table 1**). We trained this classifier by comparing our positive set of disease-causing variants to the rare-variant control sets and observed a considerable improvement in all performance measures over classical deleteriousness prediction methods. This was the case when distinguishing between disease-causing and rare control variants as well as between disease-causing variants and common polymorphisms (**Fig. 1a**,**b**, **Supplementary Fig. 2** and **Supplementary Table 2**). The performance against common polymorphisms was in line with published results for deleteriousness prediction tools because these tools were trained using common polymorphisms as controls. The performance of these tools against rare non–disease-causing variants was much poorer than that against common polymorphisms. Precision was the most improved performance measure, which is important when dealing with a prioritization task (**Supplementary Table 1** and **Supplementary Note 2**). However, performance measures obtained in retrospective benchmarks such as ours are usually optimistic estimates because of the bias of prior information for gene-prioritization predictions[8,13].

[1]Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium. [2]iMinds Future Health Department, Leuven, Belgium. [3]Laboratory of Molecular Cytogenetics and Genome Research, KU Leuven, Leuven, Belgium. [4]These authors contributed equally to this work. Correspondence should be addressed to Y.M. (yves.moreau@esat.kuleuven.be).

RECEIVED 6 MARCH; ACCEPTED 26 AUGUST; PUBLISHED ONLINE 29 SEPTEMBER 2013; DOI:10.1038/NMETH.2656

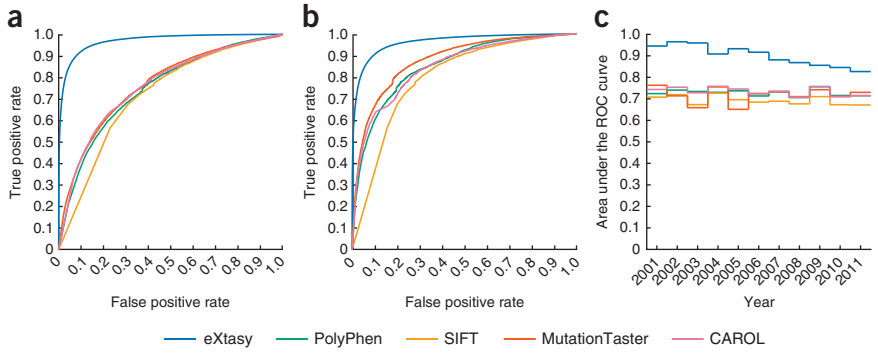

Figure 1 | Receiver operating characteristic (ROC) curves comparing eXtasy and classical deleteriousness prediction scores. (**a**–**c**) ROC curves for disease-causing versus rare control variants (**a**) and disease-causing versus common polymorphisms (**b**). In both cases, eXtasy outperforms other methods, as can be seen by an increase in the area under the curve (AUC). (**c**) We stratified disease-causing variants by year of discovery to test biases in our retrospective benchmark. More recent disease-variant associations show a decrease in AUC performance for eXtasy as biases decrease, but eXtasy always outperforms classical deleteriousness prediction scores.

This problem is difficult to address in an initial study and can be truly resolved only by long-term prospective benchmarks wherein predictions are made on the current state of knowledge and validated in future studies.

To estimate the extent of this bias in our benchmark, we assessed classification performance on the basis of the discovery date of the causative variation using data for gene prioritization before the year of discovery (**Supplementary Fig. 3**). If a bias was present, it would be less prominent in recent discoveries, as these gene-disease associations would be less likely to be directly or indirectly incorporated in the gene-prioritization data sources. Even though we saw a slight decline in performance for more recent discoveries, the method still performed considerably better than classical deleteriousness prediction scores (**Fig. 1c** and **Supplementary Tables 3** and **4**). The performance of deleteriousness prediction scores was not affected by the year of discovery of the association, which is expected because these scores do not integrate time-dependent information. Finally, we looked at feature importance by shuffling each feature across disease-causing and control variants, evaluating the increase in classification error (**Supplementary Fig. 4**). This analysis showed that all included features were informative and improved classification to some degree.

We have demonstrated an approach that integrates biochemical, evolutionary and phenotype-specific information to prioritize mutations for follow-up validation studies. This approach helps to distinguish disease-causing mutations from neutral common polymorphisms as well as rare, potentially deleterious but phenotypically unrelated variation in the coding genome. We acknowledge that performance measures are likely overestimated because of the biased nature of retrospective benchmarks, yet we see a marked improvement in prediction performance over frequently used deleteriousness prediction scores in recently published mutations. We envision that in the near future, initiatives such as the Critical Assessment of Genome Interpretation, although currently focused on single-phenotype benchmarks, will play a major role in providing unbiased prospective benchmarks to optimally assess the performance of methods such as the one described in this study.

Future research and the availability of larger public disease-causing variation data sets will likely widen the scope of the method to other types of mutations (mitochondrial, noncoding, synonymous, nonsense and splice-site mutations, as well as indels). Also, the addition of other data sources, such as locus-specific information (for example, copy number–variant prevalence and genome-wide association study–related loci), to our method and its integration into genetic association tests across multiple samples[14,15] will likely increase its power to discover the cause of genetic disease.

A web tool and stand-alone version implementing our approach are available at http://homes.esat.kuleuven.be/~bioiuser/eXtasy/ (source code is at http://github.com/asifrim/eXtasy/).

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research is supported by Research Council KU Leuven: GOA/10/09 Manet, PFV/10/016 SymBioSys, IOF 3M120274 Immunosuppressive drugs; iMinds: SBO 2013; Hercules Stichting: Hercules III PacBio RS; the Flemish Institute for Science and Technology: IWT-SB/093289, IWT-TBM Haplotyping; EU: Cost Action BM1006: NGS Data Analysis Network, FCT Neuroclinomics.

AUTHOR CONTRIBUTIONS

A.S., D.P. and Y.M. conceptually defined the project. A.S. and D.P. wrote the initial draft of the manuscript and performed the analyses. A.S. generated the data sets and developed the software tools. D.P. developed the benchmarks and trained the models. L.-C.T. and A.S. computed the Endeavour gene prioritizations. A.A. and A.S. developed the web tool. R.S. and J.A. advised on data visualization and visual analytics. P.K. advised on statistical concerns. J.R.V. advised on genetical concerns. All authors revised and proofread the paper. B.D.M. cosupervised the project. Y.M. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Adzhubei, I.A. *et al. Nat. Methods* **7**, 248–249 (2010).
2. Ng, P.C. & Henikoff, S. *Nucleic Acids Res.* **31**, 3812 (2003).
3. Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. *Nat. Methods* **7**, 575–576 (2010).
4. Kumar, S., Sanderford, M., Gray, V.E., Ye, J. & Liu, L. *Nat. Methods* **9**, 855–856 (2012).
5. Chun, S. & Fay, J.C. *Genome Res.* **19**, 1553–1561 (2009).
6. Asthana, S. *et al. Proc. Natl. Acad. Sci. USA* **104**, 12410–12415 (2007).
7. Tennessen, J.A. *et al. Science* **337**, 64–69 (2012).
8. Moreau, Y. & Tranchevent, L.-C. *Nat. Rev. Genet.* **13**, 523–536 (2012).
9. Aerts, S. *et al. Nat. Biotechnol.* **24**, 537–544 (2006).
10. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. *PLoS Genet.* **6**, e1001154 (2010).
11. Breiman, L. *Mach. Learn.* **45**, 5–32 (2001).
12. Stenson, P.D. *et al. Genome Med.* **1**, 13 (2009).
13. Myers, C.L., Barrett, D.R., Hibbs, M.A., Huttenhower, C. & Troyanskaya, O.G. *BMC Genomics* **7**, 187 (2006).
14. Yandell, M. *et al. Genome Res.* **21**, 1529–1542 (2011).
15. Ionita-Laza, I. *et al. Am. J. Hum. Genet.* **89**, 701–712 (2011).

ONLINE METHODS

Data generation. To determine the effectiveness of our approach, we assembled and evaluated a data set consisting of disease-causing variants and different types of control variants. We used a positive data set consisting of 24,454 disease-causing nSNVs from the HGMD associated with 1,142 different Human Phenotype Ontology (HPO) terms[16]. Mapping between HGMD disease descriptors and HPO terms was performed using the Phenomizer tool[17]. Phenotypic terms for which fewer than three implicated genes were known were discarded (excluding the gene in which the variant was located) so as to allow meaningful subsequent gene prioritization. For the control data sets, we considered two classes of variants on the basis of their minor allele frequency (MAF) in the 1000 Genomes Project data set: common polymorphisms (MAF > 1%) and rare variation (MAF ≤ 1%). Additionally, we compiled a third control set of rare variants using in-house–sequenced exomes of healthy individuals and retaining only high-quality calls (coverage > 30×) not present in any publicly available variant database (NHLBI EVS, dbSNP or 1000G). After randomly assigning groups of 500 control variants from each control set to each HPO term represented in the disease-causing variants, we annotated all variant-phenotype combinations with functional information. For every variant, we extracted PolyPhen-2 (ref. 1), sorting intolerant from tolerant (SIFT)[2], MutationTaster[3] and likelihood-ratio test (LRT)[5] deleteriousness prediction scores from the dbNSFP database[18]. We calculated combined annotation scoring tool (CAROL) aggregate deleteriousness scores[19] and added phyloP[20] and PhastCons[21] evolutionary conservation scores across vertebrates, primates and placental mammal subsets from the UCSC Genome Browser database. We also included precalculated gene haploinsufficiency prediction scores[10] where available. Finally, we computed gene prioritization scores with Endeavor[9] using gene-phenotype associations obtained from HPO. (For additional information on the data generation and annotation process see **Supplementary Note 1**).

Classifier benchmarks. We set up a benchmark to determine the optimal classifier for genomic data fusion. To do this, we followed published guidelines[22] applicable to these types of studies. Initially, we removed all records containing missing values and split the disease-causing variants into training (two-thirds of the total number of genes) and testing sets (one-third of the genes) (**Supplementary Fig. 5**). The data records were stratified at the gene level (the highest level of granularity) to assure that the algorithm is not overfitting the gene level information and thus does not overestimate performance. Subsequently, we randomly subsampled the negative examples in the training set so as to balance their number with the number of positive examples. We then trained eight different classifiers using the same combinations of training and test subsets for all classifiers. In addition, we applied state-of-the-art deleteriousness prediction methods (PolyPhen, SIFT, MutationTaster and CAROL scores) using their respective published or documented decision thresholds (0.85, 0.95, 0.5 and 0.98, respectively). As there were several records per single variant in the data set (because of the different phenotypes per variant), we used the maximum score across phenotypes in testing to predict the outcome for each variant. We chose the maximum as this is robust to noninformative phenotypes (because of phenotypically variable diseases), which might be present in our benchmark data set of disease-causing variants. This did not apply for methods that produce a single score per variant because they are not phenotype specific (PolyPhen, SIFT, MutationTaster and CAROL).

The procedure was repeated 100 times on different random splits of data to obtain an estimate of the variance of the resulting performance measures. The results were computed in terms of accuracy, sensitivity, specificity, positive and negative predictive value, Matthews correlation coefficient, area under the receiver operating characteristic (ROC) curve and precision-recall (PR) curves. We observed that all the classifiers built using all heterogeneous data sources performed better than state-of-the-art deleteriousness prediction methods, proving that these data sources contain additional information that facilitates better predictions.

Among all classification algorithms, Random Forests (RF) outperformed all others in terms of all performance metrics, with the exception of sensitivity and negative predictive value (NPV), for which linear discriminant analysis (LDA) performed marginally better (**Supplementary Table 1** and **Supplementary Figs. 6** and **7**). However, precision (positive predictive value or PPV) of the LDA was very low (0.35), indicating that the classifier is overly optimistic. The same observation holds for all of the state-of-the-art methods (precision between 0.20 and 0.23). Furthermore, the PR curve for the Random Forests shows that changing the default threshold (0.5) of the classifier results in a sharp increase in precision but a small loss in sensitivity. This suggests that most of the true positives are highly ranked by the method. Also, the reported accuracy and other aggregate performance measures of the state-of-the-art tools depend greatly on where the decision threshold is set and the skewness of the class distribution. Even though the AUC for some of the ROC curves is higher on one tool than on another, sometimes a point can be found where a certain measure is higher for the tool with the lower AUC.

For all classification algorithms, we use their respective Matlab 7.10 implementations: classregtree class for the decision trees, TreeBagger class for the Random Forests, function knnclassify for *k* nearest neighbors, function classify (with argument "linear" or "quadratic") for the LDA and QDA, respectively, class NaiveBayes for naive Bayes classification, and functions svmtrain and svmclassify for the feed-forward neural networks. Most of the used functions and classes are part of the Statistical Toolbox, except for KNN and SVMs (Bioinformatics Toolbox). The details on parameter settings of particular classifiers are provided in **Supplementary Table 5**.

Control-set benchmarks. We set up an additional benchmark to assess the behavior of the final model under different classification schemes with regard to the different negative outcomes. We considered two different scenarios for training and testing. In a first scenario, the Random Forests model is trained using a subset of the rare non–disease-causing variants as negatives and is tested against the rest of them (as in our standard setup). In a second scenario, the Random Forests model is trained using all of the rare non–disease-causing variants available as negatives and is tested against the data set containing common polymorphisms.

In the case of the first scenario, the validation scheme is identical to the classifier benchmark: all data, including all positive and the negative examples, have been repeatedly (100 iterations)

divided at random into training and testing subsets, grouped gene by gene. In the case of the second scenario, only the positive examples were randomly assigned, whereas the two distinct groups of negatives, rare and polymorphisms, were used respectively for training and testing (**Supplementary Fig. 8**). We made sure that during a single iteration, a split of positives stayed the same across the two scenarios.

Temporal stratification analysis. To analyze the sensitivity of the method with regard to a priori gene or disease association biases, we set up an additional benchmark to estimate the effect size of such biases. Under the assumption that recently published genes would be less likely to be biased by our gene prioritization step, we stratified our positive testing set of disease-causing variants by year of publication (2000–2012) while training the model only on data published before 2000. In this way we measure performance before and after Endeavor's data sources were last updated (the last update occurred in 2008). This threshold applies to both variant and gene level of data granularity, so variants discovered after 2000 but that are associated with genes that are part of the training (i.e., for which the gene-phenotype association was discovered before 2000) are also removed from test sets. The negative examples (non–disease-causing variants) were randomly assigned to one given year between 2000 and 2012, in numbers matching the class distribution of the whole data (given the number of positives in a particular year) and with no overlap between training and test sets (see **Supplementary Fig. 9**). As before, the splits were performed genewise. After the training phase, the classifier was used on cases from the subsequent years (2000–2011). The whole procedure of randomly assigning negatives was repeated 100 times to get stable estimates of performance metrics.

We observed a slight temporal decline in performance throughout all testing years. Nevertheless, eXtasy still significantly outperformed all classical deleteriousness prediction methods across all years. We attribute this decline in performance to the fact that over time some disease-causing genes are better described in the gene-prioritization sources (e.g., literature mining) and are therefore easier to classify. Although such effects are likely present and point to the fact that our main benchmark is an overoptimistic estimate of the real performance, this benchmark setting is a pessimistic estimate of the real performance because of various properties of the training/validation scheme. First, only a fraction of the positive training data can be used (data before 2000), leading to suboptimal learning of the features of disease-causing variants. Second, the training data themselves contain well-described genes and are thus themselves biased toward easier-to-classify examples. This leads to overly optimistic decision thresholds in the classifier and thus degrades the performance when faced with more difficult examples of less well–described genes in the testing set. Finally, owing to the genewise stratification of training and testing sets (to avoid overfitting specific genes), if gene data were published before 2000 but then later published in the light of a new phenotype, they were omitted from the test set. Well-described genes are often discovered to play important roles in new (and often related) physiological processes. This type of discovery can greatly benefit from gene-prioritization approaches but are excluded from this benchmark.

The real performance of the classifier depends greatly on the use case, which is unknown to the researcher applying the method. In the case where the cause of the phenotype is a novel mutation in a previously described gene, the performance is likely to resemble, or even exceed, that of the cross-validation benchmark. When the phenotype is caused by a novel mutation in a gene not previously associated with the phenotype, or in a gene associated with a different phenotype, the real performance likely lies between the overoptimistic cross-validation benchmark and the pessimistic temporal-analysis benchmark (the shaded light blue area in **Supplementary Fig. 3**).

Feature importance analysis. The Random Forests algorithm has the intrinsic ability to estimate the importance of features used for training[11]. This is achieved by measuring the difference between the mean-square error of the prediction on out-of-bag (OOB) examples when values of a given feature are shuffled compared to the error on undisturbed OOB data. This procedure is repeated for each and every tree in the ensemble with its corresponding OOB examples, providing a global measure of feature importance. Here we analyzed how the different features in the data contributed to the overall classification. In particular, we ran 100 simulations (in line with previously described benchmarks) per feature, during which ensembles are built using different random subsamples of the negative data. The result in terms of mean-square error increase when shuffling the feature is displayed in **Supplementary Figure 4**. From the plot, it appears that all features contribute to the classification to some degree. The increase of total mean-square error when one of them is randomized ranges from around 2% for PPI-HPRD to 12% for sequence similarity and functional annotations. Second, highly correlated features, such as various Endeavor scores or state-of-the-art methods (CAROL with SIFT or PolyPhen) usually form clusters of seemingly less important features with a low yet nonzero increase in mean-square error. This is expected, as it has been shown that feature importance measures for Random Forests are strongly affected by the presence of correlation between features[23]. In the absence of a particular feature, other correlated features partially 'take over' the role of former, reducing the impact of the shuffling on the classification error. Hence, these variables are still individually very important—especially if data records contain missing values.

16. Robinson, P.N. *et al. Am. J. Hum. Genet.* **83**, 610–615 (2008).
17. Köhler, S. *et al. Am. J. Hum. Genet.* **85**, 457–464 (2009).
18. Liu, X., Jian, X. & Boerwinkle, E. *Hum. Mutat.* **32**, 894–899 (2011).
19. Lopes, M.C. *et al. Hum. Hered.* **73**, 47–51 (2012).
20. Pertea, M., Pertea, G.M. & Salzberg, S.L. *BMC Bioinformatics* **12**, 274 (2011).
21. Siepel, A. *et al. Genome Res.* **15**, 1034–1050 (2005).
22. Vihinen, M. *BMC Genomics* **13** (suppl. 4), S2 (2012).
23. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. *BMC Bioinformatics* **9**, 307 (2008).

doi:10.1038/nmeth.2656