Internal Report CS Bioinformatics Track 17-05 October 2017

Leiden University

Computer Science

Bioinformatics Track

Predicting beneficial alleles using comparative genomics

Ramon Bettings

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Predicting beneficial alleles using comparative genomics

Ramon Bettings

October 12, 2017

In partial fulfillment of the requirements
for the Degree of Master of Science
in *Computer Science (Bioinformatics)*
Leiden: S1248936
Delft: 4528417

Supervisor: Marcel Reinders
Advisor: Marc Hulsman

Abstract

Motivation: Over the last years there have been many efforts in predicting the possible deleterious effect of genetic variants. However, due to an increase in longevity research and research into protective variants, there is an increasing need for predicting protective alleles and alleles that have a beneficial effect on human functioning.
Methods: We created an evolutionary model based on the human-derived allele frequency (DAF), i.e. the allele frequency of alleles not yet present in our last common ancestor with the chimpanzee. Based on this model we generated a training set containing three types of variants: beneficial variants, deleterious variants, and neutral variants. The idea behind the model is that beneficial alleles will be overrepresented in the variants with a high DAF since beneficial alleles fixate relatively rapidly. A balance between new alleles and negative selection pressure keeps the low DAF variants enriched for deleterious variants. Neutral variants only encountering pressures from random genetic drift can be found all over the DAF spectrum and since the extreme ends are enriched for deleterious alleles and beneficial alleles the middle part of the spectrum should be enriched for neutral alleles. After labeling all variants in ExAC according to these principals, we collected a broad range of genetic and genomic features for each of the labeled variants. Using this dataset we trained a classifier to separate the three classes and assessed its performance.
Results: The model was successful in separating the three classes in our training set. The model could also separate pathogenic variants and benign variants from the ClinVar database. When inspecting beneficially predicted variants we found that they often associate to beneficial traits in genome-wide association studies. We found that deleterious alleles are different from the other alleles because of their higher conservation, where beneficial alleles could be distinguished because of their higher mutability. Lastly, we show that according to the model a cohort of cognitively healthy centenarians is enriched for beneficial alleles.

Committee members
Erwin Bakker (Universiteit Leiden)
Henne Holstege (Vrije Universiteit Amsterdam)
Marc Hulsman (Vrije Universiteit Amsterdam)
Frank Jacobs (Universiteit van Amsterdam)
Marcel Reinders (Technische Universiteit Delft)
David Tax (Technische Universiteit Delft)

Contents

1 Introduction

Life expectancy is highly variable between individuals. Some live on to extreme ages without any severe health problems, while other die early from disease. Since longevity is in part determined by a genetic component[31, 34], these differences in longevity point towards the fact that genetic variations might protect the former of the two groups against diseases while other variants might make the latter group more susceptible to disease. This insight inspired research trying to associate variants with phenotypical traits. The most common way of investigating associations between variants and traits is through a genome-wide association study(GWAS). In a GWAS one takes a group showing a certain trait and a group not showing this trait and one tries to find variants that occur more often in one group than the other. However, the GWAS method has some limitations, the most important being statistical significance. To still be able to investigate variants that are not significant researchers started trying to predict whether a variant was likely to have an effect, solely based on the properties of a variant. This way genetic researchers have an extra indication on whether they should put more effort into certain variants or whether they are better off spending their valuable time and resources on other things. Most of the variant effect predictors focus on predicting deleterious alleles since most genetic research is aimed at finding disease-related variants.

The first effort in creating an annotation for deleteriousness was Sorting Intolerant From Tolerant (SIFT) by Kumar et al.[44, 56]. This algorithm scored amino acid substitutions as tolerable or as intolerable based on the conservation of the amino acid in different species. The intuition being that highly conserved bases are more likely to be deleterious when changed since all the alternatives were filtered out by natural selection for a reason. Genomic Evolutionary Rate Profiling (Gerp)[12] later applied the same intuition only on a nucleotide level, by looking at the expected mutation rate and comparing it to the actual mutation rate to see if the site is more conserved than expected. Some algorithms that try to predict deleterious alleles take a different approach where they predict the effect on the protein itself[5] or combine this with the conservation measures[1, 66]. Recently a new type of annotation has been introduced based on combining different types of annotation[41, 35]. Combined annotation dependent depletions (CADD) is particularly interesting in this class. The creators of this method defined an evolution-based model for defining variants that were probably filtered out by natural selection and variants that were actively selected for and used machine learning to see which properties differentiate these two classes. A big advantage of this approach is that a large data set can be created and the method also works for non-coding regions. All these methods were aimed at detecting deleterious and disease-causing alleles. However, surprisingly, there are no efforts in trying to predict the protectiveness of variants. Especially, since more and more research started looking into longevity and specific variants protecting us from disease it is becoming more important to predict these beneficial alleles.

Even more so, as this difference can be understood from evolutionary theory as pioneered by Darwin[13], Wallace[86] and, Fisher[21]; i.e. alleles encoding for improved fitness fixate[23] and those decreasing fitness will die out[23] (Figure 1A). Since disease susceptibility is a large factor in longevity[48, 15, 22] one can expect genetic variations that influence disease susceptibility to also have an effect on fitness, therefore we opted for using an evolution-

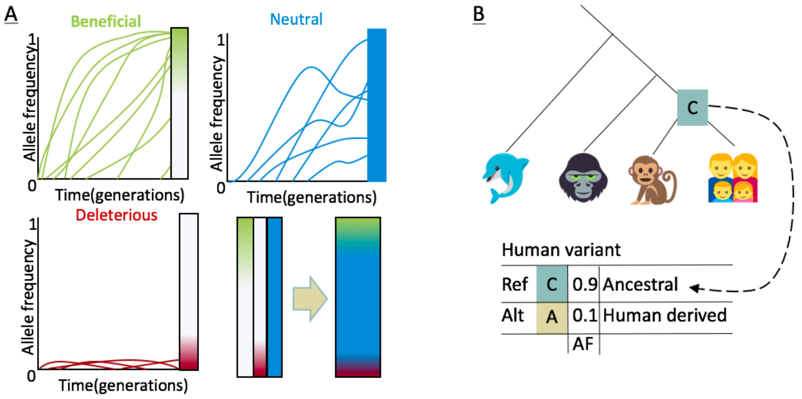


Figure 1: A) shows how the allele frequencies of the three types of variants are expected to develop over time. On the right of each type, it shows the expected distribution at the any given time. In the bottom right corner, we see what the distribution would be if we overlay them. B) A visualization of the human-derived allele frequency. The allele of the common ancestor is compared to the reference and alternative allele of a human variant to see which of the two is ancestral and which is human-derived. The allele frequency of the human-derived allele is used.

based model for predicting neutral, deleterious, but also protective variants. Evolution itself does not care much for longevity since it is only influenced by reproductive fitness. However, lowering disease susceptibility will increase not only longevity but also reproductive fitness, especially when the disease starts before the end of the reproductive age range. Even if this is not the case, one can still often observe the 'grandmother effect', that hypothesizes that based on a kin-fitness principle children whose grandparents are still alive and vital are better integrated into society and therefore have an evolutionary advantage[74].

The frequency at which relatively new alleles occur in the population is indicative of the effect that it has on the evolutionary fitness in humans. Therefore, we can simulate three different types of variants based on their allele frequency: beneficial alleles, neutral alleles, and deleterious alleles. We might be able to learn how to predict the effect of new alleles without any allele frequency information, by learning about the different properties of these three types of alleles.

In other words, one could look at the frequency of alleles and use these for training a model. More specifically, use the human-derived allele frequency (DAF), i.e. the frequency of alleles not yet present in the common ancestor humans and other primates(figure 1B).

Based on evolutionary arguments, one might expect beneficial variants to have a DAF of almost 1 since they fixate relatively rapidly and one might expect deleterious alleles to have a low frequency since the balance between new mutations and negative selection pressure keeps them at a low frequency[23](Figure 1A). The fate of neutral alleles is mainly determined by genetic drift[39, 58, 23], therefore, neutral alleles can be found with all kinds of DAF-values (Figure 1A). Since beneficial variants gravitate towards fixation and deleterious alleles variants keep a low-frequency one can expect the frequencies between the extreme ends of the spectrum to be enriched for neutral variants (Figure 1A). By creating a training set of variants with a high DAF (beneficial), a medium-high DAF (neutral) and low DAF (deleterious) we wanted to try to predict for newly introduced alleles where they might end up on the DAF-spectrum and thereby predict their effect on fitness and disease. Like for deleterious variants, we would like to be able to predict whether a newly detected variant will be protective or not. Or, to understand what makes it that one variant is protective while the other is pathogenic. Based on numerous genetic and genomic features we set out to find these differences.

2 Results

A training set was created based on variants extracted from the ExAC[49] exome database, containing all the variation in the human exomes of 60,000 alleles (See section 4.1). We devised a labeling approach based on the intuitions as laid out in the introduction using the DAF of each of the variants in ExAC. The DAF was calculated by using the ancestral genome of the chimp and humans that was inferred in the Ensembl Compara 75 release[30]. The DAF was calculated as follows:

$$DAF = \begin{cases} 1 - \text{AF}, & \text{if anc} = \text{alt} \\ AF, & \text{otherwise} \end{cases} \tag{1}$$

An extremely high DAF (> 0.999) is indicative for a beneficial allele. An extremely low DAF (< 0.001) is indicative for deleterious alleles. Variants not encountering any selection pressure (neutral variants) are enriched in DAFs between the extremes ($0.2 < \text{DAF} < 0.8$). A way of inspecting how much sense the labeling makes is comparing the probability density function for a random variant belonging to a certain allele frequency to the principles of the Neutral evolution model of as proposed by Fumio Tajima[79]. If we look at figure 2 we see a couple of things that stand out. First, we see a part in the middle that is pretty much flat, which is what one would expect for neutral variants based on the neutral model of evolution[79]. This model states that the chance of observing a neutral variant scales with $\frac{1}{i}$, where i is the allele count. To create a probability density function, one would have to multiply this probability with the observed alleles, which is i resulting in $\frac{1}{i} \times i = 1$. This means that the probability density function in a neutral model scales as a constant,

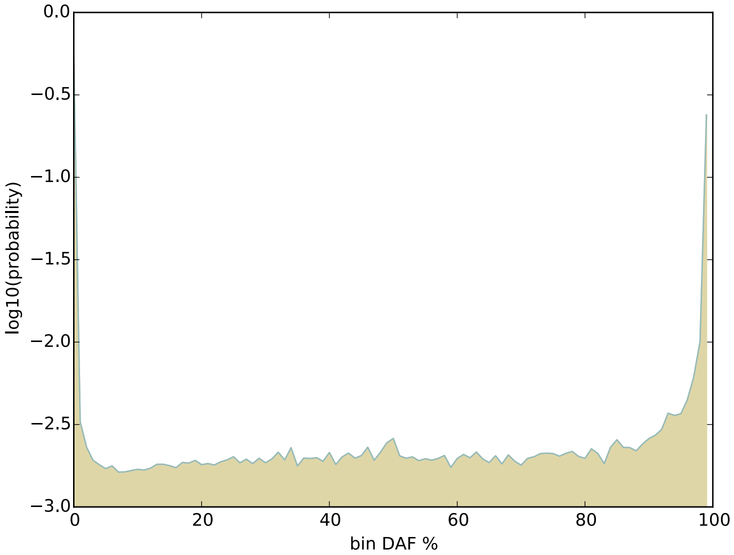


Figure 2: Probability density distribution for a random variant belonging to a certain human derived allele frequency bin in the ExAC dataset.

just as the middle part of our function. It can be seen that the 'flat part' of the spectrum corresponds to the part being labeled neutral($0.2 < \text{DAF} < 0.8$). The model assumes a stable population size. If the number of rare variants is larger than expected by Tajimas neutral evolution model, that means a recent population expansion. This observation is in agreement with the fact that the world population has drastically increased in recent times[3]. The weak selection pressure of the last ages caused by this population explosion might cause deleterious alleles to not be filtered out as efficiently as they would be otherwise, therefore the peak on the left is enriched for deleterious alleles. On the right we see a drastic increase compared to the middle part, that can be explained by the recent decrease in population size that forewent the population increase previously mentioned (i.e. a bottleneck)[3], in that case, these would be the positively selected variants[79]. Based on these principles of population genetics our labeling also is sound, since the thresholds for labeling overlaps with the observations of figure 2.
For all labeled variants, we collected features. The same features as used in CADD[41]. The created training set was used for training a multinomial logistic regression classifier for prediction.

2.1 Variant categories can be separated with good accuracy

Next, we were interested in whether we could differentiate between the three classes. For that, we trained a multinomial logistic regression. To predict the variant class we made use of a large collection of features (see table 7). First, we wanted to assess the test performance of the classifier, i.e. see how well the classifier could separate the different labels in a dataset it was not trained on. An area under the receiver operator characteristics curve (AUROC) analysis was used for this (See section 4). Since an AUROC can only be defined for the classification of two classes we trained a one-versus-rest classifier for all classes and averaged these, this gives an estimation of the actual multinomial performance[45]. Figure 3 shows the AUROC performance in separating the different classes. With an average AUC of 0.74, the model performs better than random classification(AUC=0.50), showing that the three classes are separable. It can be seen that the neutral variants are less separable from the other classes, indicating overlap of the neutral class with the other two classes.
We expected the beneficial and deleterious class to be heavily contaminated with neutral variants, so we wanted to see what the performance would be if we were to minimize this contamination. Figure 4 shows the ROC-performance of the posteriors for the beneficial and deleterious class for all ExAC variants when the variants are grouped based on their neutral posterior probability (See section 4). This figure shows that the lower the posterior for neutral, the higher the classification performance. With results ranging from good performance (AUC of 0.74 for the 0.41-1.0 neutral posterior bin) to almost perfect classification for the (0.98 for the 0.0-0.18 bin). The lower the neutral posterior the less contaminated we expect the group to be with neutral variants, therefore these results indicate that the contamination of neutral variants does, in fact, have a large effect on classification.

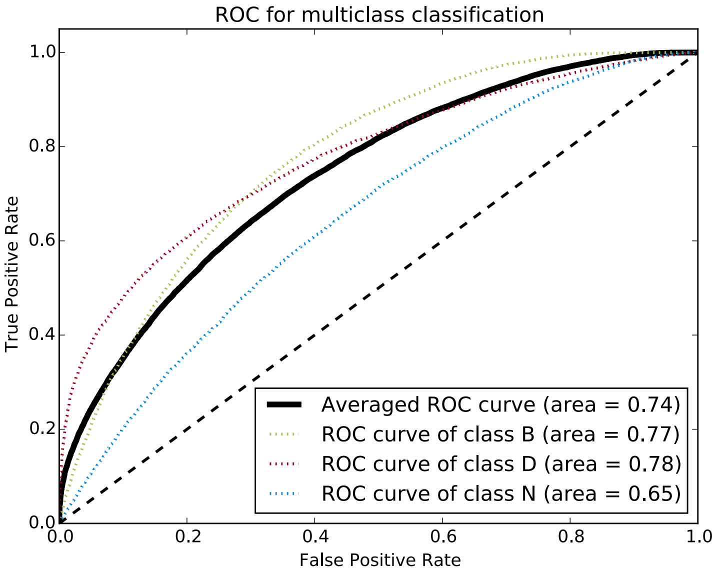


Figure 3: The ROC performance of the one-versus-rest classifiers for all the different classes, and their average ROC

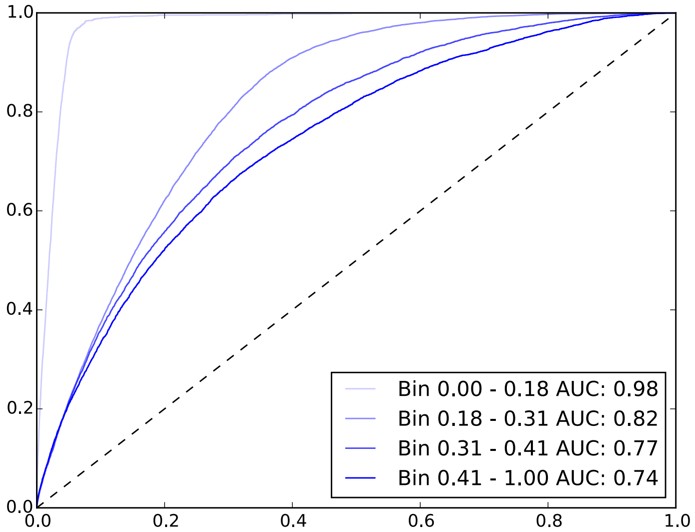


Figure 4: ROC-curves for differentiating beneficial and deleterious variants in different neutral posterior probability categories.

2.2 Validation

The fact that the classifier can separate the different classes is encouraging, however, that does not yet say anything about how the model translates to the classification of real clinically relevant variants. To investigate this, we applied the model to benign and pathogenic variants in the ClinVar[46] database.

2.2.1 ClinVar pathogenic variants predicted with high precision

The distribution of the posterior probabilities for the benign variants and the pathogenic variants in ClinVar can be seen in figure 5. Pathogenic variants score extremely high on the posterior for deleteriousness when compared to the benign ClinVar variants. The benign variants in ClinVar generally have a higher posterior for the neutral class than in the pathogenic variants in ClinVar. This could be expected since the benign variants had no known effect, just like the neutral class is defined as having no effect. Table 1 shows the classification of variants in the ClinVar database using a simple classification based on highest posterior of the multinomial model. Pathogenic variants are classified accurately in the deleterious category. For the benign variants one would hope that they get overwhelmingly

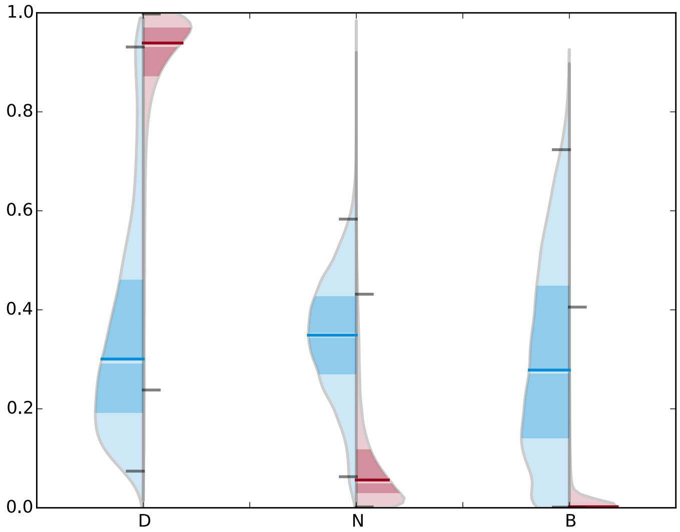


Figure 5: A split violin plot showing the posterior probability of the different classes for the benign variants (blue) and pathogenic variants (red) in ClinVar.

	Benign	Pathogenic	Protective
Number of variants	18835	26263	25
Beneficial	6731 (35,7%)	604 (2,3%)	12 (48%)
Neutral	7210 (38,3%)	602 (2,3%)	9 (36%)
Deleterious	4894 (26,0%)	25057 (95,4%)	4 (16%)

Table 1: Classification of the ClinVar database

classified as Neutral, however, although this is the most occurring classification, the misclassification rate was still 62%. Note that the benign variants are variants that did show a phenotypical effect, but an effect that was non-pathogenic. Consequently, it might be that our predictor still is right in classifying an effect, either deleterious or beneficial. Since the number of protective variants in ClinVar was low and the quality of the protective traits was poor, it was decided to use separate methods for validating the beneficial class.

2.2.2 Beneficially predicted alleles associate to beneficial traits

In order to do a validation of the beneficial variants, we inspected the traits that associate to some of the most beneficially predicted variants and for contrast we also showed the traits associated to some of the most deleteriously predicted variants in GWAS studies (See section 4). Table 5 and table 6 show the hits in the GWAS catalog[51] that have a posterior probability of 0.8 or higher for the beneficial class and the deleterious class respectively present in ExAC. The manual classification by the authors of the trait is given in the first column. Table 2 shows the number of variants after each filtering step and the percentage of truly beneficial traits and deleterious traits for both predicted classes. Both classes associate mostly to their appropriate effect. However, note that the number of associations is relatively low, especially for the beneficial class (N=9). It can be seen that the GWAS catalog had more results for the deleterious variants than the beneficial traits, probably having to do with the fact that more variants had a deleterious posterior of 0.8 or higher than there were variants with a beneficial posterior of 0.8 and higher or the fact that deleterious phenotypes are more of the topic of research. If we look at the ratios of it can be seen that most of the beneficial variants found were also judged to be beneficial whereas most of the deleterious variants were also associated to deleterious traits. These results suggest that the beneficial class as defined in the model does predominantly code for beneficial traits.

	Beneficial variants	Deleterious variants	Neutral variants
Posterior>0.8	130,344	472,982	321,496
RS numbers	20,301	46,209	51,264
In GWAS catalog	23	55	51
Manually assigned classification	9	35	-
Assigned beneficial	6 (67%)	3 (33%)	-
Assigned deleterious	2 (6%)	33 (94%)	-

Table 2: Number of variants for each step in the GWAS catalog analysis.

2.2.3 Beneficially predicted variants often located at selection sweep sites

Selective sweeps occur during a bottleneck, where the population rapidly decreases, increasing selection pressure. This leaves a pattern in the genomes of a population for all positively selected variants. These patterns can be recognized by algorithms specialized in detecting these selective sweeps. The algorithm used for this analysis is called SweepFinder[57, 62]. Figure 6 shows the probability density function for the likelihood of being located on a selection sweep site for beneficially, deleteriously, and neutrally predicted variants in ExAC as estimated by the multinomial classifier(See methods). What we wanted to find out by creating this plot is if beneficially predicted variants would be on locations of which we also see a higher likelihood of positive selection. As can be seen in the figure especially the higher likelihoods seem to be enriched for variants that were also predicted to be beneficial and the lower likelihoods seem to be enriched for deleterious alleles. A simple Mann-Whitney-U test showed that the beneficial alleles are not significantly different from the deleterious class (P=0.16), however, the plot still shows the what one would hope to see if the classifier were to work correctly.

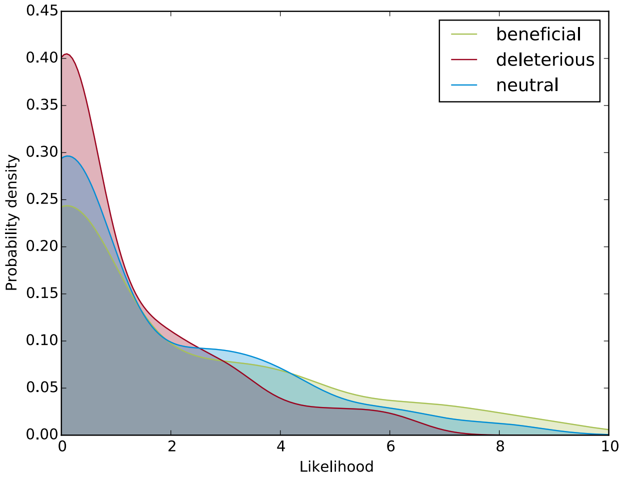


Figure 6: Probability density function with respect to the likelihood of the variant being in a selective sweep region for the three different predicted classes.

2.3 Deleterious variants overwhelmingly protein changing

After validating the model we were interested in inspecting the differences in properties between the three class labels in our training set, starting with variant types. Figure 7 shows the counts of different variant types in the training set for each class. One can see that the deleterious class exists of a disproportionate amount of nonsynonymous variants, where the beneficial class and intronic variants are mostly neutral. This could be expected from intronic variants, because of their extremely weak effect. Most of the loss of function mutations (stop gained, stop lost, canonical splice site, nonsynonymous) show an enrichment of deleterious mutations. Because of their enormous effect on protein function, this could be expected. Although there also is some literature based on empirical evidence suggesting that protective variants should consist mostly of loss of function variants[28]. All the other variant consequences show an enrichment for both beneficial and neutral variants indicating that at least based on variant consequences neutral variants and beneficial variants are similar. The intergenic variants are enriched for neutral variants. Quite interestingly the synonymous variants seem to make up a higher fraction of the beneficial variant than the neutral variants, which would not be expected based on the fact that synonymous variants have such little

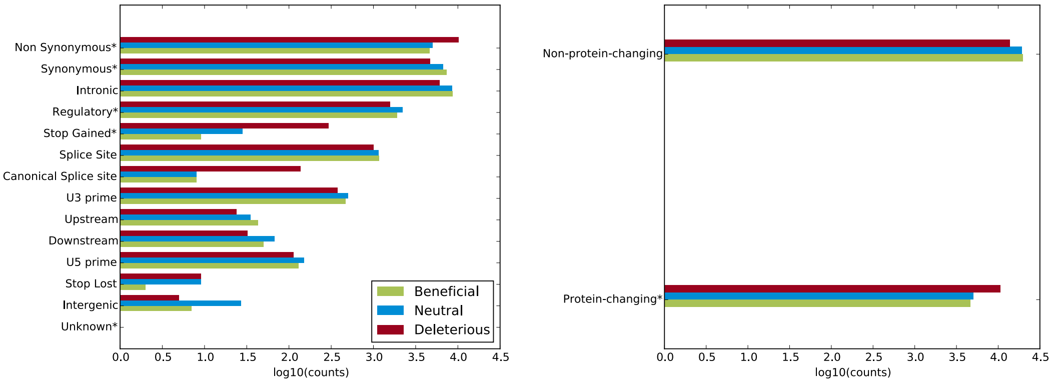


Figure 7: Shows counts of each variant type in the training set separated based on class. Green shows the beneficial class and red the deleterious class and blue the neutral class. The variant types are ordered on the difference between the beneficial and the deleterious count. The variant types with a '*' were all significantly different(See methods). The right side panel shows an aggregation of all protein-changing variant types and all non-protein-changing variant types.

effect. Although they can have quite a large effect on expression and regulation[68].
Another interesting observation is the fact that the first two components of the principal component analysis (PCA) of the training set perfectly separates different variant types. Figure 8 shows the first two principal components of the features in the training set. Clear clusters form. The clusters were color coded and numbered. The variants clustered according

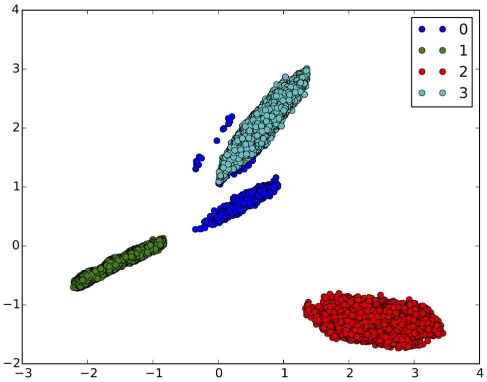


Figure 8: The first two principal components of the training set. The different clusters that formed are color coded.

to variant type: Cluster 0 was entirely made up of synonymous variants, cluster 1 was made up of intronic variants, cluster 2 was made up of stop gained, stop lost and non-synonymous variants and cluster 3 was made up of all other variants types. This indicates that the variants types account for the most variance in the features is actually caused by different variants types. It is interesting to see that all variant types so perfectly separate in different clusters. Especially, if we consider that the indicator variables for the variant types have such a low factor in the linear combination for the principal components.

2.4 Properties that differentiate

We wanted to know how informative different features were in differentiating all the classes at ones, but also how informative variants are in distinguishing between two of the three classes. We trained a classifier on all possible combination of two classes, using just one feature at the time. For each of the different classifiers, the AUROC was determined to see how well that certain feature differentiated the corresponding classes. Figure 9 shows an overview of the differentiating power of all the different feature. It can be seen that all features related to conservation are doing well in differentiating between the beneficial and deleterious class and also in differentiating the deleterious and neutral class. It was already well known that deleterious variants tend to be separable through conservation [41, 56, 1], and this finding confirms that view. It can be seen that not many features are performing well in differentiating between beneficial and neutral variants, indicating that these classes are very alike in many perspectives. On the other hand, mutability, DNA access, and transversions seem to separate beneficial alleles when compared to neutral variants. It can also be seen that the combination of an indicator value for nonsynonymous SNPs and feature values perform well in general, this indicates that non-synonymous variants point towards the deleterious class (See also figure 9). Interestingly, the nucleotide substitution, the amino acid substitution, the expression values, the chromatin state and the functional genomics data add little to no information that can be used for prediction.

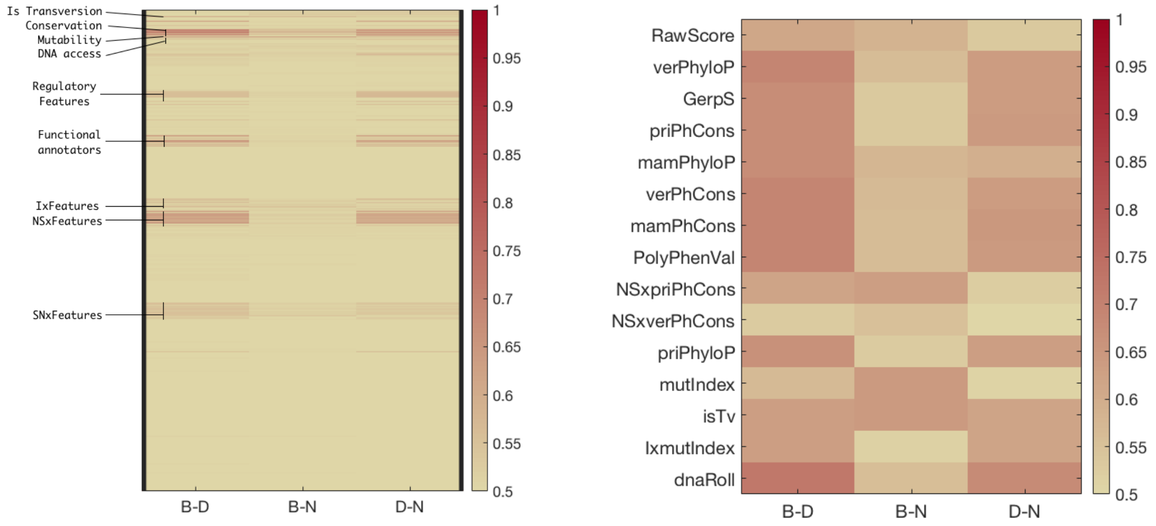


Figure 9: Heat map shows the AUROC scores of all features used in one versus one classification. The AUROC was generated training for the two corresponding classes on the feature alone and testing its AUROC performance. The left figure shows an over of all features, where the right side shows only the features that were in the top 10 for at least one of the differentiations.

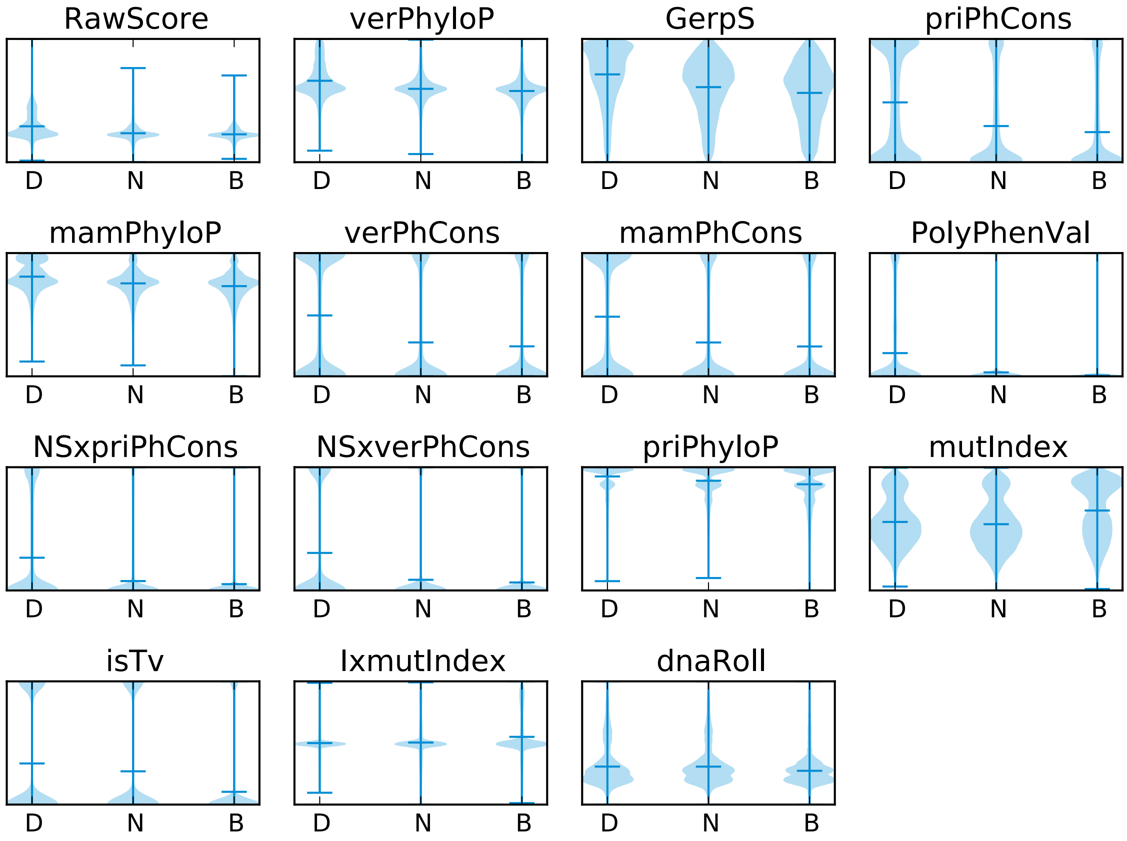


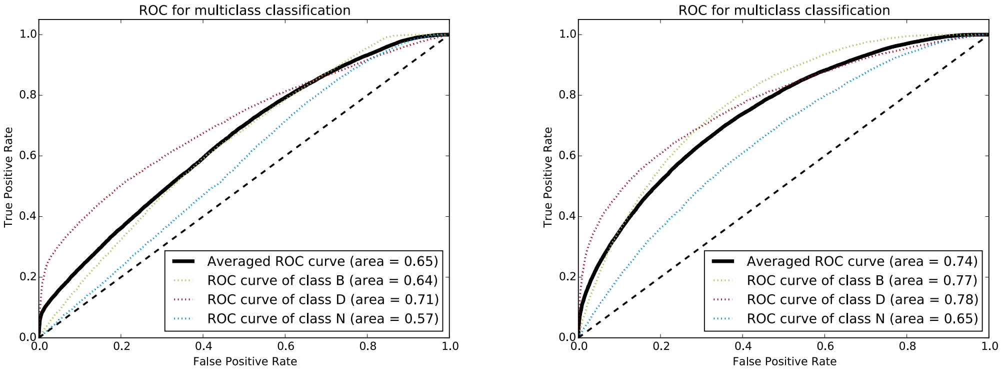
Figure 10: This plot shows the distribution of most informative features from figure 9 values among the different classes: D is the deleterious class, N is the neutral class, and B is the beneficial class. The top features were defined as scoring as one of the 10 most informative features in at least one of the different differentiations. The outer lines show the extremes, horizontal line in the middle shows the mean feature value for that class.

Figure 10 show the violin plots for the feature values for some of the most important features in differentiating the classes. This shows that the CADD score is relatively higher for the deleterious class than for the beneficial class, which makes sense since the CADD training set was relatively similar to our training set. It can also be seen that all the conservation scores show higher values for the deleterious class than for the beneficial class indicating that beneficial variants occur in less conserved regions than the deleterious variants. This could be due to the fact that conserved regions are conserved because of their important function, altering this functions might decrease the fitness of its carrier. It can also be seen that the beneficial variants differ from the other two classes in their mutability. Indicating that more mutable regions are actually more likely to carry a beneficial variants, which could be explained by the fact that more mutable regions have a higher turnover of variants and if we see the selection process as a trial and error process it makes sense that regions with a higher turnover tend to yield more variants that make it to fixation. Lastly, it can be seen that a

combination of the indicator value for non-synonymous variants combined with conservation scores also is a predictor for beneficialness pointing towards the fact that non-synonymous variants actually differ the most of all the possible variant consequences.
Since many of the top features seem to have a lot of overlap we also applied a greedy feature selection algorithm where features get selected iteratively based on how much information they add on top of the previously selected features. The 10 first selected features for each possible differentiation can be found in the supplementary data. It can be seen that the selected features do not differ much from the top performing features. These results again indicate differentiation of the deleterious class through conservation features and differentiation of the beneficial class through mutability.

2.5 The importance of the full feature matrix

To see if our training and labeling scheme had any advantage over just using one of the existing predictors for this three-class problem, we created a separate classification of the variants by taking the most informative features (the CADD RawScore) and try to separate our training data just using this feature. This resulted in figure 11. Just as expected the classifier using the full range of features performs far better than the just using the CADD score. Note, that we are mostly interested in classifications with a low false positive rate, and it is exactly this range where full feature matrix makes a huge difference in the classification for neutral and beneficial variants. To quantify this difference we investigated the partial AUROC (pAUROC) this is the AUROC integrated from a false positive rate of 0 to 0.1. Here we see quite a difference, the CADD raw score alone scored a pAUROC of 0.013 where the full feature matrix scored a 0.019. It's interesting to see that the same problem could not be resolved by just thresholding an existing feature, this indicates that it really is the combination of features that create the classification accuracy.



(a) The AUROC performance of the classifier based on just the CADD raw score

(b) The AUROC performance of the classifier using all features.

Figure 11: Comparison of the AUROC performances

2.6 Comparison to Combined Annotation Dependent Depletion

Because the model in this study bears many similarities to the CADD-algorithm[41] created by Kircher et al, it might be interesting to see how the two compare. The CADD algorithm is also based on a training set that finds its intuition in evolution. The algorithm differentiates two different types of variants: simulated variants and observed variants. Simulated variants were variants generated using a simulation approach that simulates variants that are expected to be in the population by now however that are not found in population studies. The intuition is that these are actively selected against and therefore are deleterious. The observed variants are variants that are not observed in the last common ancestor of the human and chimpanzee, however, are now totally fixated. The idea being that these could at least not be deleterious since otherwise, they would not be able to fixate.

The difference with the labeling of the training set in this study compared to the CADD algorithm are small making it interesting to see if this small change makes a difference. One difference is that the CADD algorithm focuses on variants that were completely filtered out or completely fixated (DAF = 0/DAF = 1.0, respectively), where we only looked at variants that occurred that still occurred in the human population(DAF $<$ 0.001/DAF $>$ 0.999, respectively). Our motivation for doing this was that otherwise, we would train too much on variants that made us different from the other primates, instead of variants that made humans differ from each other. Another aspect of our motivation was the fact that the 'simulated variants' also contained a lot of lethal variants, we want to focus on non-lethal variants since lethal variants would not be witnessed in the human population anyway.

For quantifying the difference between our definition of classes with respect to that of CADD, we trained a logistic regression between them and assessed the area under the ROC curve. Both deleterious classes (CADD and ours), could be differentiated with an AUROC of 0.59. The beneficial classes could be differentiated with an AUROC of 0.64 AUROC.

To inspect the difference in properties, we plotted the mean feature value for beneficial variants according to CADD (B*) and according to our definition (B), and similarly for deleterious variants according to CADD (D*) and to our definition (D). From figure 12 it can be seen that the deleterious classes follow the diagonal more tightly than the beneficial classes, which would be expected based on the AUROCs. Consequently, for the deleterious class, most differences are extremely small. However, for the beneficial class that is not the case. It can be seen that the mutability is far higher in our beneficial class than in the 'observed variants', that the beneficial class contains fewer transversions than the observed variants, and the reference nucleotides are more likely to be G/C in our beneficial class and the alternative is more likely to be A/T. The transversions can be explained by the fact that transversions have a bigger impact and therefore they are more likely to create a larger functional change of that locus. If that change is beneficial, this will mean quicker fixation. The higher G/C's in our beneficial class is also quite interesting, it might have to do with the lower rate of transversions, since G>A and C>T are transitions. The higher mutability in our model can be explained by the fact that highly mutable sites are more likely to be a variant in the human population since even after complete fixation they will still mutate.

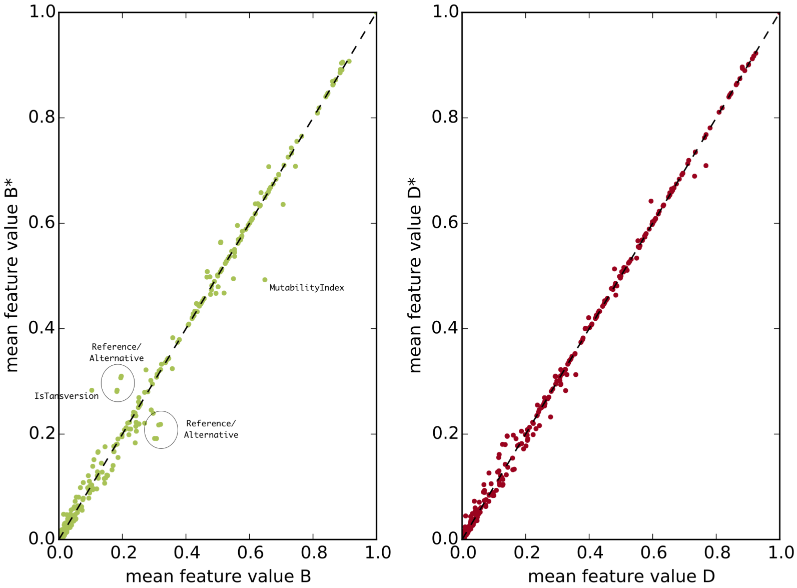


Figure 12: The differences of the closely related class of our model and CADD plotted. Each of the dots represents a feature. The closer to the dotted line the more they are alike in both class definitions.

2.7 Cognitively healthy centenarians enriched with beneficial alleles

In order to apply the model to a real use case, a data set of variants was created by comparing a cohort of early onset Alzheimer's disease (EOAD) patients to a cohort of proven cognitively healthy centenarians (CHC) (See section 4). The variants were called in both cohorts and three different sets of variants were made: 1) a set containing variants unique to the CHCs, 2) a set containing variants unique to the EOAD and, 3) a set of variants present in both cohorts. Figure 13 shows the distribution of the posterior probabilities for the different classes in the model for the variants unique to that EOAD cohort (red) and the variants unique to the CHC cohort (green). It can be seen that the mean for the neutral posterior in both groups is the same. For the deleterious posterior the mean is slightly higher for the EOAD variants, where mean posterior for the beneficial class is slightly higher for the CHCs than the EOAD. The difference for the deleterious class is just above significance and the difference for the beneficial class is just below significance. These results suggest that the variants of the two cohorts have a slightly higher posterior for the appropriate class. Table 3 shows the classification of the different groups of variants. To see if any of these results were enriched or depleted, we did a thousand random permutation procedure on the CHC and EOAD labels to create new groups. We scored these new groups to create a distribution of all classifications for all groups. The 95% extremes of these distributions were used as a normal range any value below that was considered significantly depleted and value above that was considered significantly enriched. The EOAD group is almost significantly enriched for deleterious variants and the CHC seems to be enriched for beneficial variants.

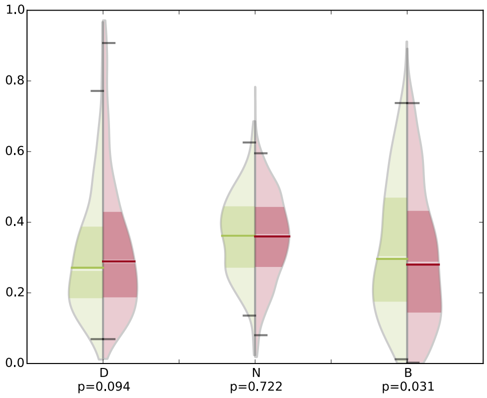


Figure 13: The distribution of the posterior probability of all the classes for the scored variants unique to the AD cohort (red) and unique to the 100plus cohort (green). The p-value for the MannWhitney U test between the left side and right side distribution is given under the distribution.

	Unique to EOAD	normal range	Unique to CHC	normal range	In both groups	normal range
Total	1502		314		201663	
Deleterious	421 (28%)	[22%-28%]	71 (22%)	[20%-23%]	47939 (24%)	[22%-25%]
Neutral	591 (39%)	[37%-41%]	128 (41%)	[39%-42%]	87224 (43%)	[41%-45%]
Beneficial	490 (33%)	[31%-36%]	115 (37%)	[32%-36%]	66500 (33%)	[31%-35%]

Table 3: Classification of the EOAD-CHC comparison variants. The 95% confidence interval for a 'normal population' was determined by using a random permutation procedure on the CHC and AD subjects.

3 Discussion

In this study, we created a model based on evolution to predict the effect of new alleles. We validated that the model often translates to clinically relevant predictions. When looking at the difference in properties between the different variants we see a clear pattern emerge, where deleterious variants are more conserved than other variants, which agrees with most research on deleterious alleles[44, 41]. However, what this research has added is a distinction in the rest of the variants between beneficial variants and neutral variants. Between these classes we see a pattern of mutability determining the fate of the variants, where beneficial variants tend to be located in the more mutable areas to the genome with better DNA-accessibility(Figure 14). What the causality is for mutability being predictive of beneficial variants remains a question. There are two possible explanations: 1) either the beneficial variants are usually located on more mutable areas, or 2) the higher mutability causes a higher chance of fixation because of a higher "genomic turnover", i.e. a faster trial and error process.

The difference seen in posteriors between the CHCs and the EOAD cohort is very small but still suggestive. However, since most of the unique variants will be private variants unrelated to cognition, it is quite interesting to see that there was some signal picked up here. When looking at the classifications itself we actually see a significant enrichment in beneficially

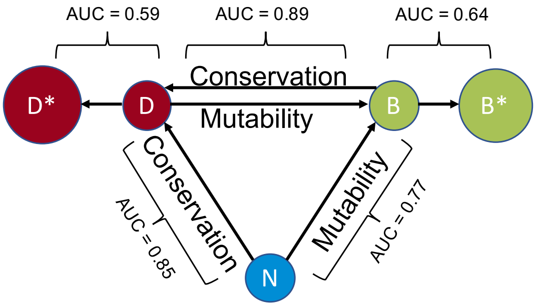


Figure 14: schematic overview of the most important findings. It shows all different classes of the model as well as the 'simulated variants' (D*) and the 'observed variants' (B*) of the CADD model. The AUC for a regression trained between just the two classes is given as well as the most important feature for differentiation.

predicted alleles in the CHCs and an almost significant depletion in deleterious alleles.

We found the deleterious variants to be overwhelmingly protein changing, where the beneficial class and neutral class were more regulatory and non-coding. This goes against the 'less is more' hypothesis[26]. This hypothesis states that evolution is mainly carried out by a loss of gene function. The question is whether these results really contradict each other or whether the loss of function variants in the beneficial class, even though they are rare, have the highest impact on evolution.

Looking at the differences in the classes between CADD and our model, one might wonder if the choice to only look at variants occurring in current populations makes much of a difference, especially when looking at the deleterious class. Looking back we still are of the opinion that our assumption is sound and, even though the difference is very small, the labeling done this way is more elegant than the complex model, and the many assumptions that come with that, needed for creating the CADD classes. A point that needs to be made with regards to this comparison is the fact that our algorithm was only trained on the exome, whereas CADD can score the whole genome because of their different labeling approach.

One disadvantage of the model is that since the model is based on evolution it will not always directly translate to clinical classifications. For example in a case where a variant protects against disease but has a second negative effect that is larger than the fitness gain of the protections, this variant will be negatively selected against. However, as we saw in the validation using the GWAS catalog most of the very certain classifications seem to associate to appropriate traits.

Another point of criticism could be that we could expect the class labels to be contaminated with the other classes. Mostly because of the fact that random genetic drift causes a lot of neutral variants to end up in fixated or be filtered out[39, 40]. To what extent this influences the model performance is hard to figure out and will remain in part unknown for now. The neutral class and the other two classes seem to differ considerably as can be seen in figure 3, which indicates that the other two classes are mostly made up of variants different from the neutral class. What figure 4 shows, however, is that filtering out the variants that look like neutral variants from the other classes improved the classification. This finding shows that there is at least some contamination of neutral variants in the other classes.

Since the number of variants in each of the classes was skewed to an unrealistic degree we decided to fix this by using downsampling to set the number variants equal for each of the classes. In a perfect world, of course, we would have like to set the prior probability for each of the classes to a number that actually represents the chance of observing each of the variant types, however, this is simply not known at the moment. There are some efforts in trying to figure out what is called the 'distribution of fitness effect'[20], however, none of these efforts were detailed enough to be able to determine the distribution of our classes.

In this study, we tried to validate the model with all the possible tools available, however, the protective variants in ClinVar could not be used in a convincing way and the number of actual GWAS hits was quite low. Although the first validation results are encouraging, before any usage in a real clinical genetics environment is possible, the model should be further validated using a larger number validated protective/beneficial variants. One could, for example, turn to the GWAS catalog again and extract all variant decreasing disease risk or having a positive effect on survival use for validation, however, creating such a data set will be very labor intensive.

For now, the genome of the chimpanzee-human last common ancestor was used for inferring the selection on a variant. This means that our reference point is anywhere from 4 million[61] to 13 million[4] years away. A lot can happen in 4 million years; variants can become fixated and be filtered out again over such a large time span. Therefore, it might be interesting to try a similar approach using the genome of human subspecies, such as the Neanderthal[65] or the Denisovan[55], or even to try using ancient human genomes, such as 'Ötzi'[38] or Egyptian mummies[70]. The beneficial variants must have fixated at a higher rate when using these genomes when sticking to the DAF threshold of 0.999 for beneficial labeling, which might result in a cleaner training set.

Another way of possibly improving the model is by using another classification method that allows for the generation of meta-features. Here some manually created combination of features were used, however, this does not make much sense from a machine learning point of view since modern machine learning algorithms, such as neural networks with a deep architecture are perfectly capable of automatically combining features.

Concluding, a model for predicting beneficial alleles was successfully implemented and validated. When inspecting beneficially predicted variants we found that they often associate to beneficial traits in genome-wide association studies. We found that deleterious alleles are different from the other alleles because of their higher conservation, where beneficial alleles could be distinguished because of their higher mutability. Lastly, we showed that according to the model a cohort of cognitively healthy centenarians is enriched for beneficial alleles.

Acknowledgements

I would like to take this opportunity to thank some of the people involved in this project. First of all the people tasked with the supervision of the project. Marc Hulsman, my daily supervisor, who was the one who came up with this project and was always there to answer even some of my most trivial questions. His extremely detailed knowledge of bioinformatics and statistics is very inspiring (although it sometimes fries the brains of us mere mortals). Marcel Reinders who managed to meet with me each week, in spite of his insanely busy schedule, to give feedback and directions for which I'm extremely grateful. Henne Holstege, for letting me be part of her research group for more than a year and making me really feel part of the team, all the while being very hands-on with this project. Her passion and enthusiasm for her 100plus study will always stay with me as an inspiration for my future endeavors.

Secondly, I would like to thank Jasper, Niccolo, Sven and, again, Marc with who I shared the office at the medical faculty of the VU. A special mention goes out to Niccolo for making me pretty much fluent in Italian or well.. at least he tried.

I would like to thank the entire team of the 100plus team for the wonderful year in Amsterdam. You really made me a part of the team. I will always look back on this year with joy in my heart and a huge part of that is the people I've met in this enormously passionate team.

The Delft Bioinformatics lab, of course, could not go unmentioned, the range of different talents in this research group is truly astounding. Every BioTalk meeting gave me at least one new insight.

I also would like to thank Erwin Bakker for his feedback and even more so for mentoring me throughout the masters'.

I would like to thank the defense committee members for taking the time to read this report.

Last but not least these acknowledgments would not be complete without mentioning the two people who were responsible for pulling off the miracle of getting me into university in the first place, my lovely parents: Nico Bettings and Mirjam Henzen. Words fail me to express my gratitude for the countless hours they spend helping me overcome my dyslexia and believing in my competence even when the system did not.

This year has been incredibly valuable: I learned a lot about bioinformatics, the scientific field and most importantly I got to know a lot of wonderful people. I am sure I will always think back to this year very fondly.

References

[1] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[2] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832, 2010.

[3] William Amos and Joe I Hoffman. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1678):131–137, 2010.

[4] Ulfur Arnason, Anette Gullberg, and Axel Janke. Molecular timing of primate divergences as estimated by two nonprimate calibration points. *Journal of Molecular Evolution*, 47(6):718–727, 1998.

[5] Evan H Baugh, Riley Simmons-Edler, Christian L Mueller, Rebecca F Alford, Natalia Volfovsky, Alex E Lash, and Richard Bonneau. Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic acids research*, 44(6):2501–2513, 2016.

[6] Sonja I Berndt, Christine F Skibola, Vijai Joseph, Nicola J Camp, Alexandra Nieters, Zhaoming Wang, Wendy Cozen, Alain Monnereau, Sophia S Wang, Rachel S Kelly, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature genetics*, 45(8):868–876, 2013.

[7] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8):R90, 2010.

[8] Daniel I Chasman, Guillaume Pare, Samia Mora, Jemma C Hopewell, Gina Peloso, Robert Clarke, L Adrienne Cupples, Anders Hamsten, Sekar Kathiresan, Anders Mälarstig, et al. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS genetics*, 5(11):e1000730, 2009.

[9] Anthony G Comuzzie, Shelley A Cole, Sandra L Laston, V Saroja Voruganti, Karin Haack, Richard A Gibbs, and Nancy F Butte. Novel genetic loci identified for the pathophysiology of childhood obesity in the hispanic population. *PloS one*, 7(12):e51954, 2012.

[10] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[11] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

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

[13] Charles Darwin. On the origins of species by means of natural selection. *London: Murray*, 247, 1859.

[14] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12):e1001025, 2010.

[15] Lonneke ML De Lau and Monique MB Breteler. Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535, 2006.

[16] Chuong B Do, Joyce Y Tung, Elizabeth Dorfman, Amy K Kiefer, Emily M Drabant, Uta Francke, Joanna L Mountain, Samuel M Goldman, Caroline M Tanner, J William Langston, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. *PLoS genetics*, 7(6):e1002141, 2011.

[17] Harmen HM Draisma, René Pool, Michael Kobl, Rick Jansen, Ann-Kristin Petersen, Anika AM Vaarhorst, Idil Yet, Toomas Haller, Ayşe Demirkan, Tõnu Esko, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature communications*, 6, 2015.

[18] Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302, 2010.

[19] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

[20] Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, 8(8):610, 2007.

[21] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[22] Oscar H Franco, Ewout W Steyerberg, Frank B Hu, Johan Mackenbach, and Wilma Nusselder. Associations of diabetes mellitus with total life expectancy and life expectancy with and without cardiovascular disease. *Archives of internal medicine*, 167(11):1145–1151, 2007.

[23] Richard Frankham, David A Briscoe, and Jonathan D Ballou. *Introduction to conservation genetics*. Cambridge university press, 2002.

[24] Koldo Garcia-Etxebarria, María Alma Bracho, Juan Carlos Galán, Tomàs Pumarola, Jesús Castilla, Raúl Ortiz de Lejarazu, Mario Rodríguez-Dominguez, Inés Quintela, Núria Bonet, Marc Garcia-Garcerà, et al. No major host genetic risk factor contributed to a (h1n1) 2009 influenza severity. *PloS one*, 10(9):e0135983, 2015.

[25] R Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974.

[26] Liza Gross. When less is more: losing genes on the path to becoming human. *PLoS biology*, 4(3):e76, 2006.

[27] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[28] Andrew R Harper, Shalini Nayee, and Eric J Topol. Protective alleles and modifier variants in human health and disease. *Nature Reviews Genetics*, 2015.

[29] Meian He, Min Xu, Ben Zhang, Jun Liang, Peng Chen, Jong-Young Lee, Todd A Johnson, Huaixing Li, Xiaobo Yang, Juncheng Dai, et al. Meta-analysis of genome-wide association studies of adult height in east asians identifies 17 novel loci. *Human molecular genetics*, 24(6):1791–1800, 2014.

[30] Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J Vilella, Stephen MJ Searle, Ridwan Amode, Simon Brent, et al. Ensembl comparative genomics resources. *Database*, 2016:bav096, 2016.

[31] Anne Maria Herskind, Matthew McGue, Niels V Holm, Thorkild IA Sörensen, Bent Harvald, and James W Vaupel. The heritability of human longevity: a population-based study of 2872 danish twin pairs born 1870–1900. *Human genetics*, 97(3):319–323, 1996.

[32] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473–476, 2012.

[33] Melissa J Hubisz, Katherine S Pollard, and Adam Siepel. Phast and rphast: phylogenetic analysis with space/time models. *Briefings in bioinformatics*, 12(1):41–51, 2010.

[34] Ivan Iachine, Axel Skytthe, James W Vaupel, Matt McGue, Markku Koskenvuo, Jaakko Kaprio, Nancy L Pedersen, Kaare Christensen, et al. Genetic influence on human lifespan and longevity. *Human genetics*, 119(3):312, 2006.

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

[36] Yuan Ji, Daniel J Schaid, Zeruesenay Desta, Michiaki Kubo, Anthony J Batzler, Karen Snyder, Taisei Mushiroda, Naoyuki Kamatani, Evan Ogburn, Daniel Hall-Flavin, et al. Citalopram and escitalopram plasma drug and metabolite concentrations: genome-wide associations. *British journal of clinical pharmacology*, 78(2):373–383, 2014.

[37] Samuel E Jones, Jessica Tyrrell, Andrew R Wood, Robin N Beaumont, Katherine S Ruth, Marcus A Tuke, Hanieh Yaghootkar, Youna Hu, Maris Teder-Laving, Caroline Hayward, et al. Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. *PLoS genetics*, 12(8):e1006125, 2016.

[38] A Keller, A Graefen, M Ball, M Matzas, V Boisguerin, F Maixner, P Leidinger, C Backes, R Khairat, M Forster, et al. New insights into the tyrolean icemans origin and phenotype as inferred by whole-genome sequencing. nat. commun. 3: 698, 2012.

[39] Motoo Kimura. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetics research*, 11(3):247–270, 1968.

[40] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[41] Martin Kircher, Daniela M Witten, Preti Jain, Brian J ORoak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310, 2014.

[42] Mirna Kirin, Aman Chandra, David G Charteris, Caroline Hayward, Susan Campbell, Ivana Celap, Goran Bencic, Zoran Vatavuk, Iva Kirac, Allan J Richards, et al. Genome-wide association study identifies genetic risk underlying primary rhegmatogenous retinal detachment. *Human molecular genetics*, 22(15):3174–3185, 2013.

[43] Yohei Kirino, George Bertsias, Yoshiaki Ishigatsubo, Nobuhisa Mizuki, Ilknur Tugal-Tutkun, Emire Seyahi, Yilmaz Ozyazgan, F Sevgi Sacli, Burak Erer, Hidetoshi Inoko, et al. Genome-wide association analysis identifies new susceptibility loci for behcet's disease and epistasis between hla-b [ast] 51 and erap1. *Nature genetics*, 45(2):202–207, 2013.

[44] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.

[45] Thomas CW Landgrebe and Robert PW Duin. Approximating the multiclass roc by pairwise analysis. *Pattern recognition letters*, 28(13):1747–1758, 2007.

[46] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2014.

[47] Jacqueline M Lane, Jingjing Liang, Irma Vlasac, Simon G Anderson, David A Bechtold, Jack Bowden, Richard Emsley, Shubhroz Gill, Max A Little, Annemarie I Luik, et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nature genetics*, 49(2):274, 2017.

[48] Eric B Larson, Marie-Florence Shadlen, Li Wang, Wayne C McCormick, James D Bowen, Linda Teri, and Walter A Kukull. Survival after initial diagnosis of alzheimer disease. *Annals of internal medicine*, 140(7):501–509, 2004.

[49] Monkol Lek, Konrad Karczewski, Eric Minikel, Kaitlin Samocha, Eric Banks, Timothy Fennell, Anne O'Donnell-Luria, James Ware, Andrew Hill, Beryl Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *BioRxiv*, page 030338, 2016.

[50] Yiwei Liu, Christian A Fernandez, Colton Smith, Wenjian Yang, Cheng Cheng, John C Panetta, Nancy Kornegay, Chengcheng Liu, Laura B Ramsey, Seth E Karol, et al. Genome-wide study links pnpla3 variant with elevated hepatic transaminase after acute lymphoblastic leukemia therapy. *Clinical Pharmacology & Therapeutics*, 2017.

[51] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[52] Salome Mack, Stefan Coassin, Rico Rueedi, Noha A Yousri, Ilkka Seppälä, Christian Gieger, Sebastian Schönherr, Lukas Forer, Gertraud Erhart, Pedro Marques-Vidal, et al. A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *Journal of lipid research*, pages jlr–M076232, 2017.

[53] Hamdi Mbarek, Hidenori Ochi, Yuji Urabe, Vinod Kumar, Michiaki Kubo, Naoya Hosono, Atsushi Takahashi, Yoichiro Kamatani, Daiki Miki, Hiromi Abe, et al. A genome-wide association study of chronic hepatitis b identified novel risk locus in a japanese population. *Human molecular genetics*, 20(19):3884–3892, 2011.

[54] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):122, 2016.

[55] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare De Filippo, et al. A high-coverage genome sequence from an archaic denisovan individual. *Science*, 338(6104):222–226, 2012.

[56] Pauline C Ng and Steven Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.

[57] Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic scans for selective sweeps using snp data. *Genome research*, 15(11):1566–1575, 2005.

[58] Tomoko Ohta. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23(1):263–286, 1992.

[59] Nicholette D Palmer, Mark O Goodarzi, Carl D Langefeld, Nan Wang, Xiuqing Guo, Kent D Taylor, Tasha E Fingerlin, Jill M Norris, Thomas A Buchanan, Anny H Xiang, et al. Genetic variants associated with quantitative glucose homeostasis traits translate to type 2 diabetes in mexican americans: the guardian (genetics underlying diabetes in hispanics) consortium. *Diabetes*, page DB_140732, 2014.

[60] Guillaume Paré, Daniel I Chasman, Alexander N Parker, Robert RY Zee, Anders Mälarstig, Udo Seedorf, Rory Collins, Hugh Watkins, Anders Hamsten, Joseph P Miletich, et al. Novel associations of cps1, mut, nox4, and dpep1 with plasma homocysteine in a healthy population. *Circulation: Cardiovascular Genetics*, 2(2):142–150, 2009.

[61] Nick Patterson, Daniel J Richter, Sante Gnerre, Eric S Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103, 2006.

[62] Pavlos Pavlidis, Daniel Živković, Alexandros Stamatakis, and Nikolaos Alachiotis. Sweed: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular biology and evolution*, 30(9):2224–2234, 2013.

[63] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[64] Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics*, 48(7):709, 2016.

[65] Kay Prüfer. authors (2014). the complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 44.

[66] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–3900, 2002.

[67] Yasunori Sato, Noboru Yamamoto, Hideo Kunitoh, Yuichiro Ohe, Hironobu Minami, Nan M Laird, Noriko Katori, Yoshiro Saito, Sumiko Ohnami, Hiromi Sakamoto, et al. Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel. *Journal of Thoracic Oncology*, 6(1):132–138, 2011.

[68] Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature reviews. Genetics*, 12(10):683, 2011.

[69] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[70] Verena J Schuenemann, Alexander Peltzer, Beatrix Welte, W Paul van Pelt, Martyna Molak, Chuan-Chao Wang, Anja Furtwängler, Christian Urban, Ella Reiter, Kay Nieselt, et al. Ancient egyptian mummy genomes suggest an increase of sub-saharan african ancestry in post-roman periods. *Nature communications*, 8, 2017.

[71] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, 2011.

[72] JR Shaffer, X Wang, E Feingold, M Lee, F Begum, DE Weeks, KT Cuenco, MM Barmada, SK Wendell, DR Crosslin, et al. Genome-wide association scan for childhood caries implicates novel genes. *Journal of dental research*, 90(12):1457–1462, 2011.

[73] Gary S Shapiro, Katja Aviszus, James Murphy, and Lawrence J Wysocki. Evolution of ig dna sequence to target specific base positions within codons for somatic hypermutation. *The Journal of Immunology*, 168(5):2302–2306, 2002.

[74] J Maynard Smith. Group selection and kin selection. *Nature*, 201:1145–1147, 1964.

[75] Ravi F Sood, Anne M Hocking, Lara A Muffley, Maricar Ga, Shari Honari, Alexander P Reiner, and Nicole S Gibran. Genome-wide association study of postburn scarring identifies a novel protective variant. *Annals of surgery*, 262(4):563–569, 2015.

[76] Cassandra N Spracklen, Peng Chen, Young Jin Kim, Xu Wang, Hui Cai, Shengxu Li, Jirong Long, Ying Wu, Ya Xing Wang, Fumihiko Takeuchi, et al. Association analyses of east asian individuals and trans-ancestry analyses with european individuals reveal new loci associated with cholesterol and triglyceride levels. *Human Molecular Genetics*, 26(9):1770–1784, 2017.

[77] Karsten Suhre, Matthias Arnold, Aditya Mukund Bhagwat, Richard J Cotton, Rudolf Engelke, Johannes Raffler, Hina Sarwath, Gaurav Thareja, Annika Wahl, Robert Kirk DeLisle, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications*, 8:14357, 2017.

[78] Robert Szulkin, Robert Karlsson, Thomas Whitington, Markus Aly, Henrik Gronberg, Rosalind A Eeles, Douglas F Easton, Zsofia Kote-Jarai, Ali Amin Al Olama, Sara Benlloch, et al. Genome-wide association study of prostate cancer–specific survival. *Cancer Epidemiology and Prevention Biomarkers*, 24(11):1796–1800, 2015.

[79] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595, 1989.

[80] Hongwei Tang, Peng Wei, Ping Chang, Yanan Li, Dong Yan, Chang Liu, Manal Hassan, and Donghui Li. Genetic polymorphisms associated with pancreatic cancer survival: a genome-wide association study. *International journal of cancer*, 2017.

[81] Chikashi Terao, Takahisa Kawaguchi, Philippe Dieude, John Varga, Masataka Kuwana, Marie Hudson, Yasushi Kawaguchi, Marco Matucci-Cerinic, Koichiro Ohmura, Gabriela Riemekasten, et al. Transethnic meta-analysis identifies gsdma and prdm1 as susceptibility genes to systemic sclerosis. *Annals of the Rheumatic Diseases*, pages annrheumdis–2016, 2017.

[82] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern recognition. *IEEE Transactions on Neural Networks*, 19(2):376–376, 2008.

[83] Thorsten Thye, Fredrik O Vannberg, Sunny H Wong, Ellis Owusu-Dabo, Ivy Osei, John Gyapong, Giorgio Sirugo, Fatou Sisay-Joof, Anthony Enimil, Margaret A Chinbuah, et al. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11. 2. *Nature genetics*, 42(9):739–741, 2010.

[84] Christina G Tise, James A Perry, Leslie E Anforth, Mary A Pavlovich, Joshua D Backman, Kathleen A Ryan, Joshua P Lewis, Jeffrey R O'Connell, Laura M Yerges-Armstrong, and Alan R Shuldiner. From genotype to phenotype: nonsense variants in slc13a1 are associated with decreased serum sulfate and increased serum aminotransferases. *G3: Genes, Genomes, Genetics*, pages g3–116, 2016.

[85] Wiesje M van der Flier, Yolande AL Pijnenburg, Niels Prins, Afina W Lemstra, Femke H Bouwman, Charlotte E Teunissen, Bart NM van Berckel, Cornelis J Stam, Frederik Barkhof, Pieter Jelle Visser, et al. Optimizing patient care and research: the amsterdam dementia cohort. *Journal of Alzheimer's disease*, 41(1):313–327, 2014.

[86] Alfred Russel Wallace. *Contributions to the theory of natural selection: a series of essays*. Macmillan, 1871.

[87] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2013.

[88] Gaoqiang Xie, Phyo Kyaw Myint, Deepak Voora, Daniel T Laskowitz, Ping Shi, Fuxiu Ren, Hao Wang, Ying Yang, Yong Huo, Wei Gao, et al. Genome-wide association study on progression of carotid artery intima media thickness over 10 years in a chinese cohort. *Atherosclerosis*, 243(1):30–37, 2015.

[89] Anatoliy I Yashin, Deqing Wu, Liubov S Arbeeva, Konstantin G Arbeev, Alexander M Kulminski, Igor Akushevich, Mikhail Kovtun, Irina Culminskaya, Eric Stallard, Miaozhu Li, et al. Genetics of aging, health, and survival: dynamic regulation of human longevity related traits. *Frontiers in genetics*, 6, 2015.

[90] H Yu, H Yan, J Li, Z Li, X Zhang, Y Ma, L Mei, C Liu, L Cai, Q Wang, et al. Common variants on 2p16. 1, 6p22. 1 and 10q24. 32 are associated with schizophrenia in han chinese population. *Molecular psychiatry*, 2016.

[91] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale. *Nucleic acids research*, 41(W1):W56–W62, 2013.

4 Methods

4.1 ExAC database

Variants were extracted from the Exome Aggregation Consortium (ExAC) exome database version 0.3.1. ExAC contains the exomes of 60,706 unrelated individuals.[49] The database contains 10,195,872 variants. After applying quality control (QC) 7,140,753 were left. QC existed of passing the following conditions: passing the ExAC QC, a number of covered alleles in the population (allele number) that was sufficiently high (at least 100,000) and a high-quality ancestral base call in Ensemble compara release 75[30].

4.2 Inferring human-derived allele frequency

The allele frequency of the human-derived allele is called the human-derived allele frequency (DAF). The DAF of each variant in ExAC was calculated by taking the allele frequency and comparing the alternative allele (alt) of the variant to the ancestral allele (anc). In the following manner:

$$\mathrm{DAF} = \begin{cases} 1 - \mathrm{AF}, & \text{if anc} = \text{alt} \\ AF, & \text{otherwise} \end{cases} \tag{2}$$

The variant effect for each of the classes was labeled as follows: 1) Beneficial: DAF > 0.999, these variants are relatively new in the population and managed to fixate in the population; 2) Neutral: $0.2 <$ DAF < 0.8, these common variants are expected to have an extremely weak effect; 3) Deleterious: DAF < 0.001, these variants were not actively selected for and therefore maintained a low DAF. This resulted in 29,793 variants being labeled as neutral, 24,818 being labeled as beneficial and being labeled as 7,086,141 deleterious. Since the distribution of the three different classes is unknown and the classes were extremely skewed it was decided to stratify the class sizes using random downsampling. Resulting in 24,818 variants per class.

4.3 Feature collection

For each of the variants a large number of features was collected, mostly extracted from the Ensembl Variant Effect Predictor(VEP, Ensembl Gene annotation v68)[54]. The features included conservation based annotations such as GEPR[12], phastCons[33], phyloP[33], Grantham-score[25], SIFT[44] and, polyphen[1]; regulatory features such as DNase hyperphosphorilation [11], transcription factors[11]; and transcript features such as expression and other functional genomics measures in cell lines. Table 7 in the supplement gives a full overview of all features used in the model and their respective references. Binary features were encoded to 0,1. Non-binary features were scaled to [0,1] using linear min-max scaling. Non-real (such as amino acid substitutions and nucleotide substitution) were represented by a series of features for every obtainable value.

4.4 Model deployment

4.4.1 Training

All variants selected during downsampling were used in training a multinomial logistic regression classification with L2 regularization and a stochastic average gradient descent solver[69]. The optimal lambda-value of the L2 regularization was tested by iterating over different values each 10x bigger or smaller and scored using cross-validation accuracy. The lambda was found to be robust, with an optimal value at the default of 1.
The entire model was implemented using sci-kit-learn[63].

4.4.2 Performance testing

The performance was examined using an area under the receiver operating characteristic curve analysis(AUROC) based on splitting the data in a training set and test set. The splitting of the data set was done using random sampling of half of the variants. For the multinomial class, the average AUROC of the one versus rest classification for all three classes was used.

4.4.3 Feature analysis

To analyze the predictive power of individual features new logistic regression classifiers were trained for each feature: one multivariate to see how well the feature can differentiate between all classes, and three one-vs-one logistic regression classifiers to see how well the feature differentiates between beneficial versus deleterious, beneficial versus neutral, and deleterious versus neutral. Since we expected a lot of overlap between the informativeness of the features we also assessed which feature combinations were the best, ranking the feature combination up to 10 features by using a greedy feature selection approach[82, 27]. Meaning first the most informative feature was selected, after which the feature that was the most informative in combination with the first feature was selected, then a third that is most informative in combination with the first two. This way the algorithm kept adding features until we had the 10 most informative features.

4.5 Assessing training performance

The testing performance was assessed using an area under the receiver operating characteristic curve (AUROC). Since this is a three-class problem and the AUROC expects a two-class problem, we chose to calculate the AUROC for three different one versus rest classifiers were trained (one for each class) and these ROC-curves would be averaged [45]. The AUROCs were created by separating the data into two parts one for training and one for testing performance. For each variant, we can calculate the (estimated) posterior for all three classes (B, D, N) using the three trained classifiers. This allows us to study the performance of differentiating between the beneficial and deleterious class for different levels of neutrality. Hereto, we separated variants according to the different levels of predicted posterior for the neutral class (binned in 5 different categories of equal size based). For all variants, the posteriors of the beneficial and deleterious category were normalized to 1 and then used to create ROC-curves for every category and category-based AUROCs.

4.6 Validation

4.6.1 ClinVar

ClinVar[46] is a database that consists of clinically validated variants. All variants from version 2017-05-30 were checked to make sure that the variants were SNPs and that the variants were within the ExAC capture regions. After which, three different types of variants in the ExAC capture region were extracted from the database: Pathogenic (n=26261), Benign (n=18831) and protective (n=25). All extracted variants were scored using the trained multinomial regression.

4.6.2 GWAS data

We used a GWAS validation approach based on GWAS databases. The variants were filtered to make sure that were in the ExAC capture region and to make sure that the variants only included single nucleotide variations (SNVs). All were scored with a split approach where the variants were randomly sampled to form two groups, a multinomial regression was trained on the one group to score the other and vice versa. All variants with a posterior probability of 0.8 or higher for either the beneficial or deleterious class were extracted. Those variants that had an RS number were queried in the NHGRI GWAS catalog to look for genome-wide significant SNPs.[87, 51] For all the found associations it was manually determined whether the associated trait was beneficial, deleterious or unassigned to each of the classes.

4.6.3 Selective sweeps

All variants in ExAC were randomly split into two splits, a multinomial regression was trained on the one group to score the other and vice versa. The 10,000 variants with the highest posterior for the beneficial class, the 10,000 variants with the highest posterior for the neutral class and the 10,000 variants with the highest posterior for the deleterious class were extracted. The entire 1000 genome project phase 3[10] was scanned for selective sweeps using the SweepFinder[57, 62] algorithm. The algorithm gives a likelihood for each position for belonging to a selective sweep. The distributions of the likelihoods for selection sweeps for the different classes were compared.

4.7 Consequence barplot

For creating the barplot of different variant consequences we simply took the training set and for each of the classes we plotted the number of variants that made up that consequence type. To determine if the difference between the classes for that consequence type were significant we used the beta distribution and calculated if their distributions overlapped for more than 2.5%. The beta distribution of each class for each variants consequence types was given by:

$$f(x, \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{3}$$

, where B is the beta function, α is the number of variants in that class that have the consequence investigated and β is the number of variants in that class not having the consequence. If the all the three distributions had an overlap less than 2.5% probability we marked the consequence as significantly different.

4.8 CHC-EOAD comparison

The CHCs used for this comparison were enrolled in the 100plus study. The 100plus study had the exomes of 217 CHCs. These CHCs were self-reported to be cognitively healthy and underwent neuropsychological testing[1]. The EOAD variants were gathered from the Amsterdam Dementia Cohort(ADC)[85]. The cohort contained the exomes of 373 EOAD cases. A minor allele frequency(MAF) filter was applied to the data set to filter out all the variants with an MAF of 0.005 or lower. Out of the total set of variants three different sets of variants were created: 1) a set of variants unique to the CHCs, 2) a set of variants unique to the ADC, and 3) a set that occurred in both groups. All three groups were scored using the multinomial classifier and the resulting posterior probabilities were analyzed. To see if the cohorts were extreme cases for any of the classifications we used a random permutation procedure. In this procedure, the AD label of CHC label of subjects was randomly permutated resulting in new unique variants. The process was repeated 1000 times to generate the 2.5% extreme and the 97.5% extreme of the resulting distribution.

[1] `www.100plus.nl`

5 Supplementary data

5.1 Derived allele frequency spectrum

Figure 15 shows the DAF spectra for variants that were predicted to be beneficial, deleterious and neutral. It can be seen that the deleterious class is very prevalent in the first percentage bin. However, deleterious variants are depleted in the rest of the spectrum. As expected neutrally predicted variants are enriched in the middle part of the spectrum and depleted in the extreme ends, and beneficial variants are enriched for the high DAF region.

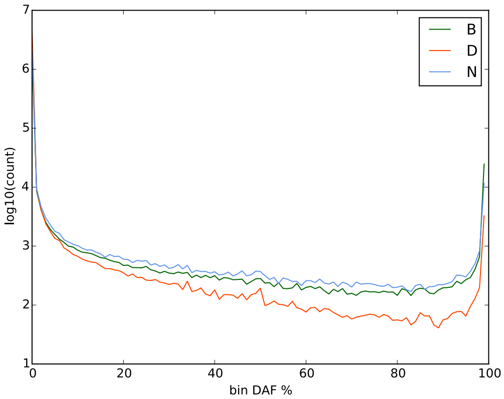


Figure 15: The derived allele frequency spectra for the predicted beneficial variants, deleterious variants, and neutral variants extracted from the ExAC exome database.

5.2 Precision recall

All the single feature analyses in the results were based on the AUROC, however, the AUC does not tell the complete story. It only informs us on how well a single feature predicts for the entire set. Another statistic that could be used is the area under the precision-recall curve (AUPRC), which is more informative in an information retrieval sense. That is, how predictive is the feature for the relevant samples? For example, how predictive is being a nonsynonymous variant for the deleterious class? Figure 16 shows the AUPRC of each of the features for each of the possible class differences. It can be seen that the precision-recall is less clustered than the AUROC.

It is interesting to see that for example some of the amino acid substitutions and nucleotide substitutions have high recall-precision whereas their AUROC was relatively low. This indicates that these features actually contribute a lot to classifier performance for the specific instances.

5.3 Supplementary tables

Name	Type	Description

Ref	Factor	Reference allele
Alt	Factor	Alternative allele
Type	Factor	Variant type (SNV, DEL, INS)
Length	Int	Length of INDEL
IsTv	Boolean	Transversion
Consetquence	Factor	Variant consequence
GC	Num	Percent GC in a window of +/- 75bp
GpC	Num	Percent GpC in a window of +/- 75bp
priPhCons[33]	Num	Primate PhastCons conservation score (excl. human)
mamPhCons[33]	Num	Mammalian PhastCons conservation score (excl. human)
verPhCons[33]	Num	Vertebrate PhastCons conservation score (excl. human)
priPhyloP[33]	Num	Primate PhyloP score (excl. human)
mamPhyloP[33]	Num	Mammalian PhyloP score (excl. human)
verPhyloP[33]	Num	Vertebrate PhyloP score (excl. human)
GerpN[14]	Num	Neutral evolution score by GEPR++
GerpS[14]	Num	GERP++ Rejected substitions score
GerpRS[14]	Num	GERP++ element score
GerpRSpval[14]	Num	GERP++ element p-value
bStatistic	int	Background selection score
mutIndex[73]	Num	Mutability index
dnaHelT[91]	Num	Predicted local DNA structure effect on dnaHelT
dnaMGW[91]	Num	Predicted local DNA structure effect on dnaMGW
dnaProT[91]	Num	Predicted local DNA structure effect on dnaProT
dnaRoll[91]	Num	Predicted local DNA structure effect on dnaRoll
mirSVR-Score[7]	Num	mirSVR-Score
mirSVR-E[7]	Num	mirSVR-E
mirSVR-Aln[7]	Num	mirSVR-Aln
cHmmTssA[19]	Num	Proportion of 127 cell types in cHmmTssA state
cHmmTssAFlnk[19]	Num	Proportion of 127 cell types in cHmmTssAFlnk state
cHmmTx[19]	Num	Proportion of 127 cell types in cHmmTx state
cHmmTxFlnk[19]	Num	Proportion of 127 cell types in cHmmTxFlnk state
cHmmTxWk[19]	Num	Proportion of 127 cell types in cHmmTxWk state
cHmmEnh[19]	Num	Proportion of 127 cell types in cHmmEnh state
cHmmEnhG[19]	Num	Proportion of 127 cell types in cHmmEnhG state
cHmmZnfRpts[19]	Num	Proportion of 127 cell types in cHmmZnfRpts state
cHmmHet[19]	Num	Proportion of 127 cell types in cHmmHet state
cHmmTssBiv[19]	Num	Proportion of 127 cell types in cHmmTssBiv state
cHmmBivFlnk[19]	Num	Proportion of 127 cell types in cHmmBivFlnk state
cHmmEnhBiv[19]	Num	Proportion of 127 cell types in cHmmEnhBiv state
cHmmReprPC[19]	Num	Proportion of 127 cell types in cHmmReprPC state
cHmmReprPCWK[19]	Num	Proportion of 127 cell types in cHmmReprPCWK state
cHmmQuies[19]	Num	Proportion of 127 cell types in cHmmQuies state

EncExp[11]	Num	Maximum ENCODE expression value
EncH3k27Ac[11]	Num	Maximum ENCODE H3K27 acetylation level
EncH3K27Me1[11]	Num	Maximum ENCODE H3K4 methylation level
EncH3K4Me3[11]	Num	Maximum ENCODE H3K4 trimethylation level
EncNucleo[11]	Num	Maximum of ENCODE Nucelosome position track score
EncOCC[11]	Num	ENCODE open chromatin code
EncOCombPval[11]	Num	ENCODE combined p-Value (PHRED-scale)
EncOCDNasePVal[11]	Num	p-Value (PHRED-scale) of Dnase evidence for open chromatin
EncOCFairePVal[11]	Num	p-Value (PHRED-scale) of Faire evidence for open chromatin
EncOCpolIIPval[11]	Num	p-Value (PHRED-scale) of polII evidence for open chromatin
EncOCctcfPval[11]	Num	p-Value (PHRED-scale) of CTCF evidence for open chromatin
EncOmycPval[11]	Num	p-Value (PHRED-scale) of Myc evidence for open chromatin
EncOCDNaseSig[11]	Num	Peak signal for Dnase evidence of open chromatin
EncOCFaireSig[11]	Num	Peak signal for Faire evidence of open chromatin
EncOCpolIISig[11]	Num	Peak signal for polII evidence of open chromatin
EncOCctcfSig[11]	Num	Peak signal for CTCF evidence of open chromatin
EncOCmycSig[11]	Num	Peak signal for Myc evidence of open chromatin
Segway[32]	Factor	Result of genomic segmentation algorithm
tOverlapMotifs	Int	Number of overlapping predicted TF motifs
motifDist	Num	Difference in predicted overlapping motifs between ref and alt
motifEcount	Int	Total number of overlapping motifs
motifEname	String	name of overlapping motifs
motifEHIPos	bool	Position highly informative
morifEscoreChng	Num	VEP score change for the overlapping motif site
TFBS	Int	Number of different overlapping ChIP transcription binding sites
TFBSPeaks	Int	TFBS summed over different celltypes
TFBSPeaksMax	Int	Maximum TFBS across all celltypes
minDistTSS	Int	Distance to closest TSS
minDistTSE	Int	Distance to closest TSE
cDNApos	Int	Base position from transcription start
relcDNApos	Num	Relative position in protein coding sequence
CDSpos	Int	Base position from coding start
relCDSpos	Num	relative position coding sequence
protPos	Int	Amino acid position from coding start
relProtPos	Num	Relative position in protein codon
Domain	String	Domain annotation inferred from VEP
Dst2Splice	Int	Distance to splice site in 20bp

Dst2SplType	Factor	Closest splice site is ACCEPTOR or DONOR
oAA	Factor	Reference Amino Acid
nAA	Factor	Amino acid of observer variant
Grantham[25]	int	Grantham score: oAA, nAA
PolyPhenCat[1]	factor	Polyphen category of change
PolyPhenVal[1]	Num	Polyphen Score
SIFTcat[44]	Factor	SIFT category
SIFTvalue[44]	Num	SIFT score
CADDvalue*[41]	Num	CADD score

Table 7: Adapted version of the CADD feature description in the CADD release notes. '*' indicates features not used in CADD

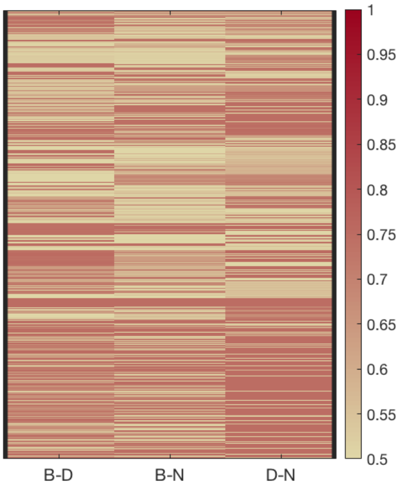


Figure 16: Heat map shows the AUPRC of all features used in one versus one classification. The AUPRC was generated training for the two corresponding classes on the feature alone and testing its AUPRC performance.

Feature	B-D	(CV)	B-N	(CV)	D-N	(CV)
Rank 1	RawScore	0.72	mutIndex	0.64	RawScore	0.70
Rank 2	priPhyloP	0.79	isTv	0.71	GerpRS	0.77
Rank 3	NSxmutIndex	0.81	RawScore	0.73	bStatistic	0.79
Rank 4	SIFTval	0.82	SNxmutIndex	0.75	NSxbStatistic	0.79
Rank 5	SNxpriPhCons	0.82	SIFTval	0.76	verPhCons	0.80
Rank 6	cHmmTssBiv	0.82	RefxG	0.76	dnaProT	0.80
Rank 7	NSxdnaProT	0.83	RefxC	0.77	IxpriPhCons	0.80
Rank 8	SNxdnaProT	0.83	NSxGerpN	0.77	RxpriPhCons	0.81
Rank 9	NSxGerpN	0.83	priPhyloP	0.77	SxpriPhCons	0.81
Rank 10	SxmutIndex	0.84	oAAxK	0.77	SIFTval	0.81

Table 4: Shows the 10 most important features of the greedy feature selection for each of the different one versus one classifications. The CV columns give the cumulative crossvalidation accuracy for the different features.

Manual assignment	Trait		OR/Beta	RS	Ref	P-value
D	Blood protein levels	(Extracellular matrix protein 1)	0.8493	13294	[77]	8,00E-102
	Height		0.06	1351394	[2]	2,00E-65
	Height		1.32	1351394	[2]	2,00E-32
B	Hepatitis B	(Viral clearance)	1.64	9277535-G	[53]	2E-21
D	Blood protein levels	(Ck-beta-8-1)	0.4715	1003645	[77]	2,00E-19
	Hip circumference	(BMI adjusted)	0.0253	1351394	[9]	5,00E-13
	Nose size		0.026	10761129	[64]	7,00E-09
B	Pancreatic cancer	(Survival)	3.3	763780-G	[80]	3,00E-08
B	Postburn scarring	(Severity)	0.23	11136645	[75]	8,00E-08
B	Non-small lung cancer	(Survival)	2.38	1656402	[67]	8,00E-08
	Obesity-related traits	(Weight)	0.03	1056513	[9]	1,00E-07
	Obesity-related traits	(Fat mass)	0.04	1056513	[9]	2,00E-07
	Obesity-related traits	(Trunk fat mass)	0.04	1056513	[9]	2,00E-07
	Obesity-related traits	(Fat free mass)	0.03	1056513	[9]	3,00E-07
	Obesity-related traits	(Lean body mass)	0.03	1056513	[9]	3,00E-07
B	Prostate cancer	(Survival)	1.15	723557	[78]	6,00E-7
D	Carotid intima media thickness	(Multiple-adjusted)		17433780	[88]	2,00E-06
	Obesity-related traits	(Hip circumference)	0.02	1056513	[9]	3,00E-06
	Obesity-related traits	(Total energy expenditure)	0.03	1056513	[9]	6,00E-06
	Obesity-related traits	(Leptin)	0.02	1056513	[9]	7,00E-06
	Obesity-related traits	(Sleep energy expenditure)	0.03	1056513	[9]	7,00E-06
B	Obesity-related traits	(Bone mineral content)	0.02	1056513	[9]	7,00E-06
	Obesity-related traits	(BMI)	0.02	1056513	[9]	8,00E-06

Table 5: All the variants from ExAC found in the GWAS catalog that had a posterior probability for the beneficial class of 0.8 or higher.

Manual assignment	Trait		OR/Beta	RS number	Ref	P-value
D	Acylcarnitine levels	(Nonaylcarnitine)	0.2057	2286963	[17]	3,00E-118
	Blood protein levels	(ERA 1)	1.142	17482078	[77]	3,00E-115
D	Homocysteine levels		0.1583	1801133	[60]	4,00E-104
	Lipoprotein (a) levels		8.63	41272114	[52]	5,00E-86
	Lipid metabolism phenotypes	(VLDL.small, whole)	4.219	676210	[8]	4,00E-64
	Metabolite levels	(C9/C10:2)	0.219	2286963	[17]	3,00E-60
	Lipid metabolism phenotypes	(VLDL.total, whole)	6.384	676210	[8]	9,00E-56
D	LDL (oxidized)		10.5	676210	[8]	3,00E-47
D	Homocysteine levels	(WGHS)	0.05	1801133	[60]	8,00E-35
D	Parkinson's disease		9.62	34637584	[16]	2,00E-28
	Lipoprotein(a)		5.73	41272114	[52]	3,00E-24
D	LDL cholesterol levels	(Trans-ethnic initial)	0.0375	1169288	[76]	2,00E-22
	HDL cholesterol levels	(Trans-ethnic initial)	0.0336	1877031	[76]	1,00E-21
D	LDL cholesterol		0.038	1169288	[76]	6,00E-21
	Gamma gluatamyl transferase levels		0.132	1169288	[76]	2,00E-18
D	Coronary heart disease		1.09	11556924	[71]	9,00E-18
D	Cholesterol, total		0.032	1169288	[76]	4,00E-17
	Blood protein levels	(EA, Cathepsin S)		267738	[77]	9,00E-17
	Response to SRIs in major depressive disorder	(S-DDCT concentration)		1065852	[36]	2,00E-16
	Response to SRIs in major depressive disorder	(S-DDCT/S-DCT ratio)		1058172	[36]	8,00E-16
	Response to SRIs in major depressive disorder	(S-DDCT/S-DCT ratio)		1065852	[36]	8,00E-16
D	LDL cholesterol		1.42	1169288	[76]	1,00E-15
	Cholesterol, total		1.45	1169288	[76]	1,00E-14
D	Glomerular filtration rate (creatinine)		0.0091	267738	[77]	1,00E-14
D	Serum alpha1-antitrypsin levels		1.79	1169288	[76]	2,00E-12
D	Behcet's disease	(Turkish cases with Uveitis)	4.56	17482078	[43]	5,00E-11
D	Coronary artery disease		1.08	11556924	[71]	5,00E-11
	Sleep traits (multi-trait analysis)			12140153	[47]	1,00E-10
D	Schizophrenia		1.17	1051061	[90]	1,00E-10
D	Celiac disease or Rheumatoid arthritis			2298428	[18]	3,00E-10
D	Coronary artery disease		10.989	11556924	[71]	3,00E-10
D	Coronary artery disease or large artery stroke			11556924	[71]	8,00E-10
	Height		0.033	2270518	[29]	8,00E-10
D	Coronary artery disease or ischemic stroke			11556924	[71]	9,00E-10
D	Lipid metabolism phenotypes	(TG.assay, fasting)	0.047	676210	[8]	2,00E-09
B	Lifespan	(females)	5.445	2229188	[89]	2,00E-08
D	Mild influenza (H1N1) infection			41529445	[24]	2,00E-08
	Plasma plasminogen levels		0.056	41272114	[52]	-3,00E-08
	Chronotype		0.039	12140153	[37]	3,00E-08
D	Behcet's disease	(All Turkish cases)	3.08	17482078	[43]	4,00E-08
D	Myocardial infarction		1.07	11556924	[71]	4,00E-08
D	Rhegmatogenous retinal detachment		1.23	267738	[42]	1,00E-07
D	Chronic lymphocytic leukemia		1.29	757978	[18]	1,00E-07
D	Celiac disease		1.13	2298428	[18]	2,00E-07
	Serum sulfate level		0.02	362272	[84]	3,00E-07
D	Excessive daytime sleepiness		0.036	12140153	[47]	7,00E-07
D	Systemic sclerosis	(EA)	0.654	35677470	[81]	9,00E-07
D	ALT levels after remission induction therapy in ALL	(AA)	7.692.308	144122212	[50]	2,00E-06
D	Chronic lymphocytic leukemia		1.46	757978	[6]	3,00E-06
B	Survival in pancreatic cancer		2.46	3795244	[80]	3,00E-06
D	ALT levels after remission induction therapy in ALL	(AA)	8.55	139242087	[50]	3,00E-06
D	Tuberculosis		1.46	1434579	[83]	4,00E-06
	Morning vs. evening chronotype		1.07	12140153	[37]	4,00E-06
	Glucose homeostasis traits	(SG)	2.09	17650440	[59]	6,00E-06
	Height		0.018	6180	[29]	8,00E-06
D	Dental caries		1.33	2302189	[72]	8,00E-06

Table 6: All the variants from ExAC found in the GWAS catalog that had a posterior probability for the beneficial class of 0.8 or higher.