TMED9 Protein
Final Report

Course: 8389385-36

Supervisor: Dr. Moran Dvela-Levitt

By: Itai Alcalai 206071110

Table of Contents

Documentation

The entirety of the project, from tools to outputs, has been documented and archived in our dedicated repository. For an insight and access to all technical components, you can visit the provided link: https://github.com/itaialcalai/TMED_project.git . Please refer to the README from the said repository, serving as a brief overview of the project's structure and contents.

Abstract

Proteinopathies provide invaluable insights into cellular protein homeostasis, exemplified by conditions like Mucin 1 Kidney Disease (MKD). Within this context, the TMED protein family stands out, especially TMED9, for its pivotal role in optimizing protein trafficking. Recent findings highlight the co-localization of mutant MUC1-fs protein with TMED9, suggesting TMED9's integral role in MKD. Compounds such as BRD4780 have further demonstrated therapeutic promise by facilitating protein degradation pathways in MKD scenarios.

Building on this foundation of understanding around TMED9's role in MKD, our bioinformatics study delved into the Gene Expression Omnibus (GEO) repositories to discern TMED9's potential implications across a broader range of diseases. Our selection emphasized conditions marked by protein misfolding, including Aging, Cystic Fibrosis, Parkinson's Disease, among others. Employing rigorous statistical analyses through tools like R and Python, and buttressed by visual aids such as scatter plots, heatmaps, and regression plots, we took a deep dive into TMED9's potential role in these diseases, with a subsequent focus on Type 2 Diabetes (T2D). Our research began with a feasibility assessment on MKD using the GSE129943 dataset, laying the groundwork for future investigations, and the preliminary findings were encouraging.

In our primary exploration of T2D, initial examinations suggested variations in TMED9's expression among T2D patients. As we delve deeper into T2D research, the pronounced expression of TMED9 became more evident, when

considered alongside its correlations with established control genes like SRR, NFKB1, and PDE4B. This relatively heightened expression underscores TMED9’s potential significance in the pathogenesis of T2D. Furthermore, TMED9's distinct behavior compared to the housekeeping gene, GAPDH, hints at a specialized function within this framework. However, the intricate web of gene interactions demands prudence.

While data tilts in favor of TMED9's participation, the precise mechanistic role, and its broader impact within T2D warrant further clarification. Additionally, the research did not provide a definitive insight concerning variations across different tissues and population origins.

In conclusion, while we observed a consistent correlation in TMED9's behavior within T2D anchored by the reference control genes, our research did reveal some anomalies in consistency. Hence, our final takeaway underscores that while TMED9's involvement is not dismissed based on our bioinformatics approaches, further nuanced investigations, especially before delving into extensive wet-lab experiments, remain paramount.

Biological background

Proteinopathies:

Proteinopathies, a group of diseases stemming from protein misfolding and accumulation within cells, are becoming a significant area of concern in the realm of medical research. The intracellular accumulation of these misfolded proteins is known to cause toxic proteinopathies—some of which currently lack specific therapies. Historically, numerous diseases, including amyotrophic lateral sclerosis (ALS), Parkinson’s disease, and retinitis pigmentosa (RP), have been linked to the intracellular build-up of these problematic proteins. In some of these diseases, such as RP, the misfolded proteins gather within the secretory pathway, specifically the endoplasmic reticulum (ER) and Golgi apparatus. In contrast, diseases like Huntington’s disease showcase misfolded protein aggregates either in the cytoplasm or nucleus.

The eukaryotic cell's early secretory pathway comprises the ER, the ER-Golgi intermediate compartment (ERGIC), and the Golgi apparatus. These organelles play a pivotal role in synthesizing, folding, and maturing nearly one-third of all cellular proteins. Any disruption within this intricate system, such as the accumulation of misfolded proteins, can trigger the unfolded protein response (UPR). The UPR acts as a cellular defense mechanism, striving to restore balance and maintain cellular homeostasis. However, when the stress is too great or prolonged, the UPR's protective capacity might fail, leading to the onset of cell death—a hallmark seen in various proteinopathies.

Mucin 1 Kidney Disease (MKD):

One particular proteinopathy of interest, the Autosomal dominant tubulo-interstitial kidney disease-mucin1 (ADTKD-MUC1 or MUC1 kidney disease, MKD), arises from a frameshift mutation in the MUC1 gene. This mutation produces a variant of the protein, referred to as MUC1-fs, which differs from the wild-type protein. Specifically, this mutant lacks the transmembrane and intracellular domains present in the conventional MUC1 protein.

The MUC1 gene encodes the glycoprotein mucin 1 (MUC1), prevalent on the apical surface of various epithelial cells spread across organs like the mammary gland, lungs, and kidneys. Within the kidneys, its presence is primarily seen in the distal convoluted tubules and collecting duct. Interestingly, post ischemia, this protein can also emerge within the proximal tubule.

A prominent feature of MKD is the consistent nature of the causative mutations across all known cases. They all produce the same frameshift in the protein, specifically via the insertion of an extra cytosine within one of the VNTR subunits. Despite understanding its genetic origins and protein effects, the exact molecular mechanism triggering MKD remains elusive. Further amplifying the concern is the absence of an effective therapy targeting this disease.

TMED protein family:

The p24 family, which includes TMED (Transmembrane Emp24 Domain-containing proteins), is a group of evolutionary conserved membrane proteins that play pivotal roles in the trafficking of proteins between the endoplasmic reticulum (ER) and the Golgi apparatus. TMED9, a member of this family, specifically functions as a cargo receptor and assists in the packaging and transport between the ER and the cis-Golgi compartments. Given its role in COPI retrograde transport from the cis-Golgi back to the ER, TMED9 is essential for ensuring that the protein traffic within the cell is seamless and efficient.

Mutant MUC1-fs protein tends to co-distribute with TMED9. This co-distribution was consistently observed across various sources including human cells, kidney organoids, kidney biopsy samples, and mouse kidney. This observation highlights the significance of the interaction between MUC1-fs and TMED9 in the pathology of Mucin 1 Kidney Disease (MKD) and further emphasizes the central role of the TMED protein family in regulating cellular protein homeostasis.

Mechanism Reversing MKD

The molecular entrapment of the mutant MUC1-fs in the early secretory compartments of the cell triggers the unfolded protein response (UPR), indicating cellular stress. This study demonstrated that the mutant MUC1-fs tends to accumulate in TMED9 cargo receptor-enriched compartments within the cell, leading to an imbalanced protein homeostasis.

To rectify this imbalance, Dr. Dvela-Levitt's research team identified a small molecule named BRD4780. The molecule showcases potential as it releases MUC1-fs from TMED9-enriched compartments, promoting its anterograde trafficking towards the lysosome for degradation. Interestingly, the deletion of TMED9 produced a similar effect to that of BRD4780, implying the possibility of targeting TMED9 as a therapeutic intervention.

In vivo experiments with mice models, when treated with BRD4780, showed a dose-dependent clearance of the mutant MUC1-fs protein from their kidneys.

Furthermore, BRD4780 was also found to effectively reduce pathways associated with ER stress and UPR, indicating its potential in restoring cellular balance.

When testing the impact of BRD4780 on human kidney organoids derived from MKD patient iPSCs, the findings mirrored the animal studies. The mutant protein was successfully cleared from intracellular compartments without affecting the wild-type MUC1 protein levels.

The exact mechanism through which BRD4780 exerts its effect was further clarified with experiments that demonstrated the compound's reliance on a functional secretory pathway and lysosomal degradation. In essence, BRD4780 promotes the trafficking of MUC1-fs through the secretory pathway to the lysosome where it is degraded. This ensures that the misfolded protein does not accumulate within the cell, potentially offering a novel therapeutic avenue for MKD and possibly other proteinopathies.

Technological background

GEO:

The Gene Expression Omnibus (GEO) is an expansive and publicly available repository that houses high-throughput genetic data. Developed and maintained by the National Center for Biotechnology Information (NCBI), GEO plays a vital role in providing researchers with a wealth of genetic datasets. This includes data from microarray experiments, next-generation sequencing, and other forms of high-throughput functional genomics data. The open nature of GEO ensures that we can freely access information regarding our project.

RNA Sequencing (RNA-Seq):

RNA sequencing (RNA-Seq) stands as one of the most transformative technologies in functional genomics. This technique deciphers the quantitative and qualitative nature of RNA molecules in a given sample. Unlike traditional methods, RNA-Seq allows for a comprehensive survey of the entire transcriptome, capturing a wide range of RNA species, including mRNAs,

non-coding RNAs, and small RNAs. The depth and breadth of RNA-Seq have enabled researchers to glean insights into differential gene expression, novel transcript identification, and alternative splicing events, and more.

Project Goals

The field of bioinformatics has created an opportunity to research cellular intricacies of diseases more easily. Given the substantial understanding of TMED9's involvement in Mucin 1 Kidney Disease (MKD) due to mutant protein interactions through Dr. Dvela-Levitt's work, emerges a question and the basis of the project: Is TMED9 implicated in other disease states?

To address this, we plan to utilize the vast repositories of the Gene Expression Omnibus (GEO). With its assistance, we aim to mine indications of TMED9's involvement in diseases that have been less explored within the context of proteinopathies.

Considering the vast array of diseases influenced by protein misfolding, our focus will be on diseases that are common, affect multiple systems, and present a pronounced proteinopathy element. The rationale behind this is to ensure we are targeting diseases that would benefit most from a pre-investigation through bioinformatics. Notably, diseases where a wet-lab confirmation of TMED9's involvement would be particularly challenging or costly will be given priority.

Disease States Prioritized for Investigation:

1. Diabetes Type 2 - Given its prevalence and the complex interplay of cellular processes involved.
2. Aging - A natural process, yet with potential links to various proteinopathies.
3. Cystic Fibrosis - A genetic condition with notable disruptions in protein processing.
4. Parkinson's Disease - Renowned for its protein aggregation, offering a potential avenue for TMED9 involvement.

5. Kreutzfeld-Jakob Disease (Prions) - A rare neurodegenerative disorder where protein misfolding plays a crucial role.
6. ALS (Amyotrophic Lateral Sclerosis) - Given its historical link to intracellular build-up of problematic proteins.
7. Alzheimer's Disease - With hallmark protein aggregates, understanding TMED9's role could offer novel insights.
8. Obesity - A pervasive condition with multifaceted cellular disruptions, where protein pathways could play a part.
9. Cancer - Given the myriad of cellular anomalies, especially in protein signaling and homeostasis.

Our project will lean heavily on already available RNA-Seq data from GEO. This data will provide insights into the differential gene expression patterns related to TMED9 across different disease states. By analyzing these patterns, we hope to draw connections and hypotheses about TMED9's role beyond its known involvement in MKD.

Given the success of molecules like BRD4780 in rectifying protein imbalances associated with MKD, a key part of our endeavor will be to determine if similar pathways can be identified and exploited for diseases where TMED9 plays a role.

In conclusion, the overarching goal of this project is to utilize a bioinformatics-driven approach to indicate TMED9's involvement across a broader spectrum of diseases. We hope to pave the way for novel research directions and treatments, enriching the collective scientific understanding of TMED9's role in human health.

Methods of Work

Statistical Considerations:

Our null hypothesis for each pairwise comparison is that there is no differential gene expression between the diseased and healthy samples in the context of TMED9 involvement. Rejections of the null hypothesis will prioritize p-values, and where possible, adjusted p-values to control for the false discovery rate.

Data Collection:

To ensure an unbiased investigation into TMED9's involvement across various disease states, we will manually search GEO for relevant public RNA-seq datasets. These datasets should contain samples from both diseased and healthy human subjects. Our emphasis will be on datasets that capture the natural state of the disease without any external interventions, such as drug treatments or gene knockouts, that might introduce bias to our results.

Data Processing and Analysis:

During the early stages of data analysis, our approach involved fitting data to pyDESeq2 and manual calculations using average expression values to gauge gene significance. However, as the project progressed and after several iterations, we refined and perfected our methodology to achieve more accurate results. This was an iterative learning process, wherein feedback from the preliminary results contributed to enhancing the final script used for analysis.

With the selected datasets, we will conduct a pairwise comparison using a custom R script. This script has been designed to perform a series of transformations, normalization, and statistical analyses using the libraries: limma, GEOquery, and BiocManager.

The workflow of the R script handling the pairwise comparison is outlined as follows:

- Data Retrieval and Preparation: The function begins by fetching the specified GEO dataset, followed by a series of data cleaning and

transformation steps. These include filtering out excluded samples, performing log2 transformations if necessary, and normalizing data between arrays.

- Design Matrix Creation: All samples receive group assignments based on a manually curated classification string of bits (0 for "sick" and 1 for "normal"). This subsequently is used to create a design matrix essential for statistical modeling.
- Linear Model Fitting and Contrast Definition: A linear model is fit to the data, and contrasts of interest are defined to determine differential expression between groups.
- Statistical Analysis: The linear model coefficients are recalculated based on the contrasts, and empirical Bayesian statistics are computed to determine significance.
- Data Extraction: Tables of significant genes are generated, and subset based on specific genes of interest, and will output tables for the top 1000 genes, control genes, and mainly the TMED gene family.
- Output Data: The findings from the analysis are recorded in local, tab-separated tables.

Data Visualization:

Post-analysis, the fold-change and p-values from the results will be visualized using scatter plots to aid in data interpretation. We will also leverage other visualization techniques based on the tables generated by the R script to gain deeper insights into TMED9's involvement.

Feasibility Test: TMED9's Involvement in MKD

To initiate our exploration of TMED9's role across diseases, we began by analyzing the GSE129943, curated by Dr. Dvela-Levitt and Kohnert E of the Broad Institute's Greka Lab, this dataset examines the gene expression disparities between MUC1 kidney disease (MKD) patients and healthy controls. Given the dataset's relevance to our research objectives it presents a propitious starting point.

By identifying significant differentials in this dataset, especially concerning TMED9, we set a foundational premise for our broader research. A positive indication of TMED9's involvement in MKD would not only validate our initial hypotheses but also solidify the credibility and direction of our subsequent analyses across other disease states.

Analysis:

Employing the previously mentioned methods, we subjected the datasets available RSEM counts file to a DESeq analysis, to discern differentially expressed genes, especially the TMED protein family. We present the DESeq2 output table with columns: baseMean, Log2FoldChage, lfcSE, stat, pvalue, padj, -log(padj). As well as raw statistical data table- the average expression and its standard deviation (± stdv):

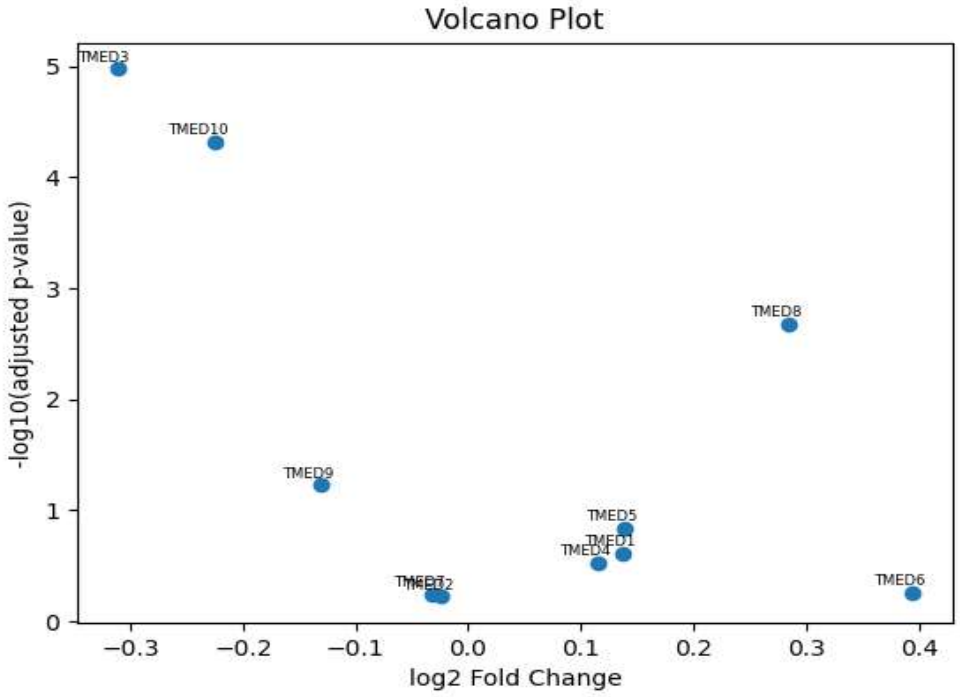
Output table

	index	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	-log10(padj)
0	TMED1	899.380422	0.137489	0.094506	1.454814	0.145721	0.242868	0.614630
1	TMED10	10617.890719	-0.224445	0.050704	-4.426580	0.000010	0.000048	4.319942
2	TMED2	6706.403298	-0.023752	0.045176	-0.525769	0.599048	0.599048	0.222538
3	TMED3	1059.780764	-0.311651	0.063860	-4.880207	0.000001	0.000011	4.974799
4	TMED4	3022.803728	0.115958	0.092280	1.256583	0.208905	0.298435	0.525150
5	TMED5	3677.872811	0.138982	0.077592	1.791194	0.073262	0.146524	0.834090
6	TMED6	9.274198	0.394829	0.512706	0.770088	0.441247	0.551559	0.258408
7	TMED7	5433.210685	-0.031330	0.047898	-0.654109	0.513042	0.570047	0.244090
8	TMED8	2521.922830	0.284681	0.083322	3.416635	0.000634	0.002113	2.675031
9	TMED9	5589.698525	-0.130402	0.057681	-2.260757	0.023774	0.059436	1.225953

Statistical Output table

TMED#	Normal ± stdv	MKD ± stdv	Diff
TMED1	923.89 ± 8.528	833.395 ± 54.596	-90.495
TMED2	6501.27 ± 1086.993	6592.87 ± 118.794	91.6
TMED3	914.5 ± 118.087	1153.5 ± 67.175	239
TMED4	3151.495 ± 16.256	2839.5 ± 12.021	-311.995
TMED5	3575.585 ± 791.189	3383.855 ± 88.197	-191.73
TMED6	10.5 ± 0.707	8.0 ± 7.071	-2.5
TMED7	5174.75 ± 955.485	5308.72 ± 0.75	133.97
TMED8	2669.5 ± 242.538	2142.5 ± 232.638	-527
TMED9	5080.0 ± 410.122	5752.0 ± 84.853	672

To visualize we present a scatter plot. On the x-axis, the log2 fold change depicts the direction and degree of gene expression differences between the two conditions. Meanwhile, the y-axis illustrates the -log10(adjusted p-value), emphasizing the statistical significance of each gene's differential expression. A higher value on the y-axis denotes greater significance.



Scatter Plot of TMED genes in MKD

The baseMean illustrates the average expression of genes across all samples, with Log2FoldChange (Log2FC) highlighting the expression difference between healthy and diseased samples. For our dataset, most genes had a Log2FC value between -0.3 and 0.3, indicating stable expression

of the TMED gene family between the conditions. Specifically, TMED9 showed a slight increase in expression in diseased samples, evidenced by a Log2FC of 0.057. The significance of expression changes was captured by the p-value, with TMED10, TMED3, and TMED8, exhibiting particularly low p-values, implying their potential relevance in the disease.

Conclusion:

TMED9 exhibited a notable upregulation in MKD patients, further supported by its low p-value, aligning with wet-lab findings linking TMED9 to MKD. However, our study had limitations: a narrow focus on the TMED family without comparison to other genes, and a small sample size that could impact statistical robustness. Still, as a feasibility test, our results affirm the value of this computational approach for future investigations into TMED9. For clarity, future analyses will correct misidentifications, like substituting TMED8 with TMED10P1 and adjusting the FC value interpretation.

Findings Type 2 Diabetes Analysis

Type 2 Diabetes (T2D) is a chronic metabolic disorder characterized by insulin resistance and the body's inability to maintain proper glucose homeostasis. This condition, being the most common form of diabetes, affects millions globally. Its emergence can be attributed to a combination of genetic factors, obesity, sedentary lifestyles, and unhealthy diets. The long-term implications of T2D are severe, ranging from cardiovascular complications to neuropathies, nephropathies, and potential retinal damage.

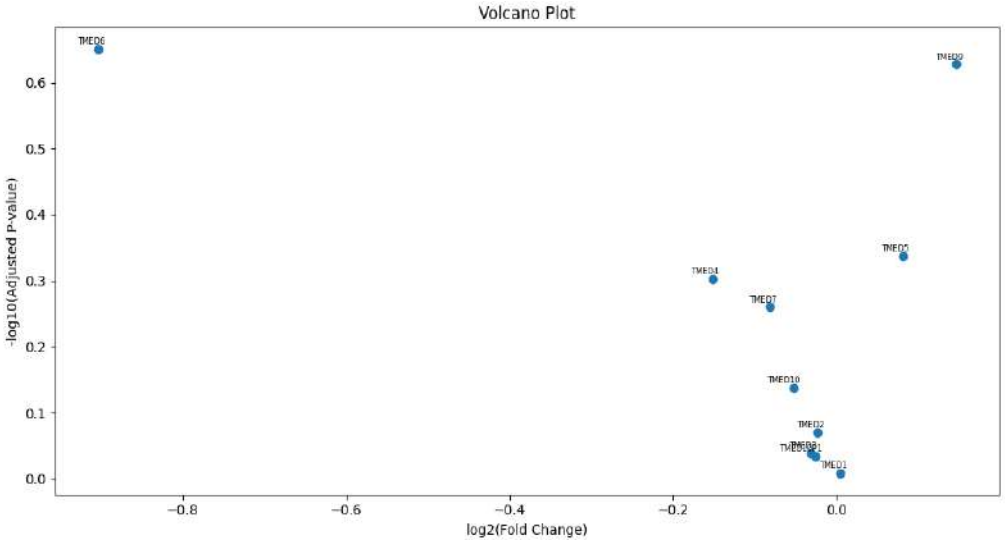
It's noteworthy that proving the involvement of mutant proteins, specifically in conditions termed as proteinopathies that might involve TMED9, is challenging using traditional wet-lab experimental techniques. These challenges amplify the need for computational methods.

Initial Testing

Our research started with an initial test, aiming to uncover potential indications of TMED9's role in T2D using datasets derived from pancreatic islets. Given the critical role these cells play in insulin production and regulation, they represent an ideal focal point for T2D studies. We curated three specific datasets for our primary analysis, all of which were sourced from North European populations. This not only ensures a consistent genetic backdrop but also aims to offer a uniform ground for our comparative assessments.

The following section presents the refined R script output table, alongside the aforementioned scatter plot for each of the three datasets:

1. <u>GSE50397: 33 T2D, 44 Normal:</u>
    - Characteristics: 36 male, 27 female. Average age: 56



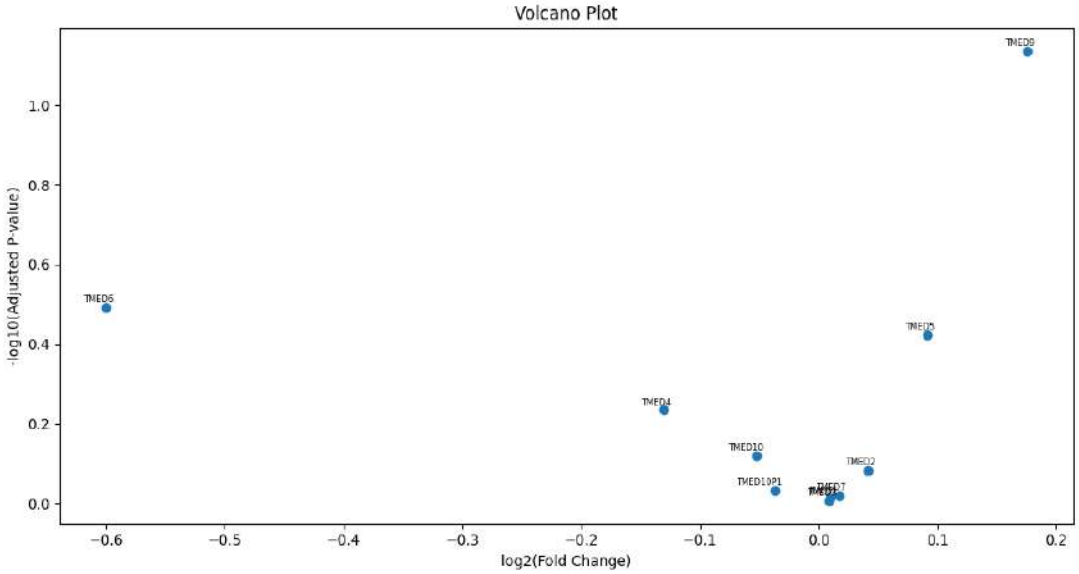
Scatter Plot of TMED genes in T2D dataset 1

Gene.symbol	logFC	t	B	P.Value
TMED6	-0.90300	-2.8100	-2.3676	0.006515
TMED9	0.14600	2.6600	-2.7114	0.009897
TMED5	0.08060	1.6900	-4.5044	0.096494
TMED4	-0.15200	-1.5600	-4.6854	0.123607
TMED7	-0.08170	-1.4100	-4.8875	0.164304
TMED10	-0.05280	-0.9050	-5.4092	0.368738
TMED2	-0.02350	-0.5600	-5.6399	0.577470
TMED3	-0.03220	-0.3540	-5.7261	0.724173
TMED10P1	-0.02610	-0.3090	-5.7400	0.758341
TMED1	0.00463	0.0904	-5.7801	0.928236

Assessing the fold change of the difference in expression per group: normal vs T2D. and corresponding p-value after adjustments – for all TMED genes in a scatter plot.

Results significant in genes: TMED6, **TMED9** with FC: 0.1, and Pval: 0.006.

2. GSE38642: 9 T2D, 54 Normal:
    - Characteristics: 44 male, 33 female. Average age: 57

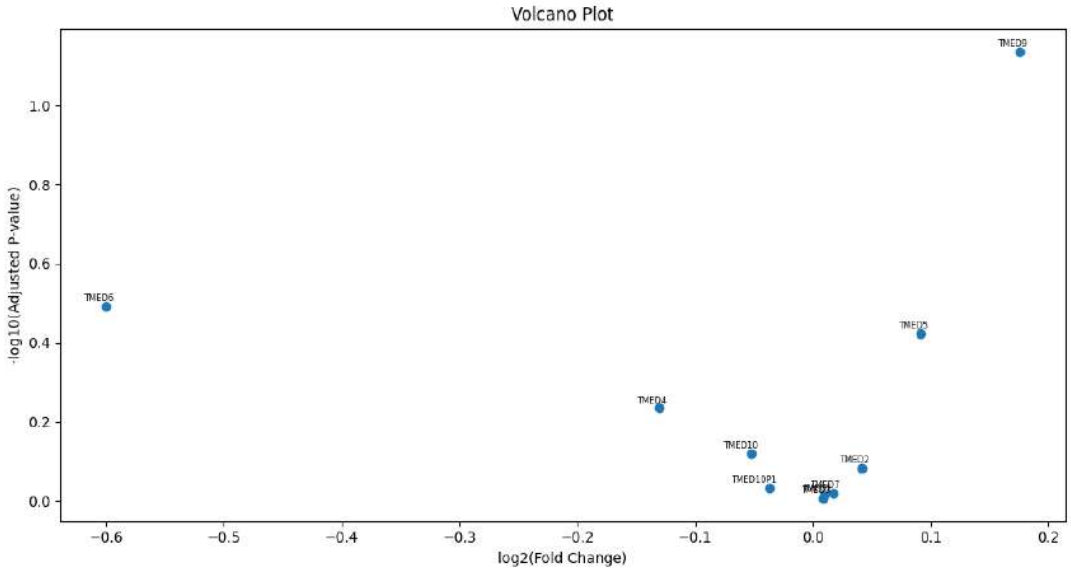


Scatter Plot of TMED genes in T2D dataset 2

Gene.symbol	logFC	t	B	P.Value
TMED9	0.175486	3.912732	0.60361	0.000191
TMED6	-0.600069	-2.444397	-3.15941	0.016700
TMED5	0.091638	2.237973	-3.57454	0.028000
TMED4	-0.130423	-1.659588	-4.56321	0.101000
TMED10	-0.052528	-1.147985	-5.20866	0.254000
TMED2	0.041467	0.926034	-5.41815	0.357000
TMED10P1	-0.037118	-0.498033	-5.69755	0.620000
TMED7	0.017108	0.344609	-5.75711	0.731000
TMED1	0.009664	0.232247	-5.78701	0.817000
TMED3	0.009105	0.116150	-5.80569	0.908000

Results significant in genes: TMED6, **TMED9** with FC: 0.146, and Pval: 0.001.

3. GSE41762: 20 T2D, 57 Normal:
    - Characteristics: 44 male, 33 female. Average age: 57



Scatter Plot of TMED genes in T2D dataset 3

Gene.symbol	logFC	t	B	P.Value
TMED9	0.175486	3.912732	0.60361	0.000191
TMED6	-0.600069	-2.444397	-3.15941	0.016700
TMED5	0.091638	2.237973	-3.57454	0.028000
TMED4	-0.130423	-1.659588	-4.56321	0.101000
TMED10	-0.052528	-1.147985	-5.20866	0.254000
TMED2	0.041467	0.926034	-5.41815	0.357000
TMED10P1	-0.037118	-0.498033	-5.69755	0.620000
TMED7	0.017108	0.344609	-5.75711	0.731000
TMED1	0.009664	0.232247	-5.78701	0.817000
TMED3	0.009105	0.116150	-5.80569	0.908000

Results significant in genes: **TMED9** with FC: 0.175, and Pval: 0.0002.

Conclusion

From our initial analysis leveraging the three datasets, we observe compelling indications of TMED involvement in the context of T2D. Despite the inherent constraints of this preliminary test, which includes the absence of a positive control and reference, as well as a limited tissue and population distribution, the results are nonetheless promising.

For instance, TMED9 consistently showcased an elevated expression in the datasets, with a notable average fold change (FC) of around 0.140 and an impressive average p-value of 0.0024, indicating statistical significance. In two of the datasets, alongside TMED9, TMED6 also emerged as significant. Furthermore, the stability in the direction of fold changes — a positive trend for TMED9 across all datasets — accentuates the potential reliability of our findings. In contrast, TMED6 displayed a consistent negative trend.

The consistent and significant pattern in its expression serves as a strong indication of its involvement. While these results provide a promising start, we acknowledge the limitations intrinsic to our initial test. Moving forward, we aim to delve deeper into the role of TMED9 addressing and overcoming the limitations faced in this preliminary phase.

General Analysis

For the next phase of our research, we procured 23 datasets, spanning diverse geographies and tissue types. The datasets originate from multiple countries including the USA, Japan, the UK, Austria, Italy, Australia, Germany, Sweden, Canada, Denmark, Netherlands, and Norway. These datasets provide us with insights into different tissue samples such as muscle, liver, brain, dermal tissue, peripheral blood mononuclear cells (PBMCs), pancreas, duodenal, adipose, and arterial tissues. The sample sizes in these datasets range from as few as 8 to as many as 77 samples, with a cumulative total of 829 samples. The broadened geographical, tissue-type, and sample size distribution compared to our initial feasibility test substantially improves the robustness of our study.

To determine the significance of TMED9 in the context of Type 2 Diabetes Mellitus (T2D), it's essential to contrast its behavior with genes already established as significant markers for the condition by the broader medical community. This comparative approach leans on the concept of positive control. If TMED9 mirrors or exhibits patterns consistent with these proven markers, it offers a stronger testament to its relevance in T2D's etiology.

Drawing from the foundational work presented by Raza et al. (School of Medical Engineering, Sanquan College of Xinxiang Medical University, Xinxiang, China; and other collaborating institutions), three genes were identified as especially pertinent to T2D: SRR, NFKB1, and PDE4B. The study hinged on a comprehensive analysis of 20 public cDNA datasets, synergizing various bioinformatics tools and methods, including Gene Ontology (GO) annotations, co-expression analysis, and pathway enrichment. Their comprehensive systems-level analysis unearthed genes associated with critical biological functions like amino acid metabolism, signal transduction, and intracellular transport. Importantly, the results underlined the enriched presence of these genes in insulin signaling and other T2D-associated pathways, such as the P13K/Akt pathway.

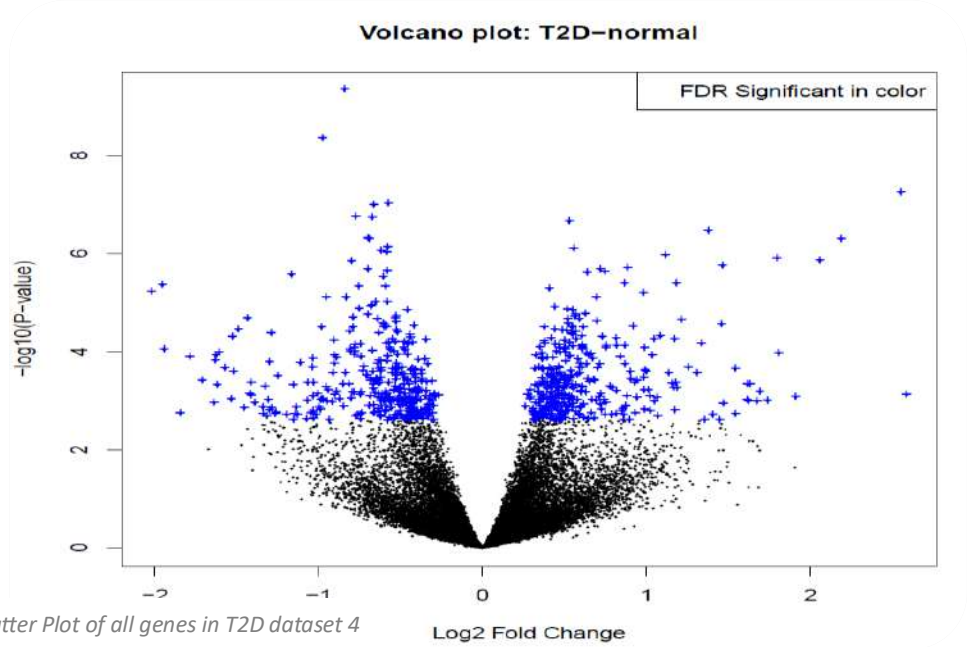
Quantitative PCR (qPCR) validation, using glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as a reference gene, revealed a significant expression level change for the genes: SRR (FC ≤ 0.12), NFKB1 (FC ≤ 1.09), and PDE4B (FC ≤ 0.9) compared to controls (FC ≥ 1.6). Their altered expression was linked to metabolic disorder manifestations and the pathophysiological development of T2D. As we pivot to our analysis, evaluating TMED9 against these benchmark genes will enable a clearer discernment of TMED9's potential role in T2D.

At this stage we deployed our finalized R script as mentioned in the work methods as part of a generalized program designed to process each dataset by its specific GSE matrix file sourced from the GEO database. This facilitated a streamlined and uniform analysis across the datasets, ensuring consistency and reproducibility in our methodological approach.

For visualization purposes, a series of plots were generated for each of the 23 studies. To elucidate the gene expression dynamics and the visualization techniques employed, we will showcase the results from one of the studies as a representative example of the outputs at this stage.

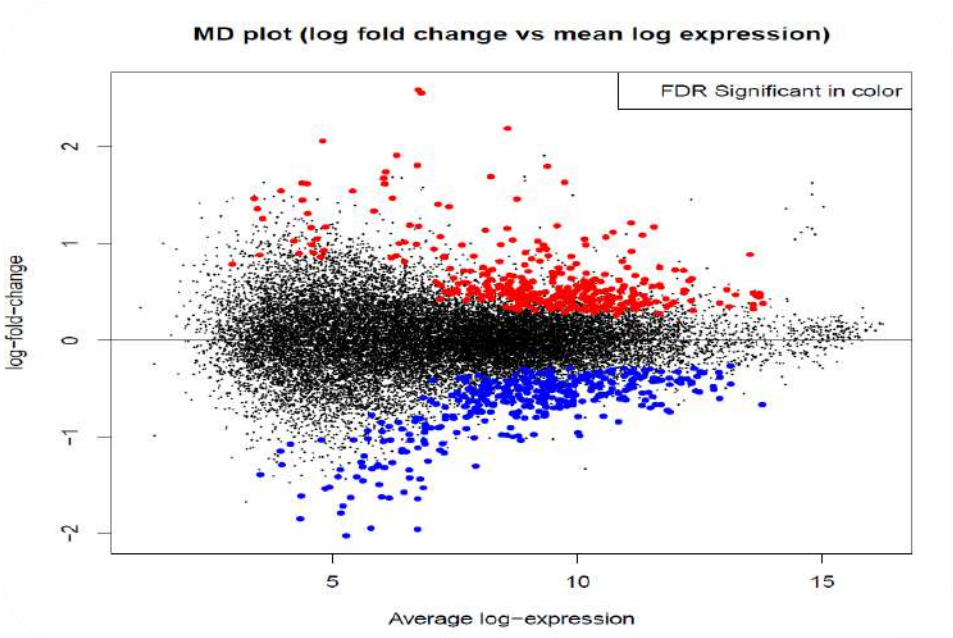
1. Global Gene Distribution Plot: plot captures the vast spectrum of genes by plotting the p-value (y-axis) against the log fold change (log FC) on the x-axis. This provided a macro view of the gene behavior

across the datasets, highlighting the genes with significant adjustments in their expression at the top corners. Genes that met the False Discovery Rate (FDR) criteria were color distinguished.



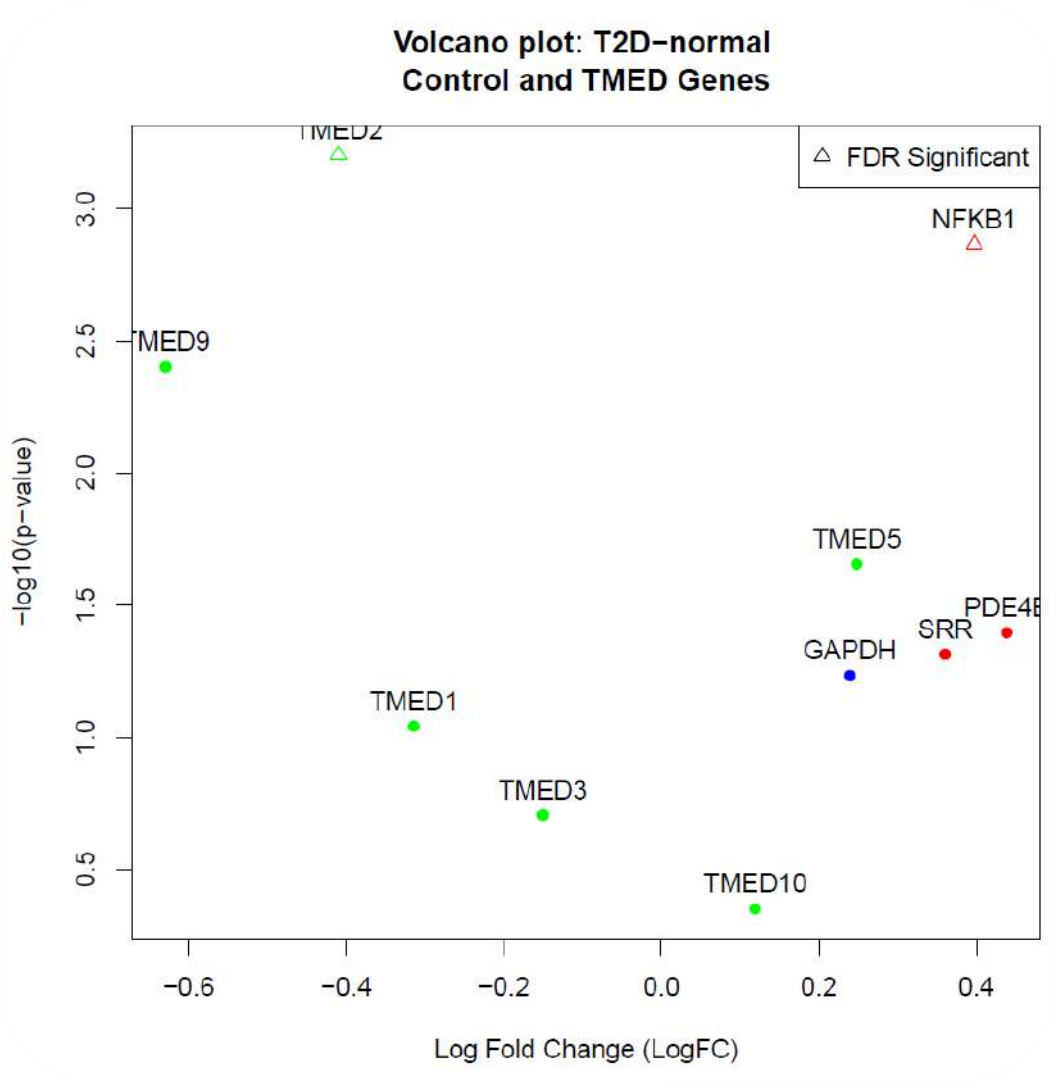
Scatter Plot of all genes in T2D dataset 4

2. Gene Expression versus Fold Change: This plots the log fold change (y-axis) against the average log expression (x-axis), providing insight into the relationship between gene abundance and the magnitude of their differential expression.



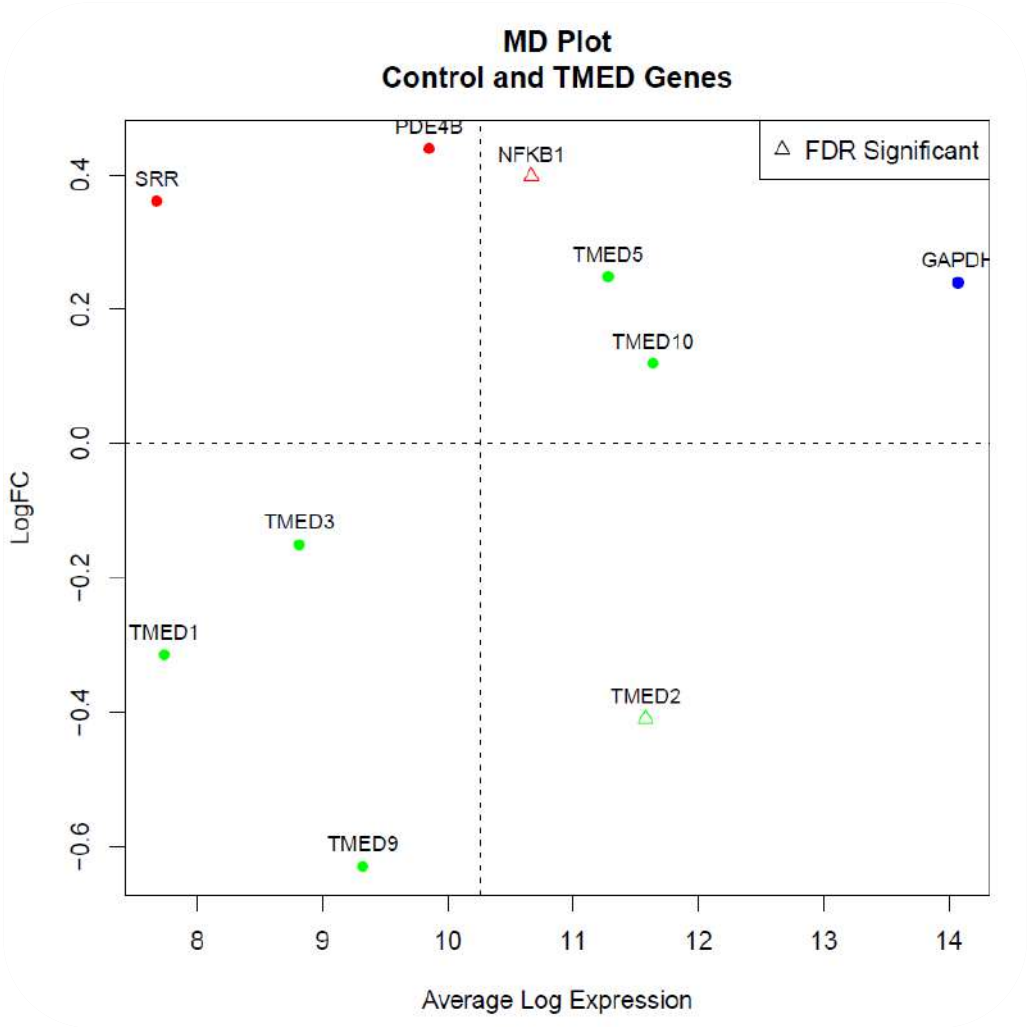
Scatter Plot of Average Expression of all genes in T2D dataset 4

3. Focused TMED Family Distribution: Zeroing in on our genes of interest, this plot specifically showcased the TMED family genes, the three control genes from the sited study, and the reference gene GAPDH. By isolating these genes, we aimed to gain a more micro view of their behavior.



Scatter Plot focused on genes of interest in T2D dataset 4

4. Focused Gene Expression versus Fold Change: Akin to the second plot but tailored for our genes of interest.



Scatter Plot of Average Expression focused on genes of interest in T2D dataset 4

Discerning clear patterns and behaviors from the results was challenging, given their dispersion across 23 distinct studies. This fragmentation underscored the necessity for a more integrated approach in our subsequent analyses. The forthcoming phase of our research seeks to consolidate these fragmented insights, aiming to provide a unified perspective on gene expression dynamics. Through the deployment of tools like heatmaps and color-coded scatter plots categorized by study, we anticipate a clearer, more coherent representation of the overarching trends and relationships embedded in our data.

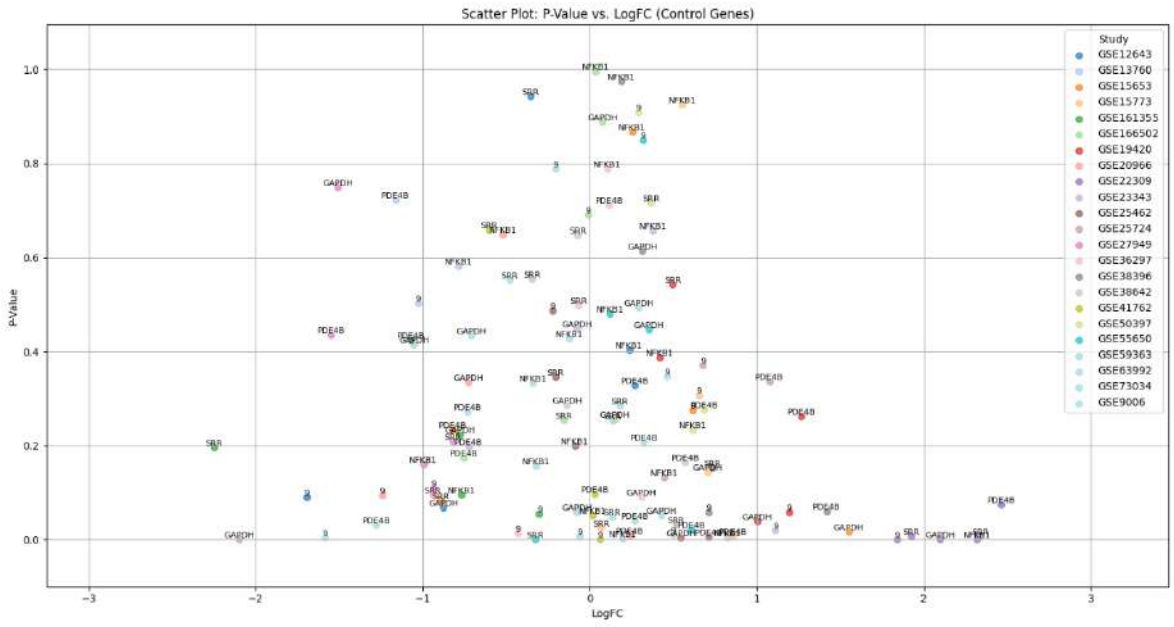
Unified Analysis Across Studies

To achieve a more comprehensive understanding of the gene expression dynamics, especially for our control, reference genes, and the gene of primary interest, TMED9, we integrated data from all 23 datasets into unified scatter plots.

Scatter plots

Scatter Plot Colored by Study and Labeled by Gene Name

The objective of this visualization was to discern the clustering behavior of data points stemming from different studies, while still highlighting the gene under observation. We anticipated the formation of multicolored clusters, where each color represents a specific study. An ideal outcome would feature these clusters, especially in the top corners of the scatter plot, indicating significant gene expressions.

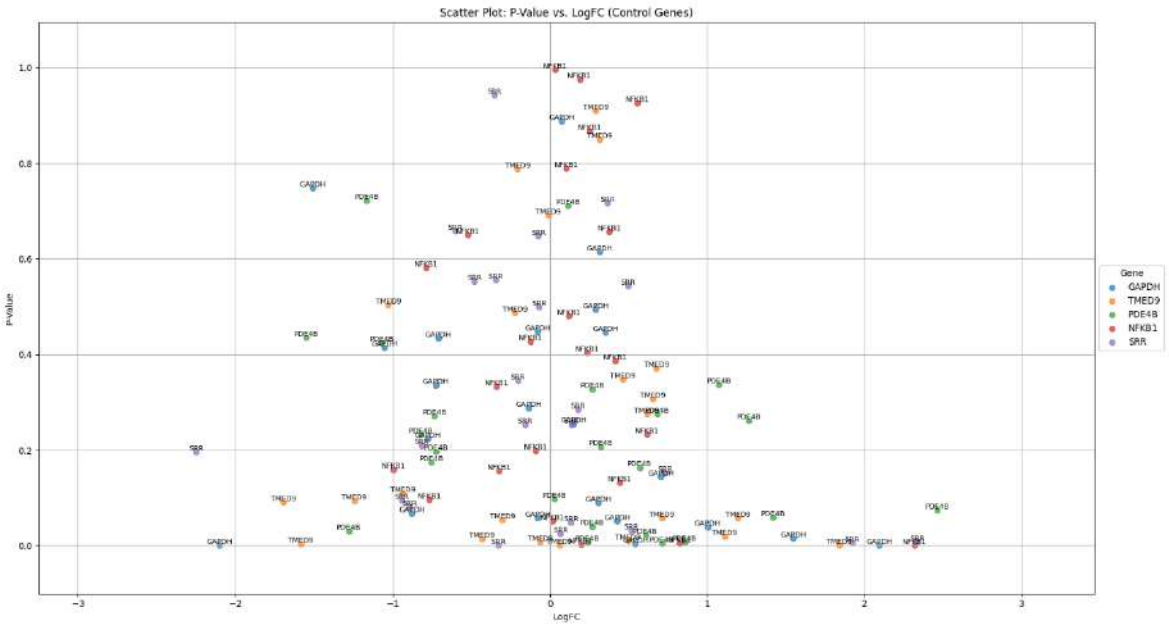


Scatter Plot by study. Focused on genes of interest in T2D 23 for datasets combined.

The data points are scarcely clustered by gene, indicating that there isn't a distinct pattern of expression that can be attributed to specific genes across the different studies. The absence of pronounced colored clusters (by study) underscores the heterogeneous nature of the data.

Scatter Plot Colored by Gene and Labeled by Gene Name:

Transitioning from the study-centric coloring, this plot provides a more direct visualization of the genes themselves. By coloring and labeling the data points by gene name, we aimed to capture the dispersion and clustering of specific genes across all studies.



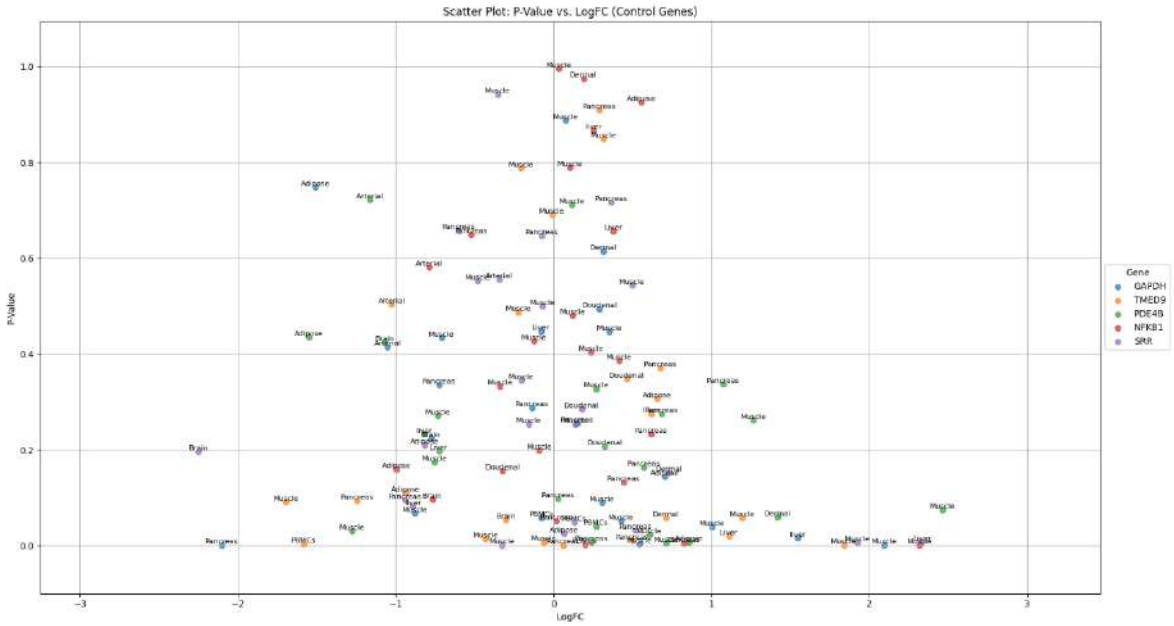
Scatter Plot by gene. Focused on genes of interest in T2D for 23 datasets combined.

Within this scatter plot, discernible clusters emerge for TMED9 exhibiting upregulation with moderate p-values, suggesting a pattern of expression for this gene across some datasets. Additionally, small clusters showcasing the downregulation of SRR with favorable p-values are evident. The clustering of these genes might be indicative of their potential biological significance or recurrent behavior in certain conditions. Contrarily, the absence of distinct clustering for other control genes aligns with our objective, highlighting the unpredictable nature of their expression. Notably, the reference gene, GAPDH, presents consistently good p-values across most studies. Its median value gravitates around 0 Fold Change, which is ideal for our analysis, attesting to its stability and appropriateness as a reference gene.

Scatter Plot Colored by Gene and Labeled by Tissue:

Building upon the previous visualizations, this scatter plot delves into the tissue-specific behavior of the genes. By labeling data points with the tissue

type, we sought to uncover any potential clustering patterns that might hint at tissue-dependent expression behaviors.



Scatter Plot by gene and tissue. Focused on genes of interest in T2D for 23 datasets combined.

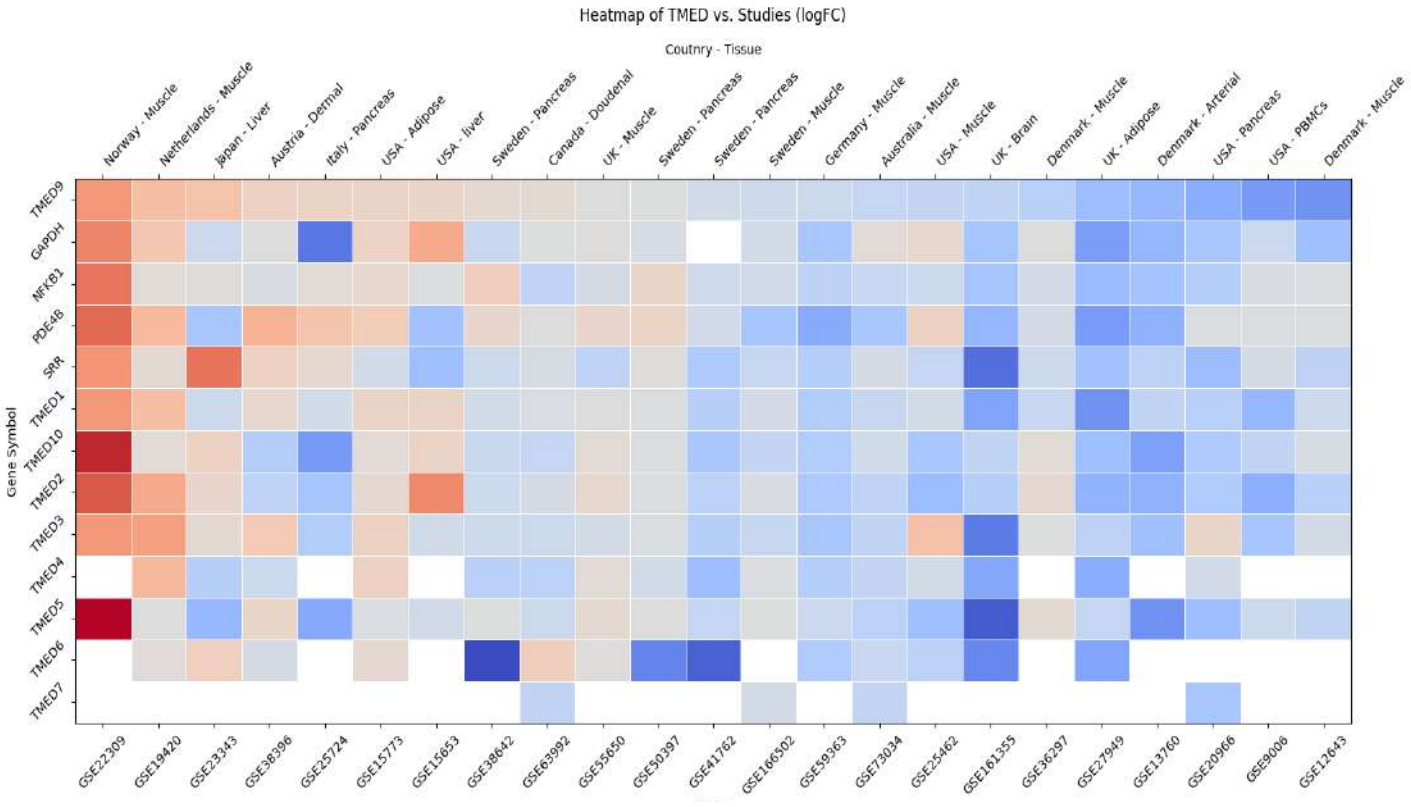
This scatter plot unfortunately did not present clear tissue-related clustering. Such an observation could indicate that the gene expression dynamics for the genes in question might be more universal across tissue types, or alternatively, that the variability across different studies might overshadow tissue-specific trends. This lack of clustering highlights the intricacies and challenges of achieving consistent gene expression results across various studies, which may be due to different experimental setups, tissue sources, or other variables inherent to each study.

It's noteworthy that the majority of absolute LogFC values fall below 1, pointing towards a subdued fold change in expression. The wide disparity in p-values across the datasets might signify the inherent variability in the studies or could result from factors like experimental design, or data processing variations.

Given the mixed observations from the scatter plots, we further refined our analysis by constructing a heatmap of the FC values.

Heatmap

Each cell color-codes the mean normalized log-fold change (logFC) of gene expression values across the studies. The genes are arrayed along the y-axis, with TMED9 listed at the top, followed immediately by reference and control genes, and then by the rest of the TMED family. Along the x-axis, studies are displayed, ordered based on the expression values of TMED9. This ordering provides a visual reference point to identify studies where TMED9 expression is particularly high (or low). Additionally, above the heatmap, corresponding country and tissue associated with each study are labeled, offering further insight into the biological context of each study.



Heatmap of the FC values for genes of interest sorted by descending TMED9 values. Labeled by study, country, and tissue in T2D for 23 datasets combined.

In analyzing the heatmap, the first point of interest is the apparent incoherence in fold change values for all genes, including the control genes, across various studies. The data does not display a distinct horizontal gradient in terms of gene expression, but there is a vertical pattern across the studies. TMED7 stands out distinctly, showing no variance in expression

across all studies, as indicated by its consistent white coloration, denoting a 0 fold change. When we consider the country of origin and tissue type associated with each study, there appears to be no discernible pattern relating to TMED9's expression. The distribution seems random, suggesting that these factors don't significantly influence TMED9's expression in the context of T2D. Crucially, both the control genes and TMED9, along with TMED1 and TMED3, seem to exhibit a moderate degree of parallelism in their expression patterns across studies. In stark contrast, the reference genes and other TMED family members do not align with the expression patterns of TMED9. This disparity is particularly intriguing as we investigate the potential significance of TMED9 . Following these observations, our next step will be to delve deeper into the relationship between TMED9 and other genes in our study using correlation tests and regression analyses.

Correlation and Regression Methods

Methods

Primary Comparisons: The first tier of our analysis involved a comprehensive comparison of TMED9 against each of the control genes. Both correlation and regression methodologies were used:

1. Correlation: The Pearson correlation was employed to decipher the linear relationship between TMED9 and each control gene. This statistic helped to discern the strength and direction of any potential associations.
2. Regression: Further, to delve deeper into the potential predictive or causal relationships between these genes, Ordinary Least Squares (OLS) regression was applied. This quantitative approach aimed to explain any variations in the control genes based on TMED9 expression.

Establishing Baselines – Negative Control Comparisons: Recognizing the importance of benchmarking our findings, additional comparisons were made:

- GAPDH and Control Genes: We subjected GAPDH to the same correlation and regression analyses with each of the control genes. Given GAPDH's established role as a housekeeping gene, these analyses served as a form of 'negative control' to set a baseline for our expectations.
- TMED9 and GAPDH: Furthermore, a direct comparison between TMED9 and GAPDH was also conducted to provide context to the other results.

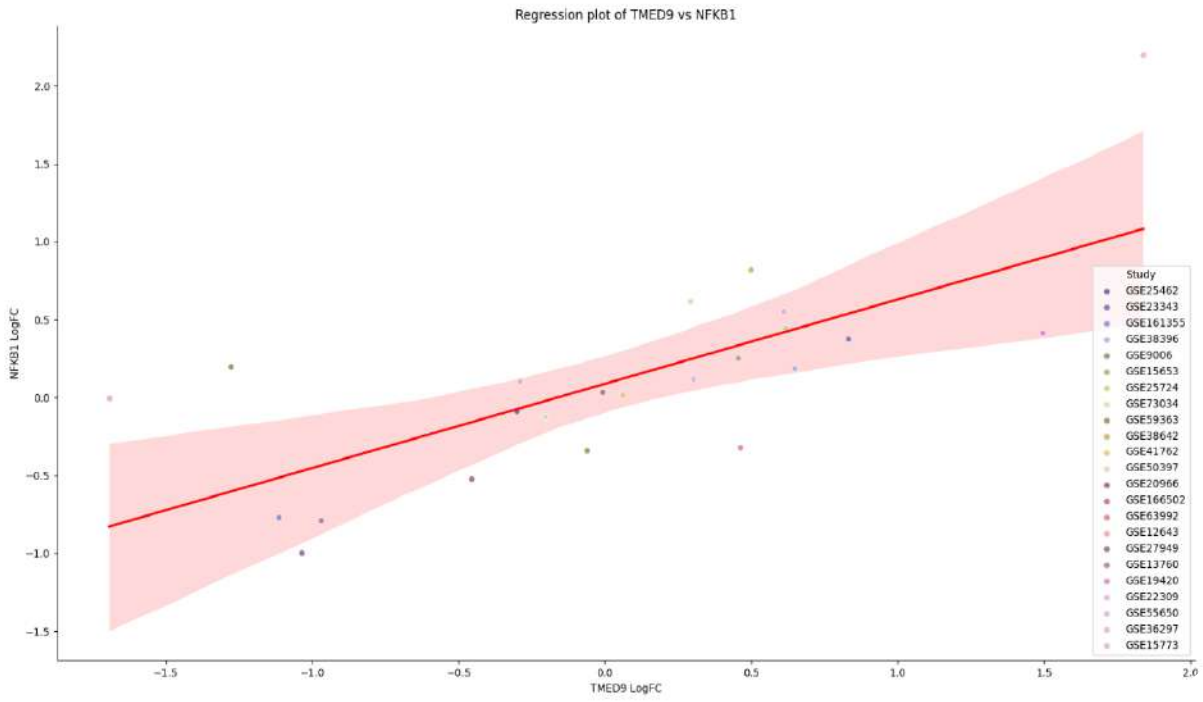
Results by Gene

NFKB1

1. Primary Comparisons Results to TMED9

Correlation: coefficient between TMED9 and NFKB1 is r = 0.7183, with a statistically significant p-value of p=0.000113. This strong positive correlation suggests a closely intertwined relationship between the expressions of TMED9 and NFKB1.

Regression: The R-squared value derived from the regression analysis stands at 0.516, indicating that approximately 51.6% of the variability in the NFKB1 gene's expression can be attributed to changes in the TMED9 gene's expression. Regression Plot:



Regression Scatter Plot for NFKB1 vs. TMED9 in T2D for 23 datasets combined.

On the x-axis, we have the normalized expression values for TMED9, and on the y-axis, those of the other gene – in this case NFKB1. Each study's data points are color-coded to distinguish between studies. A red line represents the general trend or regression between TMED9 and the gene of interest. The shaded pink area surrounding the red line indicates regions of statistical significance, with a smaller p-value suggesting a more robust relationship between the two genes.

Given that the relationships between TMED9 and all three control genes exhibited similar patterns, we've chosen to display only this representative plot for clarity, as the other two would not provide additional insights.

2. Negative Control to GAPDH

Correlation: The correlation between GAPDH and NFKB1 is 0.195, suggesting a very weak positive relationship. P-value: The p-value is 0.373, indicating that the correlation is not statistically significant at the usual significance level (e.g., 0.05).

Regression analysis: R-squared: The R-squared value is 0.038, meaning that only 3.8% of the variability in the dependent variable (presumably gene expression levels) can be explained by GAPDH.

SRR

1. Primary Comparisons Results to TMED9

Correlation: There's a positive correlation with TMED9 as indicated by the coefficient r=0.7420 and a highly significant p-value of p=5.06×10^−5.

Regression: The R-squared value for SRR stands at 0.551. This means that about 55.1% of the variability in the expression levels of the SRR gene can be explained by TMED9.

2. Negative Control to GAPDH

Correlation: The correlation between GAPDH and SRR is 0.235, suggesting a weak positive relationship. P-value: The p-value is 0.281, indicating that this correlation is not statistically significant.

Regression analysis: R-squared: The R-squared value is 0.055, indicating that only 5.5% of the variability in the dependent variable can be explained by GAPDH.

PDE4B

1. Primary Comparisons Results to TMED9

Correlation: TMED9's relationship with PDE4B is also positive as depicted by a correlation coefficient of r=0.6698, backed by a significant p-value of p=0.000472. Though this relationship is slightly weaker compared to the other genes, it still represents a meaningful connection between the two genes.

Regression: From the regression analysis, an R-squared value of 0.449 was obtained for PDE4B. This suggests that roughly 44.9% of the variability in PDE4B's expression is dictated by TMED9. The positive x1 coefficient (0.7278) points to an analogous trend where rising expression levels of TMED9 coincide with a proportional increase in PDE4B.

2. Negative Control to GAPDH

Correlation: The correlation between GAPDH and PDE4B is 0.241, suggesting a weak positive relationship. P-value: The p-value is 0.268, suggesting that the correlation is not statistically significant.

Regression: R-squared: The R-squared value is 0.058, suggesting that only 5.8% of the variability in the dependent variable can be attributed to GAPDH.

TMED9 vs GAPDH

Correlation: The correlation between GAPDH and TMED9 is 0.435, suggesting a moderate positive relationship. The p-value is 0.0382, indicating that this correlation is statistically significant at the 0.05 level.

Regression analysis: R-squared: The R-squared value is 0.189, meaning that 18.9% of the variability in the dependent variable can be explained by GAPDH.

Conclusion

In the context of the study, the relationships between GAPDH, commonly referenced as a housekeeping gene, and the control genes NFKB1, SRR, and PDE4B are tenuous and lack statistical significance. This portrays the expected dynamics between a housekeeping gene like GAPDH and other non-specialized genes. In stark contrast, TMED9 demonstrates notable correlations with these control genes, with SRR leading in strength, followed by NFKB1, and then PDE4B.

Moreover, while TMED9 exhibits a moderate and statistically significant association with GAPDH, the R-squared value of this relationship remains below 20%. This suggests an association between the two, yet it’s not as pronounced as TMED9's correlations with the control genes. The distinctiveness of these associations implies that TMED9 behaves differently from a conventional housekeeping gene, suggesting a potentially more specialized role for TMED9 within this context. The evidence from the analysis suggests an apparent correlation between TMED9 and the control genes, underscoring the hypothesis of a pronounced connection between TMED9 and the aforementioned significant control genes. Taken together, this collective evidence could imply a potential impact that TMED9 has on the diabetic disease state, aligning with the overarching goal of indicating TMED9's involvement in T2D.

Bibliography

- Dvela-Levitt, M., & Kost-Alimova, M. (2019). Small Molecule Targets TMED9 and Promotes Lysosomal Degradation to Reverse Proteinopathy. *Cell*, Volume 178(3), 521-535.e23. https://www.sciencedirect.com/science/article/pii/S009286741930741X

- Raza, W., Guo, J., Qadir, M. I., Bai, B., & Muhammad, S. A. (2022). qPCR Analysis Reveals Association of Differential Expression of SRR, NFKB1, and PDE4B Genes With Type 2 Diabetes Mellitus. *Frontiers in Endocrinology*, 12. https://www.frontiersin.org/articles/10.3389/fendo.2021.774696/full

- Aber, R., Chan, W., Mugisha, S., & Jerome-Majewska, L. A. (2019). Transmembrane emp24 domain proteins in development and disease. *Genet Res (Camb)*, 101, e14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7045115/

- Rau, A., Marot, G., & Jaffrézic, F. (2014). Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics*, 15, 91. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-91