Evaluation of Phenotype Classification Methods for Obesity using Direct to Consumer Genetic Data

Casimiro Aday Curbelo Montañez[1], Paul Fergus[1], Abir Hussain[1], Dhiya Al-Jumeily[1], Mehmet Tevfik Dorak[2], Rosni Abdullah[3]

[1]Liverpool John Moores University, UK.
c.a.curbelomontanez@2015.ljmu.ac.uk
{p.fergus, a.hussain, d.aljumeily}@ljmu.ac.uk
[2]Liverpool Hope University, UK
dorakm@hope.ac.uk
[3]Universiti Sains Malaysia, Malaysia
rosni@usm.my

Abstract. Direct-to-Consumer genetic testing services are becoming more ubiquitous. Consumers of such services are sharing their genetic and clinical information with the research community to facilitate the extraction of knowledge about different conditions. In this paper, we build on these services to analyse the genetic data of people with different BMI levels to determine the immediate and long-term risk factors associated with obesity. Using web scraping techniques, a dataset containing publicly available information about 230 participants from the Personal Genome Project is created. Subsequent analysis of the dataset is conducted for the identification of genetic variants associated with high BMI levels via standard quality control and association analysis protocols for Genome Wide Association Analysis. We applied a combination of Random Forest based feature selection algorithm and Support Vector Machine with Radial Basis Function Kernel learning method to the filtered dataset. Using a robust data science methodology our approach identified obesity related genetic variants, to be used as features when predicting individual obesity susceptibility. The results reveal that the subset of features obtained through the Random Forest based algorithm improve the performance of the classifier when compared to the top statistically significant genetic variants identified in logistic regression. Support Vector Machine showed the best results with sensitivity=81%, specificity=83% and area under the curve=92% when the model was trained with the top fifteen features selected by Boruta.

Keywords: Bioinformatics, Data Science, Machine Learning, Feature Selection, Genetics, Obesity, SNPs.

1. Introduction

The global prevalence of obesity has reached epidemic proportions [1]. According to the World Health Organization (WHO)[1], approximately 2.8 million people die each year as a consequence of being overweight or obese [2]. Obesity is a major risk for

[1] http://www.who.int/

other chronic diseases which include diabetes, cardiovascular diseases and cancer [3]. The occurrence of obesity is a common problem in high-income countries but, its frequency is also rising in low and middle-income countries [4]. In England, the National Obesity Observatory (NOO) reported that the direct cost to the National Healthcare Service (NHS) for treating overweight, obesity and related morbidities increased from £479.3 million in 1998 to £4.2 billion in 2007[2]. The effects of obesity are so grave that it reduces life expectancy on average by 3 years – in cases of severe obesity this can be between 5 and 13 years [5].

Advances in Human Genomics have provided significant opportunities and research suggests that it might be possible to quantify an individual's susceptibility to obesity from an early age and manage risk as individuals' progress through life [6]. Therefore, combining personalised medicine with genomic information and integrating it into medical care and individualised risk assessments will allow us to mitigate the long-term effects of obesity and its associated co-morbidities. This is being made possible through advances in bioinformatics [7], data science [8] and advanced machine learning algorithms [9].

This paper explores these ideas further and proposes a robust methodology to combine state-of-the-art bioinformatics and data science to investigate genetic profiling and risk factor assessment for obesity. We combined two statistical approaches for Single Nucleotide Polymorphism (SNP) evaluation. Risk-Based approach and Classification-Based Approach. The first approach is applied to identify statistically significant SNPs whilst the second is used to identify a set of SNPs appearing conjointly which can serve to predict obesity. The motivation for this research is to identify strong genetic markers for use in decision support systems. Data science is utilised to automatically build a dataset, using publicly available demographic and genetic information provided by individuals. This dataset and subsequent analysis is intended to provide a starting point for genetic variants data analysis.

2. Background

The decreased costs associated with Deoxyribonucleic acid **(**DNA) sequencing have made it easier to obtain genomic data. For example, the 100,000 Genomes Project[3], conducted by Genomics England, has sequenced 100,000 genomes from 70,000 NHS patients suffering with rare diseases. The information will be used to create a genomic medicine service for the NHS and enable new scientific discovery and medical insights. In the private sector, genetic screening services are delivered directly to consumers. Individuals provide a saliva sample to a Direct-to-Consumer Genetic Testing (DTCGT) company and obtain genetic information without any health care provider involvement [10]. Many of these DTCGT services use SNP identification to determine ancestry and genetic markers associated with specific diseases with the objective of informing clients about their health and how to change behaviours to improve it [10].

The Personal Genome Project (PGP)[4] is a non-profit organization created to promote the availability and use of personal health information and genome data to help accelerate the understanding of genetic variation in humans. While many object to

[2] https://www.noo.org.uk/

[3] http://www.genomicsengland.co.uk/

[4] http://www.personalgenomes.org/

privacy, confidentiality and anonymity issues, the PGP believes that sharing such data is fundamentally advantageous for the advances in science and society. This is a view endorsed by members of the public who understand the risks and share their personal information. The founding pilot project of the PGP was initiated by the Harvard Personal Genome Project, which now hosts publicly shared genomic and health data from thousands of participants. In 2005 information on 10 fully identified individuals was available; today, more than 4000 US participants have publicly shared their genomic information. There is also evidence that information across initiatives is being shared with genetic data from 23andMe appearing in PGP datasets [11].

Bioinformaticians routinely extract information from websites using web-scrapping techniques to obtain content originally presented for human use [12]. Collecting this data is tedious and time-consuming. Several institutions have invested heavily in data collection, gathering clinical and genetic data within different domains for decades. This has resulted in significant amounts of big data [13] and today organisations, such as the National Institute of Health (NIH), which sponsored the Database of Genomes and Phenotypes (dbGaP), are making this data available to interested parties, subject to specific terms and conditions [14]. However, to access this data, researchers must follow a data request procedure that can be restrictive to general users from other domains that want to make use of genetic data. Consequently, other organisations such as the PGP rely on a different strategy defined by publicly accessible data that anybody from diverse backgrounds can use to get started on genetic data analysis. Having access to such repositories has had a huge positive impact on the scientific community who no longer need to generate their own data for the studies that they conduct.

Approximately 99.5% of the total number of base pairs (nucleotides) in the human genome are identical for any two human individuals [15]. Hence, in genetic association studies, bases where there is variation between humans are commonly considered. Studies utilizing hypothesis-free methodologies such as genome-wide association studies (GWAS) have been used in obesity studies to identify many obesity related loci. GWAS permit the analysis of a large number of genetic variants (whole genome) for association with traits of interest. In statistical association test, logistic regression is often the preferred approach as it has been extensively developed although it is not the only one [16]. Currently, associations of common variants usually should reach threshold levels of $P < 5x10^{-8}$ to be considered significant [17]. Conversely, variants with threshold levels of $P < 10^{-5}$ are termed suggestive SNPs [18] and could be studied further. The importance of GWAS is advancing scientific understanding of disease mechanisms and providing starting points and potential opportunities for researchers to improve the development of medical treatments.

Following an open data initiative, genetic association analysis and predictive modelling strategies are conducted in this study for the analysis of obesity as a binary trait.

3. Materials & Methods

The dataset used in this paper comprises 230 participants from the PGP, which donated genetic data from Direct-to-Consumer genotyping. This data is extracted and analysed by 23andMe using microarray genotyping, which provides an efficient and cost-effective way of evaluating genetic variation in individuals and across populations [19]. In addition to genetic data, clinical information is also provided. Collected contributors

are aged between 23 and 79 years of age (average age 46.59) and are all from the United States of America. The average height, body weight and BMI of all participants is 1.74 meters, 78.97 kg, and 25.97 respectively. Of the total population, 150 (65.22%) are males and 80 (34.78%) females. All participants considered in the study reported white as ethnical background.

3.1 Data Collection and Description

During the initial data collection process, 733 observations/participants and 9 variables were scrapped from the PGP website[5]. Table 1 provides a description of the data fields extracted for each participant.

The Participant_ID is a unique participant identifier assigned in the PGP. The variable Data link provides a Uniform Resource Locator (URL) used to download the genetic profile of each participant. In addition, DoB, Gender, Weight, Height, Ethnicity, and Blood Type contain personal information for each participant. Data about the condition Type 2 diabetes (T2D) was also included, although more features based on the existing variables were subsequently incorporated to the clinical data file.

The resulting dataset contained several empty fields. Only observations with complete values for the variables in Table 1 were retained. Individuals who reported being of ethnicities other than white were excluded to avoid population stratification in our analysis. This reduced the dataset to 235 observations. The data links for five participants were incorrect so these were also discarded from the final dataset, resulting in 230 individuals.

Full genome profiles were downloaded in *txt* format using the variable Data Link identified in Table 1. Only full genome data was included in the analysis i.e., if a participant from the PGP uploaded exon and whole genome data to the PGP website, only the whole genome profiles were considered since full genome provides complete representation of the genome. The genetic profile of each participant contains four variables: *rsid*, *chromosome*, *position* and *genotype*, and several hundred thousand observations that depend on the amount of variants discovered by the genotyping process used by 23andMe [19]. The variables included in the genetic profiles represent genetic variants or SNPs.

Table 1: Variables selected in the web scraping process.

Variables	Description
Participant_ID	Participant ID
Data link	Genetic data URL
DoB	Date of Birthday
Gender	Gender
Weight	Weight in Kg
Height	Height in meters
T2D	Type 2 Diabetes
Ethnicity	Ethnical background
Blood Type	Blood Type

[5] http://www.personalgenomes.org/

Downloaded genetic profiles were converted to binary file format [20]. This type of format allows for a more efficient and convenient way of manipulating SNP data when using open source software for automated GWAS quality control (QC) and analysis, such as PLINK [20]. Subsequently, all 230 genetic profiles were merged into one main binary file (.bed, .bim, .fam). Finally, two main data frames were created – one containing the clinical information and the other containing genetic variants identified by 23andMe for the 230 participants.

Additional features were generated using information from existing columns. These include body mass index (BMI), constructed from the Weight and Height variables and calculated using the metric formula, $\text{BMI} = \frac{\text{Weight (Kg)}}{(\text{Height(m)})^2}$. A Status feature was also generated from the BMI result. Following the WHO classification for BMI[6], 5 standard weight status categories associated with BMI ranges for adults were derived. Table 2, summarizes the number of participants included in each status category. The category Normal range has the highest representation among the participants (50%) whereas Underweight is the category with the lowest representation (1.74%). The categories Overweight, Obese and Extremely obese, when grouped together, constitute 111 participants. In other words, 48.26% of the participants analysed were included in one of these three status categories. Hence, as shown in Table 2, two closely balanced classes based on the BMI were created, representing the phenotypic variable for risk prediction of obesity. The variables considered in the clinical data frame are: Participant_ID, age, gender, height (m), weight (kg), BMI, Status, T2D, Race and blood type. In the case of the genetic information, the variables considered are: SNP name (rsid), chromosome number, position in the DNA sequence and genotype.

Table 2: BMI status among participants included in the study.

Class	Status	Total number
Normal 51.74%	Underweight	4
	Normal range	115
Risk 48.26%	Overweight	65
	Obese	41
	Extremely obese	5

3.2 Data Pre-processing

Analyses were conducted using PLINK and R software[7]. After the data set construction, and prior to analysis, data QC was performed. Cases and controls in the present study are defined as risk and normal. Following protocols for genetic case-control association studies, QC was performed on individuals and then on markers, to optimise the number of SNPs remaining in the study [21].

In the per-individual QC process, 7 individuals were removed leaving 223 remaining individuals of which, 107 are cases and 116 are controls. Individuals were excluded if they showed abnormal heterozygosity, discordant sex information, were duplicated or related individuals, and individuals of divergent ancestry. Strict values for missing rate

[6] http://www.who.int/

[7] http://www.r-project.org/

were not considered since most samples in the study would be removed. This might be an indicative of poor quality DNA sample [22].

In the per-marker QC process, SNPs with minor allele frequency (MAF<4%), call rate of <98% and deviations from Hardy-Weinberg equilibrium ($p<1\times10^{-3}$) were excluded. A MAF cut point of 4% is commonly applied in small sample settings due to statistical power considerations [23].

3.3 Genetic Association Analysis

For discovery, association analysis on 107 risk cases and 116 controls was performed by testing SNPs and individuals that satisfied quality control. Logistic regression was used to identify SNPs showing a strong association with the trait of interest. However, none of the SNPs reached significance level (P-value < 5×10^{-8}) nor were suggestive of association (P-value < 1×10^{-5}) as shown in Fig. 1(a), a Manhattan plot of genome genome-wide association analysis results. The figure illustrates, in the y-axis, the level of statistical significance as measured by the negative log of the corresponding P-value, for each SNP. Significant and suggestive levels are represented in red and blue respectively. Each typed SNP is indicated by a dark-blue or orange dot. In the x-axis, SNPs are arranged by chromosomal location.

While no SNPs were identified as significant or suggestive, a subset of SNPs with P-values < 1×10^{-3} were considered for subsequent analysis as similarly performed in [24]. Consequently, a total of 261 SNPs showing the strongest association with the phenotype (risk or normal) were identified. Extracted features were ordered by statistical significance, being the most important those with lower P-values. In Fig. 1(b), SNPs with P-values lower than 10^{-3} are highlighted in green. The red line represents the significant level while the blue line indicates, this time, the new threshold considered (P-values < 1×10^{-3}). The figure displays a Manhattan plot of SNPs considered after suggestive threshold modification.

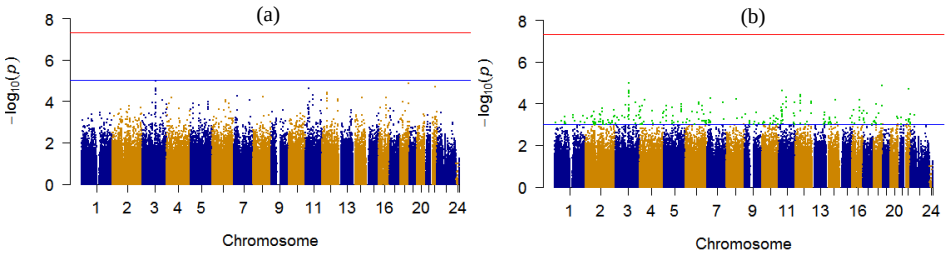

Fig.1. Manhattan Plot for GWAS: (a) suggestive threshold P-value < 1×10^{-5}, (b) after suggestive threshold modification P-values < 1×10^{-3}.

4. Feature Selection

After features were extracted, we explored feature selection to determine which features might be the most relevant when discriminating between risk and normal classes.

Some samples had missing genotypes for some individuals so we removed them, resulting in a final number of 185 SNPs considered as features for classification analysis. Additionally, age, gender and T2D were not included in the total set of features i.e., only genetic variants were considered.

Feature selection is performed using Boruta, a random forest (RF) based feature selection method, which provides unbiased and stable selection of important and non-important attributes [25]. Random Forest has been successfully used in genomic data analysis as it is highly data adaptive and accounts for correlation as well as interactions among features[26].

Features selected identified were ranked by importance and divided into three groups. The first group contained the top five most prominent features, the second group the top ten and the third group the top fifteen.

Results were compared against those reported when the top most notable features extracted from association analysis were considered, as we will discuss in the following sections.

5. Results

This section presents the classification results for normal and risk BMI status using data extracted from the PGP website.

After the QC filter process, 722,512 genetic variants and 223 people (145 males and 78 females) remained for the analyses. Subsequent genetic association analysis using logistic regression allowed us to reduce the number of variants to 261. However, missing genotypes in some of the samples caused a further reduction in the number of SNPs (185 SNPs remained).

The top features extracted after QC and association analysis and, those selected by Boruta, are used to model a Support Vector Machine with Radial Basis Function Kernel (SVM) classifier. Support Vector Machine is a well-known machine learning algorithm which provided the best results in previous experiments using similar data [27].The performance is measured using sensitivity (SE), specificity (SP) and area under the curve (AUC) values. In this study, it is important to predict risk classes, therefore SE are considered higher priority than SP.

K-fold cross validation is used as a prediction metric with 10 folds and 30 repetitions. The average performance obtained from 30 simulations is utilized. This number is considered, by statisticians, to be an adequate number of iterations to obtain an acceptable average. Support Vector Machine was designed and evaluated using appropriate training and testing sets. The selection of hyperparameters to establish an approximately optimal configuration for SVM is addressed using Caret for random search parameter tuning [28]. Tuning parameters, free parameter of the Gaussian radial basis function (sigma) and penalty cost (C), shown in Table 3 and Table 5 produced the models with the best receiver operator characteristic (ROC) curve values.

The performance of SVM when the algorithm is trained and tested with the top features identified in the association analysis ranked by P-value is shown in Table 3 and Table 4 respectively. Conversely, the performance of SVM when fed with the features selected by Boruta are organised in Table 5 for training and Table 6 for testing. Details on the SNPs extracted and selected can be found in Appendix.

Table 3. Training

Association Features	Sensitivity	Specificity	ROC	Best tuning parameters
Top 5 SNPs	0.6322	0.8476	0.7859	σ = 0.0215 C = 1.0203
Top 10 SNPs	0.7856	0.7325	0.8672	σ = 0.0125 C = 1.0203
Top 15 SNPs	0.8399	0.8578	0.9142	σ = 0.0105 C = 1.0203

Sensitivity, Specificity and ROC values for SVM performance in the training data when using extracted features from association analysis.

Table 4. Prediction

Association Features	Sensitivity	Specificity	ROC
Top 5 SNPs	0.7692	0.6897	0.8150
Top 10 SNPs	0.6923	0.9586	0.8622
Top 15 SNPs	0.8462	0.8276	0.9092

Sensitivity, Specificity and AUC values for SVM when predicting the two classes in the test data, using extracted features from association analysis.

Table 5. Training

RF Features	Sensitivity	Specificity	ROC	Best tuning parameters
Top 5 SNPs	0.7514	0.7855	0.8318	σ = 0.0183 C = 1.0203
Top 10 SNPs	0.7849	0.7704	0.8482	σ = 0.0106 C = 1.0203
Top 15 SNPs	0.8291	0.8071	0.9011	σ = 0.0120 C = 1.0203

Sensitivity, Specificity and ROC values for SVM performance in the training data when using features selected by Boruta.

Table 6. Prediction

RF Features	Sensitivity	Specificity	ROC
Top 5 SNPs	0.6154	0.7931	0.8176
Top 10 SNPs	0.7692	0.7931	0.8674
Top 15 SNPs	0.8077	0.8276	0.9231

Sensitivity, Specificity and AUC values for SVM when predicting the two classes in the test data, using features selected by Boruta.

To illustrate the performance in binary classification, it is particularly advantageous to use the ROC curve. It is a convenient way of displaying the cut-off values for the false and true positive rates. The ROC curves in Fig. 2 illustrates the SE, SP and AUC values in Table 4 and Table 6. The models with ROC curve closer to the top left corner show higher performance as the SE and SP increase. Therefore, the area under the curve increases as the curve moves away from the grey diagonal line towards top left corner of the graph.

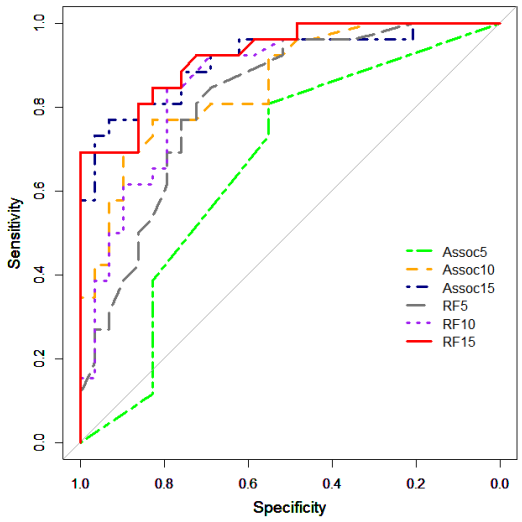

Fig. 2. ROC curves for PGP data when using the subsets of SNPs extracted by association analysis and selected by Boruta.

6. Discussion

The web-scrapping process applied in this study is susceptible to failures in the future if the PGP website structure changes. This is an issue referred to as "medieval torture" [29].

PGP dataset is pre-processed via standard QC and association analysis protocols for GWAS. Although no SNPs were identified as significant or suggestive, we included SNPs with P-values lower than 1×10^{-3} for subsequent analyses as accomplished somewhere else [24]. After QC, 722,512 SNPs were considered for association and lately reduced to 261 SNPs showing certain level of importance among all the variants, using logistic regression. These SNPs are highlighted in Fig. 1(b). Finally, 185 SNPs were considered for classification analysis.

The total 185 features are a subset with the most relevant SNPs obtained after applying QC and logistic analysis to the genetic data binary files. The top most significant features were then organised in three groups to be compared against the most relevant features selected by Boruta.

Using RF-based algorithm as a feature selection technique, three groups of the top features with the highest discriminatory capacity were selected.

Results revealed that using RF-based algorithm ranking of features resulted in an improvement in the performance of SVM when predicting the risk and normal cases. All the ROC values obtained with the three sets of SNPs in Table 7 are higher than those obtained by the set of most statistically significant features listed in Table 5. In most cases, SE were lower than SP, which is not encouraging given that predicting pathological cases is more important than those that are normal. However, when RF-based method is used, the top fifteen features produced a closely balance SE and SP values of 81% and 83% respectively.

Results in Table 7 indicate that SVM showed the best results with SE=81%, SP=83% and AUC=92% when the model was trained with the top fifteen features selected by Boruta. These features are listed in Appendix section. Reducing the number of features to five did not result in an improvement in the classifier performance.

The ROC curves from Fig. 2 shows how using the top fifteen features selected by Boruta (red ROC curve) allows the highest discrimination between the two classes considered in our study. The lowest performance was achieved with the top five SNPs extracted in association analysis, which is represented in green colour in Fig. 2.

Additionally, the most important feature reported in both approaches was rs4821758 as reported in Appendix.

7. Conclusions

This paper focuses on an approach for selecting informative SNPs from publicly available data collected using web scrapping techniques. The created dataset was built from research-grade data (that is, not for clinical use), and the conductors of the PGP stated that many types of errors are possible. Some of these include errors in the data, failure to report or discover significant genetic issues and ambiguous or false positive findings. This suggests the utilisation of a more reliable data set in future studies, for a solid discovery of genetic risk variants in complex disease prediction.

A small portion of SNPs that have main effects on obesity as binary trait, have been selected after applying QC and association analysis using logistic regression. Subsequent analysis applying a Support Vector Machine with Radial Basis Function Kernel classifier are conducted for the evaluation of the model in two scenarios. First, the algorithm was evaluated using a subset of the most statistically significant genetic variants obtained from GWAS analysis, based on a modified suggestive threshold. Then, results were compared when a subset of features were selected using the Random Forest based algorithm Boruta. Using the selected features improved the performance of SVM although the subset of fifteen SNPs achieved the highest performance.

While the results show specific genetic variants that could serve as good discriminators in the investigation of classification studies, more analysis with a higher representation of samples must be carried out. We propose a set SNPs to be used in future studies as features for the prediction of obesity and other comorbidities such as T2D. The identified genetic variants need to be validated and contrasted with other studies, particularly the SNP rs4821758, which was the most important feature in the association analysis as well as the feature selection process using Boruta. Future work will consider the discriminative capacity of the SNPs identified in this study evaluated in a more complete dataset. A comparison between various feature selection techniques will also be considered.

Appendix

Ranking of Features Considered in the Study.

	Features Extracted in Association Analysis					Features Selected by RF-based algorithm			
Rank	SNP	CHR	BP	Allele	P-Value	SNP	CHR	BP	Allele
1	rs4821758	22	38591190	C	1.980e-05	rs4821758	22	38591190	C
2	rs6768523	3	111962851	C	2.449e-05	rs10790866	11	127063240	G
3	rs9872691	3	111985107	C	2.449e-05	rs7117995	11	19947811	C
4	rs9871650	3	111914624	C	2.449e-05	rs9872691	3	111985107	C
5	rs9288938	3	111921116	A	2.449e-05	rs7574062	2	71716494	T
6	rs441703	11	29822605	T	2.458e-05	rs9871650	3	111914624	C
7	rs4682278	3	111225977	A	3.036e-05	rs9288938	3	111921116	A
8	rs6437989	3	111203691	G	3.036e-05	rs12570718	10	116416959	T
9	rs1553090	3	111180433	A	3.167e-05	rs6768523	3	111962851	C
10	rs10880063	12	41760767	T	3.940e-05	rs2159723	2	230045272	T
11	rs1000147	12	41727678	C	4.659e-05	rs7776422	6	106240162	G
12	rs1451327	11	57991093	A	5.086e-05	rs4483247	9	37045825	G
13	rs10957744	8	75899434	A	6.360e-05	rs5006218	6	133126220	G
14	rs7498886	16	62079202	G	6.526e-05	rs2207900	20	54289508	C
15	rs12579740	12	125177752	T	7.871e-05	rs9493446	6	133125643	T

Ranking of features extracted by association analysis (shown as shaded) and selected by Boruta. The SNPs are listed in order of importance. The top features extracted using logistic regression are ordered by P-value whilst the features selected by RF-based algorithm are ordered by importance. Information about the chromosome number, base pair position and allele is provided. In addition to this information, the P-values for features selected in association analysis are also listed.

References

1. James, W.P.T.: WHO recognition of the global obesity epidemic. Int. J. Obes. (Lond). 32 Suppl 7, S120–S126 (2008).
2. Poloz, Y., Stambolic, V.: Obesity and cancer, a case for insulin signaling. Cell Death Dis. 6, e2037 (2015).
3. Rao, K.R., Lal, N., Giridharan, N. V: Genetic & epigenetic approach to human obesity. Indian J. Med. Res. 140, 589–603 (2015).
4. Li, S. et al.: Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. PLoS Med. 7, 1–9 (2010).
5. Bello, A. et al.: Using linked administrative data to study periprocedural mortality in obesity and chronic kidney disease (CKD). Nephrol. Dial. Transplant. 28, iv57-iv64 (2013).
6. Loos, R.J.F.: Genetic determinants of common obesity and their value in prediction. Best Pract. Res. Clin. Endocrinol. Metab. 26, 211–226 (2012).
7. Samish, I., Bourne, P.E., Najmanovich, R.J.: Achievements and challenges in structural

bioinformatics and computational biophysics. Bioinformatics. 31, 146–150 (2014).
8. Higdon, R. et al.: Unravelling the Complexities of Life Sciences Data. Big Data. 17–23 (2012).
9. Tanwani, A.K., Afridi, J., Shafiq, M.Z., Farooq, M.: Guidelines to select machine learning scheme for classification of biomedical datasets. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 5483 LNCS, 128–139 (2009).
10. Su, P.: Direct-to-consumer genetic testing: a comprehensive view. Yale J. Biol. Med. 86, 359–65 (2013).
11. Ball, M.P. et al.: Harvard Personal Genome Project: lessons from participatory public research. Genome Med. 6, 10 (2014).
12. Glez-Pena, D., Lourenco, A., Lopez-Fernandez, H., Reboiro-Jato, M., Fdez-Riverola, F.: Web scraping technologies in an API world. Brief. Bioinform. 15, 788–797 (2014).
13. Marx, V.: Biology: The big challenges of big data. Nature. 498, 255–260 (2013).
14. Tryka, K.A. et al.: NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 42, D975–D979 (2014).
15. Gonzaga-Jauregui, C., Lupski, J.R., Gibbs, R.A.: Human Genome Sequencing in Health and Disease. Annu. Rev. Med. 63, 35–61 (2012).
16. Bush, W.S., Moore, J.H.: Chapter 11: Genome-Wide Association Studies. PLoS Comput. Biol. 8, (2012).
17. Fadista, J., Manning, A.K., Florez, J.C., Groop, L.: The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. 24, 1202–1205 (2016).
18. Zhang, Y.-B. et al.: Genome-wide association study identifies multiple susceptibility loci for craniofacial microsomia. Nat. Commun. 7, 10605 (2016).
19. Stoeklé, H.-C., Mamzer-Bruneel, M.-F., Vogt, G., Hervé, C.: 23andMe: a new two-sided data-banking market model. BMC Med. Ethics. 17, 19 (2016).
20. Purcell, S. et al..: PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
21. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., Zondervan, K.T.: Data quality control in genetic case-control association studies. Nat. Protoc. 5, 1564–73 (2010).
22. Turner, S. et al.: Quality control procedures for genome-wide association studies. Curr. Protoc. Hum. Genet. Chapter 1, Unit1.19 (2011).
23. Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M.P., Foulkes, A.S.: A guide to genome-wide association analysis and post-analytic interrogation. Stat. Med. 34, 3769–3792 (2015).
24. Gül, H., Aydin Son, Y., Açikel, C.: Discovering missing heritability and early risk prediction for type 2 diabetes: a new perspective for genome-wide association study analysis with the Nurses' Health Study and the Health Professionals' Follow-Up Study. Turkish J. Med. Sci. 44, 946–954 (2014).
25. Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. J. Stat. Softw. 36, 1–13 (2010).
26. Cordell, H.J.: Detecting gene–gene interactions that underlie human diseases. Nat. Rev. Genet. 10, 392–404 (2009).
27. Curbelo Montañez, C.A. et al: Machine Learning Approaches for the Prediction of Obesity using Publicly Available Genetic Profiles. In: 2017 International Joint Conference on Neural Networks (IJCNN). p. 8. , Anchorage, Alaska (2017).
28. Kuhn, M.: Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28, 1–26 (2008).
29. Stein, L.: Creating a bioinformatics nation. Nature. 417, 119–120 (2002).