Unraveling the Molecular Heterogeneity in type 2 Diabetes: A Potential Subtype Discovery Study Followed by Metabolic Modeling

Maryam Khoshnejat
University of Tehran Institute of Biochemistry and Biophysics
Kaveh Kavousi (kkavousi@ut.ac.ir)
Institute of Biochemistry and Biophysics(IBB), University of Tehran, Tehran, Iran.
https://orcid.org/0000-0002-1906-3912
Ali Mohammad Banaei Moghaddam
University of Tehran Institute of Biochemistry and Biophysics
Ali Akbar Moosavi-Movahedi
University of Tehran Institute of Biochemistry and Biophysics

Research article

Keywords: Type 2 Diabetes, Subtype, Classification, Clustering, Flux variability analysis, Muscle insulin resistance, Metabolic modeling

Posted Date: March 18th, 2020

DOI: https://doi.org/10.21203/rs.2.20464/v2

Version of Record: A version of this preprint was published on August 24th, 2020. See the published version at https://doi.org/10.1186/s12920-020-00767-0.

Abstract

Background: Type 2 diabetes mellitus (T2DM) is a complex multifactorial disease with a high prevalence in the world. Insulin resistance and impaired insulin secretion are the two major abnormalities in the pathogenesis of T2DM. Skeletal muscle is responsible for over 75% of the glucose uptake thus plays a critical role in T2DM. Here, we attempt to provide a better understanding of abnormalities in this tissue.

Methods: We have explored the muscle gene expression pattern in healthy and newly-diagnosed T2DM individuals using supervised and unsupervised classification along with the discovery of potential subtypes of type 2 diabetes.

Results: A machine-learning technique applied to identify a pattern of gene expression that could potentially discriminates between normoglycemic and diabetic groups. A gene set comprises of 26 genes identified which was able to discriminate healthy from diabetic individuals with the 94% accuracy after 10-fold stratified cross-validation. In addition, three distinct potential subtypes with different dysregulated genes and metabolic pathways identified in diabetic patients.

Conclusion: This study implies that it seems the disease has triggered through different cellular/molecular mechanisms. Therefore, subtyping of T2DM patients in combination with real clinical profiles of patients will provide a better understanding of abnormalities in each group and lead to the recommendation of the appropriate precision therapy for each subtype in the future.

Background

T2DM is a complex multifactorial disorder. The main cause of T2DM is impaired insulin secretion by pancreatic β-cells, usually due to having a background of reduced sensitivity to insulin in target tissues [1]. Skeletal muscle, liver, and adipose tissues are the key insulin-sensitive tissues in which skeletal muscle is responsible for over 75% of the glucose uptake thus takes the major role in lowering the blood glucose level [2, 3]. Therefore, there is an essential need to improve our understanding of the molecular mechanisms underlying insulin resistance in skeletal muscle to develop better prognostic signatures and to identify molecular drivers that can be therapeutically targeted. However, considerable experimental and computational attempts have been made to determine the molecular mechanisms involved in insulin resistance [4-8], but the exact underlying cause of it, is unclear [4] and failure of current therapies in some cases has occurred. One possible reason for this failure could be the multifactorial nature of T2DM. It is probable to find different groups of molecular mechanisms that all lead to insulin resistance and apply precision therapy for each group. This approach possibly improves the success rate of T2DM treatment.

In the present work, we attempt to provide a better understanding of diabetes and examine the hypothesis of the existence of potential subtypes in the disease. We studied expression profiles of human myocytes from healthy and newly diagnosed diabetic patients with two goals. 1) We aimed to identify a set of genes whose expression levels associated with type 2 diabetes using a machine-learning approach. 2) We used the gene expression profile from diabetic individuals with the aim of potential subtyping of the

disease. For this purpose, unsupervised classification was used to find different possible subtypes of T2DM which suggest different abnormality leading to the type 2 diabetes phenotype in order to develop effective treatments for this disease in the future. We used differential gene expression analysis and metabolic modeling to gain an in-depth insight into the molecular mechanisms underlying specific abnormalities in each cluster which potentially lead to insulin resistance. The overall study design is shown in Figure 1.

Methods

Data

Gene expression data of vastus lateralis skeletal muscle samples obtained from a sub-study of the Finland-United States Investigation of NIDDM Genetics project [9]. These subjects had glucose tolerance ranging from normal to newly diagnosed T2D in which 91 and 63 individuals were healthy and diabetics, respectively. Data are available through the repository's data access request procedures in the dbGaP database with the accession code phs001068.v1.p1. Data from healthy and diabetic individuals were downloaded and used for subsequent analysis.

Differential Gene Expression

Differential gene expression analysis was conducted using the Bioconductor R package DESeq2 [10]. A pre-filtering stage was performed which removes genes whose expression level was below a minimum cutoff level (< 5 read counts in less than 25 percentages of samples). Then, according to the DESeq2 manual, between samples normalization applied to account for differences in sequencing depth. Differentially expressed genes (DEGs) between two states (e.g. healthy vs. diabetics) were assessed based on a negative binomial distribution. Multiple testing correction by adjusting the p-values using the Benjamini–Hochberg procedure applied and genes with false discovery rate < 0.1 considered as differentially expressed. Moreover, KEGG pathway enrichment analysis of significant DEGs performed using Enrichr [11].

Feature selection method: GA–SVM

To select a near-optimal feature subset, a wrapper feature selection algorithm, which is a hybrid of genetic algorithm (GA) and Support Vector Machine (SVM), was used. GA is a global optimal search algorithm inspired by Darwin's theory of evolution. In the algorithm, the candidate solution (feature subset) encoded on a chromosome-like structure. A set of chromosomes constitute a population in which crossover and mutation can occur to generate new feature subsets. For each chromosome, a fitness value is calculated that represents how well a feature subset is adapted to the environment. The algorithm employs a competing solution in which better feature subsets have more chance to be selected for reproduction and creating the next generation. This search process will be repeated until a stopping criterion is satisfied.

In this analysis, a binary genetic algorithm was implemented in which each gene can have one of the binary values of 1 and 0 as either the presence or absence of a particular feature at the relevant chromosome, respectively. The chromosome length and the population size were set to the number of features and 500 chromosomes, respectively. The maximal number of generations was set to be 100. SVM classification accuracy calculated as the fitness score. The genetic algorithm terminated either when the fitness score was representing the accuracy of at least 95% or the maximum number of generations was reached.

Supervised Classification

Supervised machine learning methods such as SVM, k-Nearest Neighbour (KNN), Neural Network (NN), Naïve Bayes (NB) and a Random Forest (RF) classifier were considered to build classifiers using the Orange data mining toolbox [12]. The classifier was validated by 10-fold stratified cross-validation and analysis of Area Under the ROC curve (AUC), accuracy (ACC), F1 score, precision, and Recall was reported.

Unsupervised Classification

Potential subtype discovery performed to categorize the diabetic samples on the basis of similarity in their gene expression pattern. For this purpose, complete linkage hierarchical clustering based on Euclidean distance was applied using R. Normalized gene expression values using DESeq2 normalization method and pre-filtering of lowly expressed genes were applied here. This resulted in 21826 genes, which were used as the features for the clustering.

Cluster-based genome-scale metabolic modeling

To reconstruct personalized metabolic model we need a generic genome-scale metabolic model (GEM) and gene expression data. A generic human GEM is reconstructed from the all possible reactions that their relevant enzymes are encoded in the genome and can occur in different human cell types. By having gene-protein-reaction associations and mapping gene expression data to the generic metabolic model, active enzymes and subsequently active reactions are identified and a context-specific metabolic model will come out. These context-specific metabolic models can be employed for subsequent simulations to study metabolic reprogramming under specific condition. Possible minimum and maximum flux through a specific reaction can be simulated using Flux variability analysis (FVA). The readers are referred to [13] for full description of principle concept of this simulation.

Here, personalized metabolic models were reconstructed based on the Human Metabolic Reaction 2 (HMR 2) generic model [14, 15]. E-Flux method was applied for the reconstruction of the context-specific metabolic models using gene expression data [16]. Pre-processing of gene expression data including pre-filtering of low expressed genes, between-sample normalization (DESeq2 normalization method with gene length adjustment), and log2 transformation was applied. The myocyte biomass reaction was added to the model from the Bordbar model [6]. Body fluid metabolites used as media conditions here

[17]. The objective function was set to maximize flux through the production of mitochondrial ATP. In addition, to ensure the viability of the cell, the lower bound of biomass reaction was set to 0.8 of the maximum amount of biomass production in the healthy model [18]. Flux variability analysis (FVA) for each model was applied to get the minimum and maximum possible flux of each reaction using the Cobra Toolbox version 3.0 [19]. Personalized metabolic models (154 models) were categorized into the three groups based on the clustering that was obtained using the hierarchical method in the previous section. Then, to find perturbed reactions between each cluster and healthy individuals, a two-sample t-test was performed on the minimum and maximum fluxes obtained from FVA. Multiple testing correction applied by the Benjamini–Hochberg procedure and reactions with false discovery rate < 0.1 considered as perturbed reactions. Figure 2 shows the workflow for this section.

Results

Supervised classification

There are 57820 gene expression values for each individual that can be regarded as features in the classification. Using all of these genes as features are not applicable, leading to the high dimensional data and will reduce the performance of the conventional machine learning approaches. With the aim of reducing dimensionality and enhancing accuracy in the classification, it was decided to find differentially expressed genes between two states and apply them as the primary feature vector in each binary classification. Thus, DEGs between healthy and T2DM were explored which resulted in 247 differentially expressed genes. These 247 genes were used as a feature of classification and classifiers accuracy was investigated using this subset. SVM, KNN, NN, NB and RF classifiers tested and SVM showed the best performance in our analysis as are shown in table 1.

To achieve a near-optimal feature subset and improve classification accuracy, feature selection applied based on a combination of GA and SVM method. This method employs a GA as the feature selector and the SVM algorithm as the classifier. The genetic algorithm terminated either when the fitness score is represented at least 95% accuracy or the maximum number of generations is reached. Therefore, a subset of features with which SVM classifier can distinguish T2DM from normoglycemic subjects with approximately 90 percentages accuracy will be found. The GA-SVM procedure repeated 100 times and 100 feature subsets with the prediction accuracy around 95 percentages obtained. Then, features have been ranked according to the frequency with which each gene has participated in these 100 subsets. Our analysis revealed that 26 genes with at least 80% frequency can improve classification accuracy to 94 percent. These top-ranked genes were extracted and classification performance assessed with classifiers (Table 2).

In addition, to evaluate the SVM classifier with top-ranked genes, classification repeated 100 times with 10-fold cross-validation and accuracy, sensitivity and specificity were calculated. Figure S1 in Additional file 1 shows the box plot of this evaluation.

This top-ranked list comprises of important genes including CERK, FGFBP3, ETV5, E2F8, MAFB and 10 non-coding genes, which were explored for their functionality and importance in disease. The complete list of genes with Ensemble ID can be found in Additional File 2.

Unsupervised classification

T2DM is a multifactorial complex disease with different abnormalities thus it will be possible to find different groups of gene expression patterns which all lead to the insulin-resistant phenotype. Therefore, in this study, we asked the questions that 1) do the diabetic participants show different patterns of gene expression or not and 2) is it possible to categorize T2DM samples into distinct sub-groups with specific gene expression abnormalities? Thus, to answer the above-mentioned questions, unsupervised hierarchical clustering performed with the measure of Euclidean distance and complete linkage method. The top three clusters were isolated and studied (Figure 3). Cluster 1 to 3 consist of 18, 18 and 27 individuals, respectively.

To study biological differences between clusters, metabolic modeling of each cluster, as well as differentially expressed genes between each cluster and normal samples were explored. It was found that differences in gene expression and pathways between healthy and all newly diagnosed diabetic patients are low, while clustering of patients and analysis between each cluster and healthy individuals lead to the finding of more differentially expressed genes and more perturbed pathways. Results showed that each cluster has specific dysregulated genes and pathways which do not exist in the other two clusters. A heat map representation of the fold changes for dysregulated genes having absolute log2 fold change more than 0.9 in at least one cluster compared to healthy group is shown in Figure 4. In addition, pathway enrichment analysis of DEGs with absolute log2 fold change more than 0.9 in each cluster was performed. The results can be found in Table S1-3 of Additional File 1.

The analysis demonstrated that cluster 1 had the most number of perturbed pathways and dysregulated genes between the three clusters. Dysregulation of several genes in cluster 1 including down-regulation of DDIT4L, subunits of cytochrome c oxidase, several mitochondrial genes, ADIPOQ and up-regulation of several inflammatory genes such as GADD45G, TGFB1, CARD9, IGHA2, IGHG2, IGHA1, IGHD, and MIF genes were found. Down-regulation of several genes encoding mitochondrial genes and subunits of cytochrome c oxidase(COX) can reflect mitochondrial dysfunction and oxidative stress. Down-regulation of the adiponectin gene also was found in cluster 1. At the metabolic modeling level, perturbation in pathways related to Inositol phosphate metabolism, Pentose phosphate pathway, Tyrosine metabolism, Folate metabolism, Acylglycerides metabolism, Glutathione metabolism, ROS detoxification, Glycerolipid metabolism, Acyl-CoA hydrolysis, Fatty acid activation, Beta oxidation of fatty acids, Sphingolipid metabolism, Glycerophospholipid metabolism, Chondroitin/heparan sulfate metabolism, purine and pyrimidine metabolism, Carnitine shuttle, TCA, oxidative phosphorylation, Omega-3 and Omega-6 fatty acid metabolism, and Glycosphingolipid metabolism was observed.

Cluster 2 displayed no significant perturbed pathway in metabolic modeling. Although, a change in expression of various genes like as overexpression of SPP1, TNFRSF11B, FRK and down-regulation of

PRKAG3 and ATP2A1 was found here. We compared the phenotypic features of people in each cluster with healthy ones. Table 3 shows the average value of each feature in different clusters. In addition, the box plots of fasting glucose and fasting insulin values in each diabetic cluster and normoglycemic group are shown in Figures S2 and S3 of Additional file 1. This analysis revealed that this cluster is very close to a healthy state in terms of blood glucose and insulin levels.

Cluster 3 had the least number of DEGs against healthy individuals although perturbation in the expression of various important genes like as MSTN, ErbB3, EGR1, CIDEC, and HK2 was found in this cluster. At the metabolic level, the perturbation in glucose metabolism was found. Dysregulation of Branched-chain amino acids (BCAAs) metabolism, Glycolysis, pyruvate metabolism, Tricarboxylic acid cycle and glyoxylate/dicarboxylate metabolism and several exchange and transport reactions were observed. The complete list of differentially expressed genes and perturbed reactions in each cluster can be found in Additional File 2.

Discussion

Supervised classification of genes expression discriminate diabetic patients from healthyones

In this study gene expression data from newly diagnosed type 2 diabetic patients analyzed using supervised and unsupervised machine learning approaches. At the supervised level, we aimed to identify a set of genes whose expression in skeletal muscles was dysregulated in most patients and could potentially discriminate normoglycemic from type 2 diabetic individuals.

The gene set comprises of genes such as FGFBP3, CERK, ETV5, E2F8, MAFB and non-coding RNAs, which may be used to study and development of treatment strategies in the future. Noticeably, the injection of FGFBP3 has been patented as a treatment for diabetes, obesity and nonalcoholic fatty liver disease [20, 21]. This invention demonstrated that the administration of FGFBP3 with the single injection of this protein can regulate blood glucose level and keep it at the healthy stage for more than 24 hours. CERK plays an important role in inflammation-associated diseases [22]. Using CERK-null mice it was observed that CERK deficiency suppressed the elevation of obesity-mediated inflammatory cytokines, improve high-fat-diet-induced decrease of the insulin receptor, GLUT4, and adiponectin and improves glucose intolerance [23]. Studies also indicate the relationship between diet and obesity and ETV5 gene expression which participates in food intake control mechanisms [24]. Moreover, it has been found that impaired glucose tolerance in obese individuals associates with the up-regulation of E2F8 and therefore this gene possibly implicated in the progression of obesity, glucose intolerance and its complications [25]. MAFB also links to the metabolism and development of obesity and diabetes. The MAFB-deficient mice exhibited higher body weights and faster rates of body weight increase than control mice [26]. Up-regulation of MAFB expression in human adipocytes is correlated with adverse metabolic features and inflammation which may lead to the development of insulin resistance [27]. In addition to the protein-encoding genes, we found that about 40 percent of top-ranked genes comprises of the non-coding RNAs including pseudogenes and Long Non-coding RNAs. Recent studies have revealed that the deregulation

of pseudogenes and LncRNAs can relate to diabetes [28, 29]. Here, in this analysis more non-coding candidates found that strengthening the role of lncRNAs in complex diseases like diabetes. These non-coding RNA can be functionally analyzed to understand their biological roles in the pathology of T2DM.

Unsupervised classification of gene expression profile of diabetic patients reveals potential existence of molecular subtypes

The objective of analysis at the unsupervised level was to discriminate possible different gene expression patterns, which all lead to the insulin resistance and T2DM. In this part, the diabetic samples were categorized into three clusters and specific dysregulated genes and pathways in each cluster were reported.

Cluster 1: mitochondrial dysfunction, oxidative stress and inflammation

In cluster 1, perturbed pathways and dysregulated genes possibly represent perturbation of lipid and free fatty acids (FFAs) metabolism, inflammation, oxidative stress, and mitochondrial dysfunction. Perturbed pentose phosphate, folate metabolism, and glutathione metabolism as well as dysregulated genes such as IGHA1 and IGHA2, GADD45G and DDIT4 exhibit inflammation and oxidative stress. The up-regulation of IGHA1 and IGHA2 may start an inflammatory cascade involving a neutrophilic response, phagocytosis, the oxidative burst, and subsequent tissue damage. In addition, GADD45G play roles as stress sensors [30] which are overexpressed in this group. DNA damage and energy stress also can activate DDIT4 expression thus this gene implicate in the regulation of reactive oxygen species [31]. Oxidative stress may impair mitochondrial function which possibly leads to impairment of insulin sensitivity. Some evidence supports the role of oxidative stress and mitochondrial dysfunction in the pathogenesis of insulin resistance and type 2 diabetes [32]. In diabetes mellitus, mitochondria are the major source of oxidative stress [32]. Free radicals can damage lipids, proteins, and DNA and play a role in diabetes complications. Down-regulated mitochondrial genes and lowered flux in oxidative phosphorylation may demonstrate mitochondrial dysfunction in this cluster. Furthermore, MIF that is a proinflammatory cytokine is up-regulated in this cluster. A positive association between MIF Plasma levels with FFAs concentration and insulin resistance is shown. The perturbation of FFAs metabolism which leads to the increase in FFAs was observed in this cluster. Evidence showed that FFAs can induce insulin resistance in skeletal muscle. FFAs may induce insulin resistance via mitochondrial dysfunction, increased ROS production and oxidative stress and activation of inflammatory signals which was observed in this cluster [33]. Increase in FFAs associated with a decrease in adiponectin. ADIPOQ is mainly known as the adipokine but the importance of adiponectin production in muscle cells was also demonstrated [34]. This study also reported an increased level of adiponectin expression in response to rosiglitazone treatment in muscle cells and confirmed the functional role of muscle adiponectin in insulin sensitivity. Adiponectin contributes to the glucose metabolism of muscle cells via increased insulin-induced serine phosphorylation of protein kinase B and inhibition of the inflammatory response [35]. Moreover, in this cluster abnormalities in inositol phosphate metabolism with Myo-inositol deficiency were observed. Myo-inositol, one of the inositol isomers, participates in signal transduction and vesicle trafficking and

associated with glucose utilization. Clinical reports suggest that the administration of inositol supplements is a therapeutic approach in insulin resistance and improves glucose metabolism [36]. Figure S4 in Additional file 1 shows the overview of abnormalities in this cluster.

Cluster 2: ER-stress and inflammation

Surprisingly, no significant dysregulated pathway found in the second cluster. Therefore, we compared the phenotypic features of people in each cluster with healthy individuals. It was interesting that this cluster is very close to a healthy state in terms of blood glucose and insulin levels. Therefore, people of this group possibly are at the early stage of diabetes onset, and there is still no obvious change in their metabolism. However, using differential gene expression analysis, the change in expression of non-metabolic genes like as overexpression of OPN, OPG, CHAC1, ERN1, and down-regulation of SERCA1 were observed in this cluster. These genes are related to diabetes by promoting ER-stress and inflammation. OPN and OPG play roles in inflammation, insulin resistance, prediabetes, and diabetes. A recent study demonstrated that OPN and OPG level in pre-diabetic subjects are higher and alterations in OPN and OPG might be involved in the pathogenesis of prediabetes and T2DM [37, 38]. Obese mice lacking osteopontin improved whole-body glucose tolerance and insulin resistance also with decreased markers of inflammation [39]. In addition, ER-stress can induce the expression of OPN and OPG. Recent pieces of evidence support the presence and role of ER stress in muscle [40-42]. In this cluster, SERCA1 which is an intracellular membrane-bound Ca2+-transport ATPase enzyme encoded by the ATP2A1 gene was down-regulated. The Dysregulation of SERCA promotes ER Stress [37]. SERCA1 resides in the sarcoplasmic or endoplasmic reticula of muscle cells and contributes to the modulation of cellular Ca2+ homeostasis within the physiological range. Lower SERCA expression may lead to reduced Ca2+ accumulation in the ER lumen and ER dysfunction. High luminal calcium concentration is essential for proper protein folding and processing and Ca2+ depletion can result in the accumulation of unfolded proteins and trigger the unfolded protein response (UPR) and cell death [43]. High-fat diet and obesity induce ER stress in muscles and subsequently suppresses insulin signaling [44]. Antidiabetic compounds such as Azoramide and Rosiglitazone, demonstrated to induce SERCA expression and increased ER Ca2+ accumulation [45, 46]. Schematic representation of abnormalities in cluster 2 is shown Figure S5 of Additional file 1.

Cluster 3: perturbation in IRS-mediated insulin signaling

In cluster 3, the differential gene expression analysis revealed the perturbation in insulin signaling and inflammation. Results showed down-regulation of insulin-responsive genes, HK2, EGR1, and CIDEC which verify insulin resistance through deficiency of insulin signaling. Furthermore, overexpression of MSTN and ErbB3 was found. Myostatin induces insulin resistance by degrading IRS1 protein [47] and diminishing insulin-induced IRS1 tyrosine phosphorylation thus interrupting insulin signaling cascade [48]. In addition, treating Hella cells with myostatin suppressed hexokinase 2 expression [49]. Evidence revealed that stress-induced transactivation of ErbB2/ErbB3 receptors triggers a PI3K cascade leading to the serine phosphorylation of IRS proteins [50, 51]. Overexpression of ErbB3 may enhance PI3K activity and implicating ErbB proteins in stress-induced insulin resistance. Taken together, MSTN and ErbB3 can

lead to the Serine phosphorylation of IRS, reducing tyrosine phosphorylation of IRS and degradation of them. Since expressions of insulin-regulated genes are positively correlated with insulin sensitivity, Down-regulation of HK2, EGR1 and CIDEC genes in this group possibly verify insulin resistance through deficiency of insulin signaling. In addition, at the metabolic analysis, lower phosphorylation of glucose with the subsequent lowering in glycolysis and TCA fluxes was observed. Moreover, Dysmetabolism of branched-chain amino acids was observed at metabolic analysis. A proposed mechanism linking higher levels of BCAAs and T2DM involves leucine-mediated activation of the mammalian target of rapamycin complex 1 (mTORC1). This activation results in the serine phosphorylation of IRS1 and IRS2 and subsequent uncoupling of insulin signaling at an early stage [52]. A brief representation of abnormalities in this cluster is shown Figure S6 of Additional file 1.

Cluster-based study can improve understanding of T2DM

Taken together, our analysis showed that at the early stage of diabetes, associated changes at the gene expression level in skeletal muscle are low compared to healthy subjects. Moreover, the clustering of patients leads to the identification of abnormalities that are usually hidden in cohort studies. For example, dysregulation of genes such as MIF, ATP2A1, GADD45G, EEF2, EGR1, CIDEC, MSTN and several other genes and pathways like as BCAAs metabolism, folate metabolism, and pentose phosphate observed in our cluster analysis whereas analysis in cohort study between normoglycemic and all diabetic individuals did not determine them as differentially expressed genes or dysregulated pathways. In a cohort study, a sample consists of several subjects (63 diabetic individuals in the original study) is gathered and examined (Figure S7 of Additional file 1). This makes it possible to see only an approximate average of the features in the samples and as a result, some of the abnormalities are covered in this way. In a cluster-based study, a sample that has been collected in a cohort study is broken down into the sub-groups so that the members within each subgroup have the most similarity and differ from the members of the outer sub-groups. Then each sub-group will be analyzed individually (e.g. here we divided the diabetic group to three sub-groups). The cluster-based analysis in this study led to find more dysregulated genes and pathways that are specific in each cluster. Therefore, for a progressive and heterogenic disease like T2DM, applying a cluster-based study will enhance our understanding of the factors involved in the disease. Focusing on homogeneous sub-groups in a heterogenic disease such as T2DM may improve the success of therapeutic strategies.

Conclusions

Here in this study, the change in the gene expression pattern of newly-diagnosed diabetic patients analyzed using supervised and unsupervised classification. Using gene expression alone, it was possible to discriminate T2DM from healthy with approximately 90 percent of accuracy. Clustering of diabetic patients according to their gene expression pattern and subsequent deeper analysis of each cluster unraveled specific abnormalities leading to the insulin resistance in each cluster. We proposed that using unsupervised classification of gene expression in diabetic patients in combination with real clinical

profiles of patients will be helpful to find significant molecular subtypes of T2DM with specific abnormalities and subsequently discover the best target for the treatment of each subtype.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available through the repository's data access request procedures in the dbGaP database with the accession code phs001068.v1.p1.

The codes used in this article are available at https://github.com/Maryamkhn/T2DM_potential_subtyping.

Competing interests

The authors declare no conflicts of interest.

Funding

Not applicable.

Author's Contributions

MK: Conceptualization, investigation, data collection, design, implementation, analysis, interpretation, writing & editing. KK: Supervision, design, interpretation of data, review & editing. AMB-M: Supervision, design, interpretation of data, review & editing. AM: Supervision, review.

Acknowledgments

Not applicable.

Additional material

Additional file 1 (Additional_file1.docx): Additional Figures S1-7 showing bar plot of SVM classification evaluation, boxplots related to individuals characteristics in each cluster and schematic representation of abnormalities in each cluster. Additional Table S1-3 related to KEGG pathway enrichment analysis of each cluster

Additional file 2(Additional_file2.xlsx): The complete list of differentially expressed genes in each cluster, top-ranked genes with Ensemble ID, perturbed reactions results of metabolic modeling in each cluster

Abbreviations

Type 2 diabetes mellitus	T2DM
Support vector machine	SVM
Genetic algorithm	GA
K-nearest neighbour	KNN
Neural network	NN
Naïve bayes	NB
Random forest	RF
Area under the ROC curve	AUC
Accuracy	ACC
Genome-scale metabolic models	GEMs
Flux variability analysis	FVA
Differentially expressed genes	DEGs
Free fatty acids	FFAs
Unfolded protein response	UPR

References

1. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, Hu FB, Kahn CR, Raz I, Shulman GI: Type 2 diabetes mellitus. Nature reviews Disease primers 2015, 1:15019-15019.
2. Björnholm M, Zierath J: Insulin signal transduction in human skeletal muscle: identifying the defects in Type II diabetes. In.: Portland Press Limited; 2005.
3. DeFronzo RA: From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. Diabetes 2009, 58(4):773-795.
4. Yaribeygi H, Farrokhi FR, Butler AE, Sahebkar A: Insulin resistance: Review of the underlying molecular mechanisms. Journal of cellular physiology 2019, 234(6):8152-8161.
5. Nogiec C, Burkart A, Dreyfuss JM, Lerin C, Kasif S, Patti M-E: Metabolic modeling of muscle metabolism identifies key reactions linked to insulin resistance phenotypes. Molecular metabolism 2015, 4(3):151-163.
6. Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BO, Famili I: A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC systems biology 2011,

5(1):180.

7. Väremo L, Scheele C, Broholm C, Mardinoglu A, Kampf C, Asplund A, Nookaew I, Uhlén M, Pedersen BK, Nielsen J: Proteome-and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. Cell reports 2015, 11(6):921-933.
8. Väremo L, Nookaew I, Nielsen J: Novel insights into obesity and diabetes through genome-scale metabolic modeling. Frontiers in physiology 2013, 4:92.
9. Scott LJ, Erdos MR, Huyghe JR, Welch RP, Beck AT, Wolford BN, Chines PS, Didion JP, Narisu N, Stringham HM: The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nature communications 2016, 7.
10. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 2014, 15(12):550.
11. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research 2016, 44(W1):W90-W97.
12. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A: Orange: data mining toolbox in Python. The Journal of Machine Learning Research 2013, 14(1):2349-2353.
13. Orth JD, Thiele I, Palsson BØ: What is flux balance analysis? Nature biotechnology 2010, 28(3):245-248.
14. Mardinoglu A, Agren R, Kampf C, Asplund A, Nookaew I, Jacobson P, Walley AJ, Froguel P, Carlsson LM, Uhlen M: Integration of clinical data with a genome-scale metabolic model of the human adipocyte. Molecular systems biology 2013, 9(1).
15. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J: Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications 2014, 5:3083.
16. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng T-Y, Moody DB, Murray M, Galagan JE: Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS computational biology 2009, 5(8).
17. Hadi M, Marashi S-A: Reconstruction of a generic metabolic network model of cancer cells. Molecular BioSystems 2014, 10(11):3014-3021.
18. Chénard T, Guénard F, Vohl M-C, Carpentier A, Tchernof A, Najmanovich RJ: Remodeling adipose tissue through in silico modulation of fat storage for the prevention of type 2 diabetes. BMC systems biology 2017, 11(1):60.
19. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdottir HS, Keating SM, Vlasov V, Wachowiak J: Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3. 0. arXiv preprint arXiv:171004038 2017.
20. Garman KA: Fibroblast Growth Factor Binding Protein 3: A Novel Target for Glucose Intolerance and Nonalcoholic Fatty Liver Disease Treatment. Georgetown University; 2018.

21. Wellstein A: Compositions and Treatments of Metabolic Disorders Using FGF Binding Protein 3. In.: US Patent App. 15/784,730; 2018.
22. Yu WL, Sun Y: CERK inhibition might be a good potential therapeutic target for diseases. British journal of pharmacology 2015, 172(8):2165-2165.
23. Mitsutake S, Date T, Yokota H, Sugiura M, Kohama T, Igarashi Y: Ceramide kinase deficiency improves diet-induced obesity and insulin resistance. FEBS letters 2012, 586(9):1300-1305.
24. Boender AJ, Van Rozen AJ, Adan RA: Nutritional state affects the expression of the obesity-associated genes Etv5, Faim2, Fto, and Negr1. Obesity 2012, 20(12):2420-2425.
25. Minchenko O, Bashta Y, Minchenko D, Ratushna O: Glucose tolerance in obese men is associated with dysregulation of some angiogenesis-related gene expressions in subcutaneous adipose tissue. Fiziolohichnyi zhurnal (Kiev, Ukraine: 1994) 2016, 62(2):12-23.
26. Tran MTN, Hamada M, Nakamura M, Jeon H, Kamei R, Tsunakawa Y, Kulathunga K, Lin YY, Fujisawa K, Kudo T: MafB deficiency accelerates the development of obesity in mice. FEBS open bio 2016, 6(6):540-547.
27. Pettersson AM, Acosta JR, Björk C, Krätzel J, Stenson B, Blomqvist L, Viguerie N, Langin D, Arner P, Laurencikiene J: MAFB as a novel regulator of human adipose tissue inflammation. Diabetologia 2015, 58(9):2115-2123.
28. Zhang N, Geng T, Wang Z, Zhang R, Cao T, Camporez JP, Cai S-Y, Liu Y, Dandolo L, Shulman GI: Elevated hepatic expression of H19 long noncoding RNA contributes to diabetic hyperglycemia. JCI insight 2018, 3(10).
29. Chiefari E, Iiritano S, Paonessa F, Le Pera I, Arcidiacono B, Filocamo M, Foti D, Liebhaber SA, Brunetti A: Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. Nature communications 2010, 1:40.
30. Liebermann DA, Hoffman B: Gadd45 in the response of hematopoietic cells to genotoxic stress. Blood Cells, Molecules, and Diseases 2007, 39(3):329-335.
31. Ellisen LW, Ramsayer KD, Johannessen CM, Yang A, Beppu H, Minda K, Oliner JD, McKeon F, Haber DA: REDD1, a developmentally regulated transcriptional target of p63 and p53, links p63 to regulation of reactive oxygen species. Molecular cell 2002, 10(5):995-1005.
32. Asmat U, Abad K, Ismail K: Diabetes mellitus and oxidative stress—a concise review. Saudi Pharmaceutical Journal 2016, 24(5):547-553.
33. Rachek LI: Free fatty acids and skeletal muscle insulin resistance. In: Progress in molecular biology and translational science. vol. 121: Elsevier; 2014: 267-292.
34. Liu Y, Chewchuk S, Lavigne C, Brûlé S, Pilon G, Houde V, Xu A, Marette A, Sweeney G: Functional significance of skeletal muscle adiponectin production, changes in animal models of obesity and diabetes, and regulation by rosiglitazone treatment. American Journal of Physiology-Endocrinology and Metabolism 2009, 297(3):E657-E664.
35. Wang CH, Wang CC, Huang HC, Wei YH: Mitochondrial dysfunction leads to impairment of insulin sensitivity and adiponectin secretion in adipocytes. The FEBS journal 2013, 280(4):1039-1050.

36. Bevilacqua A, Bizzarri M: Inositols in insulin signaling and glucose metabolism. International journal of endocrinology 2018, 2018.
37. Daniele G, Winnier D, Mari A, Bruder J, Fourcaudot M, Pengou Z, Hansis-Diarte A, Jenkinson C, Tripathy D, Folli F: The potential role of the osteopontin–osteocalcin–osteoprotegerin triad in the pathogenesis of prediabetes in humans. Acta diabetologica 2018, 55(2):139-148.
38. Kahles F, Findeisen HM, Bruemmer D: Osteopontin: A novel regulator at the cross roads of inflammation, obesity and diabetes. Molecular metabolism 2014, 3(4):384-393.
39. Chapman J, Miles PD, Ofrecio JM, Neels JG, Joseph GY, Resnik JL, Wilkes J, Talukdar S, Thapar D, Johnson K: Osteopontin is required for the early onset of high fat diet-induced insulin resistance in mice. PloS one 2010, 5(11):e13959.
40. Deldicque L, Cani PD, Philp A, Raymackers J-M, Meakin PJ, Ashford ML, Delzenne NM, Francaux M, Baar K: The unfolded protein response is activated in skeletal muscle by high-fat feeding: potential role in the downregulation of protein synthesis. American Journal of Physiology-Endocrinology and Metabolism 2010, 299(5):E695-E705.
41. Deldicque L, Hespel P, Francaux M: Endoplasmic reticulum stress in skeletal muscle: origin and metabolic consequences. Exercise and sport sciences reviews 2012, 40(1):43-49.
42. Rayavarapu S, Coley W, Van der Meulen JH, Cakir E, Tappeta K, Kinder TB, Dillingham BC, Brown KJ, Hathout Y, Nagaraju K: Activation of the ubiquitin proteasome pathway in a mouse model of inflammatory myopathy: a potential therapeutic target. Arthritis & Rheumatism 2013, 65(12):3248-3258.
43. Arruda AP, Hotamisligil GS: Calcium homeostasis and organelle function in the pathogenesis of obesity and diabetes. Cell metabolism 2015, 22(3):381-397.
44. Koh H-J, Toyoda T, Didesch MM, Lee M-Y, Sleeman MW, Kulkarni RN, Musi N, Hirshman MF, Goodyear LJ: Tribbles 3 mediates endoplasmic reticulum stress-induced insulin resistance in skeletal muscle. Nature communications 2013, 4:1871.
45. Shah R, Gonzales F, Golez E, Augustin D, Caudillo S, Abbott A, Morello J, McDonough P, Paolini P, Shubeita H: The antidiabetic agent rosiglitazone upregulates SERCA2 and enhances TNF-α-and LPS-induced NF-κB-dependent transcription and TNF-α-induced IL-6 secretion in ventricular myocytes. Cellular Physiology and Biochemistry 2005, 15(1-4):041-050.
46. Fu S, Yalcin A, Lee GY, Li P, Fan J, Arruda AP, Pers BM, Yilmaz M, Eguchi K, Hotamisligil GS: Phenotypic assays identify azoramide as a small-molecule modulator of the unfolded protein response with antidiabetic activity. Science translational medicine 2015, 7(292):292ra298-292ra298.
47. Bonala S, Lokireddy S, McFarlane C, Patnam S, Sharma M, Kambadur R: Myostatin induces insulin resistance via Casitas B-lineage lymphoma b (Cblb)-mediated degradation of insulin receptor substrate 1 (IRS1) protein in response to high calorie diet intake. Journal of Biological Chemistry 2014, 289(11):7654-7670.
48. Liu XH, Bauman WA, Cardozo CP: Myostatin inhibits glucose uptake via suppression of insulin-dependent and-independent signaling pathways in myoblasts. Physiological reports 2018,

6(17):e13837.

49. Liu Y, Cheng H, Zhou Y, Zhu Y, Bian R, Chen Y, Li C, Ma Q, Zheng Q, Zhang Y: Myostatin induces mitochondrial metabolic alteration and typical apoptosis in cancer cells. Cell death & disease 2013, 4(2):e494.
50. Hemi R, Paz K, Wertheim N, Karasik A, Zick Y, Kanety H: Transactivation of ErbB2 and ErbB3 by tumor necrosis factor-alpha and anisomycin leads to impaired insulin signaling through serine/threonine phosphorylation of IRS proteins. Journal of Biological Chemistry 2002.
51. Hemi R, Yochananov Y, Barhod E, Kasher-Meron M, Karasik A, Tirosh A, Kanety H: p38 mitogen-activated protein kinase-dependent transactivation of ErbB receptor family: a novel common mechanism for stress-induced IRS-1 serine phosphorylation and insulin resistance. Diabetes 2011:DB_091323.
52. Lynch CJ, Adams SH: Branched-chain amino acids in metabolic signalling and insulin resistance. Nature Reviews Endocrinology 2014, 10(12):723.

Tables

Table 1. Discriminative performance of different classifiers between healthy and diabetic patients when 247 differentially expressed genes used as the features

Method	AUC	ACC	F1	Precision	Recall
SVM	0.889	0.838	0.806	0.788	0.825
NN	0.877	0.812	0.772	0.766	0.778
RF	0.837	0.766	0.71	0.721	0.698
NB	0.801	0.734	0.717	0.634	0.825
KNN	0.758	0.727	0.58	0.784	0.46

AUC, ACC, F1 score, precision, and Recall was reported.

Table 2. Performance of different classifiers when the 26 top-ranked genes used as feature subset

Method	AUC	ACC	F1	Precision	Recall
SVM	0.958	0.942	0.927	0.950	0.905
NN	0.966	0.903	0.878	0.900	0.857
NB	0.896	0.818	0.791	0.746	0.841
RF	0.836	0.799	0.735	0.796	0.683
KNN	0.829	0.721	0.538	0.833	0.397

Table 3. Subject characteristics including average fasting plasma glucose, fasting serum insulin, BMI and waist/hip ratio (WHR) in each diabetic cluster and healthy group. P values calculated by ANOVA F-test.

	Healthy	Cluster 1	Cluster 2	Cluster 3	P value
Glucose (mmol/L)	5.62 ± 0.3	7.17 ± 0.5	6.86 ± 0.5	7.39 ± 0.75	6.14e-43
Insulin (mu/l)	6.87 ± 3.3	10.19 ± 5.3	7.79 ± 3.9	12.93 ± 8.7	1.08e-06
BMI	26.35 ± 3.5	29.03 ± 5.0	28.58 ± 4.5	30.13 ± 5.5	1.96e-04
WHR	0.92 ± 0.08	0.99 ± 0.07	0.95 ± 0.06	1.02 ± 0.06	3.64e-08

Figures

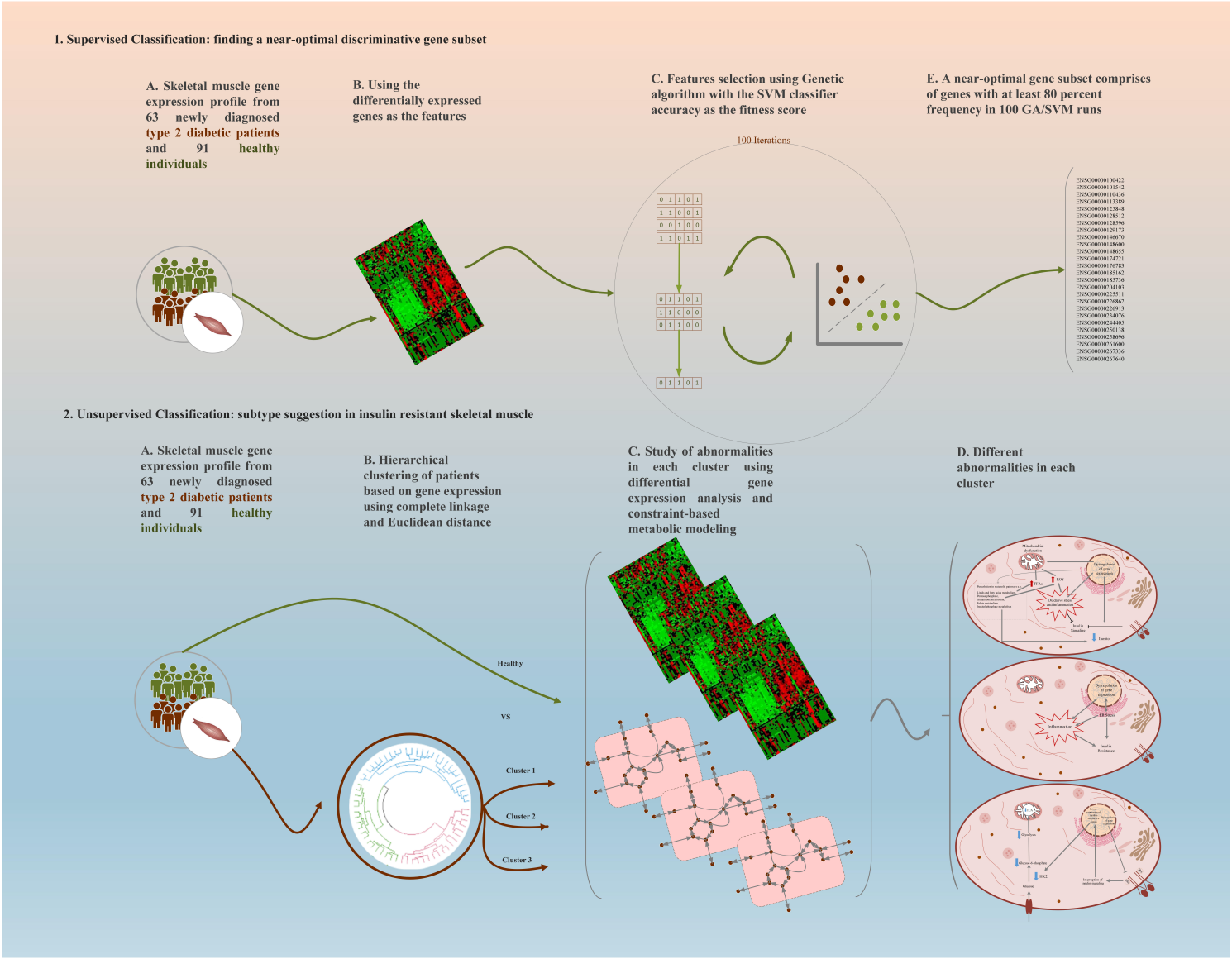

Figure 1

Graphical overview of study design. This study includes two supervised and unsupervised classification section. At the supervised classification part, we used a machine learning approach to identify a set of genes whose expression levels associated with type 2 diabetes. At the unsupervised section, clustering of T2DM patients with the aim of potential subtyping of the disease was employed.

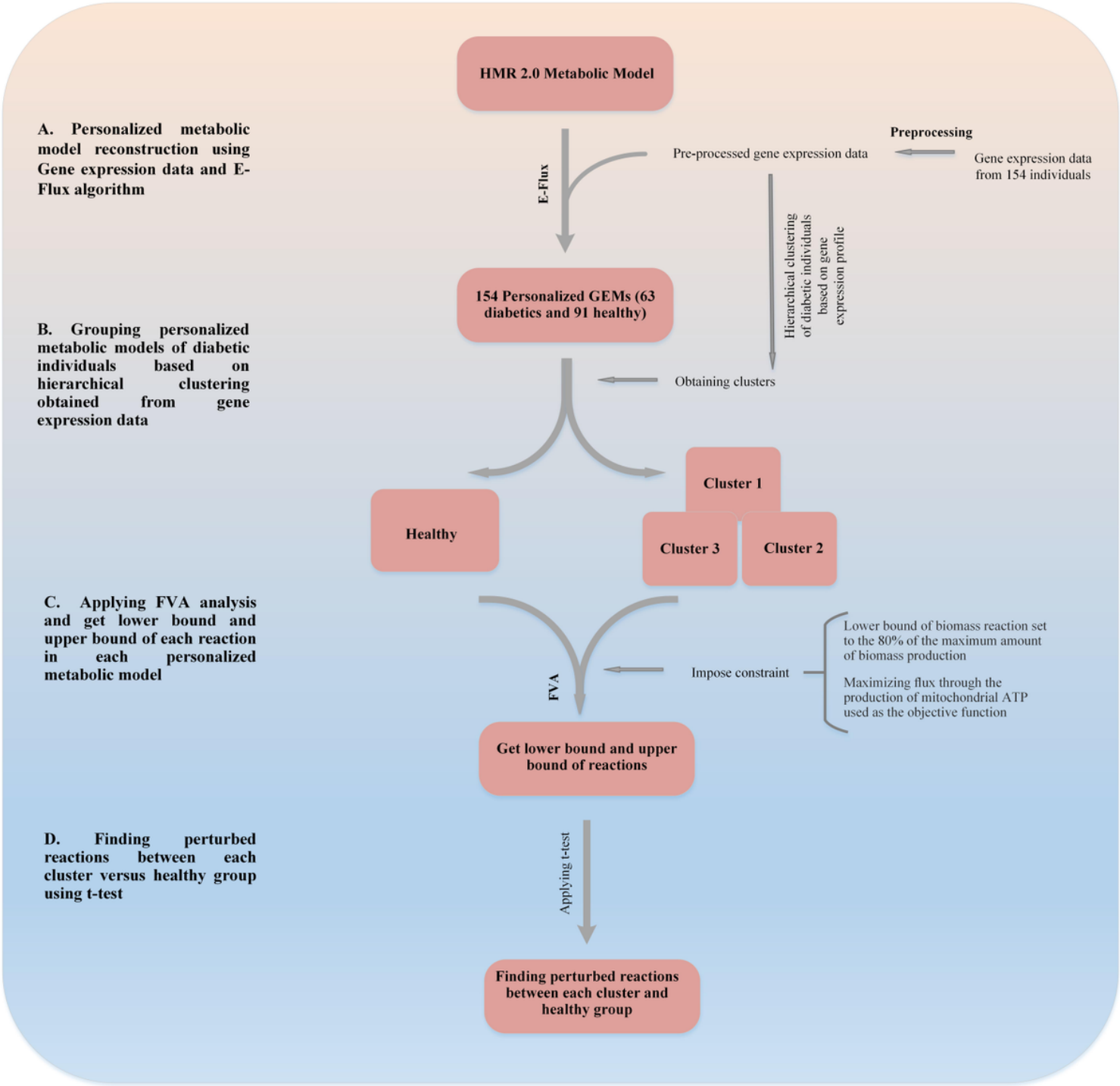

Figure 2

Workflow for cluster-based metabolic modeling. HMR2 model was used as the generic model. Personalized metabolic models of 154 individuals reconstructed by integrating gene expression data to the HMR2 using E-flux algorithm. Diabetic models categorized in three groups based on hierarchical clustering obtained from gene expression data. FVA was employed to obtain maximum and minimum possible flux in each reactions. Perturbed reactions in each cluster in compared to healthy groups identified by applying t-test on obtained fluxes.

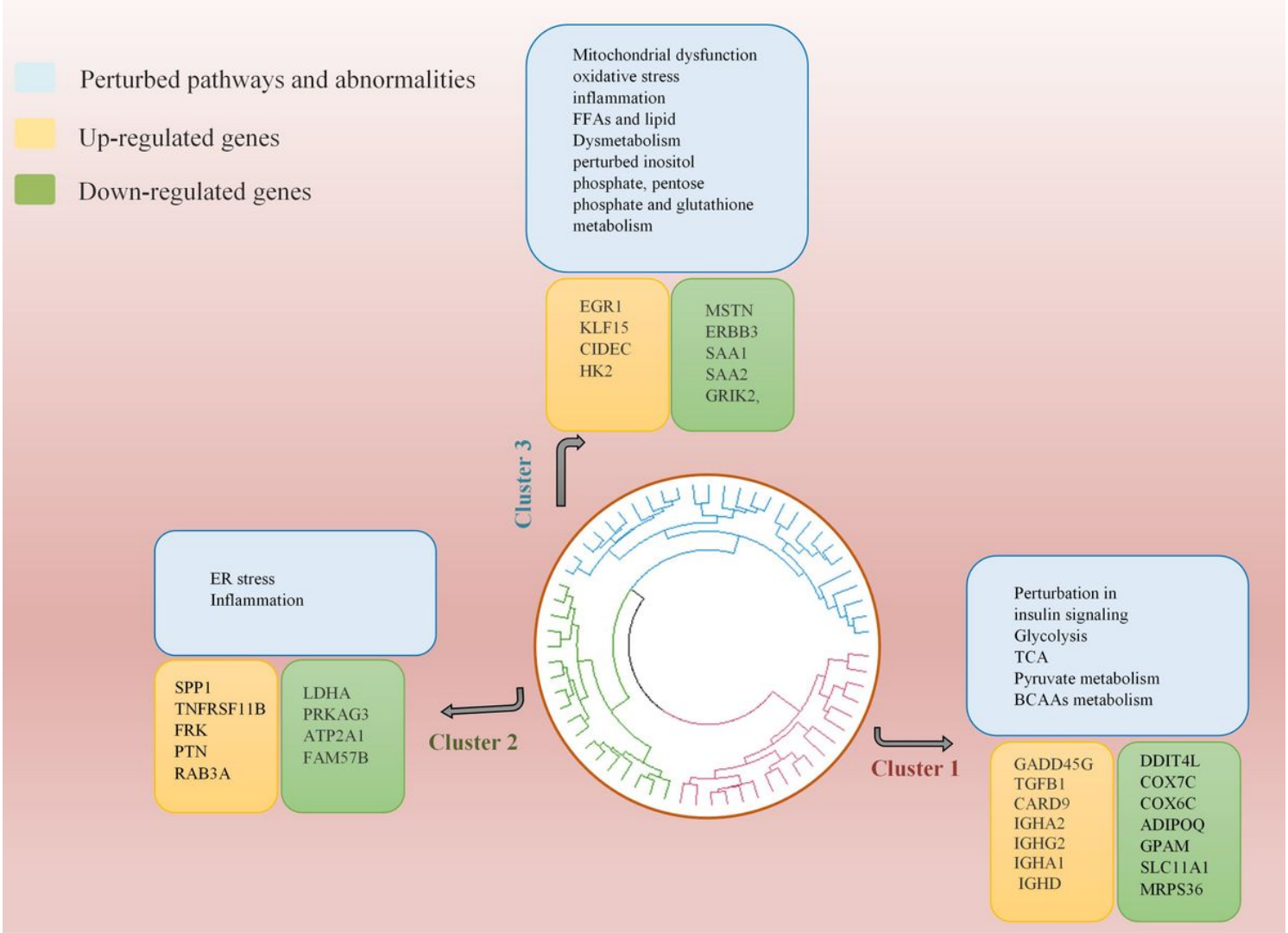

Figure 3

Hierarchical clustering of diabetic samples. Top three clusters were selected and differentially expressed genes and perturbed pathways in each cluster compared to normal samples were found. Some of the dysregulated genes and pathways in each cluster are shown in the boxes as an example. Green boxes show down-regulated genes and yellow boxes show up-regulated genes. The blue boxes show perturbed pathways and abnormality in each cluster.

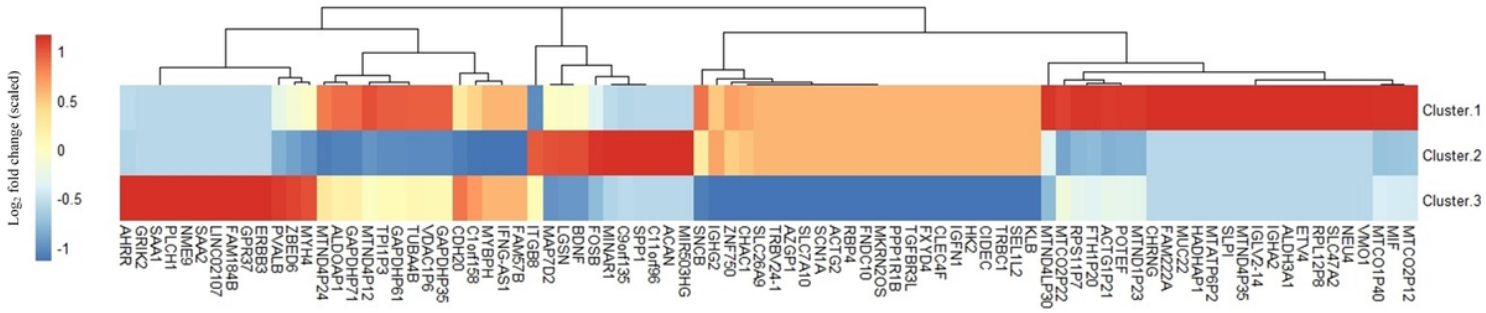

Figure 4

Hierarchical clustering of diabetic samples. Top three clusters were selected and differentially expressed genes and perturbed pathways in each cluster compared to normal samples were found. Some of the dysregulated genes and pathways in each cluster are shown in the boxes as an example. Green boxes show down-regulated genes and yellow boxes show up-regulated genes. The blue boxes show perturbed pathways and abnormality in each cluster.

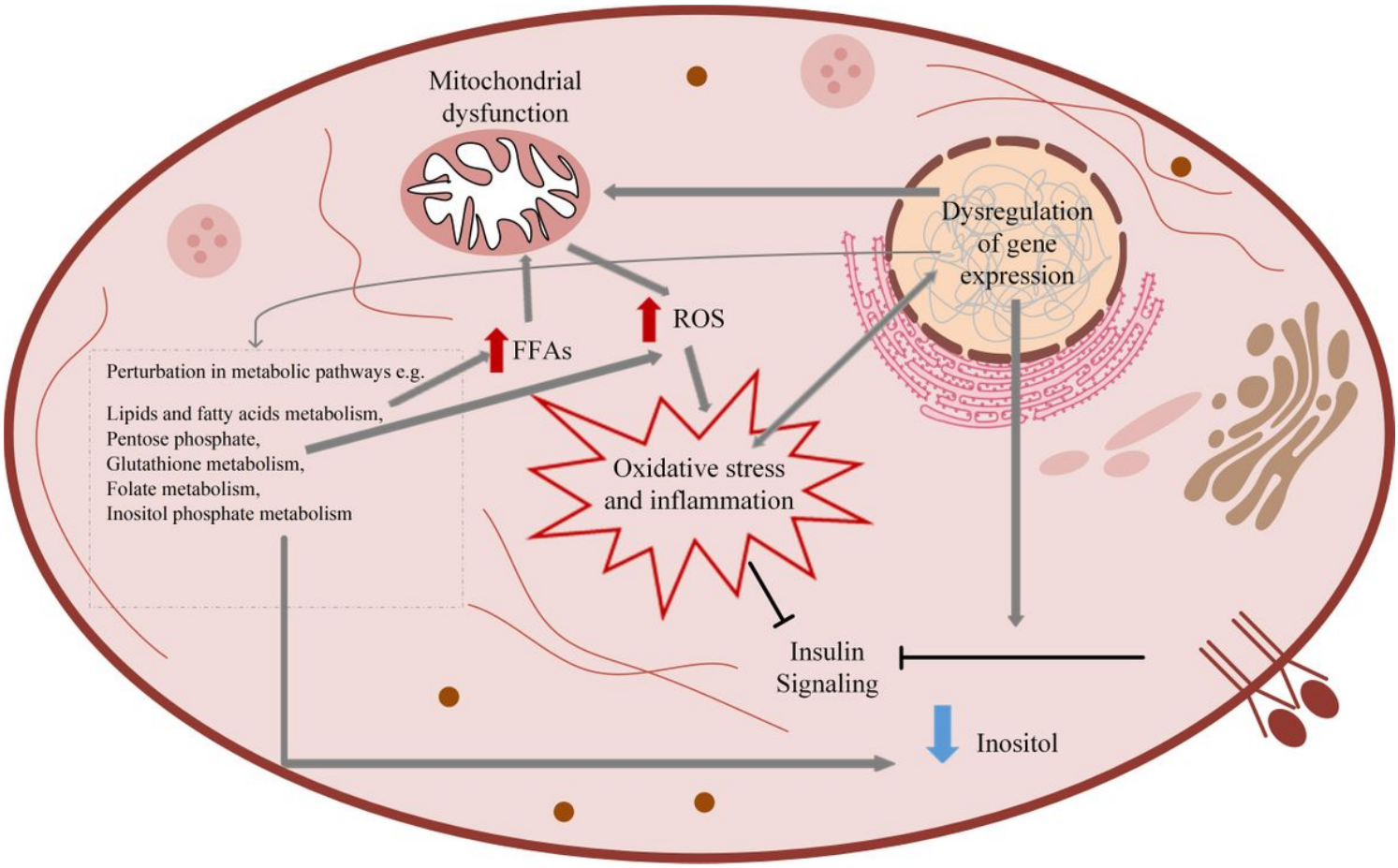

Figure 5

Schematic representation of abnormalities in cluster 1

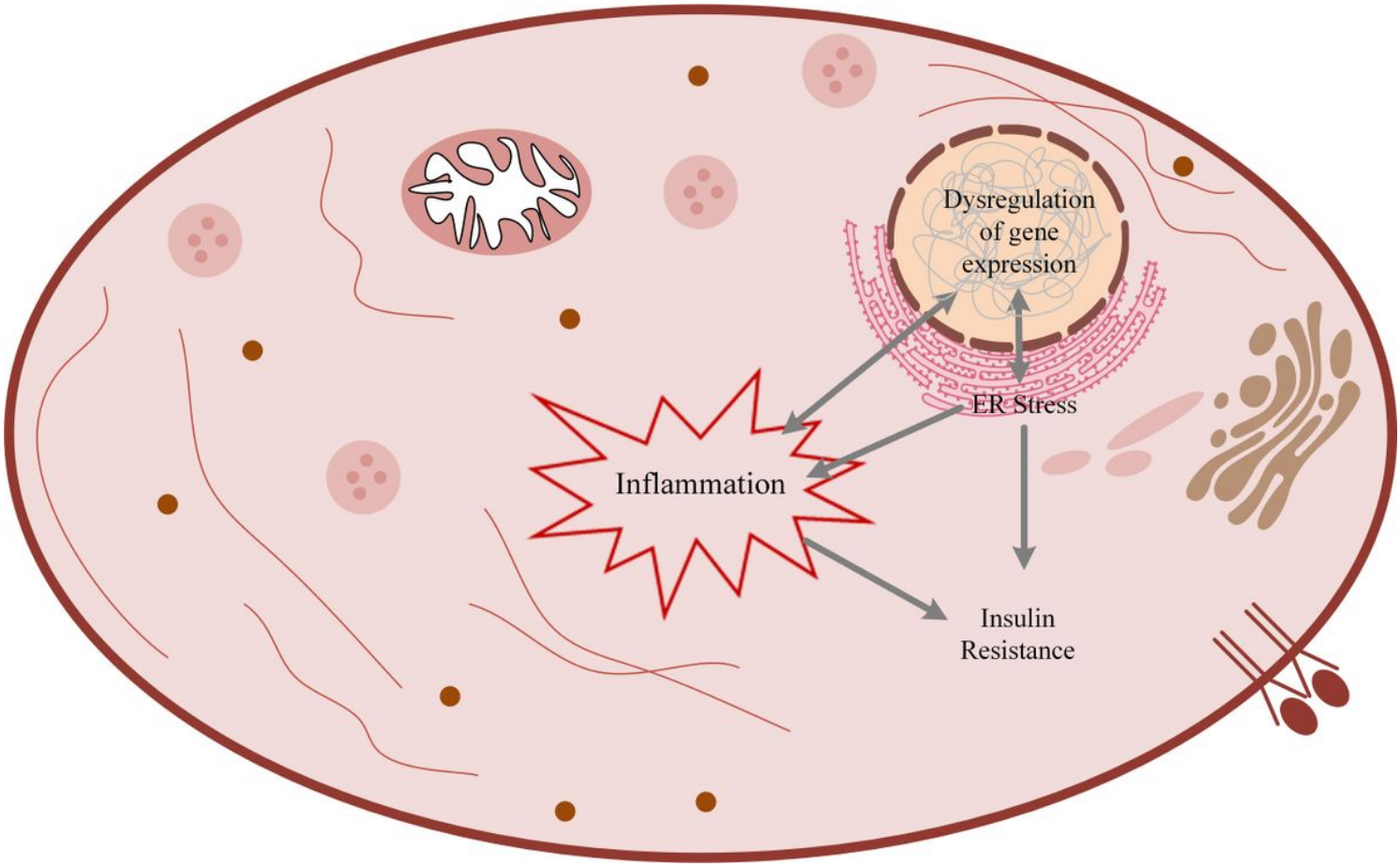

Figure 6

Schematic representation of abnormalities in cluster 2

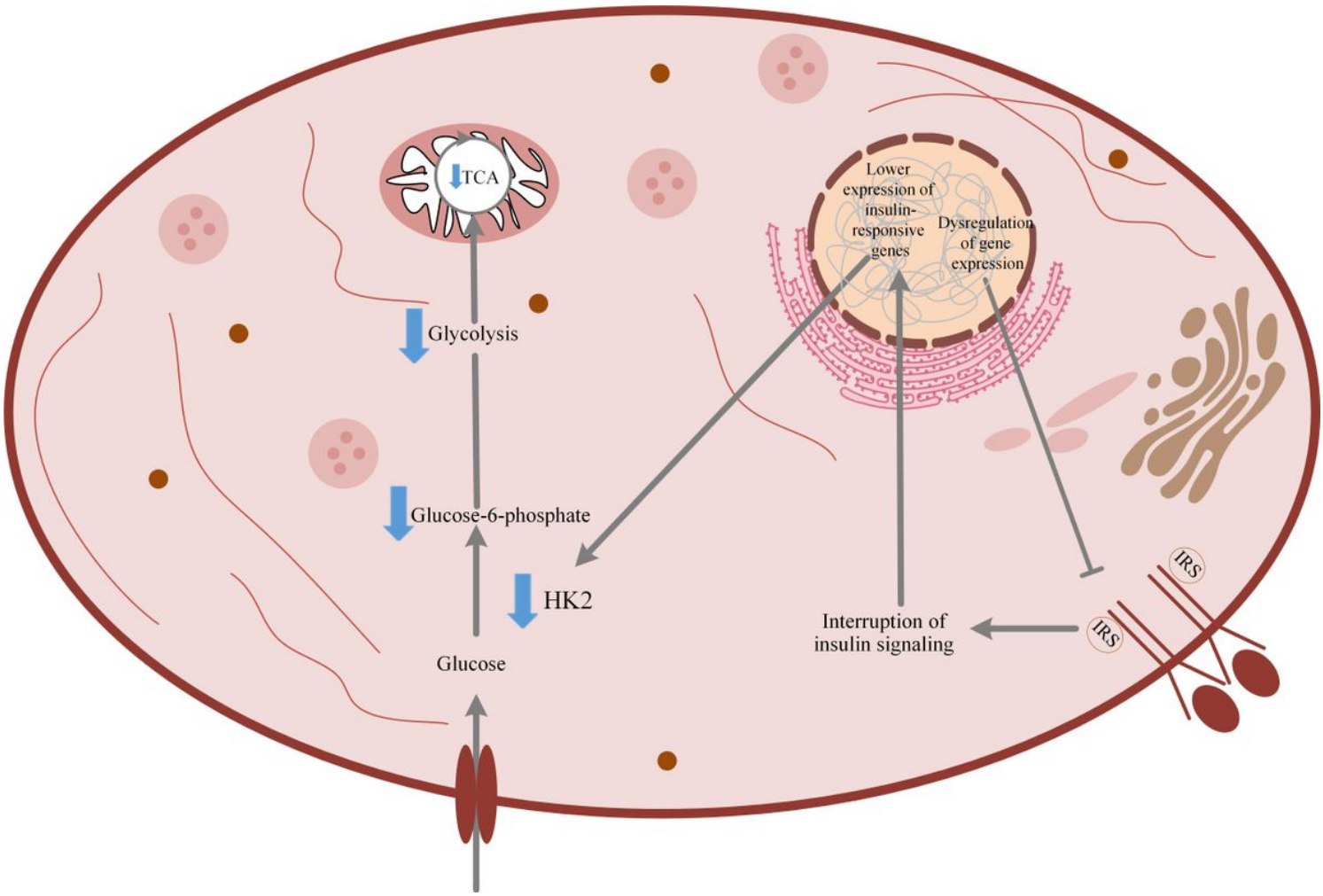

Figure 7

Schematic representation of abnormalities in cluster 3

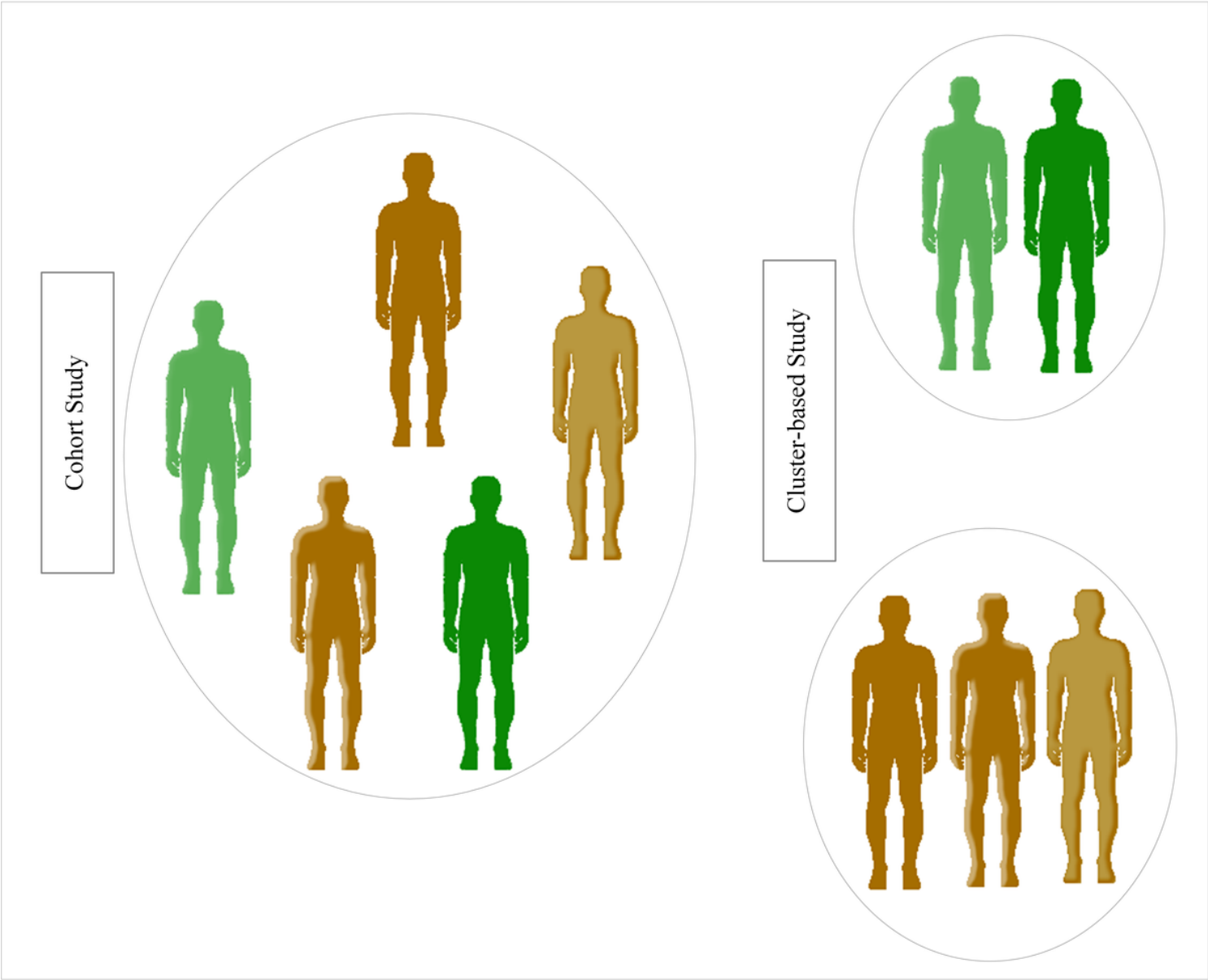

Figure 8

Cluster-based study versus cohort study. In a cohort study, a sample consists of several subjects is gathered and examined. In a cluster-based study, a sample that has been collected in a cohort study is broken down into the sub-groups so that the members within each subgroup have the most similarity and differ from the members of the outer sub-groups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2.xlsx
- Additionalfile1.docx