# Marker genes of incident type 1 diabetes in peripheral blood mononuclear cells of children: A machine learning strategy for large-p, small-n scenarios

Kushan De Silva[1], Ryan T. Demmer[2,3], Daniel Jönsson[4,5], Aya Mousa[1], Andrew Forbes[6], Joanne Enticott[1]

[1]Monash Centre for Health Research and Implementation, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Clayton, 3168, Australia

[2]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA

[3]Mailman School of Public Health, Columbia University, New York, USA

[4]Department of Periodontology, Faculty of Odontology, Malmö University, Malmö, 21119, Sweden

[5]Department of Clinical Sciences, Lund University, Malmö, 21428, Sweden

[6]Biostatistics Unit, Division of Research Methodology, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Melbourne, 3004, Australia

**Name and contact information of the corresponding author**

Kushan De Silva

**Postal Address**

Locked Bag 29

Level 1, 43-51 Kanooka Grove

Clayton VIC 3168

Australia

Monash Centre for Health Research and Implementation

School of Public Health and Preventive Medicine

Faculty of Medicine, Nursing and Health Sciences

Monash University

Australia

**Email:** kushan.ranakombu@monash.edu

Telephone: +61 3 99056824

**ORCID ID:** 0000-0003-0301-0805

**Key-words/phrases:** Biomarkers; Dimension reduction; Gene expression; High dimensionality; Machine learning; Type 1 diabetes

## ABSTRACT

**Background and objective:** Type 1 diabetes (TID) is a complex, polygenic disorder, the etiology of which is not fully elucidated. Machine learning (ML) genomics could provide novel insights on disease dynamics while high-dimensionality remains a challenge. This study aimed to identify marker genes of incident T1D in peripheral blood mononuclear cells (PBMC) of children via a ML strategy attuned to high-dimensionality.

**Methods:** Using samples from 105 children (81 with incident T1D and 24 healthy controls), we analyzed microarray transcriptomics via a workflow consisting of three sequential steps: application of dimension reduction strategies on the processed transcriptome; ML on the reduced gene expression matrix; and downstream network analyses to demarcate seed nodes (statistically significant genes) and hub genes. Sixteen dimension-reduction algorithms belonging to three groups (3 tailored; 3 regularizations; 10 classic) were applied. Four ML algorithms (multivariate adaptive regression splines, adaptive boosting, random forests, XGB-DART) were trained on the reduced feature set and internally-validated using repeated, 10-fold cross-validation. Marker genes were determined via variable importance metrics. Seed nodes were identified by the '*OmicsNet*' platform while nodes having above average betweenness, closeness, and degree in the network were demarcated as hub genes.

**Results:** The processed gene expression matrix comprised 13515 genes which was reduced to contain 1003 genes collectively selected by dimension reduction algorithms. All four ML algorithms on this reduced feature set attained perfect and uniform predictive performance on internal validation. On removal of redundancies, variable importance metrics identified 30 marker genes of incident T1D in this cohort, while Early Growth Response 2 (EGR2) was uniformly selected by all four ML algorithms as the most important marker gene. Network

analyses classified all 30 marker genes as seed nodes. Additionally, we identified 14 hub genes, 7 of which were found to be marker genes of incident T1D elucidated by ML.

**Conclusions:** We identified marker genes of incident T1D in PBMC of children via a ML analytic strategy attuned to the high dimensional structure of microarrays, with downstream analyses providing high biological plausibility. The demonstrated ML strategy would be useful in analyzing other high-dimensional biomedical data for biomarker discovery.

**Keywords: Biomarkers; Dimension reduction; Gene expression; High dimensionality; Machine learning; Type 1 diabetes**

## 1. INTRODUCTION

Type 1 diabetes (T1D) is a chronic, polygenic disorder with a multifactorial aetiology encompassing strong genetic, autoimmune and environmental components, although its natural history is not yet fully elucidated [1]. It accounts for nearly 10% of the global diabetes prevalence and is the most common form of diabetes among children [2]. Insulin secretory dysfunction and hyperglycemia being hallmarks of the disease, people with T1D require lifelong insulin therapy. Increased mortality rates and complications [3], concomitantly with reduced quality of life [4], frequently associate with T1D.

Pathogenesis of T1D is driven by T-cell mediated destruction of β-cells, whilst autoantibodies targeting insulin, glutamic acid decarboxylase 65 (GAD65), tyrosine phosphatase-related islet antigen 2 (IA-2) and zinc transporter-8 (ZNT8) tend to appear long before symptomatic onset of the disease, and are thus considered early biomarkers of T1D [5]. Current evidence suggests an increased susceptibility of individuals with human leukocyte antigen (HLA)-DR and HLA-DQ genotypes to T1D [6], while genome-wide association studies have discovered a large number of non-HLA, T1D-associated genes as well [7]. Moreover, T1D-discordant monozygotic twin studies have reinforced the impact of environmental triggers, ascribing a potentially causal role for epigenetic changes such as hypomethylated gene promoter regions, in T1D pathogenesis [8, 9]. While important advances have been made with respect to T1D along its continuum of care, further concerted efforts are required to identify biomarkers and facilitate early detection and management to prevent complications.

Precision medicine initiatives combining analytic approaches such as artificial intelligence (AI) with multi-omics data, have shown promise for yielding novel and valuable findings on diseases including diabetes [10, 11]. With rapid and continuous progress in the disciplines of

big data, omics, and AI, multidisciplinary studies amalgamating these domains to address knowledge gaps pertaining to the genetic basis of complex diseases such as T1D, are now a reality. As the classical statistical paradigm is constrained in its capability to handle the intricacies associated with high-dimensional data, machine learning (ML) offers a viable alternative for analyzing omics data.

The large-p, small-n structure (p>>>n), also coined as the "curse of dimensionality", in which the feature space (p) heavily outnumbers the observations (n), is inherent in omics data including in gene expression microarrays, microbiome compositional data, and single cell RNA-sequencing assays. Being counterfactual to the common phenomenon of large-n, small-p contexts, classical statistical models are challenged by this unique structure. Various techniques proposed to address the curse of dimensionality in omics data, are dominated by feature selection workflows [12] and regularization techniques [13], while de-novo algorithms tailored to omics data have also been developed [14, 15]. Although a single best method to deal with high-dimensionality does not exist, proposed approaches have unequivocally contributed to increasing the robustness of omics data analytics and biomarker discovery.

In a previous study, we demonstrated the viability of a strategy combining multiple feature selection algorithms with ML for elucidating the markers of prediabetes using an epidemiological cohort [16]. However, omics data such as gene expression microarrays present a formidable analytic challenge with an enormous feature space of a disproportionately smaller sample. Robustness of an analogous approach combining classic feature selection, regularization, and de-novo dimension reduction algorithms with ML upon omics data is hitherto untested.

Gene expression profiles of incident T1D could potentially consist of marker genes and reveal insights on causally-linked transcriptomic alterations associated with the onset of T1D. Identification of peripheral blood-based biomarkers of T1D can also be useful for clinical diagnostic and screening purposes. In a previous study, differential expression of interleukin 1 beta (IL1B), early growth response 3 (EGR3), and prostaglandin-endoperoxide synthase 2 (PTGS2) genes in peripheral blood mononuclear cells (PBMC) was associated with incident T1D [17]. The objective of the present study was to analyze the gene expression profile of PBMC with respect to incident T1D in children using a ML strategy amenable to large-p, small-n scenarios and to identify marker genes. Further, we determined the biological plausibility of the findings by conducting downstream analyses including constructing protein-protein interactions (PPI) networks and demarcating hub- and seed- nodes.

## 2. METHODS

The analytic workflow is illustrated in **S1 Figure**.

### 2.1. Data retrieval and processing

Data used in this study are available on the open-source National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) [18, 19] platform with the unique GEO accession ID of GSE9006. Microarray data were retrieved via the '*getGEO*' function of '*GEOquery*' R package [20]. Of note, GSE9006 contains gene expression profiles of children with both T1D and T2D, measured on diagnosis at baseline and 4 months afterwards, following treatment. As per the objectives of this study and since the sample of children with T2D was also smaller, only T1D samples were included. This comprised 105 samples in total, of which 81 were children with T1D and 24 were healthy controls. Furthermore, as gene expression profiles change with treatment, only baseline data were eligible. Observations without corresponding gene symbol annotations were removed. In case of

multiple probes hybridized to the same gene, their mean expression values were estimated across the samples to create a single row each for a given gene symbol. The gene expression matrix was transposed and three demographic features deemed important, namely, age (years), gender, and race (White/Other), were annotated to the transposed matrix. Gene expression values of this curated dataset was $\log_2$ Normalized to create symmetric distributions devoid of skewness (**S2 Figure**).

### 2.2. Dimension reduction

We applied 16 dimension-reduction algorithms belonging to three categories on the curated, transposed, and $\log_2$ Normalized gene expression matrix (**Table 1**). A brief description of these algorithms follows.

2.2.1. De-novo algorithms tailored to handling the curse of dimensionality in omics (n = 3)

- *Sparse Principal Component Analysis (SPCA)*: Incorporated in the '*mixOmics*' R-package [14], this algorithm selects features via singular value decomposition and lasso penalization on the loading vectors. The optimal number of principal components (PC) was determined via elbow method.
- *Sparse Partial Least Squares Discriminant Analysis (SPLSDA)*: Incorporated in the '*mixOmics*' R-package [14], this supervised algorithm performs partial least squares-based feature reduction administering an $L_1$ penalty on the loading vectors of the input matrix. Parameter tuning was performed to determine the optimal number of features and components by evaluating the balanced classification error rate (BER) of PLSDA against the number of features and components (**Figures 1-5**).

- *Max-Min Parents and Children (MMPC)*: Incorporated in the *MXM* R-package [15], this algorithm conducts a constraint-based feature selection, assuming a Bayesian network for input variables. The permutation option was activated (R = 999) and the max_k value was specified as: sample size/10 for optimizing the performance.

2.2.2. Regularization algorithms (n = 3)

- *Lasso*: This algorithm performs feature reduction based on $L_1$ regularization, which adds a penalty term equal to the absolute value of the magnitude of coefficients.
- *Ridge*: This algorithm performs feature space reduction based on $L_2$ penalization which adds a penalty term equal to the square of the magnitude of coefficients.
- *Elastic net*: This algorithm performs feature selection via a regularization method which linearly combines $L_1$ and $L_2$ penalties.

2.2.3. Classic feature selection algorithms (n = 10)

- *Boruta*: This is an all-relevant wrapper algorithm around Random Forests which selects features based on mean decrease accuracy, by default [21].
- *Recursive feature elimination*: A greedy algorithm which extracts the best subset of features by iteratively building models, rank ordering feature importance and removing the least important features.
- *Information gain, gain ratio, and symmetrical uncertainty*: These entropy-based algorithms, incorporated in the *FSelector* R package [22], determine feature weights based on their correlation with the target variable.
- *Linear correlation and rank correlation*: These two correlation-based algorithms, integrated into the *FSelector* [22], utilise Pearson's- and Spearman's correlation coefficients, respectively, for feature selection.

- *Random forests*: A filter algorithm which selects features based on a random forest learner, found in *FSelector* [22].
- *One R*: A simple filter algorithm which selects features by generating one rule per predictor, found in *FSelector* [22].
- *Chi Squared*: This algorithm in *FSelector* [22] attributes importance to discretised features based on a chi squared test.

Features selected by each algorithm were compiled and repeated genes were removed. Demographic features (age, gender, race) were appended and the categorical variables (gender, race) were one-hot encoded to create dummy variables.

### 2.3. Machine learning

Machine learning with hyperparameter tuning was performed on the reduced feature set using four algorithms, namely, multivariate adaptive regression splines (MARS), adaptive boosting (AdaBoost), random forests (RF), and extreme gradient boosting-dropouts meet multiple additive regression trees (XGB-DART). The MARS algorithm is a non-parametric regression technique suited for high dimensional data, and an advancement of linear models, capable of automatic accounting for non-linearities and interactions within the feature space [23]. The AdaBoost is a meta-algorithm ensemble of decision trees which generates boosted classifiers by combining weak decision tree learners into a weighted sum [24]. The RF algorithm is an ensemble meta-learner based on bootstrap aggregating (“bagging”) of a multitude of decision trees accounting for overfitting [25]. The XGB-DART is a hybrid algorithm of two learners: XGBOOST is an ensemble algorithm [26] which collates a large number of decision trees with a small learning rate, while DART incorporates dropout techniques emanating from the deep neural network algorithmic paradigm in order to overcome overfitting [27]. For each algorithm, hyperparameter tuning was performed using *tuneLength* and *tuneGrid* commands

of the *Caret* R package [28] and the metric for the optimal model selection was specified as "ROC".

Internal validation of each model was performed by repeated, k-fold cross-validation (k = 10). The cross-validation approach has been the method of choice for internally validating omics-based ML models including gene expression analytics [29, 30]. In k-fold cross-validation, data are split into k non-overlapping folds. Each fold is set aside as test data while all other folds are combined into training sets. After a total of k models are fitted, performances are evaluated on the k test datasets and the mean performance is estimated. Repeated k-fold cross validation is an advancement which reduces the noise (error) of mean performance metrics attained by k-fold cross-validation.

The predictive performance of each model was evaluated using multiple performance metrics (**Table 2, S3 Figure**) and marker genes were identified via variable importance metrics (**Table 3, Figures 6-9**).

### 2.4. Downstream analysis of marker genes

The ten most important genes identified by each ML model was compiled and repetitions were removed to generate the list of marker genes of incident T1D in PBMC of children. This list of genes was fed into *OmicsNet* [31, 32] which determines the significant genes (seeds) within the PPI network (**Figure 10**) by mapping input genes to the specified molecular interactions database.

To demarcate the hub genes, we first created the gene-gene interactions network in GENEMANIA [33], which integrates an array of functional enrichments such as co-expression, pathways, physical interactions, co-localization, genetic interactions, and protein domain similarity into the network (**S4 Figure**). Next, the network was imported to *Cytoscape* [34] and its *Centiscape* plugin [35] was used to visualize the network and estimate

topological metrics (**Figure 11**). As per Liu et al [36], nodes with above-average betweenness, closeness, and degree values were defined as hub genes (**Table 4**).

## 3. RESULTS

The unprocessed gene expression matrix of PBMC at baseline consisted of 22283 probes measured on 105 children (81 with incident T1D and 24 healthy controls). After pre-processing, 13515 probes with unique gene symbols were retained (**S1 Table**). Age(years) distribution ranged from 2 – 17, with a mean (SD) of 9.81 (3.91) and a median (IQR) of 10.00 (6.00). Samples comprised 59 female and 46 male individuals. The dichotomized covariate of self-reported race was distributed as: 60 White and 45 Other (Black or African American = 8; selected more than one race = 5; other race = 3, race not reported = 29).

### 3.1. Features selected by tailored algorithms

Genes selected by running each of the 16 dimension-reduction algorithms are presented in **S2 Table**. For the SPCA algorithm, the optimal number of PC was found to be three according to the elbow method, and the 150 features selected (50 each from the 3 PC) are shown in **S2 Table**. The optimal number of features and components minimizing BER, with respect to the SPLSDA algorithm were found to be 150 and 3, respectively (**Figures 1 & 2**), and the selected 150 features (50 each from the 3 components) are illustrated in **Figures 3 - 5**. The **S3 Table** contains the complete feature ranking produced by the MMPC algorithm while the top 100 features ranked according to p-values that were selected for ML are shown in **S2 Table**.

### 3.2. Features selected by regularization algorithms

Seven candidate genes with non-zero coefficients (CSDE1, PPP2R5E, ZNF473, ACAD10, PNP, EGR2, LPIN2) were produced by Lasso regularization (**S4 Table**). The top 100 features with the largest *absolute* coefficients were selected from the ridge regularization algorithmic

output (**S5 Table**). Elastic net algorithm converged with 377 non-zero feature coefficients (**S6 Table**).

### 3.3. Features selected by classic algorithms

Boruta algorithm selected 48 candidate genes (20 confirmed and 28 tentative) (**S7 Table**). Recursive feature elimination identified a sum of 22 variables maximizing cross-validated accuracy (**S2 Table**). Output produced by the eight algorithms incorporated in *FSelector* is given in **S8 Table**. Each of the three entropy-based algorithms (information gain, gain ratio, symmetrical uncertainty) converged with 379 non-zero coefficients. The top 100 features from the two correlation-based algorithms (rank- and linear- correlation) were selected for ML. We selected the top 100 features produced by random forests and One-R feature selection algorithms as well. Finally, the chi-squared algorithm generated 379 non-zero feature coefficients.

After removing redundancies, the collated feature set produced by all algorithms contained 1003 genes (**S9 Table**). The curated dataset containing these 1003 candidate genes and basic demographic variables (age, gender, race) used for ML is presented in **S10 Table**.

### 3.4. Machine learning

All four repeated, 10-fold cross-validated ML models converged attaining perfect and uniform predictive accuracy, classification-, and discrimination- metrics (**Table 2; S3 Figure**).

### 3.5. Variable importance

The 10 most important genes identified by the MARS model, in descending order, were: EGR2; RAP1B; RNF4; CSDE1; GAR1; MLEC; CTDSP1; LOX; POMC; CNTLN (**Figure 6**). The AdaBoost model elucidated: EGR2, PTP4A2, PNP, ZNF473, SLC35A3, KBTBD4,

FRAT2, UCHL3, IL1B, CHFR, as the top 10 marker genes, in descending order (**Figure 7**). The RF model ranked: EGR2; GNG11; ZNF473; SUMO3; LPIN2; CR2; CRYL1; UCHL3; PTP4A2; ACAD10 genes as the top 10 marker genes, in diminishing importance (**Figure 8**). The most ten important genes identified by XGB-DART model, in descending order, were: EGR2; MICA/////MICB; ZNF473; SUMO3; LPIN2; PTP4A2; RAP1B; EXOC7; MRPS10; NR1D2 (**Figure 9**). Collation of the marker genes identified by the four ML models and removal of redundancies resulted in 30 candidate genes (**S11 Table**).

### 3.6. Network modules and seed nodes

Based on the 30 input genes, *OmicsNet* constructed a single subnetwork consisting of 504 nodes, 575 edges, and 23 modules (**Figure 10**), while all 30 marker genes were identified as seeds (statistically significant nodes of the PPI network) (**S12 Table**).

### 3.7. Hub genes

Details on the gene-gene interaction network constructed by GENEMANIA from the 30 input marker genes are provided in **S13 Table**. We identified 14 hub nodes within the PPI network which had above-average betweenness, closeness, and degree as per the topological values estimated by *Centiscape* (**S14 Table**). Seven of these hub nodes were found within the set of marker genes (MLEC, NR1D2, CSDE1, UCHL3, RNF4, RAP1B, PNP) while the remaining 7 hub genes were predicted by GENEMANIA (CDV3, GTF2H3, IMP3, CTDSP2, SP3, H1-0, CUL2).

## 4. DISCUSSION

In this study, we found a range of marker genes of incident T1D in PBMC of children, via an extensive ML workflow amenable for high dimensionality inherent in transcriptomics microarrays. Biological plausibility of our findings was confirmed by downstream analyses and appraisal of contemporary evidence. We envisage that the proposed ML strategy has the

potential to integrate into various omics data types for an efficient biomarker discovery process.

### 4.1. Marker genes of incident T1D in PBMC of children: Biological plausibility

Noteworthily, the zinc finger transcription factor EGR2 was uniformly chosen by all four ML algorithms as the most important gene associated with incident T1D in this cohort. This is in tandem with previous studies in which EGR genes (both EGR2 and EGR3) were identified as highly perturbed at the onset of T1D [17]. There is emerging evidence on the role of EGR2 in the natural history of diabetes, through the development of insulin resistance which is increasingly observed in people with T1D [37] to the progression of complications [38, 39]. Recent studies which revealed crucial roles of EGR2 in the induction of T cell anergy and suppression of activities mediated by LAG3+ regulatory T cells (Tregs) are intriguing [40], given that Tregs are strongly implicated in T1D etiopathogenesis [41] and extensively investigated as therapeutic targets to ameliorate islet autoimmunity of T1D [42].

Our findings on the marker genes of incident T1D were congruent with previous studies which reported an innate inflammatory transcriptomic profile characterized by perturbations of genes such as CSDE1, EGR2, FRAT2, GNG11, IL1B, PTP4A2, SLC35A3, and UCHL3 [17, 43]. Moreover, there is topical evidence that RAP1B regulates glucose homeostasis [44], while the dysregulation of protein tyrosine phosphatases such as PTP4A2 (PRL2) is observed in several diabetes phenotypes [45]. Both MICA and MICB candidate genes belong to the HLA complex, which is collectively and consistently associated with T1D, contributing to 50% of its genetic risk [6]. An integrated analysis revealed that PNP is a strong diagnostic marker of T1D [46]. Furthermore, CR2 was found to be associated with abatacept-resistant, new-onset T1D, suggesting a likely role in B lymphocyte alterations [47]. Contemporary evidence also supports associations of POMC [48], and NR1D2 clock gene [49], with T1D.

### 4.2. Seeds and hub genes: Biological plausibility and implications

The inclusion of all 30 candidate genes as statistically significant nodes (seeds) in the PPI network was indicative of substantial gene-gene interactions between them. Further studies exploring these interactions could enhance our understanding of the genetic basis of T1D. Hub genes which appear as extensively and densely connected nodes in scale-free gene regulatory networks [50] tend to have important regulatory functions and offer value as potential therapeutic targets [51]. Therefore, hub genes identified in this analysis may be useful for guiding pharmacogenomics studies focused on T1D.

### 4.3. Methodological and analytic aspects

We observed a considerable degree of consistency with respect to the marker genes of T1D identified by ML, notwithstanding the essentially different dynamics between algorithms. However, some marker genes were uniquely identified by each algorithm, underscoring the need to train more than a single learner to gain robust and meaningful insights; a recommended practice in ML analytics [16]. It should also be noted that ML is intrinsically geared to optimizing prediction, although it is increasingly used to derive etiologic insights of diseases [52, 53]. Further studies focusing on the candidate genes identified in the current analysis, pragmatically combined with relatively larger samples are recommended, as they may provide novel insights on T1D pathogenesis. While the perfect predictive performance demonstrated by ML algorithms is encouraging, we underscore the caveats associated with the cross-validation approach, which is frequently used in omics analyses primarily owing to the large-p, small-n structure. Although we used repeated, k-fold cross-validation, which was found superior to other cross-validation techniques, they all have limitations [54]. Therefore, future research aiming to expand on this work should focus on externally validating our findings on different cohorts or internally validating on larger samples.

### 4.4. Applications

As high-dimensionality is frequently observed across different omics data typologies [55-57], we envision avenues for practical usage of the proposed ML strategy, beyond microarray transcriptomics.

## 5. CONCLUSIONS

In conclusion, we identified marker genes of incident T1D in PBMC of children via a ML analytic strategy attuned to the high dimensional structure of microarrays, with downstream analyses on seed nodes and hub genes providing high biological plausibility. The demonstrated ML strategy would have broader applications to studies aimed at biomarker discovery using high-dimensional, biomedical data.

## DECLARATIONS

**Funding:** KDS is supported by a PhD scholarship funded by the Australian Government under Research Training Program (RTP).

**Role of the Funder/Sponsor:** The funder was not involved in the design of the study; the collection, analysis, and interpretation of data; writing the report; and did not impose any restrictions regarding the publication of the report.

**Ethics approval:** Not required being a secondary analysis of publicly available, deidentified data.

**Conflicts of interest statement:** Authors declare that there are no conflicts of interest.

**Author contributions:** KDS performed data acquisition, pre-processing and curation of data, conducted the analyses and wrote this manuscript. KDS, RTD, AF and JE were responsible for study conceptualization and design, contributed to validating analyses, results and interpretation, and drafting the manuscript. DJ and AM contributed to study design,

interpretation of data and critically and comprehensively revised the manuscript for important intellectual content.

**Availability of data and material**: Data used in this study are freely available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) portal (https://www.ncbi.nlm.nih.gov/geo/), with the unique GEO accession ID of GSE9006.

## REFERENCES

1. DiMeglio LA, Evans-Molina C, Oram RA. Type 1 diabetes. *Lancet*. 2018;391(10138):2449-2462. https://doi.org/10.1016/S0140-6736(18)31320-5
2. Mobasseri M, Shirmohammadi M, Amiri T, Vahed N, Hosseini Fard H, Ghojazadeh M. Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. *Health Promot Perspect*. 2020 Mar 30;10(2):98-115. DOI: 10.34172/hpp.2020.18.
3. Bjerg L, Gudbjörnsdottir S, Franzén S, Carstensen B, Witte DR, Jørgensen ME, Svensson AM. Duration of diabetes-related complications and mortality in type 1 diabetes: a national cohort study. *Int J Epidemiol*. 2021:dyaa290. https://doi.org/10.1093/ije/dyaa290
4. Joensen LE, Almdal TP, Willaing I. Associations between patient characteristics, social relations, diabetes management, quality of life, glycaemic control and emotional burden in type 1 diabetes. *Prim Care Diabetes*. 2016;10(1):41-50. https://doi.org/10.1016/j.pcd.2015.06.007

5. Katsarou A, Gudbjörnsdottir S, Rawshani A, Dabelea D, Bonifacio E, Anderson BJ, Jacobsen LM, Schatz DA, Lernmark Å. Type 1 diabetes mellitus. *Nat Rev Dis Primers*. 2017;3:17016. https://doi.org/10.1038/nrdp.2017.16
6. Hu X, Deutsch AJ, Lenz TL, Onengut-Gumuscu S, Han B, Chen WM, Howson JM, Todd JA, de Bakker PI, Rich SS, Raychaudhuri S. Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat Genet*. 2015;47(8):898-905. https://doi.org/10.1038/ng.3353
7. Lee HS, Hwang JS. Genetic aspects of type 1 diabetes. *Ann Pediatr Endocrinol Metab*. 2019;24(3):143-148. https://doi.org/10.6065/apem.2019.24.3.143
8. Elboudwarej E, Cole M, Briggs FB, Fouts A, Fain PR, Quach H, Quach D, Sinclair E, Criswell LA, Lane JA, Steck AK, Barcellos LF, Noble JA. Hypomethylation within gene promoter regions and type 1 diabetes in discordant monozygotic twins. *J Autoimmun*. 2016;68:23-9. https://doi.org/10.1016/j.jaut.2015.12.003
9. Stefan M, Zhang W, Concepcion E, Yi Z, Tomer Y. DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. *J Autoimmun*. 2014;50:33-7. https://doi.org/10.1016/j.jaut.2013.10.001
10. Porumb M, Stranges S, Pescapè A, Pecchia L. Precision Medicine and Artificial Intelligence: A Pilot Study on Deep Learning for Hypoglycemic Events Detection based on ECG. *Sci Rep*. 2020;10(1):170. https://doi.org/10.1038/s41598-019-56927-5
11. Pearson ER. Personalized medicine in diabetes: the role of 'omics' and biomarkers. *Diabet Med*. 2016;33(6):712-7. https://doi.org/10.1111/dme.13075
12. Perez-Riverol Y, Kuhn M, Vizcaíno JA, Hitz MP, Audain E. Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS One*. 2017;12(12):e0189875. https://doi.org/10.1371/journal.pone.0189875

13. Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform.* 2021;22(1):77-87. https://doi.org/10.1093/bib/bbaa122

14. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752

15. Tsagris M, Tsamardinos I. Feature selection with the R package MXM. *F1000Res*. 2018;7:1505. https://doi.org/10.12688/f1000research.16216.2

16. De Silva K, Jönsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *J Am Med Inform Assoc*. 2020;27(3):396-406. https://doi.org/10.1093/jamia/ocz204

17. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC. Gene expression in peripheral blood mononuclear cells from children with diabetes. *J Clin Endocrinol Metab*. 2007;92(9):3705-11. https://doi.org/10.1210/jc.2007-0979

18. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991-5. https://doi.org/10.1093/nar/gks1193

19. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-10. https://doi.org/10.1093/nar/30.1.207

20. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23(14):1846-7. https://doi.org/10.1093/bioinformatics/btm254

21. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;36(11):1-13.

22. Cheng T, Wang Y, Bryant SH. FSelector: a Ruby gem for feature selection. *Bioinformatics*. 2012;28(21):2851-2. https://doi.org/10.1093/bioinformatics/bts528

23. Friedman JH. Multivariate adaptive regression splines. The annals of statistics. 1991:1-67. https://www.jstor.org/stable/2241837

24. Schapire RE. Explaining AdaBoost. In: *Empirical Inference*. Springer, Berlin, Heidelberg. 2013. https://doi.org/10.1007/978-3-642-41136-6_5

25. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324

26. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. xgboost: eXtreme Gradient Boosting. R package version 1.4.1.1. 2021;1-4. https://mran.microsoft.com/web/packages/xgboost/vignettes/xgboost.pdf

27. Vinayak RK, Gilad-Bachrach R. DART: Dropouts meet Multiple Additive Regression Trees. In: 18th International Conference on Artificial Intelligence and Statistics (AISTAT). 2015; 38:489-497. http://proceedings.mlr.press/v38/korlakaivinayak15.pdf

28. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1-26. http://www.math.chalmers.se/Stat/Grundutb/GU/MSA220/S18/caret-JSS.pdf

29. Leite DMC, Brochet X, Resch G, Que YA, Neves A, Peña-Reyes C. Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics*. 2018;19(Suppl 14):420. https://doi.org/10.1186/s12859-018-2388-7

30. Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Front Genet*. 2019;10:256. https://doi.org/10.3389/fgene.2019.00256

31. Zhou G, Xia J. OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res*. 2018;46(W1):W514-W522. https://doi.org/10.1093/nar/gky510

32. Zhou G, Xia J. Using OmicsNet for Network Integration and 3D Visualization. *Curr Protoc Bioinformatics*. 2019;65(1):e69. https://doi.org/10.1002/cpbi.69

33. Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, Morris Q. GeneMANIA update 2018. *Nucleic Acids Res*. 2018;46(W1):W60-W64. https://doi.org/10.1093/nar/gky311

34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504. DOI: 10.1101/gr.1239303.

35. Scardoni G, Tosadori G, Faizan M, Spoto F, Fabbri F, Laudanna C. Biological network analysis with CentiScaPe: centralities and experimental dataset integration. *F1000Res*. 2014;3:139. https://doi.org/10.12688/f1000research.4477.2

36. Liu Y, Yi Y, Wu W, Wu K, Zhang W. Bioinformatics prediction and analysis of hub genes and pathways of three types of gynecological cancer. Oncol Lett. 2019;18(1):617-628. https://doi.org/10.3892/ol.2019.10371

37. Lu L, Ye X, Yao Q, Lu A, Zhao Z, Ding Y, Meng C, Yu W, Du Y, Cheng J. Egr2 enhances insulin resistance via JAK2/STAT3/SOCS-1 pathway in HepG2 cells treated with palmitate. *Gen Comp Endocrinol*. 2018;260:25-31. https://doi.org/10.1016/j.ygcen.2017.08.023

38. Reddy MA, Das S, Zhuo C, Jin W, Wang M, Lanting L, Natarajan R. Regulation of Vascular Smooth Muscle Cell Dysfunction Under Diabetic Conditions by miR-504. *Arterioscler Thromb Vasc Biol*. 2016;36(5):864-73. https://doi.org/10.1161/ATVBAHA.115.306770

39. Fan Y, Yi Z, D'Agati VD, Sun Z, Zhong F, Zhang W, Wen J, Zhou T, Li Z, He L, Zhang Q, Lee K, He JC, Wang N. Comparison of Kidney Transcriptomic Profiles of Early and Advanced Diabetic Nephropathy Reveals Potential New Mechanisms for Disease Progression. *Diabetes*. 2019;68(12):2301-2314. https://doi.org/10.2337/db19-0204

40. Okamura T, Yamamoto K, Fujio K. Early Growth Response Gene 2-Expressing CD4+LAG3+ Regulatory T Cells: The Therapeutic Potential for Treating Autoimmune Diseases. *Front Immunol*. 2018;9:340. https://doi.org/10.3389/fimmu.2018.00340

41. Bettini M, Bettini ML. Function, Failure, and the Future Potential of Tregs in Type 1 Diabetes. *Diabetes*. 2021:dbi180058.

42. Volfson-Sedletsky V, Jones A 4th, Hernandez-Escalante J, Dooms H. Emerging Therapeutic Strategies to Restore Regulatory T Cell Control of Islet Autoimmunity in Type 1 Diabetes. *Front Immunol*. 2021;12:635767. https://doi.org/10.3389/fimmu.2021.635767

43. Cabrera SM, Chen YG, Hagopian WA, Hessner MJ. Blood-based signatures in type 1 diabetes. *Diabetologia*. 2016;59(3):414-25. https://doi.org/10.1007/s00125-015-3843-x

44. Kaneko K, Lin HY, Fu Y, Saha PK, De la Puente-Gomez AB, Xu Y, Ohinata K, Chen P, Morozov A, Fukuda M. Rap1 in the VMH regulates glucose homeostasis. *JCI Insight*. 2021;6(11):142545. https://doi.org/10.1172/jci.insight.142545

45. He RJ, Yu ZH, Zhang RY, Zhang ZY. Protein tyrosine phosphatases as potential therapeutic targets. *Acta Pharmacol Sin*. 2014;35(10):1227-46. https://doi.org/10.1038/aps.2014.80

46. Gao L, Sun N, Xu Q, Jiang Z, Li C. Comparative analysis of mRNA expression profiles in Type 1 and Type 2 diabetes mellitus. *Epigenomics*. 2019;11(6):685-699. https://doi.org/10.2217/epi-2018-0055

47. Linsley PS, Greenbaum CJ, Speake C, Long SA, Dufort MJ. B lymphocyte alterations accompany abatacept resistance in new-onset type 1 diabetes. *JCI Insight*. 2019;4(4):e126136. https://doi.org/10.1172/jci.insight.126136

48. Martin RJ, Savage DA, Carson DJ, McKnight AJ, Maxwell AP, Patterson CC. Association analysis of proopiomelanocortin (POMC) haplotypes in type 1 diabetes in a UK population. *Diabetes Metab*. 2011;37(4):298-304. https://doi.org/10.1016/j.diabet.2010.11.021

49. Jiao X, Lu D, Pei X, Qi D, Huang S, Song Z, Gu J, Li Z. Type 1 diabetes mellitus impairs diurnal oscillations in murine extraorbital lacrimal glands. *Ocul Surf*. 2020;18(3):438-452. https://doi.org/10.1016/j.jtos.2020.04.013

50. Buchberger E, Bilen A, Ayaz S, Salamanca D, Matas de Las Heras C, Niksic A, Almudi I, Torres-Oliva M, Casares F, Posnien N. Variation in Pleiotropic Hub Gene Expression Is Associated with Interspecific Differences in Head Shape and Eye Size in Drosophila. *Mol Biol Evol*. 2021;38(5):1924-1942. https://doi.org/10.1093/molbev/msaa335

51. Liu H, Qu Y, Zhou H, Zheng Z, Zhao J, Zhang J. Bioinformatic analysis of potential hub genes in gastric adenocarcinoma. *Sci Prog*. 2021;104(1):368504211004260. https://doi.org/10.1177/00368504211004260

52. Atabaki-Pasdar N, Ohlsson M, Viñuela A, Frau F, Pomares-Millan H, Haid M et al. Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts. *PLoS Med*. 2020;17(6):e1003149. https://doi.org/10.1371/journal.pmed.1003149

53. Hosoda Y, Miyake M, Yamashiro K, Ooto S, Takahashi A, Oishi A, Miyata M, Uji A, Muraoka Y, Tsujikawa A. Deep phenotype unsupervised machine learning revealed the significance of pachychoroid features in etiology and visual prognosis of age-related macular degeneration. *Sci Rep*. 2020;10(1):18423. https://doi.org/10.1038/s41598-020-75451-5

54. Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. 2009. https://doi.org/10.1007/978-0-387-39940-9_565

55. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*. 2020;36(2):347-355. https://doi.org/10.1093/bioinformatics/btz565

56. Vidman L, Källberg D, Rydén P. Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PLoS One*. 2019;14(12):e0219102. https://doi.org/10.1371/journal.pone.0219102

57. Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform*. 2021;22(1):77-87. https://doi.org/10.1093/bib/bbaa122

## TABLES

**Table 1: Summary of the 16 algorithms applied for dimension reduction and the outputs**

| Algorithm | Description | Number of features selected |
|---|---|---|
| **Tailored** | | |
| *MXM*: MMPC | Max-Min Parents and Children (MMPC) in the *MXM* package is a constraint-based feature selection algorithm which assumes a Bayesian network for input variables. Performs robustly with high dimensional feature space | Top 100 features according to p-values. |

| | | |
|---|---|---|
| | such as omics data. Permutation option was incorporated (R = 999) and max_k was specified as: sample size/10 for optimizing performance. | |
| *mixOmics*: SPCA | Sparse Principal Components Analysis (SPCA) algorithm in *mixOmics* package, which is a sparse version of the classical PCA technique, was used. Optimal number of principal components (PC) as per the elbow method was 3. Number of variables for each PC was set at 50. | 50 each from the 3 PC amounting to 150 features. |
| *mixOmics*: SPLSDA | Sparse Partial Least Squares Discriminant Analysis (SPLSDA) is a supervised algorithm which applies $L_1$ penalty on the loading vectors of the input matrix. Amenable for high dimensional biomedical data mining which aims to identify biomarkers. The optimal *ncomp* argument was set to 3, as identified by evaluating the balanced classification error rate (BER) of PLSDA against the number of components. | 50 each from the 3 components amounting to 150 features. |
| **Classical** | | |
| Boruta | An all-relevant wrapper algorithm around Random Forests which selects features based on mean decrease accuracy, by default. | 48 features were selected (20 confirmed & 28 tentative) |
| Recursive feature elimination (RFE) | A wrapper algorithm which applies a backward selection process to extract the optimal subset of features by iteratively building models, rank ordering feature importance and removing the least important features. | 22 features included. |
| *FSelector*: Chi Squared | Importance of discretised features based on a chi squared test | 379 non-zero features included. |
| *FSelector*: Gain Ratio | An entropy-based filter algorithm which selects features guided by gain ratio (ratio of information gain to the intrinsic information). | 379 non-zero features included |
| *FSelector*: Information Gain | A filter algorithm which selects features based on the reduction of entropy. | 379 non-zero features included. |
| *FSelector*: Symmetrical Uncertainty | A filter algorithm which selects features based on a modified information gain criterion. | 379 non-zero features included. |
| *FSelector*: Linear Correlation | A filter algorithm which selects features according to Pearson's correlation coefficients. | Top 100 features included. |
| *FSelector*: Rank Correlation | A filter algorithm which selects features according to Spearman's correlation coefficients. | Top 100 features included. |
| *FSelector*: One-R | A filter algorithm which selects features by generating one rule per predictor. | Top 100 features included. |
| *FSelector*: Random Forests | A filter algorithm which selects features based on a random forest learner. | Top 100 features selected. |
| **Regularizations** | | |
| Lasso | Dimension reduction based on $L_1$ penalization (sum of absolute values of the coefficients) | Seven features with non-zero coefficients selected. |
| Ridge | Algorithm is based on $L_2$ penalization (sum of squared coefficients) | 100 features with the largest *absolute* coefficients included. |
| Elastic Net | Regularization which linearly combines $L_1$ and $L_2$ penalizations. | 377 features with non-zero coefficients included. |

**Table 2: Predictive performance assessment metrics of internally-validated machine learning models. All four machine learning models achieved perfect performance on internal validation with uniform metrics given below.**

| Metric | Formula/Derivation | Value |
|---|---|---|
| Sensitivity | TPR = TP / (TP + FN) | 1.0000 |
| Specificity | SPC = TN / (FP + TN) | 1.0000 |
| Precision | PPV = TP / (TP + FP) | 1.0000 |
| Negative Predictive Value | NPV = TN / (TN + FN) | 1.0000 |
| False Positive Rate | FPR = FP / (FP + TN) | 0.0000 |

| | | |
|---|---|---|
| False Discovery Rate | FDR = FP / (FP + TP) | 0.0000 |
| False Negative Rate | FNR = FN / (FN + TP) | 0.0000 |
| Accuracy | ACC = (TP + TN) / (P + N) | 1.0000 |
| F1 Score | F1 = 2TP / (2TP + FP + FN) | 1.0000 |
| Matthews Correlation Coefficient | TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)) | 1.0000 |
| AUROC | Discrimination metric estimated as the area under the curve of sensitivity against (1-specificity). | 1.0000 |

ACC = accuracy; AUROC = area under the receiver operating characteristic curve; FN = number of false negative observations; FDR = false discovery rate; FNR = false negative rate; FP = number of false positive observations; FPR = false positive rate; N = total number of negative observations (true negative and false negative); NPV = negative predictive value; P = total number of positive observations (true positive and false positive); PPV = positive predictive value; SPC = specificity; TN = number of true negative observations; TP = number of true positive observations; TPR = true positive rate.

**Table 3: Variable importance metrics of machine learning models**

| MARS | | ADABOOST | | RF | | XGB-DART | |
|---|---|---|---|---|---|---|---|
| **Gene** | **Importance** | **Gene** | **Importance** | **Gene** | **Importance** | **Gene** | **Importance** |
| EGR2 | 100.000 | EGR2 | 0.8940 | EGR2 | 0.6183 | EGR2 | 0.339923 |
| RAP1B | 77.974 | PTP4A2 | 0.8400 | GNG11 | 0.6083 | MICA/////MICB | 0.102461 |
| RNF4 | 67.503 | PNP | 0.8313 | ZNF473 | 0.4648 | ZNF473 | 0.060776 |
| CSDE1 | 56.900 | ZNF473 | 0.8076 | SUMO3 | 0.4578 | SUMO3 | 0.053481 |
| GAR1 | 46.791 | SLC35A3 | 0.8071 | LPIN2 | 0.4025 | LPIN2 | 0.031212 |
| MLEC | 40.009 | KBTBD4 | 0.8050 | CR2 | 0.2653 | PTP4A2 | 0.031015 |
| CTDSP1 | 32.086 | FRAT2 | 0.8025 | CRYL1 | 0.2569 | RAP1B | 0.025142 |
| LOX | 26.531 | UCHL3 | 0.8020 | UCHL3 | 0.2535 | EXOC7 | 0.016295 |
| POMC | 22.024 | IL1B | 0.7989 | PTP4A2 | 0.2489 | MRPS10 | 0.015143 |
| CNTLN | 4.588 | CHFR | 0.7989 | ACAD10 | 0.2458 | NR1D2 | 0.014148 |

**Table 4: Marker genes demarcated as hub nodes in the biological interaction network**

| **Gene** | **Betweenness** | **Closeness** | **Degree** |
|---|---|---|---|
| | **mean = 70.938775510204** | **mean = 0.00869472439680752** | **mean = 7.83673469387755** |
| MLEC | 95.9111111111112 | 0.00961538461538461 | 9 |
| NR1D2 | 102.273835804718 | 0.0104166666666666 | 15 |
| CSDE1 | 131.437541053717 | 0.0104166666666666 | 17 |
| UCHL3 | 217.048665713371 | 0.0112359550561797 | 23 |
| RNF4 | 342.601101186394 | 0.0105263157894736 | 18 |
| RAP1B | 88.3532635664988 | 0.010204081632653 | 11 |
| CDV3* | 133.400057948587 | 0.010752688172043 | 11 |
| GTF2H3* | 96.9161763073529 | 0.010204081632653 | 10 |
| IMP3* | 151.888667704844 | 0.0103092783505154 | 16 |
| CTDSP2* | 124.136457333516 | 0.00961538461538461 | 11 |
| SP3* | 323.10581085581 | 0.0103092783505154 | 13 |
| H1-0* | 120.752832135185 | 0.01 | 10 |
| CUL2* | 255.441218748571 | 0.010752688172043 | 16 |
| PNP | 132.481177156177 | 0.00925925925925925 | 9 |

*Predicted by GENEMANIA

## FIGURES

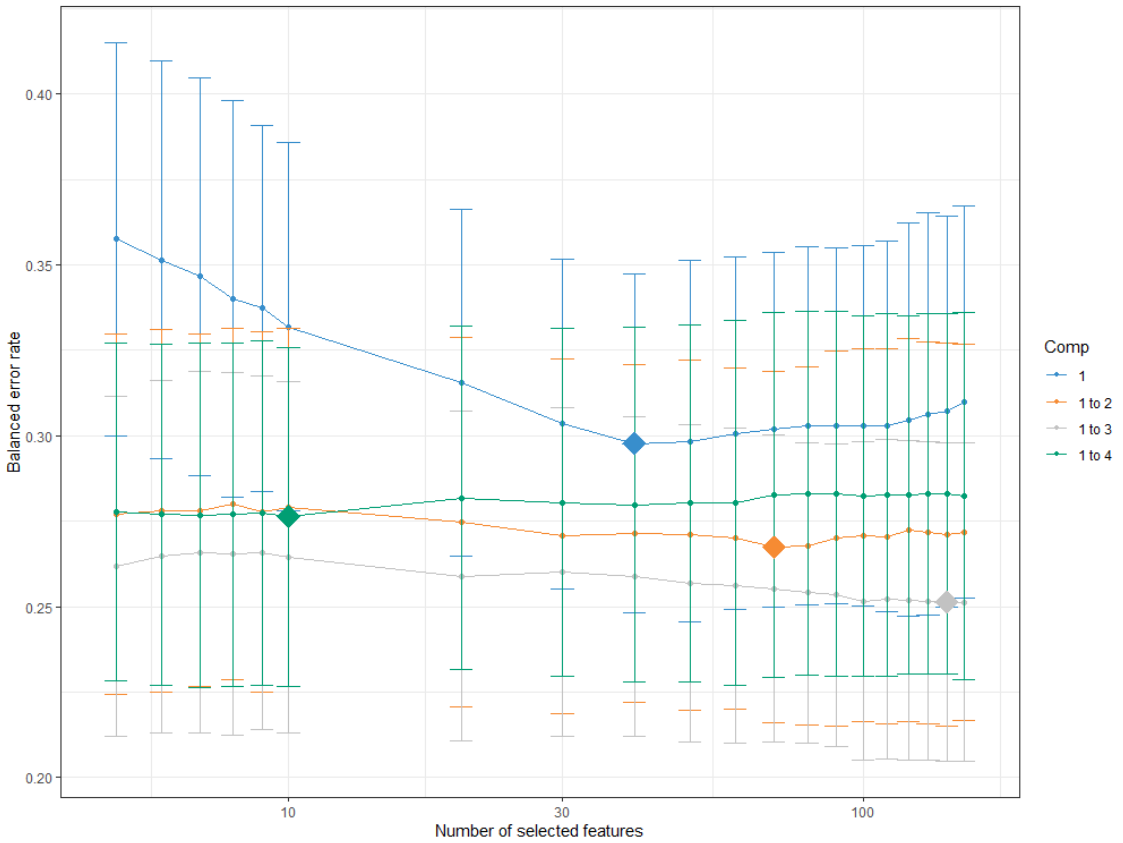


**Figure 1: Parameter tuning to determine the optimal number of features and components for SPLSDA. Plot of the number of selected features against balanced error rate identified 150 features per component and ncomp = 3 as optimal.**

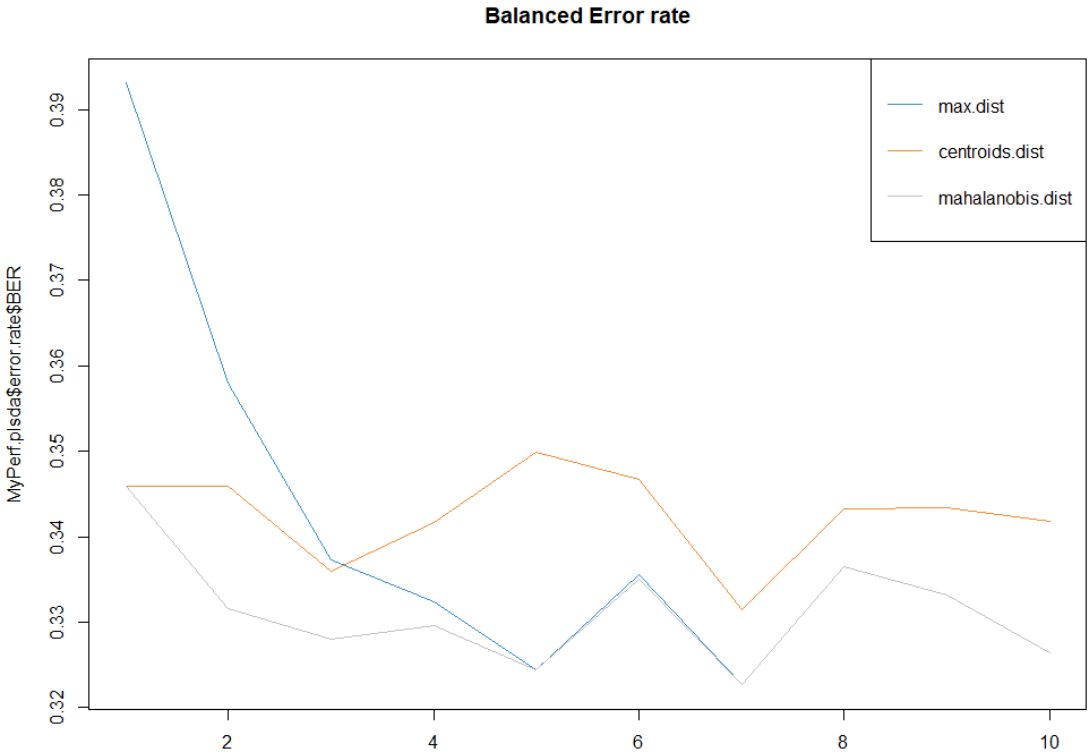


**Figure 2: Parameter tuning to determine the optimal *ncomp* for SPLSDA: Plot of balanced classification error rate (BER) versus the number of components as per three prediction distances. The ncomp = 3 was selected as a reasonable criterion minimizing BER.**

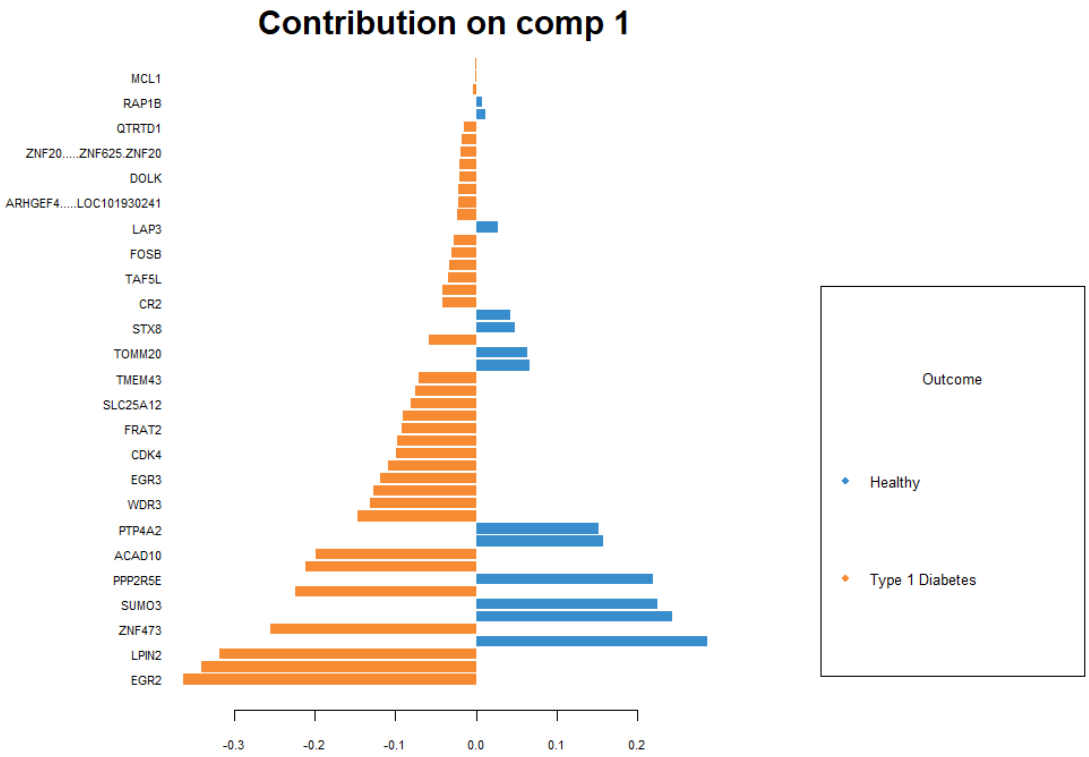


**Figure 3: Features of the first component of SPLSDA output distributed between the two classes of the target variable.**

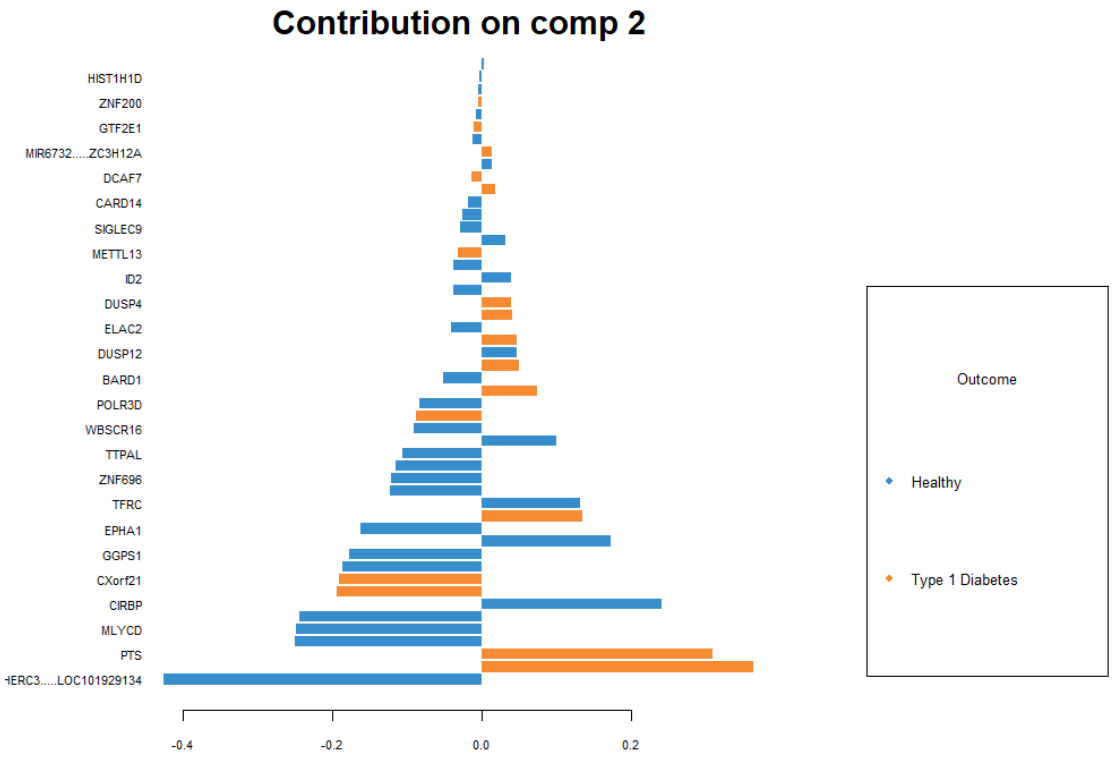


**Figure 4: Features of the second component of SPLSDA output distributed between the two classes of the target variable.**

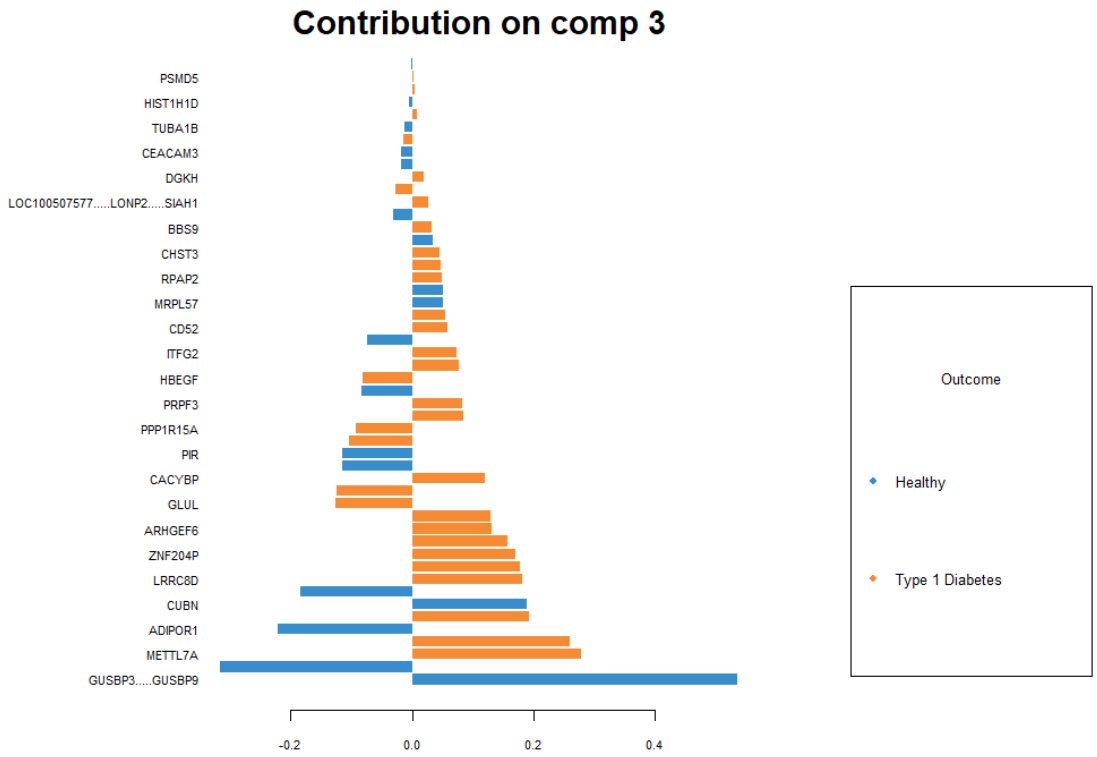


**Figure 5: Features of the third component of SPLSDA output distributed between the two classes of the target variable.**

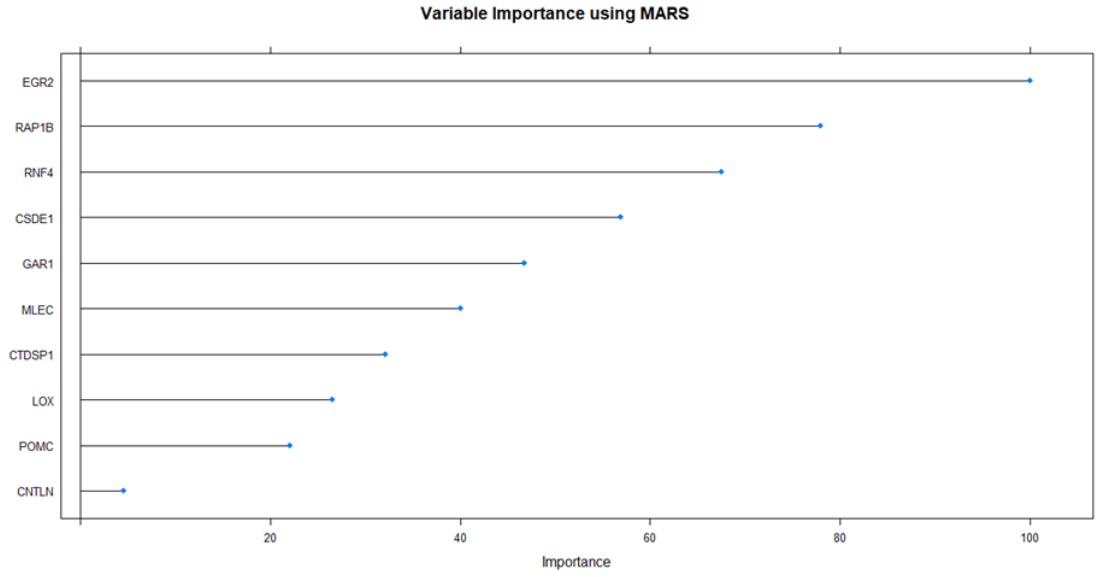


**Figure 6: Variable importance plot of MARS model visualizing the top 10 most important features.**

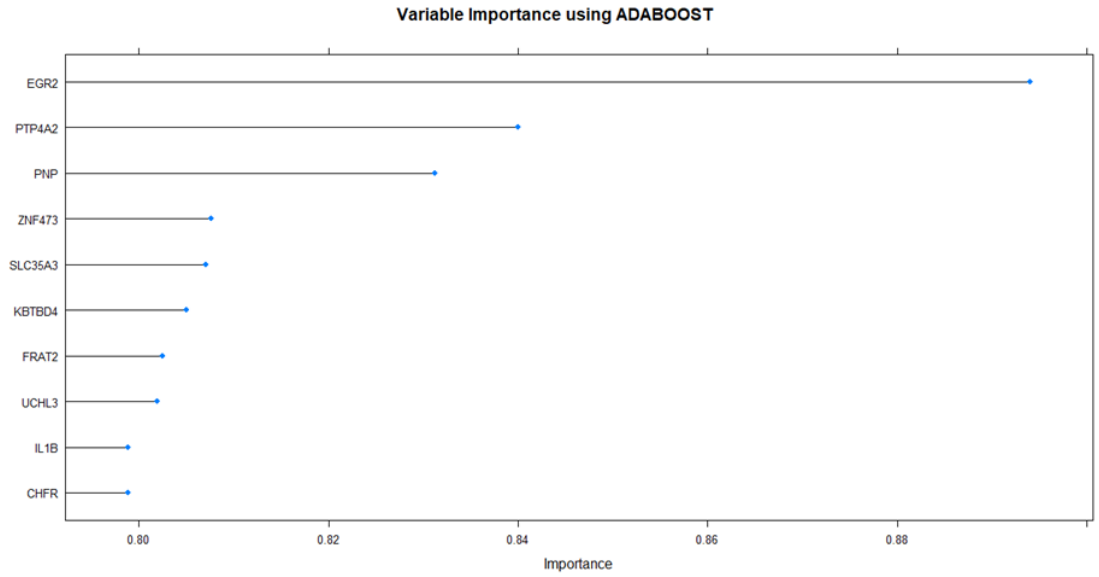


**Figure 7: Variable importance plot of ADABOOST model visualizing the top 10 most important features.**

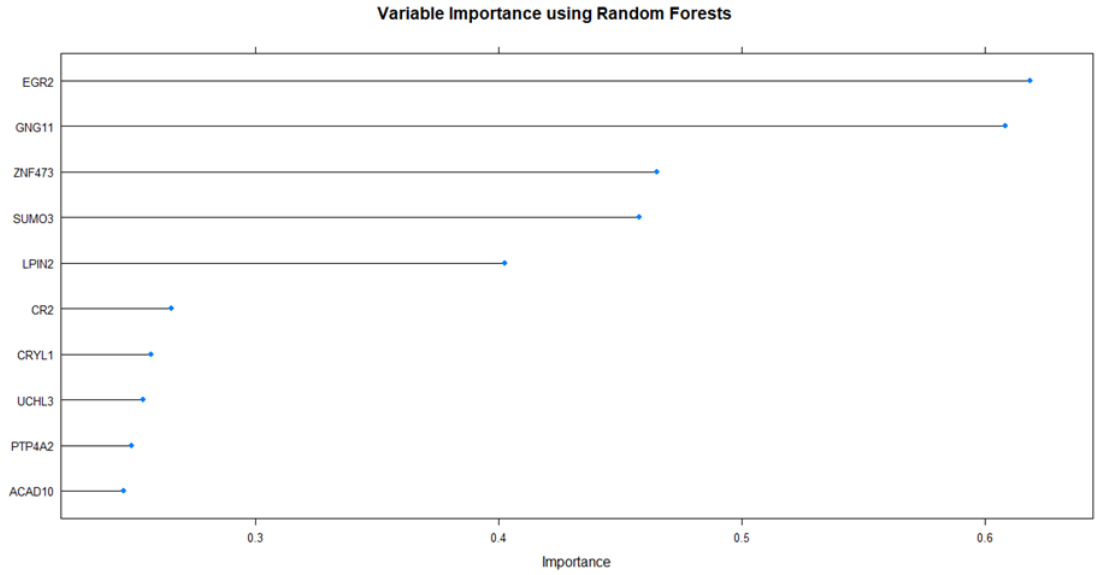


**Figure 8: Variable importance plot of RF model visualizing the top 10 most important features.**

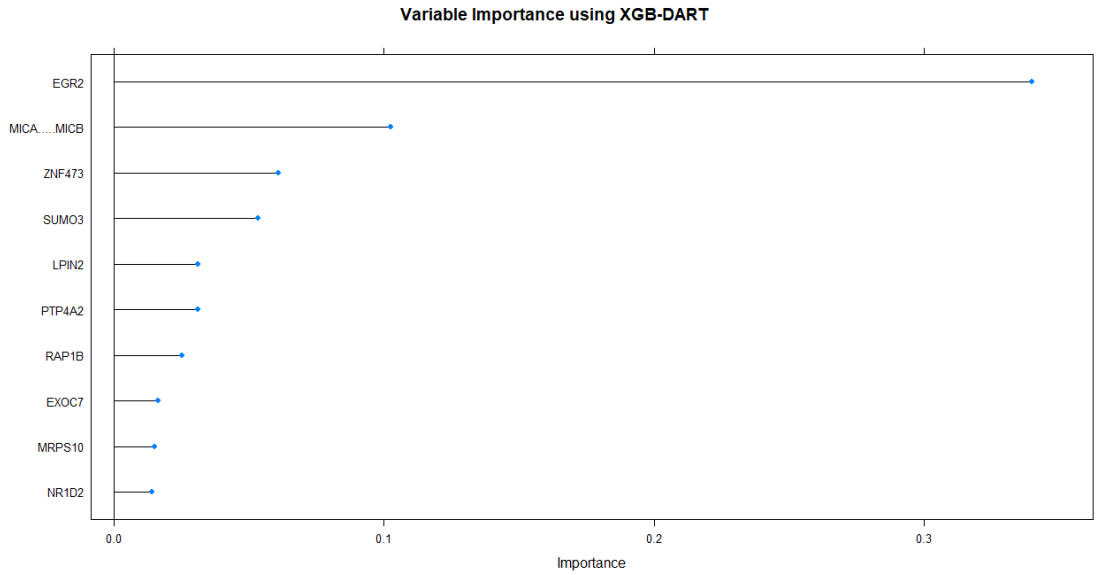


**Figure 9: Variable importance plot of XGB-DART model visualizing the top 10 most important features.**

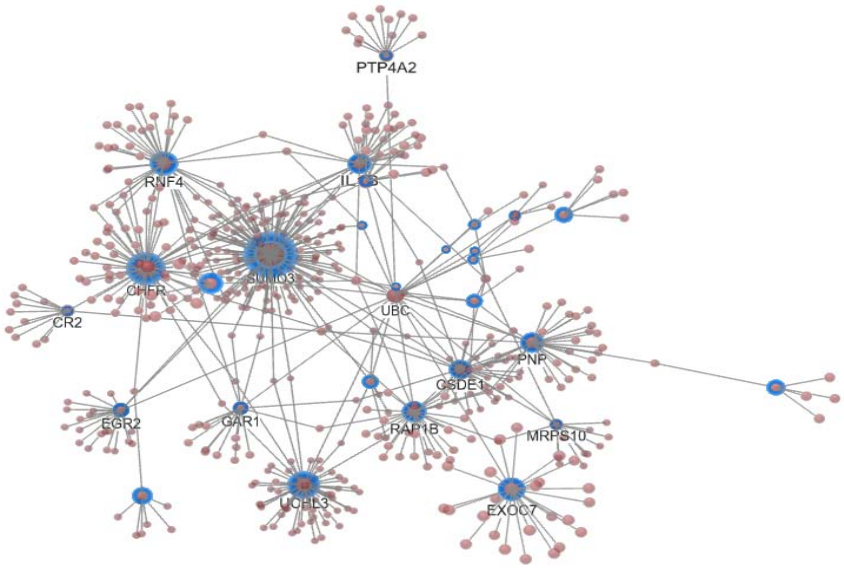


**Figure 10: Protein-protein interactions network generated by *OmicsNet*. Seed nodes are highlighted (n = 30)**

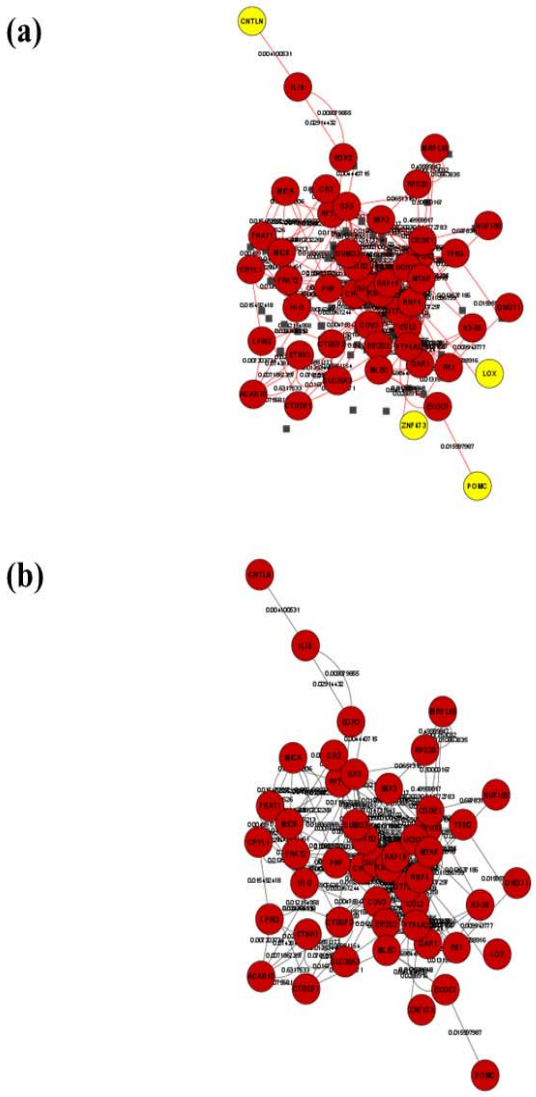


**Figure 11: Two different presentations of the protein-protein interactions network visualized in *Cytoscape*.**