UVIC
UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

Master of Science in Omics Data Analysis

Master Thesis

Machine learning methods in personalized medicine: application to genomic data in Alzheimer's disease.

by

Mireia Ballestà López

Supervisor: Juan Ramón González, ISGlobal

Academic tutor: Josep M. Serrat, University of Vic

Department of Systems Biology

University of Vic – Central University of Catalonia

September 2018

Abstract

The main goal of this project is to validate and compare machine learning methods to perform GWAS analysis. This study worked with genomic data on Alzheimer's disease (AD). The data obtained was imputed by the Michigan Imputation Server and pre-processed by a quality control at both SNPs and individual's level. In order to reduce the dimensionality, SNPs were filtered using different Linkage-Disequilibrium (LD) thresholds (0.2, 0.4 and 0.6). Filtered data was then analysed by five machine learning statistical methods: logistic regression, random forest, k-nearest neighbours, Gradient Boosting Machine and, deep neural networks. The model performance were compared using AUC, sensitivity, specificity and F-measure to evaluate the predictive capacity or reliability of the models. In addition, best models were validated using KEGG pathways. Our conclusion is that best results are obtained when applying a LD threshold of 0.2. From all the five algorithms performed, GBM with a LD threshold 0.2 was seen to be the best model to predict AD based on AUC, sensitivity, specificity, F-measure and validating the results with KEGG pathways.

Keywords: Machine learning, GWAS, Alzheimer's disease.

Introduction

Biological data has exponentially grown, as can be seen in the GenBank database for instance (NCBI, 2008). Nowadays, with the increase of data obtainment, transforming the data in knowledge has become more important (Larrañaga *et al.,* 2006). Machine learning has become useful to obtain such knowledge from large and complex datasets. The algorithms covered in machine learning allows to classify, predict and discover patterns from large and complex amounts of data by building statistical models (Krumholz, 2014).

Machine learning is described as the field of artificial intelligence in computer science which uses statistics to have the ability to "learn" (Sajda, 2006). Learning in the context of machine learning means building a model based on statistics and past experience to infer from data (Kononenko, 2001; Larrañaga *et al.,* 2006, Libbrecht & Noble, 2015; Sajda, 2006). Machine learning can be used to develop genomic medicine by interpreting the genome and predicting phenotypes for instance. In such case, the phenotype is predicted from a set of relevant biomarkers in the sequence of DNA, and hence, a supervised learning problem (Leung *et al*, 2016).

In medical diagnosis, machine learning has an increased importance for its possible capability to diagnose diseases with better sensitivity and/or specificity helping at the decision-making process (Sajda, 2006). In biomedicine, machine learning has some important requirements. First of all, the algorithm requires at least better performance than physicians and other techniques which in some cases is difficult. In other words, the model has to be able to generalise. Machine learning has to be able to deal with noisy data and to subset the important attributes (Baldi & Brunak, 2001; Kononenko, 2001; Libbrecht & Noble, 2015).

Different approaches to implement machine learning on single-nucleotide polymorphism (SNP) data has been made (Cruz & Wishart, 2006; Krishnan & Westhead, 2003; Long *et al.*, 2007; Nguyen *et al*., 2015; Szymczak *et al.*, 2009). Genome-wide association studies (GWAS) get genetic effects to complex diseases or interactions. In consequence, GWAS identify the genetic variants or SNPs that have an effect on the risk to have a certain studied disease. In other words, the carrier of a SNP associated to a disease means a greater chance to suffer from such disease (Kruppa *et al.*, 2012). Machine learning could allow to analyse multiple SNPs simultaneously (Han *et al*., 2012; Nicodemus & Malley, 2009; Szymczak *et al.*, 2009) which is helpful on classification for the risk disease because not all strong associated SNP are good classifiers (Kruppa *et al.*, 2012).

In GWAS, a pre-filtering of SNPs is performed with the quality control. Quality control at SNP's level is based on call rate, minor allele frequency and Hardy-Weinberg equilibrium (Pongpanich *et al.* 2010). The quality control at individual's level are performed on sex

anomalies, relatedness, and population substructure for instance (Turner *et al.*, 2011). In addition, data can be filtered by linkage disequilibrium (LD) between SNPs (Szymczak *et al.*, 2009) refers to the correlation or inheritance of an allele of one SNP to the allele of another SNP within the population due to contiguous position in the genome (Bush & Moore, 2012). Thus, SNPs in linkage disequilibrium can add noise to the analysis by decreasing the variable importance of the true risk SNP (Liu *et al.*, 2013; Meng *et al.*, 2009).

The traditional methods used in statistics for GWAS are regression approaches, Random Forest (RF), Gradient Boost Machine (GBM), K-Nearest Neighbours (KNN), and Artificial Neural Networks (ANN) of one layer (Kruppa *et al.*, 2012; Szymczak *et al.*, 2009). However, other novel statistical methods with better predictive capacity had been emerging as Deep Neural Networks (DNN) which is not yet extended for GWAS purposes. A review on the application of machine learning in GWAS studies was made by Kruppa *et al.* (2012). First of all, the general aims that they found in the published studies were to identify interesting regions based on variable's importance, to find GWAS associations, and to find gene - gene interactions. There were found studies using the entire GWAS data, while other projects filtered by GWAS association for instance.

In this project, different supervised machine learning approaches were used to perform GWAS using Alzheimer's disease (AD) genomic data. Dimensionality reduction was performed based on different LD thresholds. Several machine learning algorithms were studied to predict AD. Some already used methods as GBM, RF and KNN were applied. In addition, this project also assessed the usefulness of DNN. The methodology would allow to evaluate the effects of filtering considering different LD thresholds on model building and the performance of different machine learning algorithms. The comparison between methodologies was performed based on AUC, sensitivity, specificity and F-measure. In order to validate the methodologies, the most reliable models found were analysed in KEGG pathways where AD item was expected as a result. This first insight on AD is problematic because of the difficulty to model the complex relationship between a disease and the whole genotype (Leung *et al*, 2016).

Materials and methods

Alzheimer data

Data was obtained from dbGAP with the study accession phs000168.v2.p2 (NCBI, 2015). The study published in dbGAP was performed by the NIA-LOAD Family Study. The clinical and genotyping data were obtained from individuals with the late-onset form of the AD and, from unrelated individuals of similar age and ethnic background without dementia. Data was composed by 3007 cases and 585082 SNPs.

Data imputation

Genotyped SNPs were first imputed by the Michigan Imputation Server (imputationserver.sph.umich.edu). In order to use the Michigan Imputation Server, the data was converted from Human Genome version 18 (hg18) to hg19 with liftOverPlink (Fujita *et al.*, 2011) and, data was also converted from plink binary files format to Variant Call Format (vcf) with plink (Purcell *et al.*, 2007). Once imputed, the data obtained was processed to obtain the SNPs ID for each position based on the Haplotype Reference Consortium (The Haplotype Reference Consortium, 2018) with bcftools (Li *et al.*, 2009). In addition, the imputed data was transformed from vcf format to plink binary format with plink. The code is accessible in GitHub (2018) in the "Imputation" file.

Data pre-processing

Quality control of SNPs was performed. SNPs with a call rate lower than 95% were removed. In addition, markers with the minor allele frequency lower than 5% and, SNPs not following Hardy-Weinberg equilibrium were also removed. In addition, a quality control at individual's level was also performed. Individuals with a call rate lower than 95% and, individuals with outlying heterozygosity were removed. The heterozygosity was considered outlying when the F-statistic was higher than 10% in absolute terms. Finally, an identity-by-descent (IBD) analysis was applied using a linkage disequilibrium (LD) threshold of 0.2, 0.4 and, 0.6. This step was performed using R version 3.5.0 (2018 – 04 - 23) using snpStats (Clayton, 2015), SNPRelate (Xiuwen *et al*., 2012) and, SNPassoc (Gonzalez & Moreno, 2017) packages. The code is available in GitHub (2018) in the "data_preprocessing" file.

Statistical methods

Two thirds of data were used for training models and the other third, the test set, was used to validate the model. The test set is important to ensure that the model performs the same with new data and not only with the original dataset (Kruppa *et al.*, 2012). The variable case-control was the factor defined to predict by the models. The algorithms used were Logistic Regression (LR), Random Forest (RF), k-nearest neighbours (KNN), Gradient Boosting Machine (GBM) and, Deep Neural Network (DNN). In order to perform the different machine learning study on the data, H2O was used (H2O, 2018), except for KNN which was performed with the R package class (Venables & Ripley, 2015).

From the machine learning algorithms used, LR works as the multiple linear regression obtaining the odds ratio but with a binomial response variable (Sperandei, 2014). RF and GBM are ensemble methods. On one hand, RF uses many decision trees created by bootstrap of the samples (Touw *et al.*, 2012). On the other hand, GBM minimizes loss function by gradient descent and trees (Szymczak *et al.*, 2009). Another algorithm used in the project were KNN which classifies unlabelled samples to the class of the most similar labelled samples (Zhang, 2016). The last algorithm applied in the data was DNN. DNN are mathematical models inspired in biological neuronal systems which classifies numerical data by extracting linear combinations of inputs and modelling as non-linear function (Hastie *et al*., 2001; Nicholson, 2018). In the case of DNN, some previous exploration was required to obtain the best model for the data. The exploration was performed with Random Grid Search (H2O, 2018) to find the best values for the parameters l1, l2, input dropout ratio and hidden dropout ratios. Finally, several models were

built with different number of nodes and hidden layers in order to try to find the global maximum of the system. The script is accessible in GitHub (2018) in the file “Machine_learning”.

Machine learning comparison

The evaluation of the model should be complemented with more than one parameter (Kruppa *et al.*, 2012). Hence, the comparison between the different models obtained from machine learning was performed based on the Area Under the receiver operating characteristic Curve (AUC), sensitivity, sensitivity and, F-measure. Sensitivity (or recall) is described as the proportion of true positives predicted from the real positives, while specificity is the proportion of true negatives predicted from the real negatives. AUC is the area under the curve which represents the trade-off between sensitivity and specificity (Florkowski, 2008). AUC has been seen to be a good parameter to compare machine learning algorithms. For instance, Bradley (1997) in his study concluded the good use of AUC as a single parameter to assess machine learning algorithms. F-measure is the weighted average of precision and sensitivity, being precision the proportion of true positives to the total correctly predicted (Hripcsak & Rothschild, 2005). The script is available in GitHub (2018) in the file “Machine_learning”.

To evaluate if a certain value of AUC is a good result or not, this project based on the traditional point system which goes from 0.5 to 1. An AUC of 0.5 is considered a fail because it means that the result obtained is equal to a random classification, while above 0.7 is considered a good result (Mandrekar, 2010). Nevertheless, when working with genomic data and complex diseases as AD, the genetic epidemiology of the disease has to be taken into account. Wray *et al*. (2010) described heritability and disease prevalence as the main factors that alter the maximum AUC achievable by the perfect predictor model when using genetic predictors. Maximum AUC above 99 % were obtained by diseases with high heritability and low prevalence. In the case of AD, the disease prevalence to know the maximum AUC depended mainly on age. The data for this project contained individuals with a mean age in the group between 75 – 84 years, which had a described prevalence of 17 %. The maximum AUC for 17 % of prevalence in AD is 84% (Escott-Price *et al.*, 2017). In consequence, the resultant AUC was scaled to 84 %.

Enrichment analysis

The enrichment analysis was performed to check if the most important SNPs for the models were associated to a disease with the Kyoto Encyclopedia of Genes and Genomes (KEGG) (KEGG, 2018). This last step of the project would serve as a validation of the methodology for GWAS analysis because AD would have been the expected result (Kanehisa *et al.*, 2009). The validation with enrichment analysis was performed with the variables from the models with an AUC above 70 % for each of the three LD thresholds used. In order to select the most important variables from each model, six different percentile measures were used as a threshold (10 %, 5 %, 1 %, 0.5 %, 0.1 % and, 0.05). The enrichment analysis was performed entirely in R software. The packages used to change the annotation of the SNPs for the analysis and for the enrichment analysis itself were biomaRt (Durinck *et al*., 2005), clusterProfiler (Yu *et al*., 2015), and GOstats (Falcon & Gentleman, 2007). A Disease Ontology Semantic and Enrichment analysis (DOSE) was performed with the same methodology as explained previously with KEGG analysis but using the R package DOSE (Yu *et al*., 2015). The code is available in GitHub (2018) in the file “Enrichment_analysis”.

Results

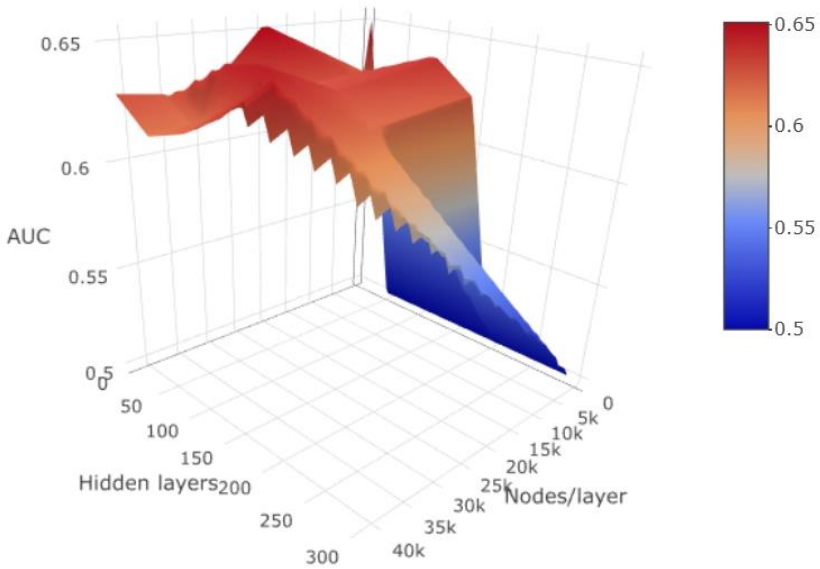
Data pre-processing

In the quality control for SNPs, no SNPs were filtered due to a low call rate nor low minor allele frequency. However, 238 SNPs were filtered because of not following Hardy-Weinberg equilibrium. The quality control at individual's level filtered a total of 15 individuals. The 15 individuals were removed due to outlying heterozygosity, while none individuals were removed due to low call rate. In the IBD analysis, when using a LD threshold of 0.2, the SNPs were reduced to 62807. In the case of LD 0.4 and 0.6, SNPs were decreased to 119922 and 197734 respectively. Finally, a total of 447 samples were removed due to missing phenotype. In other words, 2545 samples were analysed, where 1696 were used to train the models and 849 samples were used to test the model.

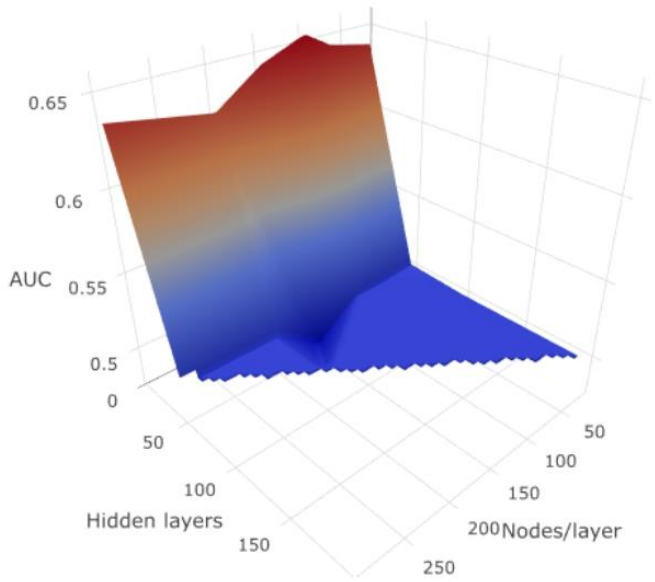
Statistical methods

The Random Grid Search for the different LD threshold obtained different results. In LD 0.2, the Random Grid Search obtained the default parameters as the best ones to predict. In the case of LD 0.4, the Random Grid Search led to an AUC of 56.9 %, while the default parameters gave an AUC of 63.7 %. For LD 0.6, the Random Grid Search was not possible to perform with the resources available. In Figure 1 the results from the different number of hidden layers and nodes was shown. There was a lack of capacity as the machine was found limited at 300 nodes in the case of LD 0.4, and 200 nodes in the case of LD 0.6.

A)



B)



C)

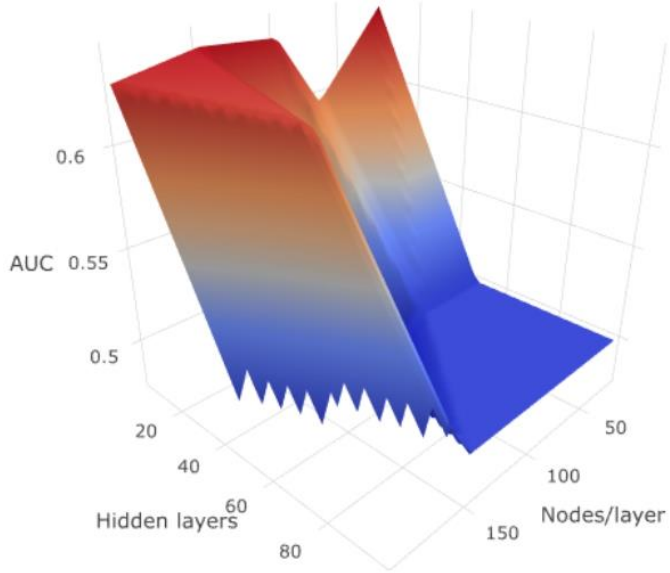


Figure 1. 3D plots showing the AUC of DNN for the different number of hidden layers and nodes in each hidden layer for data filtered by (A) a LD threshold of 0.2, (B) a LD threshold of 0.4 and, (C) a LD threshold of 0.6.

Machine learning comparison

The AUC obtained for the different LD threshold in the IBD analysis and machine learning algorithms were shown in Figure 2 - A. It can be seen that in all the LD thresholds, DNN was one of the two better models to predict. In the case of LD 0.2, the best AUC was obtained from GBM (67.2 %) and DNN (65.2 %). For LD 0.4, the best AUC was obtained from DNN (65.5 %) and LR (64.1 %). LD 0.6 obtained the best AUC from LR (67.0 %) and DNN (63.6 %).

For the DNN, the highest AUC when using data filtered by LD 0.2 was obtained for 20000 nodes and 5 hidden layers. In the case of the data obtained by LD 0.4, the best AUC was found with 100 nodes and 5 hidden layers. For LD 0.6, the higher AUC was found with 150 nodes and 5 hidden layers (Figure 1).

It could be seen different possible trends in which GBM and RF decreased AUC with a rise of LD threshold, while LR increased the resultant AUC with the LD threshold. In general, KNN and DNN seemed to maintain the AUC for the different LD thresholds.

In Figure 2 - B, the corrected AUC could be seen based in a maximum achievable of 84 % AUC due to the heritability and the prevalence of AD. It could be seen that GBM reached 80 % when using an LD threshold of 0.2. LR and DNN were seen to get good models (above 70 %) for all the LD thresholds, while KNN got under 70 % for all LD thresholds. RF obtained good AUC above 70 % for LD thresholds 0.2 and 0.4.

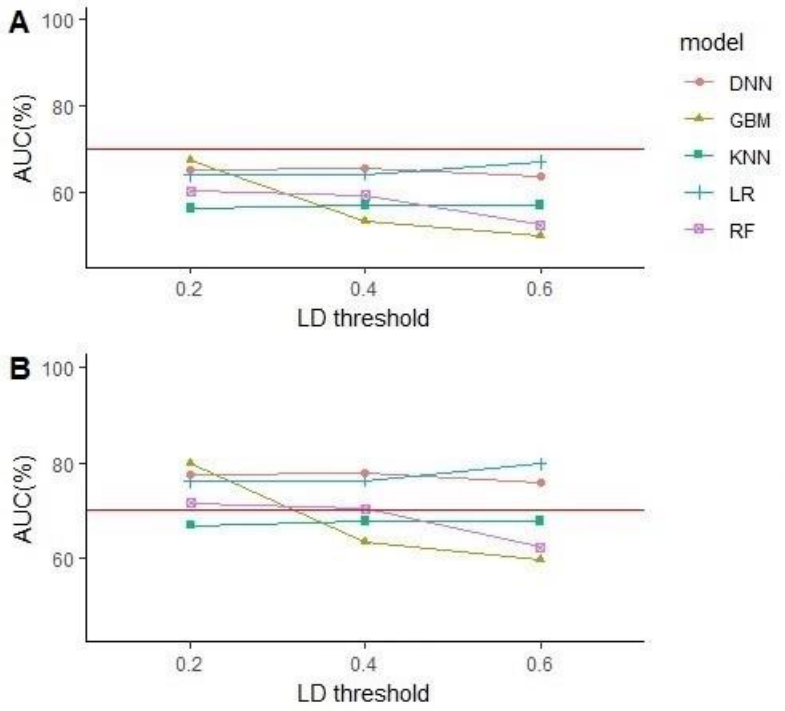


Figure 2. AUC results from the different LD threshold and models used before (A) and after (B) the correction due to heritability and prevalence of AD. The red line indicates the threshold of 70 %.

Sensitivity results were obtained and can be seen in Table S3 in Supplementary materials. The highest sensitivity seen was of 100 % in DNN with LD threshold 0.2, RF with LD threshold 0.6 and, GBM with LD threshold of 0.4 and 0.6 (Figure 3). The models with lowest sensitivity were from KNN which were between 54.50 % and 56.64 %. RF and DNN got a sensitivity above 94.00 % in all the models. While LR and GBM got a sensitivity above 80.00 %. It could be seen that LR decreased sensitivity with LD threshold, while GBM increased sensitivity with LD threshold.

The results of specificity obtained from the models for each LD threshold are presented in Figure 3 (exact values available in Table S4 in Supplementary materials). The highest specificity obtained was 58.08 % from KNN with a LD threshold 0.6 and 0.4 respectively. The lowest specificity was 0.00 % from DNN with a LD threshold 0.2 and GBM with a LD threshold 0.4. KNN got the best specificity in all the models with results between 57.00 % and 59.00 %. GBM got 42.39 % specificity with LD threshold 0.2, nevertheless with LD threshold 0.4 and 0.6 got approximately 0.00 % specificity. LR, RF and DNN got specificities between 0.00 % and 30.00 %. Some trends could be seen for specificity in relation to LD threshold. LR increased specificity with LD threshold, while RF and GBM decreased specificity.

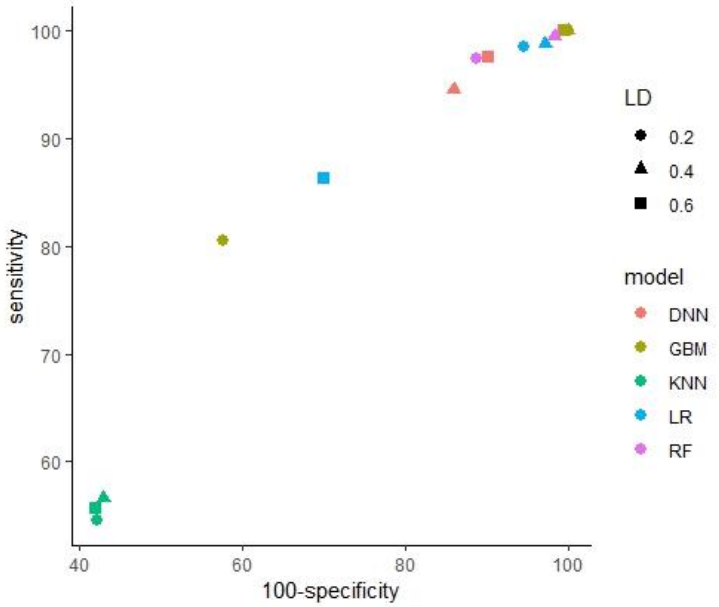


Figure 3. Results of sensitivity and 100 - specificity for each model and LD threshold.

LR, RF, GBM and DNN got an F-measure between 66.00 % and 68.00 % in all the LD thresholds (**¡Error! No se encuentra el origen de la referencia.**). KNN got F-measure between 55.00 % and 57.00 % being the lowest results of the system. All the values of F-measure obtained are available in Table S5 in Supplementary materials.

Enrichment analysis

In order to validate the models, KEGG was used to search AD as the expected result. KEGG was performed for the models with an AUC above 70 % for each LD threshold after correction by heritability and the prevalence of AD (Figure 2 - B). Different percentiles were used to select the SNPs for the analysis based on their importance on the model (In Fig. S1 from Supplementary materials, bar plots of the percentage of importance of the twenty most important SNPs were shown). In Table 1, the appearance of AD in the KEGG's results were shown for different percentiles. AD just appeared as a result for the variables obtained from LD 0.2 except for RF which got validated with a LD threshold of 0.4. AD specifically was a result for GBM with 0.1, 0.05, 0.001, and 0.0005 percentile, and for DNN with 0.01, 0.001 and, 0.0005 percentile. In addition, RF obtained AD as a result with percentile 0.1 with a LD threshold 0.2 and, with percentile 0.1, 0.05, 0.01 and, 0.005 with a LD threshold 0.4. The most successful models were GBM at LD threshold 0.2 and RF at LD threshold 0.4 with four successful results from six percentiles tried. All complete results from KEGG analysis can be found on Table S6 from Supplementary materials.

Table 1. The appearance of AD in the KEGG's results for different percentiles.

LD	Algorithm	0.1	0.05	0.01	0.005	0.001	0.0005
0.2	LR	YES	YES	YES	NO	NO	NO
	RF	YES	NO	NO	NO	NO	NO
	GBM	YES	YES	NO	NO	YES	YES
	DNN	NO	NO	YES	NO	YES	YES
0.4	LR	NO	NO	NO	NO	NO	-
	RF	YES	YES	YES	YES	NO	NO
	DNN	NO	NO	NO	NO	NO	NO
0.6	LR	NO	NO	NO	NO	NO	NO
	DNN	NO	NO	NO	NO	NO	NO

The DOSE analysis for the models with an AUC above 70 % for each LD threshold (Figure 2) found no association with diseases.

Discussion

In this project, data from dbGAP (NCBI, 2015) with 7003 cases and controls of AD and 585082 SNPs were used to explore the best filtering method and machine learning algorithm to predict the AD. The success when modelling depends on several steps or choices like the applied learner, the hyper-parameter optimization, and the data pre-processing (Bischl *et al.*, 2016). The differential filtering method was in the quality control at individual's level based in the identity-by-descent (IBD) analysis using a linkage disequilibrium (LD) threshold of 0.2, 0.4 or, 0.6. After the quality control, 2545 samples were analysed with 62807 SNPs when using a LD threshold 0.2, 119922 SNPs with a LD threshold of 0.4 and, 197734 SNPs with a LD threshold of 0.6. The machine learning algorithms compared were LR, RF, KNN, GBM, and DNN.

In DNN, a hyper-parameter optimization had to be done. The most used strategies are random or manual search. Random search was found the best strategy to optimize the hyper-parameters (Bergstra *et al.*, 2011; Bergstra & Bengio, 2012). The Random Grid Search was used to optimize in DNN the parameters l1, l2, input dropout ratio, and hidden dropout ratios. For the LD 0.2, the Random Grid Search gave the default parameters as optimal. In the case of LD 0.4, the values performed for such parameters got results worse than the results obtained with the parameters in the default value. However, with LD 0.6, the Random Grid Search was not possible to perform mainly due to the machine's capacity. Hence, Random Grid Search seemed to lead to the default parameters. Once the hyper-parameters were set, the number of hidden layers and the number of nodes in each hidden layer were searched. In DNN, the number of hidden layers and number of nodes in each hidden layer is important in order to obtain the maximum accuracy without overfitting the model (Hastie *et al.*, 2001; Karsoliya, 2012). There is no standard formula to calculate the number of hidden layers and number of nodes in each layer. Some rule-of-thumb had been described to suggest the proper number of nodes. For instance, the number of nodes should be 2/3 of the size of the input layer. Another described rules are that the number of nodes should not pass twice the number of nodes in the input layer or, that the number of nodes should be between the size of the input and output layer (Karsoliya, 2012). Nevertheless, in this project, the capacity of the machine made impossible to analyse with the supposed size of hidden layers corresponding to our data. Hence, the peaks in AUC found for DNN in the different filtering could be a local peak of the system instead of the global

AUC is defined as the statistic used to measure the efficacy of classification of a model to predict a phenotype (Wray *et al.*, 2010). In all the three filters, DNN got one of the two best AUC. GBM performed better in LD 0.2, while LR was one of the best models when using LD 0.4 and 0.6 (Figure 2- A). However, due to lack of the correct resources as explained before, the resultant AUC for DNN could be higher. In consequence, results obtained from DNN should be used to orientate as the possibility to obtain a better AUC with this data still exists. As described in material and methods, an AUC above 70 % was considered a good result (Mandrekar, 2010) and it could be seen that none of the results obtained were above 70 %. Nevertheless, taking into account the heritability and the disease prevalence of AD as described by Wray et al. (2010), the maximum AUC achievable by the perfect predictor model when using genetic predictors for AD was 84.00 % (Escott-Price *et al*., 2017). The results based on the maximum of 84.00 % were shown in Figure 2 - B. Based on Mandrekar (2010), an excellent AUC was obtained when using GBM with an LD threshold of 0.2. The models obtained by LR and DNN for all LD thresholds, and the models with RF for LD threshold of 0.2 and 0.4 were acceptable. However, models obtained by KNN for all LD thresholds, RF for 0.6 LD threshold, and GBM for 0.4 and 0.6 threshold were bad models.

When using different machine learning algorithms, different behaviours in relation to the level of filtering was seen (Figure 2- A). Increasing the LD threshold (decreasing the level of filtering), RF decreased AUC, LR increased AUC, and KNN and DNN remained constant. This fact shows that KNN and DNN could be not affected by the filtering performed. On one hand, KNN was defined as a bad method to classify our data. On the other hand, DNN was one of the two best classifiers of the data for all the filters applied. DNN could have maintained AUC with

the increase of LD better than other algorithms due to its weighting the variable importance's property (Nicholson, 2018).

In general, all the algorithms got a good sensitivity and a bad specificity with the exception of KNN (Figure 3). This means that in LR, RF, GBM and DNN, the individuals with AD were well predicted to have dementia, while the control individuals were badly predicted. In other words, there were low number of false negatives but the number of false positives was abundant (Parikh *et al.*, 2008). In the case of KNN, results were approximately the same with a mean of 56.00 %. In the case of F-measure, it quantifies the well predicted samples having into account all the wrongly classified samples (Hripcsak, & Rothschild, 2005). The values from the F-measure obtained were between 55.00 % and 68.00 % due to the high sensitivity and the low specificity. The lowest F-measure could be seen on those models with 100 % sensitivity and less than 1 % specificity. An inversely proportional relationship between sensitivity and sensitivity was seen (Parikh *et al*., 2008). As LR decreased sensitivity with the LD threshold, the specificity increased, and the inverse trend happened with GBM. In global, GBM using a LD threshold of 0.2 got the best specificity (excluding KNN from the system due to the low sensitivity compared to the other algorithms), a high sensitivity and, the best AUC of the system. This would make GBM with LD threshold 0.2 the best model for the data based on AUC, sensitivity and specificity. However, GBM with LD threshold 0.4 and 0.6 got extremely low specificity results.

With the resultant models with an AUC above 70 %, the variable's importance were used to validate the methods with KEGG. It would be expected that the most important variables to classify cases from controls would lead to AD. With this goal, different percentiles were applied to the variable's importance of the different models and LD thresholds in order to explore which model obtained AD as a result from KEGG pathways. AD mainly resulted in the variables obtained from models using data filtered with LD threshold 0.2 (exception of RF). This could be expected because the best AUC results were obtained with 0.2 as LD threshold (Figure 2). In addition, the fact that most of the results that succeeded were obtained with a low LD threshold makes sense because SNPs in linkage disequilibrium could lead to a decrease in variable importance of the true risk SNPs (Meng *et al*., 2009). GBM with an LD threshold of 0.2 was one of the most successful models as expected based on AUC, sensitivity and specificity (Table 1). The fact that GBM got no successful results with LD threshold 0.4 and 0.6 could be due to their extremely low specificity. LR was expected to have bad results because of the rule of thumb where there should be more samples than variables, which in the context of GWAS analysis was difficult (Szymczak *et al.*, 2009). However, LR got validated when using data filtered with a LD threshold of 0.2. RF was the only method that performed better with a LD 0.4 than with a LD 0.2. Nevertheless, RF did not obtain AD as a result with LD 0.6 possibly due to an extremely low specificity (Figure 3) or to the high LD threshold. For the same reason, DNN was expected to have bad results at LD threshold 0.2 because of an extremely low value of specificity. Nevertheless, DNN got high AUC and obtained successful results from KEGG. No conclusions could be performed about DOSE as there were no results. The most important SNPs of the models did not relate with any disease from Disease Ontology (DO), Network of Cancer Gene (NCG) nor DisGeNET databases (Yu *et al.*, 2015).

In summary, machine learning had been proven successful to predict AD from genomic data when applied a low LD threshold when performing the IBD analysis. In addition, GBM was the algorithm which better predicted AD in such conditions, followed by RF with a LD threshold of 0.4. Nevertheless, DNN also obtained good results and should not be diminished because they were indicative due to the available resources. Further work should be done to explore DNN with genomic data to check its potential. In future work, related individuals should be removed and it could also be improved by bootstrapping or cross-validation to reduce possible bias on the performance (Kruppa *et al*., 2012). In addition, it could be interesting to extend the project by adding other types of data to the most important SNPs found. For instance, some work had been successfully done for AD with genomic data and imaging data such as MRI (Magnetic Resonance Imaging) (Li *et al*., 2007; Liu *et* al., 2014; Moradi *et al*., 2015, Zhang & Wang, 2015). Zhang *et al.* (2014).

Conclusion

In this project, machine learning methods were validated to perform GWAS analysis. In general, the best results were obtained using a LD threshold of 0.2. GBM applying a LD threshold 0.2 was seen to be the best model to predict AD based on AUC, sensitivity, specificity and F-measure. To reinforce the idea, GBM got four out of six successful results from KEGG pathways. As successful results with KEGG means that AD appeared in the analysis of the most important variables for the model. Secondly, RF with a LD threshold 0.4 also obtained four out of six successful results from KEGG. However, the AUC and specificity was significantly lower in RF than GBM. In addition, DNN and LR got better AUC and specificity than RF. Both algorithms got AUC above 70 % for all the LD filters, but they obtained three out of six successful results from KEGG in LD threshold 0.2. Nevertheless, DNN should be further studied due to lack of better exploration with the available resources. KNN was the worst algorithm for this data.

Acknowledgements

I would like to thank to Prof. Juan R. González Ruiz from ISGlobal for giving me the chance to perform this project and guiding the project from the beginning. To Carlos Ruiz and Dietmar Fernández also from ISGlobal for their help along my work. To Dr. Josep M. Serrat for his support as academic supervisors form Universitat de Vic.

References

GitHub (2018) *master_thesis/machine_learning*. [online]. Available at: https://github.com/mballesta/master_thesis/tree/master/machine_learning [Accessed 27 August 2018]

Baldi, P. & Brunak, S. (2001). *Bioinformatics: the machine learning approach*. MIT press.

Bergstra, J. S., Bardenet, R., Bengio, Y. & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (pp. 2546-2554).

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13 (Feb): 281-305.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. & Jones, Z. M. (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research*, 17: 5938-5942.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30: 1145-1159.

Bush, W. S. & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8:e1002822. http://doi.org/10.1371/journal.pcbi.1002822

Cruz, J. A. & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2: 59-77.

Clayton, D. (2015). snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.28.0.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21: 3439–3440.

Escott-Price, V., Shoai, M., Pither, R., Williams, J. & Hardy, J. (2017). Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiology of aging*, 49: 214-e7.

Falcon, S. & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23: 257-8.

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl. 1): S83.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D. & Kent, W. J. (2010). The UCSC genome browser database: update 2011. *Nucleic acids research*, *39*(Suppl. 1), D876-D882.

González, J. R. & Moreno, V. (2017). SNPassoc: SNPs-based whole genome association studies. R package version 1.9-6. http://brge.isglobal.org

H2O (2018) *Overview.* [online]. Available at: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html [Accessed 15 March 2018]

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol 1, pp. 337-387). New York: Springer series in statistics.

Han, B., Chen, X., Talebizadeh, Z. & Xu, H. (2012). Genetic studies of complex human diseases: Characterizing SNP-disease associations using Bayesian networks. *BMC Systems Biology*, 6 (Suppl. 3), S14.

Hripcsak, G. & Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12: 296–298. http://doi.org/10.1197/jamia.M1733

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. (2009). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, *38*(Suppl. 1), D355-D360.

Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering Trends and Technology*, 3: 714-717.

KEGG (2018) *KEGG: Kyoto Encyclopedia of Genes and Genomes.* [online] Available at: https://www.genome.jp/kegg/. [Accessed 5 August 2018].

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23: 89-109.

Krishnan, V. G. & Westhead, D. R. (2003). A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*, 19: 2199-2209.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs*, 33: 1163-1170.

Kruppa, J., Ziegler, A. & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. *Human genetics*, 131: 1639-1654.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A. & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*, 7: 86-112.

Leung, M. K., Delong, A., Alipanahi, B. & Frey, B. J. (2016). Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE*, 104: 176-197.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25: 2078-9.

Li, S., Shi, F., Pu, F., Li, X., Jiang, T., Xie, S. & Wang, Y. (2007). Hippocampal shape analysis of Alzheimer disease based on machine learning methods. *American Journal of Neuroradiology*, 28: 1339-1345.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R. & Feng, D. (2014). Early diagnosis of Alzheimer's disease with deep learning. In *Biomedical Imaging: From Nano to Macro,* pp. 1015-1018.

Libbrecht, M. W. & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16: 321.

Liu, J., Wang, K., Ma, S. & Huang, J. (2013). Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. *Statistics and Its Interface*, 6: 99–115. http://doi.org/10.4310/SII.2013.v6.n1.a10

Long, N., Gianola, D., Rosa, G. J., Weigel, K. A. & Avendano, S. (2007). Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *Journal of animal breeding and genetics*, 124: 377-389.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5: 1315-1316.

Meng, Y. A., Yu, Y., Cupples, L. A., Farrer, L. A. & Lunetta, K. L. (2009). Performance of random forest when SNPs are in linkage disequilibrium. *BMC bioinformatics*, 10: 78.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J. & Alzheimer's Disease Neuroimaging Initiative. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*, 104: 398-412.

NCBI (2008). *GenBank Statistics.* [online] Available at: https://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/ [Accessed 22 August 2018].

NCBI (2015). dbGAP [online]. Available at: https://www.ncbi.nlm.nih.gov/gap [Accessed 28 August 2018].

Nguyen, T.-T., Huang, J. Z., Wu, Q., Nguyen, T. T. & Li, M. J. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, 16 (Suppl. 2): S5.

Nicholson, S. C. V. (2018). *A Beginner's Guide to Neural Networks and Deep Learning.* [online] Deeplearning4j.org. Available at: https://skymind.ai/wiki/neural-network [Accessed 22 August 2018].

Nicodemus, K. K. & Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25: 1884-1890.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56: 45–50.

Pongpanich, M., Sullivan, P. F. & Tzeng, J. Y. (2010). A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics*, 26: 1731-1737.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. & Sham, P. C. (2007). PLINK: a toolset for whole-*genome association and population-based linkage analysis. American Journal of Human Genetics,* 81.

Sajda, P. (2006). Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.,* 8: 537-565.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24: 12–18. http://doi.org/10.11613/BM.2014.003

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H. & Sun, Y. V. (2009). Machine learning in genome-wide association studies. *Genetic epidemiology*, 33 (Suppl. 1): S51-S57.

The Haplotype Reference Consortium (2018) *The Haplotype Reference Consortium.* [online] Available at: http://www.haplotype-reference-consortium.org/site [Accessed 20 May 2018].

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M. & van Hijum, S. A. (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. *Briefings in bioinformatics*, 14: 315-326.

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvick, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., Mirel, D. B., Paschall, J. E., Pugh, E. W., Rasmussen, L. V., Wilke, R. A., Zuvich, R. L. & Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, 68: 1-19.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS genetics*, 6: e1000864.

Yu, G., Wang, L., Han, Y. & He, Q. (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." *OMICS: A Journal of Integrative Biology*, 16: 284-287. doi: 10.1089/omi.2011.0118.

Yu, G., Wang, L., Yan, G. & He, Q. (2015). DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis. *Bioinformatics*, 31: 608-609. doi: 10.1093/bioinformatics/btu684, http://bioinformatics.oxfordjournals.org/content/31/4/608.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4: 218. http://doi.org/10.21037/atm.2016.03.37

Zhang, Z., Huang, H. & Shen, D. (2014). The Alzheimer's Disease Neuroimaging Initiative. Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction. *Frontiers in Aging Neuroscience*. 6: 260. (doi:10.3389/fnagi.2014.00260).

Zhang, Y. & Wang, S. (2015). Detection of Alzheimer's disease by displacement field and machine learning. *PeerJ*, 3: e1251.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C. & Weir, B. S. (2012). A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics*; doi: 10.1093/bioinformatics/bts606

Supplementary materials

Table S1. Resultant AUC for the different LD threshold and machine learning algorithms

LD	LR	RF	KNN	GBM	DNN
0.2	63.97	60.12	56.17	67.21	65.21
0.4	64.10	59.12	56.89	53.14	65.50
0.6	67.00	52.28	56.88	49.94	63.64

Table S2. AUC results based on a maximum of 84 % AUC for AD with a prevalence of 17 %. The AUCs above 0.7 are marked in bold.

LD	LR	RF	KNN	GBM	DNN
0.2	**76.16**	**71.57**	66.87	**80.02**	**77.63**
0.4	**76.31**	**70.38**	67.72	63.26	**77.98**
0.6	**79.76**	62.23	67.72	59.45	**75.76**

Table S3. Results of sensitivity for each model and LD threshold.

LD	LR	RF	KNN	GBM	DNN
0.2	98.58 %	97.39 %	54.50 %	80.57 %	100.00 %
0.4	98.82 %	99.53 %	56.64 %	100.00 %	94.55 %
0.6	86.26 %	100.00 %	55.69 %	100.00 %	97.63 %

Table S4. Results of specificity for each model and LD threshold.

LD	LR	RF	KNN	GBM	DNN
0.2	5.39 %	11.24 %	57.85 %	42.39 %	0.00 %
0.4	2.81 %	1.64 %	57.14 %	0.00 %	14.05 %
0.6	29.98 %	0.47 %	58.08 %	0.23 %	9.84 %

Table S5. F1 scores obtained from the confusion matrix of the models tested for each LD threshold.

LD	LR	RF	KNN	GBM	DNN
0.2	66.99 %	67.82 %	55.29 %	67.46 %	66.40 %
0.4	66.51 %	66.56 %	56.64 %	66.40 %	67.17 %
0.6	67.10 %	66.51 %	56.22 %	66.46 %	67.60 %

Fig. S1. Percentage of importance to build the model for the 20 SNPs most important for the models obtained with an AUC score above 70 % for each LD threshold used.

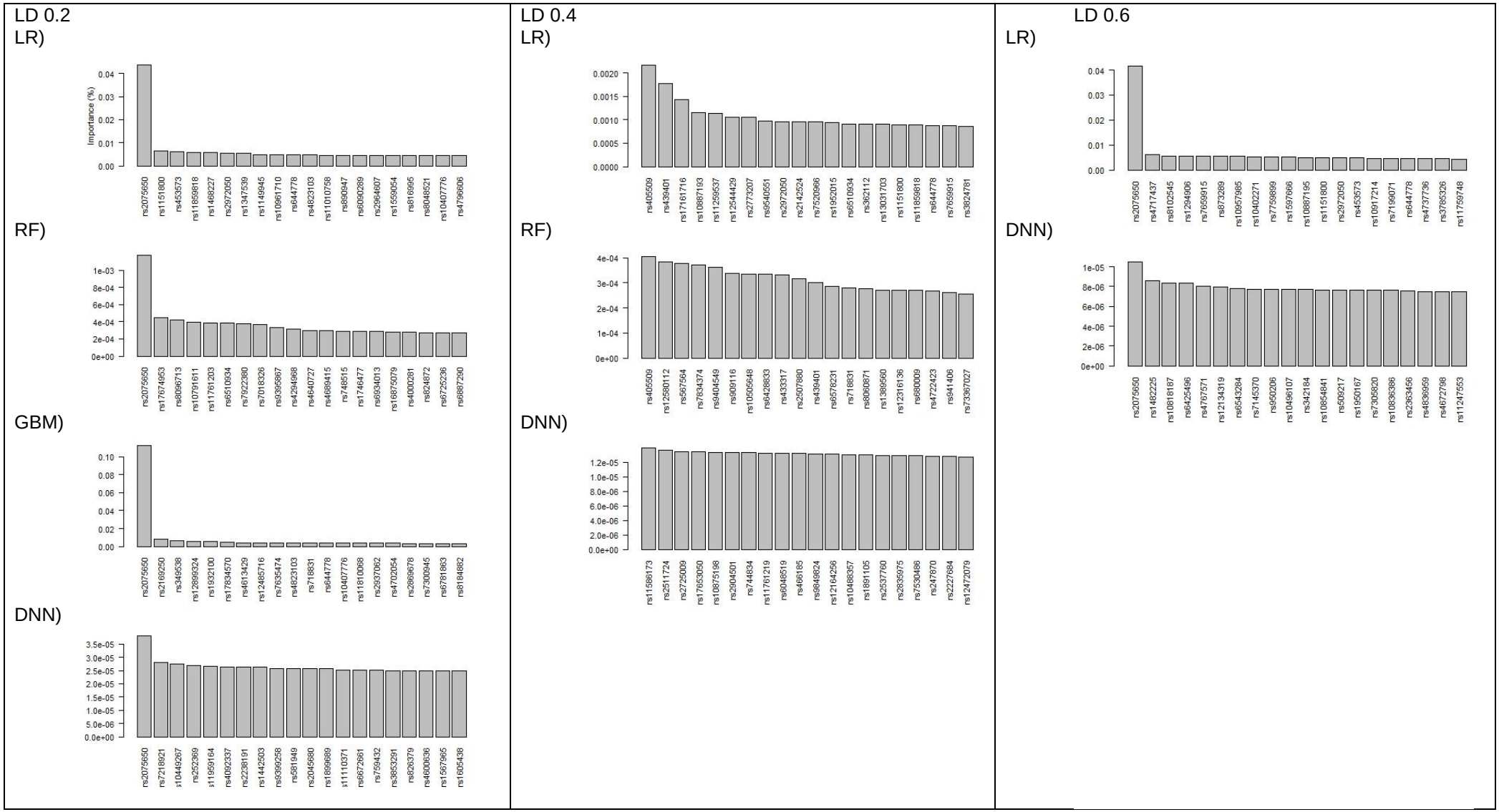


Table S6. Results from the KEGG pathways when using the models with an AUC score above 70 % for each LD threshold and the different percentiles checked.

LD 0.2
LR
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04970	0.0006392	4.906211	1.9903288	8	49	Salivary secretion
04270	0.0060360	3.270576	2.8027079	8	69	Vascular smooth muscle contraction
04962	0.0075323	6.099010	0.8123791	4	20	Vasopressin-regulated water reabsorption
03440	0.0138353	7.264706	0.5280464	3	13	Homologous recombination
05210	0.0173333	3.707576	1.5435203	5	38	Colorectal cancer
05010	0.0242754	2.945454	2.2746615	6	56	Alzheimer's disease
04070	0.0283348	3.213158	1.7466151	5	43	Phosphatidylinositol signaling system
00770	0.0300300	9.611650	0.2843327	2	7	Pantothenate and CoA biosynthesis
04730	0.0309655	3.129487	1.7872340	5	44	Long-term depression
04972	0.0328640	2.722783	2.4371373	6	60	Pancreatic secretion
04140	0.0389924	8.006472	0.3249516	2	8	Regulation of autophagy
00230	0.0414654	2.355186	3.2495164	7	80	Purine metabolism

LD 0.2
LR
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04970	0.0006392	4.906211	1.9903288	8	49	Salivary secretion
04270	0.0060360	3.270576	2.8027079	8	69	Vascular smooth muscle contraction
04962	0.0075323	6.099010	0.8123791	4	20	Vasopressin-regulated water reabsorption
03440	0.0138353	7.264706	0.5280464	3	13	Homologous recombination
05210	0.0173333	3.707576	1.5435203	5	38	Colorectal cancer
05010	0.0242754	2.945454	2.2746615	6	56	Alzheimer's disease
04070	0.0283348	3.213158	1.7466151	5	43	Phosphatidylinositol signaling system
00770	0.0300300	9.611650	0.2843327	2	7	Pantothenate and CoA biosynthesis
04730	0.0309655	3.129487	1.7872340	5	44	Long-term depression
04972	0.0328640	2.722783	2.4371373	6	60	Pancreatic secretion
04140	0.0389924	8.006472	0.3249516	2	8	Regulation of autophagy
00230	0.0414654	2.355186	3.2495164	7	80	Purine metabolism

LD 0.2
LR
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04970	0.0007860	5.394737	1.5733075	7	49	Salivary secretion
04962	0.0032333	7.867089	0.6421663	4	20	Vasopressin-regulated water reabsorption

04270	0.0058767	3.624788	2.2154739	7	69	Vascular smooth muscle contraction
04972	0.0114277	3.532468	1.9264990	6	60	Pancreatic secretion
00770	0.0192620	12.330864	0.2247582	2	7	Pantothenate and CoA biosynthesis
04140	0.0251523	10.271605	0.2568665	2	8	Regulation of autophagy
05210	0.0318143	3.675354	1.2201161	4	38	Colorectal cancer
05010	0.0320188	3.080694	1.7980658	5	56	Alzheimer's disease
05016	0.0365505	2.962022	1.8622824	5	58	Huntington's disease
00230	0.0412971	2.556687	2.5686654	6	80	Purine metabolism
00561	0.0441370	4.227273	0.8027079	3	25	Glycerolipid metabolism
03430	0.0464224	6.839506	0.3531915	2	11	Mismatch repair
04070	0.0471708	3.197663	1.3806576	4	43	Phosphatidylinositol signaling system

LD 0.2
LR
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00770	0.0061474	23.036364	0.1245648	2	7	Pantothenate and CoA biosynthesis
00561	0.0092234	7.982030	0.4448743	3	25	Glycerolipid metabolism
04970	0.0103641	5.278307	0.8719536	4	49	Salivary secretion
04350	0.0249899	5.298097	0.6406190	3	36	TGF-beta signaling pathway
05146	0.0256100	3.934921	1.1388781	4	64	Amoebiasis
04510	0.0257581	2.921371	2.3133462	6	130	Focal adhesion
04270	0.0326757	3.624908	1.2278530	4	69	Vascular smooth muscle contraction
05100	0.0350835	4.591799	0.7295938	3	41	Bacterial invasion of epithelial cells

LD 0.2
LR
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00604	0.0230096	57.11111	0.0232108	1	10	Glycosphingolipid biosynthesis - ganglio series
04977	0.0275581	46.69091	0.0278530	1	12	Vitamin digestion and absorption

LD 0.2
LR
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00604	0.0153932	95.25926	0.0154739	1	10	Glycosphingolipid biosynthesis - ganglio series
00561	0.0381490	35.51389	0.0386847	1	25	Glycerolipid metabolism
05014	0.0471399	28.34444	0.0479691	1	31	Amyotrophic lateral sclerosis (ALS)

LD 0.2
RF
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0022428	2.069172	17.609284	29	80	Axon guidance
04080	0.0067171	1.640272	31.256480	44	142	Neuroactive ligand-receptor interaction
04020	0.0126744	1.667795	23.772534	34	108	Calcium signaling pathway
05412	0.0141599	2.013797	11.666151	19	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04614	0.0156763	5.948582	1.760928	5	8	Renin-angiotensin system
04960	0.0158393	2.629749	5.723017	11	26	Aldosterone-regulated sodium reabsorption
05410	0.0159888	2.025408	11.005803	18	50	Hypertrophic cardiomyopathy (HCM)
04070	0.0161650	2.131404	9.464990	16	43	Phosphatidylinositol signaling system
04730	0.0203855	2.054250	9.685106	16	44	Long-term depression
02010	0.0256489	2.264386	6.823598	12	31	ABC transporters
05010	0.0259629	1.847715	12.326499	19	56	Alzheimer's disease
04970	0.0275468	1.909420	10.785687	17	49	Salivary secretion
04520	0.0312743	1.915371	10.125338	16	46	Adherens junction
05223	0.0476430	2.071968	6.603482	11	30	Non-small cell lung cancer
04530	0.0478919	1.614521	15.628240	22	71	Tight junction

LD 0.2
RF
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05410	0.0027876	2.740741	6.363636	14	50	Hypertrophic cardiomyopathy (HCM)
04960	0.0033736	3.704228	3.309091	9	26	Aldosterone-regulated sodium reabsorption
02010	0.0036472	3.336319	3.945454	10	31	ABC transporters
00561	0.0096219	3.282390	3.181818	8	25	Glycerolipid metabolism
04080	0.0099602	1.747858	18.072727	28	142	Neuroactive ligand-receptor interaction
04520	0.0101179	2.473928	5.854546	12	46	Adherens junction
04974	0.0109815	2.338608	6.618182	13	52	Protein digestion and absorption
05412	0.0129275	2.279114	6.745455	13	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04530	0.0141455	2.045658	9.036364	16	71	Tight junction
00640	0.0147183	3.791162	2.163636	6	17	Propanoate metabolism
04260	0.0224921	2.509875	4.327273	9	34	Cardiac muscle contraction
00564	0.0257644	2.195058	5.854546	11	46	Glycerophospholipid metabolism
04020	0.0279071	1.699843	13.745454	21	108	Calcium signaling pathway
05222	0.0397271	2.019033	6.236364	11	49	Small cell lung cancer
04512	0.0496420	1.815063	8.018182	13	63	ECM-receptor interaction

LD 0.2
RF
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00564	0.0022813	4.423724	1.8862669	7	46	Glycerophospholipid metabolism
04520	0.0022813	4.423724	1.8862669	7	46	Adherens junction
00561	0.0029074	6.086634	1.0251451	5	25	Glycerolipid metabolism
04530	0.0075981	3.130548	2.9114120	8	71	Tight junction
04510	0.0157462	2.296329	5.3307544	11	130	Focal adhesion
04614	0.0396801	7.926282	0.3280464	2	8	Renin-angiotensin system
04512	0.0420952	2.549474	2.5833656	6	63	ECM-receptor interaction
04360	0.0433184	2.330428	3.2804642	7	80	Axon guidance
05222	0.0480077	2.739649	2.0092843	5	49	Small cell lung cancer

LD 0.2
RF
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00564	0.0027724	5.949307	0.9965184	5	46	Glycerophospholipid metabolism
04510	0.0060734	3.288251	2.8162476	8	130	Focal adhesion
00561	0.0158039	6.450257	0.5415861	3	25	Glycerolipid metabolism
00400	0.0428658	45.963636	0.0433269	1	2	Phenylalanine, tyrosine and tryptophan biosynthesis
04512	0.0457334	3.220339	1.3647969	4	63	ECM-receptor interaction

LD 0.2
RF
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04666	0.0102860	16.84667	0.1609284	2	52	Fc gamma R-mediated phagocytosis
04530	0.0187143	12.11594	0.2197292	2	71	Tight junction

LD 0.2
RF
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04962	0.0306080	44.94737	0.0309478	1	20	Vasopressin-regulated water reabsorption
05014	0.0471399	28.34444	0.0479691	1	31	Amyotrophic lateral sclerosis (ALS)
03015	0.0471399	28.34444	0.0479691	1	31	mRNA surveillance pathway

LD 0.2
GBM
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05414	0.0000953	4.778038	2.8943907	11	58	Dilated cardiomyopathy
05410	0.0001260	5.075630	2.4951644	10	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0002103	4.715654	2.6448743	10	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05010	0.0057966	3.316804	2.7945841	8	56	Alzheimer's disease
04260	0.0058290	4.229965	1.6967118	6	34	Cardiac muscle contraction
05222	0.0097300	3.297814	2.4452611	7	49	Small cell lung cancer
00531	0.0149182	7.285714	0.5489362	3	11	Glycosaminoglycan degradation
04010	0.0154179	2.076726	7.4854932	14	150	MAPK signaling pathway
04020	0.0170025	2.267080	5.3895551	11	108	Calcium signaling pathway
04720	0.0183035	3.189189	2.1458414	6	43	Long-term potentiation
04070	0.0183035	3.189189	2.1458414	6	43	Phosphatidylinositol signaling system
00240	0.0248346	3.374583	1.6967118	5	34	Pyrimidine metabolism
04810	0.0275717	2.087248	5.7887814	11	116	Regulation of actin cytoskeleton
04970	0.0329253	2.737379	2.4452611	6	49	Salivary secretion
04971	0.0359348	2.674058	2.4951644	6	50	Gastric acid secretion
00562	0.0376877	3.540364	1.2974855	4	26	Inositol phosphate metabolism
05146	0.0379279	2.414869	3.1938104	7	64	Amoebiasis
05145	0.0379279	2.414869	3.1938104	7	64	Toxoplasmosis
05110	0.0478277	3.242667	1.3972921	4	28	Vibrio cholerae infection

LD 0.2
GBM
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05414	0.0000953	4.778038	2.8943907	11	58	Dilated cardiomyopathy
05410	0.0001260	5.075630	2.4951644	10	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0002103	4.715654	2.6448743	10	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05010	0.0057966	3.316804	2.7945841	8	56	Alzheimer's disease
04260	0.0058290	4.229965	1.6967118	6	34	Cardiac muscle contraction
05222	0.0097300	3.297814	2.4452611	7	49	Small cell lung cancer
00531	0.0149182	7.285714	0.5489362	3	11	Glycosaminoglycan degradation
04010	0.0154179	2.076726	7.4854932	14	150	MAPK signaling pathway
04020	0.0170025	2.267080	5.3895551	11	108	Calcium signaling pathway
04720	0.0183035	3.189189	2.1458414	6	43	Long-term potentiation
04070	0.0183035	3.189189	2.1458414	6	43	Phosphatidylinositol signaling system
00240	0.0248346	3.374583	1.6967118	5	34	Pyrimidine metabolism
04810	0.0275717	2.087248	5.7887814	11	116	Regulation of actin cytoskeleton
04970	0.0329253	2.737379	2.4452611	6	49	Salivary secretion
04971	0.0359348	2.674058	2.4951644	6	50	Gastric acid secretion
00562	0.0376877	3.540364	1.2974855	4	26	Inositol phosphate metabolism
05146	0.0379279	2.414869	3.1938104	7	64	Amoebiasis
05145	0.0379279	2.414869	3.1938104	7	64	Toxoplasmosis
05110	0.0478277	3.242667	1.3972921	4	28	Vibrio cholerae infection

LD 0.2
GBM
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04810	0.0037232	3.053289	3.7694391	10	116	Regulation of actin cytoskeleton
05410	0.0049694	4.295454	1.6247582	6	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0066426	4.016367	1.7222437	6	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05414	0.0102907	3.622781	1.8847195	6	58	Dilated cardiomyopathy
05222	0.0199439	3.534235	1.5922631	5	49	Small cell lung cancer
00240	0.0229011	4.118333	1.1048356	4	34	Pyrimidine metabolism
05200	0.0322426	2.040379	5.9466151	11	183	Pathways in cancer
00531	0.0474443	6.753387	0.3574468	2	11	Glycosaminoglycan degradation
04070	0.0489435	3.156410	1.3972921	4	43	Phosphatidylinositol signaling system

LD 0.2
GBM
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05410	0.0095168	5.423913	0.8510638	4	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0116612	5.085714	0.9021277	4	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05414	0.0158870	4.605556	0.9872340	4	58	Dilated cardiomyopathy
04512	0.0209754	4.206780	1.0723404	4	63	ECM-receptor interaction
05145	0.0221007	4.135000	1.0893617	4	64	Toxoplasmosis
04350	0.0222201	5.560976	0.6127660	3	36	TGF-beta signaling pathway
04010	0.0391666	2.628290	2.5531915	6	150	MAPK signaling pathway
04810	0.0448148	2.806653	1.9744681	5	116	Regulation of actin cytoskeleton
05222	0.0492344	3.968717	0.8340426	3	49	Small cell lung cancer

LD 0.2
GBM
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05145	0.0015347	17.662764	0.2475822	3	64	Toxoplasmosis
00190	0.0062810	21.208333	0.1237911	2	32	Oxidative phosphorylation
05222	0.0143774	13.446808	0.1895551	2	49	Small cell lung cancer
05410	0.0149455	13.161458	0.1934236	2	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0167081	12.372549	0.2050290	2	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05010	0.0185568	11.671296	0.2166344	2	56	Alzheimer's disease
05414	0.0198362	11.245536	0.2243714	2	58	Dilated cardiomyopathy
04512	0.0231949	10.303279	0.2437137	2	63	ECM-receptor interaction
05146	0.0238936	10.133064	0.2475822	2	64	Amoebiasis
04145	0.0267762	9.503788	0.2630561	2	68	Phagosome

LD 0.2
GBM
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05010	0.0044212	31.18519	0.1083172	2	56	Alzheimer's disease
04973	0.0474656	26.62500	0.0483559	1	25	Carbohydrate digestion and absorption

LD 0.2
DNN
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04020	0.0003021	2.115028	24.190329	40	108	Calcium signaling pathway
04976	0.0006887	2.692627	11.423211	22	51	Bile secretion
04971	0.0013812	2.565628	11.199226	21	50	Gastric acid secretion
04970	0.0027003	2.439085	10.975242	20	49	Salivary secretion
04916	0.0032759	2.321573	11.871180	21	53	Melanogenesis
00532	0.0074658	4.897552	2.687814	7	12	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
00920	0.0077619	8.728223	1.567892	5	7	Sulfur metabolism
04350	0.0183352	2.234594	8.063443	14	36	TGF-beta signaling pathway
04720	0.0190191	2.083021	9.631335	16	43	Long-term potentiation
05414	0.0226793	1.852933	12.991103	20	58	Dilated cardiomyopathy
04270	0.0229656	1.762590	15.454932	23	69	Vascular smooth muscle contraction
04540	0.0254972	1.856647	12.319149	19	55	Gap junction
04972	0.0327254	1.758497	13.439072	20	60	Pancreatic secretion
04070	0.0409501	1.878799	9.631335	15	43	Phosphatidylinositol signaling system
00562	0.0468674	2.185852	5.823598	10	26	Inositol phosphate metabolism

LD 0.2
DNN
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04020	0.0004028	2.338727	14.0379110	27	108	Calcium signaling pathway
00920	0.0006057	16.971299	0.9098646	5	7	Sulfur metabolism
00532	0.0021435	6.796970	1.5597679	6	12	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
04976	0.0041518	2.599295	6.6290135	14	51	Bile secretion
04270	0.0056279	2.251567	8.9686654	17	69	Vascular smooth muscle contraction
04970	0.0077489	2.474114	6.3690522	13	49	Salivary secretion
04070	0.0182770	2.344904	5.5891683	11	43	Phosphatidylinositol signaling system

04971	0.0229902	2.154971	6.4990329	12	50	Gastric acid secretion
05143	0.0284579	3.127273	2.4696325	6	19	African trypanosomiasis
04520	0.0296549	2.141011	5.9791103	11	46	Adherens junction
05414	0.0314083	1.971242	7.5388781	13	58	Dilated cardiomyopathy
00561	0.0351269	2.637116	3.2495164	7	25	Glycerolipid metabolism
05412	0.0352901	1.994580	6.8889749	12	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04744	0.0356755	3.382175	1.9497099	5	15	Phototransduction
04350	0.0362308	2.265036	4.6793037	9	36	TGF-beta signaling pathway
02010	0.0396553	2.360551	4.0294004	8	31	ABC transporters
00562	0.0429031	2.497200	3.3794971	7	26	Inositol phosphate metabolism
04720	0.0444174	2.059862	5.5891683	10	43	Long-term potentiation
04540	0.0457263	1.900086	7.1489362	12	55	Gap junction

LD 0.2
DNN
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05410	0.0021200	5.181818	1.3733075	6	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0028701	4.845172	1.4557060	6	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05414	0.0045413	4.370414	1.5930368	6	58	Dilated cardiomyopathy
04520	0.0077460	4.569475	1.2634429	5	46	Adherens junction
04260	0.0130079	4.943284	0.9338491	4	34	Cardiac muscle contraction
04350	0.0158563	4.630597	0.9887814	4	36	TGF-beta signaling pathway
05010	0.0174663	3.658645	1.5381044	5	56	Alzheimer's disease
04130	0.0348577	8.067633	0.3021277	2	11	SNARE interactions in vesicular transport
04971	0.0465649	3.203115	1.3733075	4	50	Gastric acid secretion

LD 0.2
DNN
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04520	0.0059949	6.268170	0.7473888	4	46	Adherens junction
05412	0.0099093	5.357680	0.8611219	4	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04260	0.0168244	6.233251	0.5524178	3	34	Cardiac muscle contraction
04510	0.0169901	3.251344	2.1121857	6	130	Focal adhesion
05146	0.0189045	4.356140	1.0398453	4	64	Amoebiasis
05143	0.0370081	7.429412	0.3087041	2	19	African trypanosomiasis
05222	0.0438090	4.175585	0.7961315	3	49	Small cell lung cancer
05410	0.0460901	4.085106	0.8123791	3	50	Hypertrophic cardiomyopathy (HCM)
00730	0.0479733	30.987805	0.0487427	1	3	Thiamine metabolism

LD 0.2
DNN

Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04260	0.0044754	26.51042	0.1052224	2	34	Cardiac muscle contraction
05410	0.0095322	17.56250	0.1547389	2	50	Hypertrophic cardiomyopathy (HCM)
05412	0.0106728	16.50980	0.1640232	2	53	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05010	0.0118719	15.57407	0.1733075	2	56	Alzheimer's disease
05414	0.0127034	15.00595	0.1794971	2	58	Dilated cardiomyopathy

LD 0.2
DNN
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04260	0.0016385	53.08333	0.0657640	2	34	Cardiac muscle contraction
05010	0.0044212	31.18519	0.1083172	2	56	Alzheimer's disease
00520	0.0456023	27.79348	0.0464217	1	24	Amino sugar and nucleotide sugar metabolism

LD 0.4
LR
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0000144	2.550671	26.001217	45	89	Axon guidance
04070	0.0001250	2.866982	15.776020	29	54	Phosphatidylinositol signaling system
00512	0.0001345	4.641098	7.595861	17	26	Mucin type O-Glycan biosynthesis
04520	0.0002339	2.647465	16.944613	30	58	Adherens junction
04730	0.0002367	2.765150	15.483871	28	53	Long-term depression
04510	0.0002728	1.851702	43.237979	63	148	Focal adhesion
05412	0.0004053	2.470398	18.113208	31	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04512	0.0017014	2.133271	20.158247	32	69	ECM-receptor interaction
00561	0.0037197	2.777448	9.348752	17	32	Glycerolipid metabolism
04270	0.0048169	1.849026	24.540475	36	84	Vascular smooth muscle contraction
00532	0.0067353	4.393060	4.090079	9	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
05410	0.0083315	1.950823	17.821059	27	61	Hypertrophic cardiomyopathy (HCM)
04530	0.0096218	1.719189	26.293366	37	90	Tight junction
05100	0.0136217	1.997078	14.315277	22	49	Bacterial invasion of epithelial cells
02010	0.0179173	2.097361	11.393792	18	39	ABC transporters
05414	0.0183589	1.736004	20.450396	29	70	Dilated cardiomyopathy
00562	0.0203468	2.173258	9.933049	16	34	Inositol phosphate metabolism
00534	0.0213261	2.684510	6.135119	11	21	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
00564	0.0246893	1.781663	16.652465	24	57	Glycerophospholipid metabolism
00592	0.0266228	3.249300	4.090079	8	14	alpha-Linolenic acid metabolism
04972	0.0276334	1.653811	21.034693	29	72	Pancreatic secretion
00565	0.0285924	2.267009	7.888010	13	27	Ether lipid metabolism

04514	0.0292263	1.540280	28.046257	37	96	Cell adhesion molecules (CAMs)
04974	0.0308018	1.728507	16.944613	24	58	Protein digestion and absorption
04540	0.0315115	1.700832	17.821059	25	61	Gap junction
04666	0.0321233	1.676096	18.697505	26	64	Fc gamma R-mediated phagocytosis
05146	0.0334527	1.615516	21.326841	29	73	Amoebiasis
04930	0.0342605	2.035273	9.640901	15	33	Type II diabetes mellitus
04810	0.0463691	1.377978	41.485088	51	142	Regulation of actin cytoskeleton
04970	0.0468489	1.609907	18.405356	25	63	Salivary secretion

LD 0.4
LR
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00512	0.0000367	5.424899	4.676202	14	26	Mucin type O-Glycan biosynthesis
04730	0.0001596	3.065954	9.532258	21	53	Long-term depression
04360	0.0012673	2.147512	16.006999	28	89	Axon guidance
04270	0.0024256	2.092218	15.107730	26	84	Vascular smooth muscle contraction
04514	0.0086771	1.821924	17.265977	27	96	Cell adhesion molecules (CAMs)
04520	0.0105298	2.085078	10.431528	18	58	Adherens junction
04970	0.0120688	2.001311	11.330797	19	63	Salivary secretion
04530	0.0133283	1.787170	16.186853	25	90	Tight junction
05412	0.0213245	1.892670	11.150943	18	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04070	0.0240904	1.945629	9.712112	16	54	Phosphatidylinositol signaling system
05217	0.0246180	2.304310	5.935180	11	33	Basal cell carcinoma
04330	0.0248766	2.424133	5.215764	10	29	Notch signaling pathway
04972	0.0253404	1.780277	12.949483	20	72	Pancreatic secretion
05200	0.0382341	1.368359	40.467133	51	225	Pathways in cancer
04912	0.0414622	1.688307	12.769629	19	71	GnRH signaling pathway

LD 0.4
LR
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05210	0.0016012	4.190153	2.281802	8	46	Colorectal cancer
04270	0.0078397	2.692987	4.166768	10	84	Vascular smooth muscle contraction
00512	0.0079512	4.674503	1.289714	5	26	Mucin type O-Glycan biosynthesis
04514	0.0192116	2.308101	4.762021	10	96	Cell adhesion molecules (CAMs)
04520	0.0231072	2.702866	2.877054	7	58	Adherens junction
05146	0.0265231	2.428189	3.621120	8	73	Amoebiasis
04730	0.0452647	2.501152	2.629032	6	53	Long-term depression
04144	0.0468818	1.829649	7.043822	12	142	Endocytosis

LD 0.4

LR
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05217	0.0141140	4.792725	0.9540475	4	33	Basal cell carcinoma
05213	0.0156490	4.631502	0.9829580	4	34	Endometrial cancer
04514	0.0198104	2.772472	2.7754108	7	96	Cell adhesion molecules (CAMs)
05216	0.0242073	5.443936	0.6360316	3	22	Thyroid cancer
04520	0.0248688	3.289308	1.6768107	5	58	Adherens junction
00512	0.0376245	4.491493	0.7516738	3	26	Mucin type O-Glycan biosynthesis
05210	0.0423947	3.295657	1.3298844	4	46	Colorectal cancer

LD 0.4
LR
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05213	0.0229525	9.616071	0.2379793	2	34	Endometrial cancer

LD 0.4
LR
Percentile 0.0005

No results

LD 0.4
RF
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0000000	3.436277	26.542909	52	89	Axon guidance
04020	0.0000025	2.365945	37.577602	62	126	Calcium signaling pathway
04972	0.0000870	2.545887	21.472915	37	72	Pancreatic secretion
04070	0.0005198	2.579185	16.104686	28	54	Phosphatidylinositol signaling system
04912	0.0009073	2.204045	21.174681	34	71	GnRH signaling pathway
04730	0.0023301	2.300412	15.806452	26	53	Long-term depression
04930	0.0024967	2.857796	9.841753	18	33	Type II diabetes mellitus
04010	0.0028129	1.587675	51.892879	69	174	MAPK signaling pathway
05412	0.0033162	2.100405	18.490566	29	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04720	0.0034876	2.290544	14.613512	24	49	Long-term potentiation
04270	0.0035957	1.885855	25.051735	37	84	Vascular smooth muscle contraction
04971	0.0036428	2.149423	16.999391	27	57	Gastric acid secretion
04062	0.0040222	1.759687	31.016433	44	104	Chemokine signaling pathway
05410	0.0053914	2.025847	18.192331	28	61	Hypertrophic cardiomyopathy (HCM)

05414	0.0067359	1.898814	20.876445	31	70	Dilated cardiomyopathy
00512	0.0086830	2.770531	7.754108	14	26	Mucin type O-Glycan biosynthesis
04664	0.0089094	2.042439	15.508217	24	52	Fc epsilon RI signaling pathway
04260	0.0181015	2.052597	12.227632	19	41	Cardiac muscle contraction
04080	0.0197676	1.425897	50.401704	63	169	Neuroactive ligand-receptor interaction
04510	0.0198550	1.458498	44.138770	56	148	Focal adhesion
04742	0.0238342	2.371251	7.754108	13	26	Taste transduction
00910	0.0371406	3.310791	3.578819	7	12	Nitrogen metabolism
04514	0.0395638	1.494311	28.630554	37	96	Cell adhesion molecules (CAMs)
04310	0.0397699	1.484633	29.525259	38	99	Wnt signaling pathway
05010	0.0452515	1.510478	25.349970	33	85	Alzheimer's disease

LD 0.4
RF
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0000000	3.546935	16.061169	38	89	Axon guidance
04020	0.0000077	2.458489	22.738284	43	126	Calcium signaling pathway
04070	0.0000715	3.203919	9.744979	22	54	Phosphatidylinositol signaling system
05412	0.0000898	2.946998	11.188679	24	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
05414	0.0002896	2.589984	12.632380	25	70	Dilated cardiomyopathy
04912	0.0003726	2.532723	12.812842	25	71	GnRH signaling pathway
04270	0.0004770	2.333628	15.158856	28	84	Vascular smooth muscle contraction
05410	0.0005695	2.621941	11.008217	22	61	Hypertrophic cardiomyopathy (HCM)
04260	0.0012153	2.959307	7.398965	16	41	Cardiac muscle contraction
04720	0.0013986	2.688135	8.842666	18	49	Long-term potentiation
04730	0.0014553	2.588697	9.564516	19	53	Long-term depression
04970	0.0056537	2.151061	11.369142	20	63	Salivary secretion
04010	0.0088941	1.580104	31.400487	44	174	MAPK signaling pathway
04971	0.0089841	2.130301	10.286366	18	57	Gastric acid secretion
04514	0.0090808	1.814104	17.324407	27	96	Cell adhesion molecules (CAMs)
00512	0.0112550	2.869854	4.692027	10	26	Mucin type O-Glycan biosynthesis
00603	0.0116248	4.577513	2.165551	6	12	Glycosphingolipid biosynthesis - globo series
00562	0.0122563	2.507589	6.135727	12	34	Inositol phosphate metabolism
05010	0.0128395	1.819931	15.339318	24	85	Alzheimer's disease
04540	0.0186600	1.929221	11.008217	18	61	Gap junction
04510	0.0188647	1.547962	26.708460	37	148	Focal adhesion
04916	0.0206903	1.863518	11.910530	19	66	Melanogenesis
04614	0.0213721	4.571429	1.804626	5	10	Renin-angiotensin system
04910	0.0242821	1.712309	15.339318	23	85	Insulin signaling pathway
04150	0.0251550	2.578446	4.511564	9	25	mTOR signaling pathway
04930	0.0252003	2.294674	5.955265	11	33	Type II diabetes mellitus
00532	0.0273271	3.430579	2.526476	6	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
04742	0.0325137	2.425866	4.692027	9	26	Taste transduction
04664	0.0368343	1.862901	9.384054	15	52	Fc epsilon RI signaling pathway
05143	0.0417190	2.441482	4.150639	8	23	African trypanosomiasis

|04972 | 0.0486735| 1.648807| 12.993305| 19| 72|Pancreatic secretion |

LD 0.4
RF
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0000237	4.146563	4.4418746	15	89	Axon guidance
04020	0.0004178	2.960197	6.2884967	16	126	Calcium signaling pathway
04912	0.0023934	3.258463	3.5435180	10	71	GnRH signaling pathway
04972	0.0026621	3.204860	3.5934267	10	72	Pancreatic secretion
04730	0.0042009	3.506553	2.6451613	8	53	Long-term depression
04971	0.0066101	3.216117	2.8447961	8	57	Gastric acid secretion
04270	0.0081762	2.674623	4.1923311	10	84	Vascular smooth muscle contraction
04720	0.0098940	3.269639	2.4455265	7	49	Long-term potentiation
04970	0.0120210	2.859674	3.1442483	8	63	Salivary secretion
00534	0.0183868	4.566177	1.0480828	4	21	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
04950	0.0192442	6.445135	0.5989044	3	12	Maturity onset diabetes of the young
04012	0.0218232	2.739363	2.8447961	7	57	ErbB signaling pathway
04930	0.0222234	3.474843	1.6469872	5	33	Type II diabetes mellitus
05010	0.0242788	2.327165	4.2422398	9	85	Alzheimer's disease
04540	0.0305100	2.533145	3.0444309	7	61	Gap junction
04510	0.0313435	1.904881	7.3864881	13	148	Focal adhesion
05412	0.0329999	2.486277	3.0943396	7	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
00510	0.0480489	3.227083	1.3974437	4	28	N-Glycan biosynthesis

LD 0.4
RF
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04950	0.0024262	14.217778	0.2848448	3	12	Maturity onset diabetes of the young
05412	0.0031470	4.690476	1.4716981	6	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04930	0.0071292	5.925443	0.7833232	4	33	Type II diabetes mellitus
04510	0.0076194	2.879887	3.5130858	9	148	Focal adhesion
00534	0.0125790	7.088889	0.4984784	3	21	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
04270	0.0137788	3.344017	1.9939136	6	84	Vascular smooth muscle contraction
04912	0.0254580	3.260689	1.6853317	5	71	GnRH signaling pathway
04972	0.0268631	3.211000	1.7090688	5	72	Pancreatic secretion
00510	0.0275225	5.092800	0.6646379	3	28	N-Glycan biosynthesis
05100	0.0277608	3.799399	1.1631163	4	49	Bacterial invasion of epithelial cells
04720	0.0277608	3.799399	1.1631163	4	49	Long-term potentiation
04730	0.0357467	3.484832	1.2580645	4	53	Long-term depression
05010	0.0497188	2.678082	2.0176506	5	85	Alzheimer's disease

LD 0.4
RF
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04930	0.018233	10.981324	0.2108947	2	33	Type II diabetes mellitus
04664	0.042582	6.768421	0.3323189	2	52	Fc epsilon RI signaling pathway

LD 0.4
RF
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04614	0.0359727	32.979798	0.0365186	1	10	Renin-angiotensin system
04380	0.0376903	7.595238	0.3140596	2	86	Osteoclast differentiation

LD 0.4
DNN
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04020	0.0000353	2.136949	35.967133	57	126	Calcium signaling pathway
04730	0.0000499	3.089292	15.129032	29	53	Long-term depression
04070	0.0000788	2.964444	15.414486	29	54	Phosphatidylinositol signaling system
04360	0.0002767	2.190171	25.405356	41	89	Axon guidance
04270	0.0003184	2.220171	23.978089	39	84	Vascular smooth muscle contraction
04540	0.0011560	2.308993	17.412660	29	61	Gap junction
04972	0.0012012	2.158861	20.552648	33	72	Pancreatic secretion
04970	0.0021633	2.171294	17.983567	29	63	Salivary secretion
05412	0.0036157	2.094118	17.698113	28	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04720	0.0102476	2.064613	13.987219	22	49	Long-term potentiation
04976	0.0118457	1.920907	16.556300	25	58	Bile secretion
05414	0.0132526	1.795138	19.981741	29	70	Dilated cardiomyopathy
02010	0.0140347	2.168012	11.132684	18	39	ABC transporters
05410	0.0238546	1.758549	17.412660	25	61	Hypertrophic cardiomyopathy (HCM)
04080	0.0261363	1.403732	48.241631	60	169	Neuroactive ligand-receptor interaction
04514	0.0339645	1.521951	27.403530	36	96	Cell adhesion molecules (CAMs)
04012	0.0356040	1.710768	16.270846	23	57	ErbB signaling pathway
04971	0.0356040	1.710768	16.270846	23	57	Gastric acid secretion
04512	0.0361137	1.627254	19.696287	27	69	ECM-receptor interaction
00512	0.0417851	2.160444	7.421789	12	26	Mucin type O-Glycan biosynthesis
05200	0.0438662	1.301204	64.227024	76	225	Pathways in cancer
04010	0.0467198	1.339168	49.668898	60	174	MAPK signaling pathway

LD 0.4
DNN
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04730	0.0000086	3.715491	9.290323	23	53	Long-term depression
04020	0.0000533	2.276987	22.086427	40	126	Calcium signaling pathway
04360	0.0001403	2.468802	15.600730	30	89	Axon guidance
04972	0.0002990	2.570761	12.620816	25	72	Pancreatic secretion
04540	0.0010246	2.525676	10.692635	21	61	Gap junction
05412	0.0013009	2.463151	10.867925	21	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04270	0.0016563	2.161505	14.724285	26	84	Vascular smooth muscle contraction
04970	0.0040122	2.231052	11.043214	20	63	Salivary secretion
02010	0.0045219	2.675445	6.836275	14	39	ABC transporters
04720	0.0071596	2.317749	8.589166	16	49	Long-term potentiation
04062	0.0100226	1.770841	18.230067	28	104	Chemokine signaling pathway
04971	0.0149943	2.029964	9.991479	17	57	Gastric acid secretion
04070	0.0191751	2.009023	9.465612	16	54	Phosphatidylinositol signaling system
05414	0.0284655	1.778470	12.270237	19	70	Dilated cardiomyopathy
04976	0.0368124	1.814966	10.166768	16	58	Bile secretion
00230	0.0489183	1.531016	18.054778	25	103	Purine metabolism

LD 0.4
DNN
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0048322	2.749096	4.5231284	11	89	Axon guidance
00230	0.0053521	2.576108	5.2346318	12	103	Purine metabolism
04270	0.0092534	2.620933	4.2690201	10	84	Vascular smooth muscle contraction
04972	0.0097819	2.763110	3.6591601	9	72	Pancreatic secretion
04020	0.0105493	2.245604	6.4035301	13	126	Calcium signaling pathway
05412	0.0121433	2.855812	3.1509434	8	62	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
02010	0.0127018	3.485037	1.9820450	6	39	ABC transporters
04730	0.0164753	2.922690	2.6935484	7	53	Long-term depression
04976	0.0260066	2.631863	2.9476567	7	58	Bile secretion
00532	0.0310199	5.168514	0.7115033	3	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
04720	0.0359040	2.665896	2.4902617	6	49	Long-term potentiation
04530	0.0372619	2.136428	4.5739501	9	90	Tight junction
04970	0.0387618	2.392969	3.2017651	7	63	Salivary secretion
04742	0.0401007	3.454546	1.3213634	4	26	Taste transduction
00380	0.0452726	3.303281	1.3721850	4	27	Tryptophan metabolism
00920	0.0455494	7.549091	0.3557517	2	7	Sulfur metabolism

LD 0.4
DNN

Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00380	0.0035666	7.383188	0.6491175	4	27	Tryptophan metabolism
04360	0.0050162	3.705115	2.1396835	7	89	Axon guidance
04972	0.0070520	3.911582	1.7309799	6	72	Pancreatic secretion
04530	0.0199822	3.055773	2.1637249	6	90	Tight junction
00300	0.0240414	Inf	0.0240414	1	1	Lysine biosynthesis
04730	0.0372142	3.437279	1.2741935	4	53	Long-term depression
04062	0.0374067	2.607492	2.5003043	6	104	Chemokine signaling pathway
00532	0.0430791	6.915584	0.3365794	2	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
04971	0.0467391	3.173837	1.3703591	4	57	Gastric acid secretion
00780	0.0475119	41.102564	0.0480828	1	2	Biotin metabolism
04976	0.0493143	3.114074	1.3944005	4	58	Bile secretion
04270	0.0498823	2.675334	2.0194766	5	84	Vascular smooth muscle contraction

LD 0.4
DNN
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
02010	0.0148321	12.482625	0.1898965	2	39	ABC transporters
04610	0.0220106	10.012422	0.2337188	2	48	Complement and coagulation cascades
04730	0.0265115	9.016807	0.2580645	2	53	Long-term depression
04971	0.0303606	8.350649	0.2775411	2	57	Gastric acid secretion
00920	0.0336204	36.266667	0.0340840	1	7	Sulfur metabolism
04972	0.0466082	6.530612	0.3505782	2	72	Pancreatic secretion

LD 0.4
DNN
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04610	0.0055323	23.42029	0.1168594	2	48	Complement and coagulation cascades
00920	0.0169334	77.90476	0.0170420	1	7	Sulfur metabolism
04360	0.0182656	12.22605	0.2166768	2	89	Axon guidance
00532	0.0336153	35.87912	0.0340840	1	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat

LD 0.6
LR
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term

04270	0.0007409	3.820141	3.0174326	10	96	Vascular smooth muscle contraction
04970	0.0011474	4.332733	2.1373481	8	68	Salivary secretion
05414	0.0079942	3.358209	2.3259377	7	74	Dilated cardiomyopathy
00532	0.0085536	8.595611	0.4400423	3	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
00230	0.0111142	2.670726	3.7089276	9	118	Purine metabolism
04260	0.0126164	3.976974	1.4144216	5	45	Cardiac muscle contraction
04972	0.0137449	2.993333	2.5773904	7	82	Pancreatic secretion
05412	0.0144079	3.303936	2.0116218	6	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04977	0.0175003	6.296552	0.5657686	3	18	Vitamin digestion and absorption
05410	0.0190097	3.087354	2.1373481	6	68	Hypertrophic cardiomyopathy (HCM)
05210	0.0209065	3.452517	1.6030111	5	51	Colorectal cancer
05200	0.0243285	1.929677	7.8893291	14	251	Pathways in cancer
04744	0.0266439	5.242816	0.6600634	3	21	Phototransduction
04930	0.0275439	3.830303	1.1629688	4	37	Type II diabetes mellitus
00920	0.0305237	8.937729	0.2828843	2	9	Sulfur metabolism
05323	0.0389424	2.880383	1.8858954	5	60	Rheumatoid arthritis
04742	0.0465245	4.097451	0.8172213	3	26	Taste transduction
04974	0.0492900	2.682129	2.0116218	5	64	Protein digestion and absorption

LD 0.6
LR
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04270	0.0007409	3.820141	3.0174326	10	96	Vascular smooth muscle contraction
04970	0.0011474	4.332733	2.1373481	8	68	Salivary secretion
05414	0.0079942	3.358209	2.3259377	7	74	Dilated cardiomyopathy
00532	0.0085536	8.595611	0.4400423	3	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
00230	0.0111142	2.670726	3.7089276	9	118	Purine metabolism
04260	0.0126164	3.976974	1.4144216	5	45	Cardiac muscle contraction
04972	0.0137449	2.993333	2.5773904	7	82	Pancreatic secretion
05412	0.0144079	3.303936	2.0116218	6	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04977	0.0175003	6.296552	0.5657686	3	18	Vitamin digestion and absorption
05410	0.0190097	3.087354	2.1373481	6	68	Hypertrophic cardiomyopathy (HCM)
05210	0.0209065	3.452517	1.6030111	5	51	Colorectal cancer
05200	0.0243285	1.929677	7.8893291	14	251	Pathways in cancer
04744	0.0266439	5.242816	0.6600634	3	21	Phototransduction
04930	0.0275439	3.830303	1.1629688	4	37	Type II diabetes mellitus
00920	0.0305237	8.937729	0.2828843	2	9	Sulfur metabolism
05323	0.0389424	2.880383	1.8858954	5	60	Rheumatoid arthritis
04742	0.0465245	4.097451	0.8172213	3	26	Taste transduction
04974	0.0492900	2.682129	2.0116218	5	64	Protein digestion and absorption

LD 0.6
LR
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04270	0.0007409	3.820141	3.0174326	10	96	Vascular smooth muscle contraction
04970	0.0011474	4.332733	2.1373481	8	68	Salivary secretion
05414	0.0079942	3.358209	2.3259377	7	74	Dilated cardiomyopathy
00532	0.0085536	8.595611	0.4400423	3	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
00230	0.0111142	2.670726	3.7089276	9	118	Purine metabolism
04260	0.0126164	3.976974	1.4144216	5	45	Cardiac muscle contraction
04972	0.0137449	2.993333	2.5773904	7	82	Pancreatic secretion
05412	0.0144079	3.303936	2.0116218	6	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04977	0.0175003	6.296552	0.5657686	3	18	Vitamin digestion and absorption
05410	0.0190097	3.087354	2.1373481	6	68	Hypertrophic cardiomyopathy (HCM)
05210	0.0209065	3.452517	1.6030111	5	51	Colorectal cancer
05200	0.0243285	1.929677	7.8893291	14	251	Pathways in cancer
04744	0.0266439	5.242816	0.6600634	3	21	Phototransduction
04930	0.0275439	3.830303	1.1629688	4	37	Type II diabetes mellitus
00920	0.0305237	8.937729	0.2828843	2	9	Sulfur metabolism
05323	0.0389424	2.880383	1.8858954	5	60	Rheumatoid arthritis
04742	0.0465245	4.097451	0.8172213	3	26	Taste transduction
04974	0.0492900	2.682129	2.0116218	5	64	Protein digestion and absorption

LD 0.6
LR
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04270	0.0007409	3.820141	3.0174326	10	96	Vascular smooth muscle contraction
04970	0.0011474	4.332733	2.1373481	8	68	Salivary secretion
05414	0.0079942	3.358209	2.3259377	7	74	Dilated cardiomyopathy
00532	0.0085536	8.595611	0.4400423	3	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
00230	0.0111142	2.670726	3.7089276	9	118	Purine metabolism
04260	0.0126164	3.976974	1.4144216	5	45	Cardiac muscle contraction
04972	0.0137449	2.993333	2.5773904	7	82	Pancreatic secretion
05412	0.0144079	3.303936	2.0116218	6	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04977	0.0175003	6.296552	0.5657686	3	18	Vitamin digestion and absorption
05410	0.0190097	3.087354	2.1373481	6	68	Hypertrophic cardiomyopathy (HCM)
05210	0.0209065	3.452517	1.6030111	5	51	Colorectal cancer
05200	0.0243285	1.929677	7.8893291	14	251	Pathways in cancer
04744	0.0266439	5.242816	0.6600634	3	21	Phototransduction
04930	0.0275439	3.830303	1.1629688	4	37	Type II diabetes mellitus
00920	0.0305237	8.937729	0.2828843	2	9	Sulfur metabolism
05323	0.0389424	2.880383	1.8858954	5	60	Rheumatoid arthritis
04742	0.0465245	4.097451	0.8172213	3	26	Taste transduction
04974	0.0492900	2.682129	2.0116218	5	64	Protein digestion and absorption

LD 0.6

LR
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
00532	0.0059123	20.788889	0.1183307	2	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
04744	0.0131505	13.105263	0.1774960	2	21	Phototransduction
04970	0.0189515	5.871087	0.5747491	3	68	Salivary secretion
04972	0.0309286	4.812309	0.6930798	3	82	Pancreatic secretion
00561	0.0346271	7.517172	0.2958267	2	35	Glycerolipid metabolism
04530	0.0472996	4.027880	0.8198627	3	97	Tight junction

LD 0.6
LR
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05412	0.0224623	9.973118	0.2366614	2	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)

LD 0.6
DNN
Percentile 0.1

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04360	0.0000175	2.363296	34.493661	55	101	Axon guidance
00053	0.0000247	9.741784	6.147385	15	18	Ascorbate and aldarate metabolism
04020	0.0000305	2.073520	45.422345	68	133	Calcium signaling pathway
04510	0.0000789	1.885599	53.960380	77	158	Focal adhesion
00860	0.0005508	4.138421	8.538035	17	25	Porphyrin and chlorophyll metabolism
04974	0.0012742	2.217157	21.857369	34	64	Protein digestion and absorption
04540	0.0028578	2.080031	21.857369	33	64	Gap junction
00640	0.0041037	3.239437	8.196514	15	24	Propanoate metabolism
00524	0.0046225	Inf	1.707607	5	5	Butirosin and neomycin biosynthesis
02010	0.0058443	2.404551	12.977813	21	38	ABC transporters
00500	0.0058443	2.404551	12.977813	21	38	Starch and sucrose metabolism
05412	0.0060464	1.951626	21.857369	32	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
00512	0.0064188	2.827081	9.221078	16	27	Mucin type O-Glycan biosynthesis
04972	0.0076492	1.772004	28.004754	39	82	Pancreatic secretion
00040	0.0084021	3.154785	7.171949	13	21	Pentose and glucuronate interconversions
04971	0.0091555	1.889139	21.515848	31	63	Gastric acid secretion
04970	0.0094386	1.839320	23.223455	33	68	Salivary secretion
04270	0.0107896	1.653692	32.786054	44	96	Vascular smooth muscle contraction
00983	0.0110205	2.308821	11.953249	19	35	Drug metabolism - other enzymes
05146	0.0139462	1.685295	28.004754	38	82	Amoebiasis
00534	0.0142036	2.803125	7.513471	13	22	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
00340	0.0157097	2.909836	6.830428	12	20	Histidine metabolism

04512	0.0178576	1.731959	23.223455	32	68	ECM-receptor interaction
00980	0.0193588	1.870753	17.417591	25	51	Metabolism of xenobiotics by cytochrome P450
00561	0.0256978	2.056194	11.953249	18	35	Glycerolipid metabolism
00982	0.0289017	1.794508	17.076070	24	50	Drug metabolism - cytochrome P450
04012	0.0320336	1.630531	23.223455	31	68	ErbB signaling pathway
00830	0.0321951	1.854310	14.685420	21	43	Retinol metabolism
04670	0.0362581	1.533101	28.687797	37	84	Leukocyte transendothelial migration
00010	0.0366213	1.849400	14.002377	20	41	Glycolysis / Gluconeogenesis
04912	0.0374524	1.548544	26.980190	35	79	GnRH signaling pathway
04080	0.0405713	1.327724	62.498415	74	183	Neuroactive ligand-receptor interaction
04062	0.0446825	1.400762	41.665610	51	122	Chemokine signaling pathway
04514	0.0458964	1.444207	34.493661	43	101	Cell adhesion molecules (CAMs)
04930	0.0475018	1.838266	12.636292	18	37	Type II diabetes mellitus

LD 0.6
DNN
Percentile 0.05

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04020	0.0000020	2.424378	29.403328	53	133	Calcium signaling pathway
05412	0.0000809	2.800577	14.148970	28	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04972	0.0008156	2.185520	18.128368	31	82	Pancreatic secretion
04360	0.0011453	1.994123	22.328843	36	101	Axon guidance
05414	0.0014182	2.184232	16.359746	28	74	Dilated cardiomyopathy
04510	0.0016579	1.723407	34.930269	51	158	Focal adhesion
04270	0.0017658	1.971598	21.223455	34	96	Vascular smooth muscle contraction
00053	0.0020084	4.445284	3.979398	10	18	Ascorbate and aldarate metabolism
00640	0.0023745	3.560000	5.305864	12	24	Propanoate metabolism
00512	0.0024214	3.307472	5.969097	13	27	Mucin type O-Glycan biosynthesis
04540	0.0034916	2.146863	14.148970	24	64	Gap junction
00534	0.0036047	3.556901	4.863708	11	22	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
00860	0.0036622	3.285035	5.526941	12	25	Porphyrin and chlorophyll metabolism
04970	0.0039445	2.080708	15.033281	25	68	Salivary secretion
04912	0.0043123	1.966698	17.465135	28	79	GnRH signaling pathway
02010	0.0043761	2.592847	8.400951	16	38	ABC transporters
04971	0.0061687	2.054883	13.927892	23	63	Gastric acid secretion
04730	0.0069029	2.067872	13.264659	22	60	Long-term depression
04974	0.0076705	2.004075	14.148970	23	64	Protein digestion and absorption
05146	0.0077409	1.855514	18.128368	28	82	Amoebiasis
00524	0.0097845	14.156062	1.105388	4	5	Butirosin and neomycin biosynthesis
00561	0.0126472	2.371810	7.737718	14	35	Glycerolipid metabolism
00410	0.0128733	3.547648	3.537243	8	16	beta-Alanine metabolism
04976	0.0158359	1.868361	14.148970	22	64	Bile secretion
00340	0.0189013	2.903162	4.421553	9	20	Histidine metabolism
00532	0.0198748	3.544578	3.095087	7	14	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfat
05100	0.0201594	1.875282	12.822504	20	58	Bacterial invasion of epithelial cells
04070	0.0243665	1.826570	13.043582	20	59	Phosphatidylinositol signaling system

00040	0.0267492	2.660326	4.642631	9	21	Pentose and glucuronate interconversions
00500	0.0272297	2.073208	8.400951	14	38	Starch and sucrose metabolism
00983	0.0306286	2.099018	7.737718	13	35	Drug metabolism - other enzymes
04720	0.0311639	1.829827	11.717116	18	53	Long-term potentiation
05410	0.0320784	1.703548	15.033281	22	68	Hypertrophic cardiomyopathy (HCM)
05014	0.0341176	1.989599	8.622029	14	39	Amyotrophic lateral sclerosis (ALS)
04670	0.0362449	1.597985	18.570523	26	84	Leukocyte transendothelial migration
05110	0.0421802	1.912422	8.843106	14	40	Vibrio cholerae infection
04930	0.0478416	1.922785	8.179873	13	37	Type II diabetes mellitus
05214	0.0481196	1.776492	10.611727	16	48	Glioma

LD 0.6
DNN
Percentile 0.01

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04270	0.0016990	2.650223	6.4912837	15	96	Vascular smooth muscle contraction
04020	0.0032217	2.245890	8.9931326	18	133	Calcium signaling pathway
05412	0.0032546	2.945476	4.3275225	11	64	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
04510	0.0042831	2.083026	10.6835711	20	158	Focal adhesion
04912	0.0061211	2.541962	5.3417855	12	79	GnRH signaling pathway
04514	0.0073004	2.289446	6.8293714	14	101	Cell adhesion molecules (CAMs)
04930	0.0106093	3.279786	2.5018489	7	37	Type II diabetes mellitus
00230	0.0122562	2.070862	7.9788695	15	118	Purine metabolism
04720	0.0243417	2.498208	3.5837295	8	53	Long-term potentiation
05414	0.0257359	2.201474	5.0036978	10	74	Dilated cardiomyopathy
04540	0.0266219	2.302172	4.3275225	9	64	Gap junction
04976	0.0266219	2.302172	4.3275225	9	64	Bile secretion
04974	0.0266219	2.302172	4.3275225	9	64	Protein digestion and absorption
04742	0.0276651	3.328590	1.7580560	5	26	Taste transduction
00051	0.0321404	3.176385	1.8256735	5	27	Fructose and mannose metabolism
00512	0.0321404	3.176385	1.8256735	5	27	Mucin type O-Glycan biosynthesis
04970	0.0377012	2.143622	4.5979926	9	68	Salivary secretion
04512	0.0377012	2.143622	4.5979926	9	68	ECM-receptor interaction
05100	0.0394149	2.245161	3.9218172	8	58	Bacterial invasion of epithelial cells
00524	0.0397404	9.257218	0.3380877	2	5	Butirosin and neomycin biosynthesis
04730	0.0468837	2.157568	4.0570523	8	60	Long-term depression
04330	0.0481595	2.792829	2.0285261	5	30	Notch signaling pathway
05213	0.0497062	2.467765	2.7047015	6	40	Endometrial cancer

LD 0.6
DNN
Percentile 0.005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04270	0.0035953	3.245009	3.1188590	9	96	Vascular smooth muscle contraction

04976	0.0042965	3.817604	2.0792393	7	64	Bile secretion
04720	0.0068453	3.945445	1.7218700	6	53	Long-term potentiation
04742	0.0091117	5.563025	0.8446910	4	26	Taste transduction
04020	0.0104263	2.546946	4.3209192	10	133	Calcium signaling pathway
00512	0.0104326	5.319693	0.8771791	4	27	Mucin type O-Glycan biosynthesis
04930	0.0306323	3.697479	1.2020602	4	37	Type II diabetes mellitus
00230	0.0366165	2.246957	3.8335975	8	118	Purine metabolism
05320	0.0368893	4.553750	0.7472266	3	23	Autoimmune thyroid disease
04912	0.0419875	2.521953	2.5665610	6	79	GnRH signaling pathway
04062	0.0433184	2.165675	3.9635499	8	122	Chemokine signaling pathway
04514	0.0445006	2.291177	3.2812995	7	101	Cell adhesion molecules (CAMs)

LD 0.6
DNN
Percentile 0.001

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04742	0.0006734	21.192817	0.1785526	3	26	Taste transduction
00770	0.0033772	28.401515	0.0892763	2	13	Pantothenate and CoA biosynthesis
05320	0.0105020	14.837302	0.1579503	2	23	Autoimmune thyroid disease
00780	0.0136895	150.360000	0.0137348	1	2	Biotin metabolism
05142	0.0155563	6.408696	0.5356577	3	78	Chagas disease (American trypanosomiasis)
05014	0.0287619	8.385135	0.2678288	2	39	Amyotrophic lateral sclerosis (ALS)
05416	0.0487509	6.183333	0.3571051	2	52	Viral myocarditis

LD 0.6
DNN
Percentile 0.0005

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
05014	0.0021020	40.45405	0.0721078	2	39	Amyotrophic lateral sclerosis (ALS)
03410	0.0328352	36.88235	0.0332805	1	18	Base excision repair
05320	0.0417904	28.46212	0.0425251	1	23	Autoimmune thyroid disease
04742	0.0471293	25.02667	0.0480718	1	26	Taste transduction