# Personal Pathway Analysis as a means to integrate metabolomic and transcriptomic data in T2D intervention studies.

REVANT GUPTA

Master of Science in Molecular Techniques in Life Science

Date: May 31, 2018
Supervisor: Lukas Käll
Science for Life Laboratory, Stockholm

# Abstract

Expression *omics* data can be analysed for differential expression with respect to biological conditions like disease status. An intuitive method to understand the functional implication of differential expression is pathway analysis. Pathways are networks of interacting biomolecules and represent discrete biological functions. Current methods of pathway analysis are based on the relative difference in expression of either genes or metabolites between biological states. These methods are classified as over-representation and functional class scoring [1]. A widely used method is gene set enrichment analysis. This analysis when, applied to metabolites is called metabolite set enrichment analysis. The three main drawbacks of these methods are, first, they are unable to quantify differential pathway regulation at a sample level as they are based on summary statistics (for example, mean). It is not possible to explore individual biological variation between samples. Second, they assume equal contribution of biomolecular expression to pathway regulation. Third, they have not been applied in the analysis of integrated expression data. A method to perform personal pathway analysis using concatenated multi-*omics* expression data was developed to address these shortcomings. Co-variance of molecules in each pathway was modelled by factor analysis with one factor representing the pathway. Factor scores obtained, quantified pathway regulation at the sample level. The contribution of each molecule to pathway regulation was quantified by the loadings. The method was used to analyse transcriptomic and metabolomic expression of skeletal muscle tissue from a type II diabetes (T2D) study. Differential pathway regulation analysis was performed using quantified pathway regulation values and relevant clinical variables. The analysis identified several differentially regulated pathways that are linked to T2D pathology of skeletal muscle tissue. These included *Defects in biotin (Btn) metabolism* and *Branched-chain amino acid catabolism*, among others. Results were compared with IMPaLA, which is a tool for p-value integration of gene and metabolite set enrichment analyses. Dynamic visualisation dashboards are provided to explore biological heterogeneity of individuals' pathway regulation and contribution of biomolecules to pathway regulation. This analysis advances personalised *omics* research and can potentially be adapted to analysis of single cell expression.

# Contents

**1 Introduction** **1**

**2 Methodology** **6**

**3 Results** **9**

3.1 Data description and processing . . . . . . . . . . . . . . . . . . 9

3.2 Differential pathway regulation analysis . . . . . . . . . . . . . . 12

3.3 Comparison with IMPaLA . . . . . . . . . . . . . . . . . . . . . . 15

**4 Discussion** **18**

4.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

4.2 Future work . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

4.3 Ethical reflection . . . . . . . . . . . . . . . . . . . . . . . . . 23

4.4 Acknowledgement . . . . . . . . . . . . . . . . . . . . . . . . . 24

**Bibliography** **25**

**A Appendix** **30**

**B Computing environment** **36**

# 1. Introduction

Differential expression (DE) analysis is used to compare biomolecular expression between samples. DE analysis on *omics* expression datasets is based on multiple samples from different phenotypes and, expression values from a large number of biomolecules. Therefore, it can identify several biomolecules with altered expression between phenotypes. This is the fundamental to understanding biological mechanisms and their broader effects. However, DE analysis of biomolecules, between biological states, cannot be comprehensively linked to broader functional changes, without understanding the biological interactions they participate in. The ability to gather mechanistic insight from biomolecular expression has tremendous application, for instance in studying disease, biological variation between gender & age groups and identifying targets for therapy.

Pathways are networks of interacting biomolecules and represent discrete biological functions. Pathway databases are used to define pathways. These databases offer a high degree of coverage and categorisation of biological functions. They aggregate published information, are curated and, updated regularly [2]. Several open access pathway databases exist. Examples of commonly used databases include the Kyoto encyclopedia of genes and genomes (KEGG), Reactome, WikiPathways etc.

Pathway-based DE analysis is an intuitive method to investigate changes in biological functions due to differential expression of biomolecules. Pathway-based differential expression analysis has been categorised into, over-representation analysis (ORA), functional class scoring (FCS) and pathway topology-based analysis. For ORA and FCS methods, pathways are equivalent to the sets of biomolecules associated with them and topology information is discarded [1].

In ORA, differential expression of biomolecules is analysed. Biomolecules are classified as differentially expressed based on a significance threshold. Pathways are considered differentially regulated if the number of differentially expressed biomolecules is greater than expected by random. Most commonly, either Fisher's exact test or the hyper-geometric test is used to calculate the significance of overlap between biomolecules in the pathway, and differentially expressed biomolecules [1].

Functional class scoring methods do not require a threshold for differential expression. Unlike ORA, all biomolecules in the analysis are ranked by differential expression between phenotypes. The ends of this list represent differentially expressed biomolecules, which are either up or down-regulated between phenotypes. Gene set enrichment analysis (GSEA) is a widely used FCS method for differential gene expression analysis at a functional level rather than gene level. In this method, an enrichment score for each pathway is calculated. In high scoring pathways, associated genes are enriched at the ends of the ranked list of differentially expressed genes. Significance values for each pathway are calculated relative to a null model obtained by permuting phenotype labels [3].

Another set enrichment analysis methodology is WGSEA, based on the Wilcoxon signed rank test [4]. Mean expression values, correlation coefficients or fold changes of each gene for each phenotype category can be used as input. For each pathway, genes are ranked on differential expression. The distribution of ranks is used to calculate the significance of differential pathway regulation with the null model that the rank distribution is symmetric around zero.

Other variants of Set enrichment analysis (SEA) also exist [5]. These analyses are generally performed on one type of expression dataset and thus SEA on metabolites is referred to as metabolic set enrichment analysis. In a study by Cavill et al. [6] metabolomic and transcriptomic data from the NCI60 cell line panel was analysed using WGSEA. A method to integrate transcriptomic and metabolomic analyses was proposed. SEA was performed on transcriptomic and metabolomic data separately. The two sets of p-values were combined to obtain joint p-values. Comparison of the integrated analysis to separate transcriptomic and metabolomic analyses found pathways that were significantly differentially regulated only in the joint analysis.

A web server application of this methodology called Integrated Molecular Pathway Level Analysis (IMPaLA) is available. IMPaLA uses Fisher's method to combine p-values per pathway for the transcriptomic and metabolomic sets [7]. While this is an obvious step forward from separate analyses of different types of expression data, IMPaLA only integrates the p-values and not the data itself.

Integration of biological data of different types and sources allows more complex analyses of biological interactions and functions. However, it is a challenging task. The variety of techniques and experiments used to generate biological data, lead to differences in scale and distribution. Furthermore, for certain datasets, it is not possible to establish relationships based on biological knowledge. Integration of transcriptomic and metabolomic data is an example. While protein and transcript levels can be attributed to gene expression, metabolites levels are affected by more complex interactions that cannot be modelled by a linear flow from genes to phenotype.

An existing method of pathway-based differential expression analysis that also integrates data is PAthway Representation and Analysis by Direct Reference on Graphical Models (PARADIGM). This method utilises pathway topology. It integrates data at the gene level by means of a probabilistic factor graph. While sophisticated, this is a complex model that requires quantifying gene level processes that affect expression levels. For example, the model can use copy number variation data to infer if over-expression results from amplification or differential expression of an upstream promoter [8]. PARADIGM reports individual pathway activity scores for each sample which quantify pathway regulation.

This report presents a simple and flexible method that performs pathway-based multi-*omics* expression data integration. The main features of this analysis address the shortcomings of currently used pathway analysis methods like SEA. Data is integrated by modelling the co-variance between expression values of individual molecules. For this confirmatory factor analysis is performed on concatenated expression data for each pathway, with one latent variable representing the pathway. Hereafter this method is referred to as Integrated pathway analysis (IPaA). IPaA does not focus on modelling the pathway network itself. The focus of the method is on the integration of expression data to quantify pathway regulation at an individual level.

IPaA has three notable features, first is the ability to quantify pathway regulation at a sample level. The continuous values obtained are more informative than categorical labels (i.e. up or down-regulated). These values can be correlated to clinical variables. Continuous clinical variables themselves are more informative than labels like disease status since they represent the biology of the individual. In contrast, labels like disease status are subjective interpretations based on biological measurements.

Statistical association of pathway regulation, quantified at a sample level, with clinical variables allows the investigation of phenotype variation at an individual, personalised level. This makes it possible to explore the biological variance of pathways or identify individuals with aberrant biological behaviour. This information is lost while using labels like disease status, as is the case in set enrichment, since information of individuals is collapsed into a summary statistic, usually the mean.

Second, not all biomolecules contribute equally to pathway regulation. Current methods are unable to identify the importance of biomolecules in a pathway. While there have been some attempts to use thresholds based on differential expression, the selected subset's biomolecules are still treated as equal [3]. The factor analysis model used in this method provides each biomolecule's contribution to pathway regulation. This is based on the amount of each molecule's co-variance explained by the latent variable representing the pathway.

Third, this method offers true integration of expression data from multiple types of *omics* expression datasets. All biomolecules are considered together when analysing pathway regulation, therefore, the statistical power of the analysis is increased compared to separate analyses of different types of expression datasets. More importantly, biomolecular interactions in a pathway are independent of categorisation into transcripts and metabolites. IPaA takes into account the interactions between biomolecules and so the information in integrated data analysis cannot be obtained from analysing the expression datasets separately.

Transcriptomic and metabolomic expression data of skeletal muscle tissue from the thigh was analysed using IPaA. The samples were taken from patients and controls participating in a balanced type II diabetes study. Identifying functional differences between biological states is the sole focus of most commonly used differential expression analysis methods. Comprehensive *omics* expression data from a high number of samples has been used to support the findings. However, these datasets present an opportunity to understand an important, yet poorly investigated aspect of differential expression, *biological heterogeneity*.

It is widely believed that personalised *omics* profiles will contribute to the development of personalised healthcare. This has been prompted by the influence of genetic and environmental heterogeneity in the development and progression of disease. Diabetes is a highly heterogeneous disease that is influenced by a variety of genetic and environmental factors [9].

While extensive research into the molecular mechanisms and risk factors for type II diabetes has been published, translating the knowledge gained in the lab to the clinic is a pressing challenge [10].

Since IPaA quantifies pathway regulation at a sample level, it allows individual subjects to be tracked across pathways. Clinical variables, like fasting plasma glucose, which represent individual biological conditions are used to assess differential regulation. So the overall biological state of each individual can be explored while maintaining the statistical significance of differential regulation.

# 2. Methodology

Confirmatory Factor Analysis (CFA) is a matrix decomposition method that is related to the widely used Principal Component Analysis (PCA). In PCA, components are linear combinations of variables. The first component accounts for maximum variance observed in the data and so on.

In CFA, original variables of the data are expressed as linear combinations of the components. Also, CFA models the covariance rather than the variance. The first CFA component represents the maximum covariance of the observed variables in the data. CFA components are also orthogonal and are referred to as factors. An obvious assumption is that the observed variables must be linearly correlated to some extent [11].

Factors are also called latent variables when FA is used to quantify non-observable, often abstract concepts. For example, CFA models are used extensively in psychology to test for abstract concepts like intelligence. The observed data, in this case, could be the scores on standardised tests. It is assumed that that observed performance is driven by the unobserved variable i.e. intelligence.

In this analysis, the collective expression of biomolecules in a pathway is assumed to be responsible for the function represented by the pathway. Thus, the latent variable in this analysis is the unobserved biological function which cannot be directly measured as a quantity.

For a given data matrix *X* containing p biomolecules' expression values as variables and n samples, the data matrix is decomposed in the following way,

$$X - \mu = LF + \epsilon$$

where, $\mu$ is a p × 1 matrix containing the means of each variable, *L* is the loading matrix of shape p × k, F are the factor scores of shape k × n and $\epsilon$ is the noise matrix of shape p × n.

The loading matrix *L* contains loadings that reflect the amount of covariance explained by each factor for each variable.

The noise matrix $\epsilon$ contains residuals that are also called unique factors. The variance-covariance matrix of $\epsilon$,

$$\psi = \epsilon\epsilon^T$$

represents noise variance in the data. It is constrained to be a diagonal matrix. Thus, the entire covariance in the data is modelled in *LF*. In general for each individual variable $x_p$ and $l_{pk} \in L$

$$x_p - \mu_p = l_{p1}F_1 + \cdots + l_{pk}F_k + \varepsilon_p.$$

where $F_1 ... F_k$ contain the factor scores for each latent variable. However, only one factor is considered, so the expression reduces to,

$$x_p - \mu_p = l_{p1}F_1 + \varepsilon_p.$$

In the context of a pathway, loadings ($l_{p1}$) can be interpreted as the importance or contribution of a biomolecule to the pathway's regulation.

For biomolecules' contribution to pathways, relative signs between the contributions are important. The sign itself is not relevant in this analysis as factor analysis suffers from sign indeterminacy. It is possible for signs of the loadings to reverse, though the relative signs between the loadings will remain the same. A negative loading, simply implies that the variable is inversely correlated to the latent variable.

The continuous latent variable values obtained, quantify pathway regulation. These pathway regulation values are then used to explore associations between clinical variables and biological functions that the pathways represent.

Pathway regulation values obtained are analogous to expression values. However, by itself, pathway regulation cannot be used to ascertain whether a pathway is up-regulated or down-regulated as is the case with expression data. However, when the regulation values are placed in context with the loadings and expression values of associated biomolecules, it is possible to infer the direction of pathway regulation.

There are several estimation methods that can be used to build this model. The `sklearn.decomposition.FactorAnalysis` function from the python package scikit-learn was used to perform factor analysis [12]. In this implementation, the maximum likelihood method is used to estimate the loading matrix *L* by expectation-maximization with the constraint that $\psi$ be a diagonal matrix [13].

Analysis of co-variance (ANCOVA) test is used to assess the significance of differential regulation of each pathway model. Pathway regulation is the dependent variable. Clinical variables, both continuous and categorical, can be used as independent variables in this test. If only categorical variables are used then the test would be ANOVA. The test was implemented using the `statsmodels.stats.anova.anova_lm` function from the python package statsmodels [14].

Blocking variables are used to account for the variance introduced due to nuisance variables that are not relevant to the analysis. The study described in the results section involved the collection of samples over a time interval during which an intervention was administered. The time of sample collection is included as a blocking variable to account for variance introduced by it.

Since ANCOVA tests are performed for each pathway in the analysis, the p-values obtained need to be corrected for multiple hypothesis testing. Therefore, q-values are calculated using ranked p-values as per the procedure described by Storey and Tibshirani [15].

# 3. Results

## 3.1 Data description and processing

The data consisted of gene expression (transcriptomic), metabolomic and clinical profile data from 49 males of which 25 are diagnosed type II diabetics (T2D) and 24 are normal glucose tolerant (NGT). The subjects were administered a hyperinsulinemic-euglycemic clamp which is considered a gold standard measurement for insulin sensitivity [16].

Muscle biopsies from the thigh were collected in a fasting state (basal) and 30 minutes (30-min) after the clamp was administered. Gene expression was measured with microarrays and metabolite expression with LC-MS.

Clinical data on the study participants includes basic information like age, anthropometrics and clinical variables. A summary of the clinical data is shown in table A.6. The expression data contains 20949 genes and 117 metabolites. Of these only 9006 genes and 17 metabolites were mapped to pathways in the analysis.

The factor model described above was run using four combinations of the two datasets, i.e. transcriptomic and metabolomic. The expression values were either $log_{10}$ transformed or untransformed. Figure 3.1 shows that using the $log_{10}$ transformed metabolomic expression values and untransformed transcriptomic values was appropriate.

Expression values were reconstructed using the loading matrix $L$ and factor matrix $F$. Residuals were the noise matrix $\epsilon$. Reconstructed expression values are clearly correlated with the residuals for untransformed metabolomic data. Log transformation of the data leads to an approximate transformation of a multiplicative model to an additive linear model [17]. The use of base 10 for the $log$ transformation ensured that the scales of the two sets were comparable.

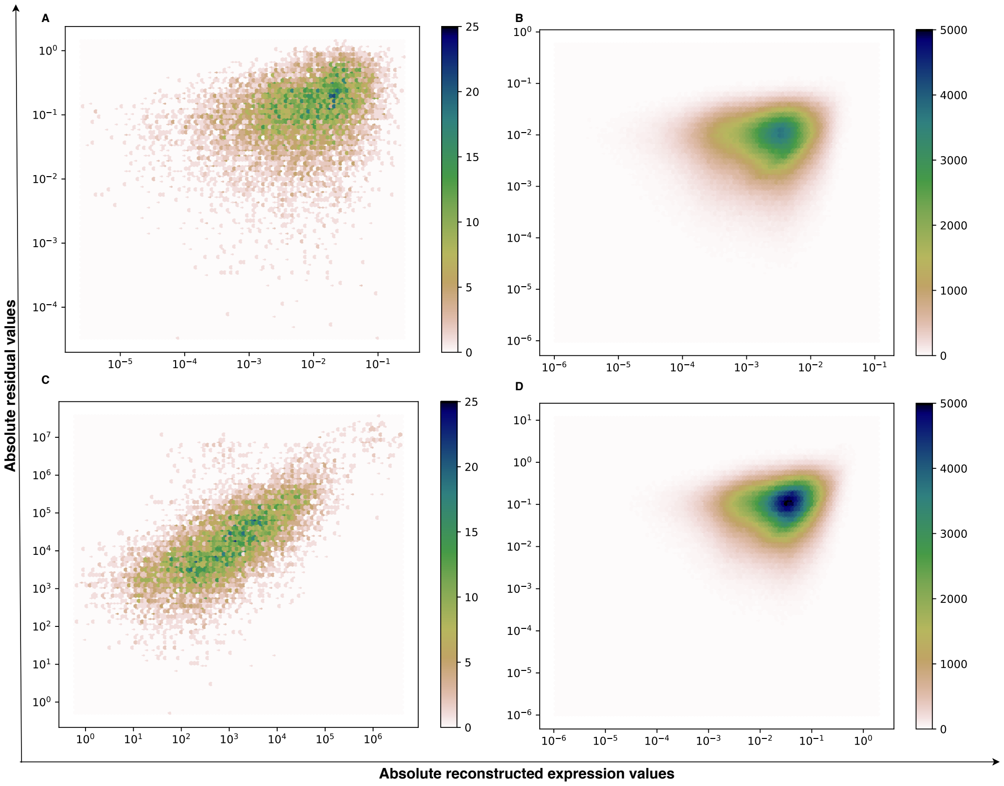

**Figure 3.1:** Hexbin 2D histograms for residuals vs reconstructed expression values from factor analysis model using **A.** $log_{10}$ transformed metabolomic data, **B.** $log_{10}$ transformed transcriptomic data, **C.** untransformed metabolomic data and **D.** untransformed transcriptomic data. Colourbar indicates the counts. Based on the distribution and scale in each histogram, the analysis was performed using $log_{10}$ transformed metabolomic and untransformed transcriptomic data.

Factor analysis models were constructed for each pathway using concatenated transcriptomic and $log_{10}$ transformed metabolomic data. Goodness of fit statistics for the pathway models are shown in figure A.2. Root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR) and explained variance-covariance ratio were reported [11]. Bartlett corrected likelihood ratio test was used to derive the RMSEA [18].

Factor analysis is quite sensitive to outliers in the data. Average hierarchical clustering using squared Euclidean pairwise distances was performed to identify any outlier samples [19]. Based on the results shown in figure 3.2 samples *37_basal, 37_30-min* and *577_30-min* were dropped from the analysis.

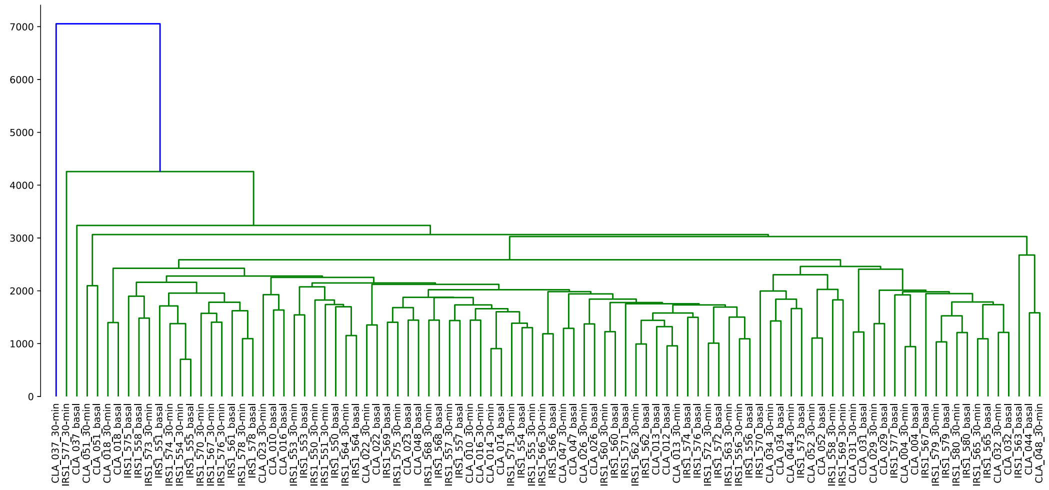

**Figure 3.2:** Average linkage clustering of samples in the integrated dataset using squared Euclidean pairwise distances. Based on this, samples *37_basal*, *37_30-min* and *577_30-min* were considered outliers and dropped from the analysis.

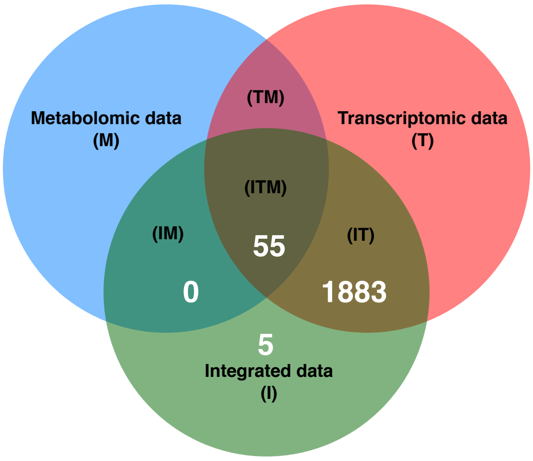

**Figure 3.3:** Pathways included in the analysis using integrated (I), transcriptomic (T) or metabolomic (M) data. A pathway is included in the analysis if two or more biomolecules are mapped to it from the expression data.

Pathway definitions from the current version of Reactome (v64) were used to generate sets for the analysis [20][21]. The database contains 2195 pathway definitions. Pathways that had at minimum two biomolecules mapped from the expression data, were considered in the analysis. Figure 3.3 shows the pathways included in the analysis for integrated dataset as well as separate analyses for the two expression datasets.

## 3.2 Differential pathway regulation analysis

Clinical diagnosis and monitoring of diabetes is based on clinical variables that assess glycaemia, glycaemic variability, insulin resistance and blood glucose levels. Compared to a binary label like disease status, these variables are more informative of the biological state of an individual. Differential pathway regulation was assessed with respect to the following clinical variables, M-index, glycated haemoglobin (HbA1c), homeostatic model assessment for insulin resistance (HOMA-IR), fasting glucose level and serum C peptide level.

It has been suggested that the risk and development of diabetes occur due to glucose fluctuations rather than sustained hyperglycaemia [22]. The M-index is calculated from multiple glucose measurements taken over a period of 24 hours. It was proposed by Schlichtkrull, Munck, and Jersild [23]. The M index is a measure of glycaemic variability and magnitude [24].

Glycated haemoglobin forms non-enzymatically due to elevated blood glucose level and accumulates over time. It is therefore used to assess the presence of elevated blood glucose levels till 3 months prior to the point of sample collection [25]. Fasting plasma glucose level along with glycated haemoglobin is an important parameter of glycaemic control [22].

C peptide is produced in equal amounts as insulin and thus is a good measure of insulin expression. Serum C peptide level is used to assess the endogenous expression of insulin in diabetics and aids in the classification of the diabetic state [26].

HOMA is a model proposed by Matthews et al. [27] based on a regulatory feedback loop between the liver and pancreatic $\beta$ cells. It is calculated using the fasting plasma glucose level and immuno-reactive insulin level. The feedback loop regulates protein synthesis, glucose storage and lipolysis in response to changing glucose levels. HOMA-IR is used to assess insulin sensitivity, though in diabetics with severe glycaemic dys-regulation it may not be as accurate [28].

| Pathways | M-index | HbA1c | HOMA-IR | S. C-peptide | F.P. glucose |
|---|---|---|---|---|---|
| Defects in biotin (Btn) metabolism | * | * | * | * | - |
| Hydrolysis of LPE | * | - | * | * | * |
| Defective HLCS causes multiple carboxylase deficiency | * | * | * | * | - |
| Branched-chain amino acid catabolism | * | - | * | * | - |
| Lysine catabolism | * | - | * | * | - |
| VEGF binds to VEGFR leading to receptor dimerization | * | - | * | * | - |
| Utilization of Ketone Bodies | * | - | * | * | - |
| Regulation of gene expression by Hypoxia-inducible Factor | * | - | * | * | - |
| Activation of PPARGC*A (PGC-*alpha) by phosphorylation | * | - | * | * | - |
| Amino acid transport across the plasma membrane | * | - | * | * | - |
| VEGF ligand-receptor interactions | * | - | * | * | - |
| Propionyl-CoA catabolism | * | * | - | - | - |
| Methylation | * | * | - | - | - |
| Glutathione synthesis and recycling | * | - | - | * | - |
| Calcineurin activates NFAT | * | - | - | * | - |
| Misspliced LRP5 mutants have enhanced beta-catenin-dependent signaling | - | - | * | * | - |
| CaMK IV-mediated phosphorylation of CREB | * | - | - | * | - |
| Amino acid synthesis and interconversion (transamination) | * | - | - | * | - |
| Defects in vitamin and cofactor metabolism | * | * | - | - | - |
| Sialic acid metabolism | * | - | - | * | - |
| Reduction of cytosolic Ca++ levels | * | - | - | - | - |
| Mitochondrial translation | * | - | - | - | - |
| Activation of Na-permeable kainate receptors | * | - | - | - | - |
| CHL* interactions | * | - | - | - | - |
| PI3K events in ERBB2 signaling | * | - | - | - | - |

**Table 3.1:** Differential regulation analysis was performed with respect to five clinical variables indicative of T2D status. The table shows 25 most frequently differentially regulated pathways considering all the analyses, at a q value threshold of 0.001. * indicates the occurrence of the pathway in the analysis at the given significance threshold.

Each clinical variable is indicative of a particular aspect of the diabetic state and all variables are indicative of disease status. This can be seen in figures 4.1 and A.1. Therefore, pathways that are significantly differentially regulated in multiple analyses are possibly more involved in the disease pathology. See table 3.1.

Pathways that are differentially regulated in each analysis are also relevant to the aspect of diabetes indicated by the particular clinical variable. Tables A.1-A.5 list the top differentially regulated pathways in each analysis.

Several pathways in table 3.1 are intimately connected to diabetes, hyperglycaemia, insulin resistance and other metabolic dys-regulation that occurs in the diabetic state. These include, *Utilization of Ketone Bodies, Branched-chain amino acid catabolism, Regulation of gene expression by Hypoxia-inducible Factor*. Pathways that reflect specific signalling and metabolic processes involved in the pathogenesis of type II diabetes in the list include, *Activation of PPARGC1A (PGC-1alpha) by phosphorylation, Defects in biotin (Btn) metabolism, Defects in vitamin and cofactor metabolism, VEGF ligand-receptor interactions, Propionyl-CoA catabolism, Hydrolysis of LPE* [29][30][31][32].

Skeletal muscle is overwhelmingly responsible for insulin-mediated glucose uptake, thus insulin resistance in skeletal muscle tissue plays a major role in the pathogenesis of type II diabetes. Type II diabetes is characterised by high plasma levels of free fatty acids driven by lipolysis in adipose tissue [33]. Skeletal muscle suffers from metabolic inflexibility, i.e. the ability to switch between fatty acid and carbohydrate oxidation due to disruption of insulin-mediated signalling networks. This can occur due to the accumulation of lipids inside muscle cells and decreased mitochondrial oxidative capacity [34][35]. The reduction in glucose uptake by skeletal muscle leads to hyperglycaemia.

The unavailability of glucose as an energy source prompts the up-regulation of ketone body production by the liver. Ketone bodies are used as energy source in peripheral tissue including skeletal muscle [36]. Obesity is a major risk factor for the development of diabetes and is characterised by metabolic disorders that overlap with diabetes. Genes, regulated by the hypoxia-inducible factor, are involved in regulating pro-inflammatory pathways and insulin signalling. Adipose tissue faces hypoxia in obesity. The oxidative stress and inflammation can lead to the development of diabetes [37].

PGC-1$\alpha$ is part of a family of transcriptional co-activators and is known to control cellular energy metabolism. Studies have shown that it is down-regulated in skeletal muscle of diabetic patients. PGC-1$\alpha$ in skeletal muscles promotes mitochondrial bio-genesis, remodelling of muscle fibres that enable fatty acid oxidation and expression of insulin-sensitive glucose transporter GLUT4. Mutant versions of the PGC-1$\alpha$ gene, that have lower expression are risk factors for type II diabetes [38]. Skeletal muscles are the primary tissue for glucose uptake after ingestion, it has been suggested that the depression of PGC-1$\alpha$ expression leads to mitochondrial insufficiency and reduced glucose uptake, both of which contribute to insulin resistance. However, it has also been suggested that in the case of a high fatty diet, insulin resistance may occur due to over-expression of PGC-$\alpha$ [39].

Increase in circulating branched chain amino acid (BCAA) levels have consistently been found to be predictive of insulin resistance development and diabetes [40]. It has been suggested that the accumulation of BCAA in plasma is driven by a decrease in BCAA catabolism in adipose tissue, which occurs due to an abundance of glucose thus removing the need for BCAA as an energy source. The excess BCAA enter catabolic pathways in skeletal muscle leading to the production of compounds like propionyl CoA and succinyl CoA, which cause insulin resistance by interfering with fatty acid oxidation [41].

## 3.3 Comparison with IMPaLA

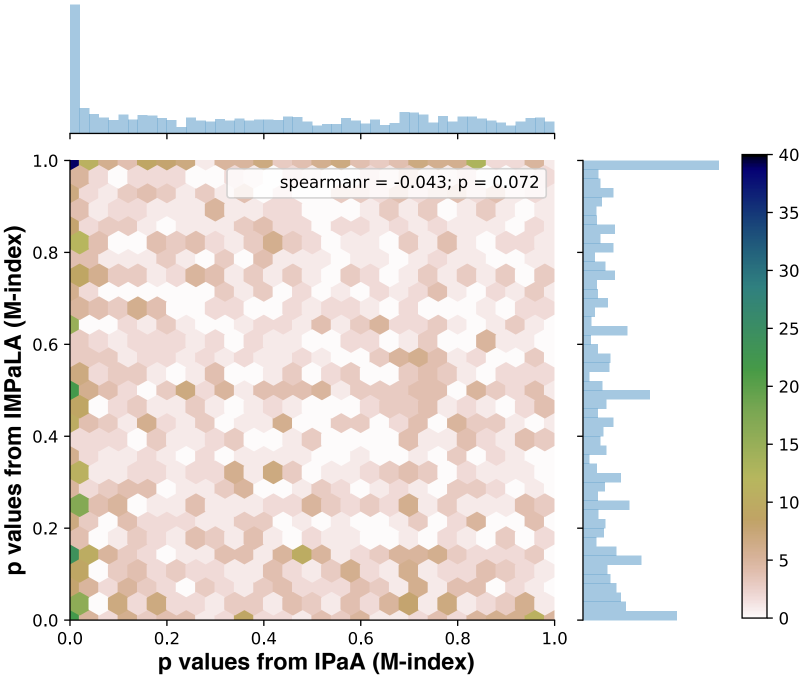

**Figure 3.4:** Hexbin 2D histogram comparing p-values for differential pathway regulation from IPaA w.rt. M-index with IMPaLA using Spearman correlation of expression values to M-index as input. Spearman correlation between the two p-values indicates that there is very low correlation between rankings of differentially regulated pathways from the analyses. Colourbar indicates the counts

IMPaLA is a web server tool that is able to perform over-representation and set enrichment analysis on transcriptomic or metabolomic data. The Wilcoxon signed rank test on gene or metabolite sets to calculate a p-value for differential pathway expression. It is also able to integrate the result of transcriptomic and metabolomic differential pathway analysis by combining the p-values from either analysis using Fisher's method. It relies upon pathway definitions from many pathway databases including KEGG, Reactome, WikiPathways and more. The definitions can be found at ConsensusPathDB which is a meta-database of pathway databases [7]. For this comparison, only pathway definitions from Reactome v64 were used.

| Pathway description | measured/actual | | Integrated analysis | | Transcriptomic analysis | | Metabolomic analysis | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Metabolism | 1627/1966 | 25/1384 | 2.84E-10 | 1.38E-07 | 2.07E-06 | 2.96E-03 | 5.25E-06 | 2.21E-02 |
| Initial triggering of complement | 62/106 | 0/14 | 1.68E-06 | 2.96E-03 | 1.68E-06 | 2.96E-03 | 1.00E+00 | 1.00E+00 |
| Immune System | 1457/1840 | 0/136 | 2.06E-06 | 2.96E-03 | 2.06E-06 | 2.96E-03 | 1.00E+00 | 1.00E+00 |
| Thyroid hormone signaling pathway - Homo sapiens (human) | 98/116 | 0/11 | 2.62E-06 | 2.96E-03 | 2.62E-06 | 2.96E-03 | 1.00E+00 | 1.00E+00 |
| Metabolism of amino acids and derivatives | 271/338 | 12/285 | 2.15E-05 | 3.48E-03 | 1.02E-03 | 7.80E-02 | 1.46E-03 | 1.00E+00 |
| Antigen activates B Cell Receptor (BCR) leading to generation of second messengers | 65/97 | 0/8 | 4.93E-06 | 3.53E-03 | 4.93E-06 | 3.53E-03 | 1.00E+00 | 1.00E+00 |
| Creation of C4 and C2 activators | 59/97 | 0/13 | 5.42E-06 | 3.53E-03 | 5.42E-06 | 3.53E-03 | 1.00E+00 | 1.00E+00 |
| Cytoplasmic Ribosomal Proteins | 73/88 | 0/0 | 5.46E-06 | 3.53E-03 | 5.46E-06 | 3.53E-03 | 1.00E+00 | 1.00E+00 |
| Adaptive Immune System | 564/732 | 0/54 | 9.15E-06 | 4.60E-03 | 9.15E-06 | 4.60E-03 | 1.00E+00 | 1.00E+00 |
| Innate Immune System | 839/1077 | 0/99 | 1.85E-05 | 8.40E-03 | 1.85E-05 | 8.40E-03 | 1.00E+00 | 1.00E+00 |
| Classical antibody-mediated complement activation | 52/89 | 0/6 | 2.34E-05 | 9.61E-03 | 2.34E-05 | 9.61E-03 | 1.00E+00 | 1.00E+00 |
| Ribosome - Homo sapiens (human) | 117/154 | 0/0 | 2.95E-05 | 9.99E-03 | 2.95E-05 | 9.99E-03 | 1.00E+00 | 1.00E+00 |
| Eukaryotic Translation Termination | 80/104 | 0/10 | 2.98E-05 | 9.99E-03 | 2.98E-05 | 9.99E-03 | 1.00E+00 | 1.00E+00 |
| Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC) | 81/106 | 0/2 | 3.09E-05 | 9.99E-03 | 3.09E-05 | 9.99E-03 | 1.00E+00 | 1.00E+00 |
| FCERI mediated MAPK activation | 70/109 | 0/5 | 4.30E-05 | 1.28E-02 | 4.30E-05 | 1.28E-02 | 1.00E+00 | 1.00E+00 |
| Formation of a pool of free 40S subunits | 88/115 | 0/0 | 4.51E-05 | 1.28E-02 | 4.51E-05 | 1.28E-02 | 1.00E+00 | 1.00E+00 |
| Transcriptional regulation by RUNX1 | 158/198 | 0/9 | 5.72E-05 | 1.44E-02 | 5.72E-05 | 1.44E-02 | 1.00E+00 | 1.00E+00 |
| Regulation of actin dynamics for phagocytic cup formation | 97/141 | 0/12 | 6.02E-05 | 1.44E-02 | 6.02E-05 | 1.44E-02 | 1.00E+00 | 1.00E+00 |
| Selenocysteine synthesis | 81/104 | 0/20 | 6.03E-05 | 1.44E-02 | 6.03E-05 | 1.44E-02 | 1.00E+00 | 1.00E+00 |
| Hemostasis | 544/668 | 0/67 | 6.97E-05 | 1.58E-02 | 6.97E-05 | 1.58E-02 | 1.00E+00 | 1.00E+00 |
| Signaling by the B Cell Receptor (BCR) | 93/127 | 0/14 | 7.80E-05 | 1.68E-02 | 7.80E-05 | 1.68E-02 | 1.00E+00 | 1.00E+00 |
| Peptide chain elongation | 77/101 | 0/25 | 1.14E-04 | 2.34E-02 | 1.14E-04 | 2.34E-02 | 1.00E+00 | 1.00E+00 |
| L13a-mediated translational silencing of Ceruloplasmin expression | 95/125 | 0/1 | 1.47E-04 | 2.90E-02 | 1.47E-04 | 2.90E-02 | 1.00E+00 | 1.00E+00 |
| Fcgamma receptor (FCGR) dependent phagocytosis | 111/159 | 0/20 | 1.54E-04 | 2.90E-02 | 1.54E-04 | 2.90E-02 | 1.00E+00 | 1.00E+00 |
| Eukaryotic Translation Elongation | 78/106 | 0/25 | 1.70E-04 | 3.06E-02 | 1.70E-04 | 3.06E-02 | 1.00E+00 | 1.00E+00 |
| FCERI mediated Ca+2 mobilization | 72/108 | 0/12 | 1.76E-04 | 3.06E-02 | 1.76E-04 | 3.06E-02 | 1.00E+00 | 1.00E+00 |
| Regulation of actin cytoskeleton - Homo sapiens (human) | 172/208 | 0/6 | 2.06E-04 | 3.45E-02 | 2.06E-04 | 3.45E-02 | 1.00E+00 | 1.00E+00 |
| Response to metal ions | 13/15 | 0/4 | 2.44E-04 | 3.95E-02 | 2.44E-04 | 3.95E-02 | 1.00E+00 | 1.00E+00 |
| TYROBP Causal Network | 53/60 | 0/0 | 2.56E-04 | 4.00E-02 | 2.56E-04 | 4.00E-02 | 1.00E+00 | 1.00E+00 |
| RNA Polymerase I Promoter Opening | 51/65 | 0/2 | 2.71E-04 | 4.09E-02 | 2.71E-04 | 4.09E-02 | 1.00E+00 | 1.00E+00 |

**Table 3.2:** Top thirty differentially regulated pathways from IMPaLA with respect to T2D status. The analysis was performed using Spearman correlation of expression values with the M-index .

First, Spearman correlations of each biomolecule to M-index, for T2D and NGT groups were used as input for IMPaLA. In figure 3.4 p-values obtained were compared with p-values from IPaA with M-index. There was a complete lack of commonality between the pathways considered differentially regulated between the analyses.

Differentially regulated pathways from IMPaLA as shown in table 3.2 did not contain any pathways that represent specific pathogenic changes in type II diabetes. Rather the pathways, in general, did not appear to be specifically linked to diabetes or hyperglycaemia.

A second IMPaLA was performed using expression values as used in IPaA, i.e. untransformed transcriptomic and $log_{10}$ transformed metabolomic data. The pathways considered differentially regulated in this analysis are shown in table 3.3. Several pathways in this list including *Metabolism, Translation and Oxidative phosphorylation* can be considered relevant in the diabetic state. However, these pathways are also expected to be perturbed in a variety of disease conditions and do not appear to represent any specific pathological change due to diabetes.

| Pathway description | measured/actual |  | Integrated analysis |  | Transcriptomic analysis |  | Metabolomic analysis |  |
|---|---|---|---|---|---|---|---|---|
|  | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Metabolism of proteins | 1652/2008 | 5/269 | 2.52E-39 | 1.22E-36 | 8.53E-41 | 3.86E-37 | 3.12E-01 | 1.00E+00 |
| Metabolism | 1627/1966 | 25/1384 | 7.70E-32 | 1.87E-29 | 1.33E-33 | 3.01E-30 | 7.51E-01 | 1.00E+00 |
| Translation | 252/310 | 5/78 | 5.14E-29 | 8.33E-27 | 2.34E-30 | 3.53E-27 | 3.12E-01 | 1.00E+00 |
| The citric acid (TCA) cycle and respiratory electron transport | 137/173 | 0/56 | 2.92E-18 | 3.31E-15 | 2.92E-18 | 3.31E-15 | 1.00E+00 | 1.00E+00 |
| Parkinson_s disease - Homo sapiens (human) | 109/142 | 0/15 | 6.91E-16 | 6.26E-13 | 6.91E-16 | 6.26E-13 | 1.00E+00 | 1.00E+00 |
| Mitochondrial translation | 87/95 | 0/47 | 1.46E-15 | 1.10E-12 | 1.46E-15 | 1.10E-12 | 1.00E+00 | 1.00E+00 |
| Oxidative phosphorylation - Homo sapiens (human) | 100/133 | 0/16 | 2.07E-15 | 1.18E-12 | 2.07E-15 | 1.18E-12 | 1.00E+00 | 1.00E+00 |
| Ribosome - Homo sapiens (human) | 117/154 | 0/0 | 2.08E-15 | 1.18E-12 | 2.08E-15 | 1.18E-12 | 1.00E+00 | 1.00E+00 |
| Huntington_s disease - Homo sapiens (human) | 158/193 | 0/3 | 3.19E-15 | 1.61E-12 | 3.19E-15 | 1.61E-12 | 1.00E+00 | 1.00E+00 |
| Metabolism of amino acids and derivatives | 271/338 | 12/285 | 1.10E-14 | 1.34E-12 | 4.68E-15 | 2.12E-12 | 6.40E-02 | 1.00E+00 |
| Mitochondrial translation termination | 81/89 | 0/25 | 1.33E-14 | 5.02E-12 | 1.33E-14 | 5.02E-12 | 1.00E+00 | 1.00E+00 |
| Mitochondrial translation elongation | 81/89 | 0/26 | 1.33E-14 | 5.02E-12 | 1.33E-14 | 5.02E-12 | 1.00E+00 | 1.00E+00 |
| Mitochondrial translation initiation | 81/89 | 0/8 | 1.54E-14 | 5.25E-12 | 1.54E-14 | 5.25E-12 | 1.00E+00 | 1.00E+00 |
| Thermogenesis - Homo sapiens (human) | 182/229 | 0/25 | 1.62E-14 | 5.25E-12 | 1.62E-14 | 5.25E-12 | 1.00E+00 | 1.00E+00 |
| Respiratory electron transport_ ATP synthesis by chemiosmotic coupling_ and heat production by uncoupling proteins. | 93/123 | 0/23 | 3.01E-14 | 8.70E-12 | 3.01E-14 | 8.70E-12 | 1.00E+00 | 1.00E+00 |
| Alzheimer_s disease - Homo sapiens (human) | 139/171 | 0/7 | 3.07E-14 | 8.70E-12 | 3.07E-14 | 8.70E-12 | 1.00E+00 | 1.00E+00 |
| Respiratory electron transport | 74/100 | 0/15 | 1.85E-13 | 4.94E-11 | 1.85E-13 | 4.94E-11 | 1.00E+00 | 1.00E+00 |
| Non-alcoholic fatty liver disease (NAFLD) - Homo sapiens (human) | 123/149 | 0/2 | 3.86E-12 | 9.19E-10 | 3.86E-12 | 9.19E-10 | 1.00E+00 | 1.00E+00 |
| Mitochondrial protein import | 49/65 | 0/4 | 6.18E-12 | 1.40E-09 | 6.18E-12 | 1.40E-09 | 1.00E+00 | 1.00E+00 |
| Electron Transport Chain | 77/103 | 0/12 | 1.06E-11 | 2.28E-09 | 1.06E-11 | 2.28E-09 | 1.00E+00 | 1.00E+00 |
| Post-translational protein modification | 1141/1383 | 1/176 | 2.38E-11 | 2.32E-09 | 8.27E-13 | 2.08E-10 | 1.00E+00 | 1.00E+00 |
| Oxidative phosphorylation | 43/61 | 0/5 | 8.44E-11 | 1.74E-08 | 8.44E-11 | 1.74E-08 | 1.00E+00 | 1.00E+00 |
| Complex I biogenesis | 39/55 | 0/2 | 3.20E-10 | 6.30E-08 | 3.20E-10 | 6.30E-08 | 1.00E+00 | 1.00E+00 |
| Regulation of Complement cascade | 79/130 | 0/8 | 4.42E-10 | 8.34E-08 | 4.42E-10 | 8.34E-08 | 1.00E+00 | 1.00E+00 |
| Complement cascade | 89/141 | 0/15 | 4.88E-10 | 8.84E-08 | 4.88E-10 | 8.84E-08 | 1.00E+00 | 1.00E+00 |
| Metabolism of RNA | 467/586 | 0/72 | 4.25E-09 | 6.94E-07 | 4.25E-09 | 6.94E-07 | 1.00E+00 | 1.00E+00 |
| SRP-dependent cotranslational protein targeting to membrane | 97/124 | 0/1 | 1.20E-08 | 1.75E-06 | 1.20E-08 | 1.75E-06 | 1.00E+00 | 1.00E+00 |
| Selenoamino acid metabolism | 103/129 | 1/61 | 2.57E-08 | 2.08E-06 | 1.19E-09 | 2.07E-07 | 1.00E+00 | 1.00E+00 |

**Table 3.3:** Top thirty differentially regulated pathways from IMPaLA with respect to T2D status using expression values.

Also, since the samples analysed in the study are skeletal muscle tissue, pathways representing biological functions specific to muscle tissue were expected to be differentially regulated. Neither IMPaLA analysis met this expectation as can be seen in table 3.2-3.3. In contrast, IPaA met these expectations very well in all five analyses as shown in table 3.1.

# 4. Discussion

Differential expression analysis assumes that changes in expression values of biomolecules, drive phenotypes observed in different biological states. While this is an obvious assumption, modelling molecular interactions is a challenging task. Existing differential pathway analysis methods based on ORA or FCS are limited to reporting differentially regulated pathways between biological states. They work with the average expression values and treat all biomolecule as equal contributors to pathway regulation.

The comparison in section 3.3 revealed the lack of commonality between IPaA and IMPaLA. From a perspective of hypothesis generation, the ideal goal is to generate pathways that reflect differences between phenotypes. The pathways considered differentially regulated by IPaA are specific to the sample being analysed and aid in the understanding of disease mechanisms. Pathways considered differentially regulated in IMPaLA could be perturbed in a variety of disease conditions and so do not aid in the understanding of type II diabetes pathology.

Though IMPaLA and IPaA, both seek to identify differentially regulated pathways, they use very different statistical hypotheses to assess which pathways are differentially expressed. Therefore, it is not surprising that the pathways assessed as differentially expressed by either analysis are very different. Furthermore, the study utilised a factorial design, therefore, expression values of the samples are affected by the administration of the hyperinsulinemic-euglycemic clamp. This information cannot be incorporated into IMPaLA's model.

The factor analysis model used in IPaA is based on the co-variation of biomolecule expression values. IMPaLA is based on the difference in distributions of average expression values per biomolecule in each group. Therefore, it may be possible that IMPaLA is affected by the number of differentially regulated biomolecules present in the pathway. This could explain the occurrence of broad pathways like *Metabolism* and *Translation* in tables 3.2-3.3.

IPaA would consider a pathway differentially regulated if differentially expressed biomolecules also had high co-variation.

Differential pathway regulation analysis in IPaA is based upon a set of values for each individual rather than a summary statistic. Since pathway regulation is a variable analogous to biomolecule expression but quantifying an entire pathway, differential pathway regulation can be performed, similar to differential expression analysis. So, differential pathway regulation analysis can utilise any number of variables that are of interest. A consequence of this is that pathways relevant to different aspects of the disease can be investigated. Investigation of differential regulation of pathways between individuals is also possible. Figure 4.1 C-D showcase the additional levels of analysis possible with IPaA.

Figure 4.1 C can be used to study the dependence of pathway regulation of clinical variables. Since the clinical variables are representative of the biology of individuals, this can aid in the identification of biological outliers as well as the study of the overall variability of pathway regulation among individuals.

Figure 4.1 D reflects the contributions of biomolecules to pathway regulation based on the loadings. This information can be used to analyse the specific role of differentially expressed biomolecules between phenotypes. It can also be used to identify markers for specific biological conditions.

The dependence of pathway regulation on clinical variables and assessment of biomolecular contribution can be done irrespective of the number of biological states being compared, if any comparison is being made at all. This makes IPaA a valuable tool to investigate pathway mechanisms and biological variation in a general population as well. This is completely unlike other pathway analysis methods where the comparison of phenotypes is fundamental to their operation.

Another interesting aspect of personal analysis is that the samples could represent any entity, not necessarily individual organisms. So, if expression data from single cells is available, it can be treated with IPaA to obtain pathway regulation values for single cells. This analysis can be extremely potent in analysing single-cell heterogeneity and comparison of cellular states under different biological conditions.

Table 4.1 compares the number of pathways found significantly differentially expressed at selected q-value thresholds in the analyses of integrated dataset and separate analyses of the expression datasets.The integration of data increases the number of pathways reported as differentially regulated as shown in figure 3.3.

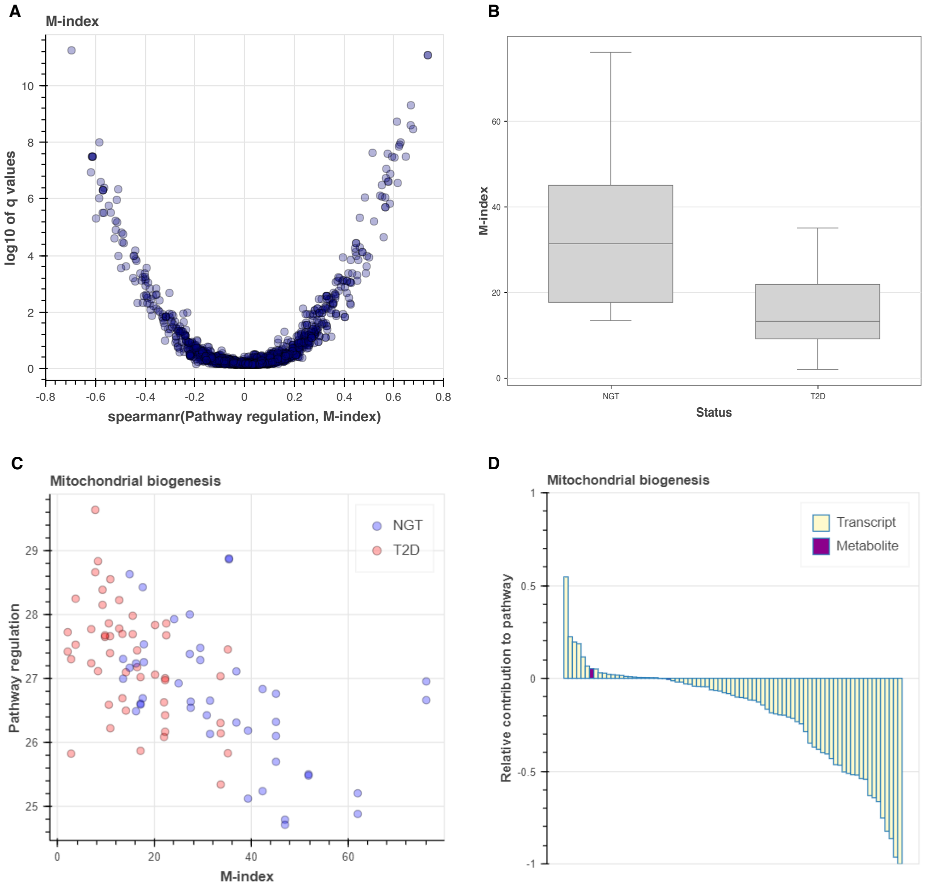

**Figure 4.1:** **A.** q-values vs Spearman correlation for differential pathway regulation analysis with respect to M-index , **B.** M-index vs T2D status. The box and whiskers are quartiles. **C.** Pathway regulation values per sample vs M-index **D.** Contribution of each biomolecule to pathway regulation.

From table 4.1 it is clear that the addition of metabolomic data leads to a greater increase in the number of pathways found significant at the chosen q-value thresholds. This indicates that the combined information of both expression datasets adds statistical power to the analysis.

The addition of metabolomic data also affects the co-variation pattern of pathways. The goal of IPaA is to model pathways based on co-variation of associated biomolecules. Therefore integration of data is essential to obtain more informative pathway models with respect to co-variance structure. Analysis of integrated data generates information that cannot be found with separate analyses alone. This is better visualised by comparing the transcriptomic and integrated analyses.

| Clinical variable | q value | | Analysis | |
|---|---|---|---|---|
| | threshold | Integrated | Transcriptomic | Metabolomic |
| M | 0.01 | 170 | 167 | 6 |
| | 0.05 | 251 | 244 | 6 |
| HbA1c | 0.01 | 59 | 59 | 0 |
| | 0.05 | 128 | 124 | 0 |
| HOMA-IR | 0.01 | 49 | 44 | 22 |
| | 0.05 | 93 | 85 | 22 |
| F.P. glucose | 0.01 | 19 | 19 | 20 |
| | 0.05 | 68 | 65 | 23 |
| S. C-peptide | 0.01 | 52 | 45 | 6 |
| | 0.05 | 112 | 109 | 20 |

**Table 4.1:** Counts of differentially regulated pathways at selected q value thresholds in analysis using integrated, transcriptomic and metabolomic datasets.

Dynamic html dashboards were developed using data from differential pathway regulation analysis with respect to five clinical variables for each dataset. They can be accessed at the following https://github.com/aron0093/integrated_pathway_analysis/tree/master/Dashboards.

The dashboards can be downloaded or viewed using an online viewer, https://htmlpreview.github.io/. These dashboards can be used to look at the variation in pathway regulation. While it is possible for outliers to occur in a single analysis, by tracking samples across pathways interesting biological outliers can be identified.

In retrospect, this project reveals that biological heterogeneity is a crucial aspect of understanding biological function. Here factor analysis is used as a model to explore this phenomenon but there are many methods that can be applied to model the interaction of biomolecules. Non-linear models based on deep learning like auto-encoders, stacked restricted Boltzmann machines or alternatively, models based on Bayesian inference could prove superior in the tasks performed by IPaA.

Another aspect that, in retrospect, needed more consideration is the mapping of molecules to pathways. While this applies to both genes and metabolites, metabolites can exist in several easily inter-convertible forms. The LC-MS processing itself can lead to subtle changes in attached groups, while functional forms of detected metabolites could also be different. This leads to challenges in mapping the identifiers for detected metabolite to identifiers of metabolites in pathways. Access to secondary identifiers for metabolites that represent these changes should result in a better mapping.

## 4.1 Summary

Biological heterogeneity is inescapable in biological analyses. While this fact is appreciated, most analyses only seek to mitigate the effect. However, most biological states, including diseases affect individuals differently. The investigation of this variation is crucial to understand the functional range of diseases, the interdependencies of many interacting processes and ultimately for better clinical outcomes for patients.

IPaA can be used as an exploratory tool to identify expression patterns at a functional level, as a comparative tool to identify differences between biological states or as an investigative tool to verify outcomes of known interventions. The unique features of this analysis are, first, the quantification of pathway regulation at a personalised level and second the quantification of biomolecular involvement in each biological process.

By considering the co-variance of biomolecular expression, IPaA models pathways as interacting biomolecules. Therefore, the integration of expression data allows IPaA to generate a complete picture of pathway regulation that cannot be modelled with either dataset alone.

## 4.2 Future work

A problem in the current analysis is that only 43 % of the genes and only 14.5 % of the metabolites are mapped to pathways in the Reactome database. Using a broader database like ConsensusPathDB would increase the pathway definitions used in the analysis. ConsensusPathDB draws pathway definitions from multiple pathway databases, each of which define pathways based on different types of molecular interactions. The interconversion of molecular identifiers is difficult. Using more varied pathway databases that are based on different molecular identifiers will reduce the amount of unused data.

Pathways are not the same as sets of biomolecules. Even though IPaA takes into consideration the inter-dependence of molecules in a pathway, the current analysis does not utilise the pathway topology. This information could be useful in modelling downstream and upstream effects for individual biomolecules. One way to do this would be to assign weights to individual nodes.

Information from other types of biological datasets like copy number variation or biological understanding of particular genes & other biomolecules can be incorporated by adjusting the weights on the biomolecules. This can be a valuable tool in assessing the broader impact of specific alterations.

Biological studies will often have samples that can be considered outliers. Factor analysis is a technique that is sensitive to outliers and there are methods that aim to make the analysis robust to outlying samples. Outlier masking can also be a major challenge for detecting outliers. This occurs when outliers have similar values and so are not detected. A technique has been proposed to make factor analysis robust to outliers by Pison et al. [42] Pison et al. have only provided tests cases for this technique on non-biological data with relatively few variables. For biological data the exclusion of samples that are not outliers but rather represent natural biological variation. Therefore, development of an outlier detection framework is required.

## 4.3 Ethical reflection

Personalised analysis of biomedical data has great promise in advancing human health by enabling individuals to receive customised care. It will also enable the exploration of the biological variation behind the sets of symptoms called diseases. Taken further personalised pathway analysis is a powerful tool that allows the measurement of biological functions at all levels in a quantifiable manner.

Ultimately these methods are only successful due to the huge amount of data that is being made available, so much so that personalised *omics* profiles are becoming a reality. With this great amount of information come the fairly standard concerns about data privacy and the increasingly in-depth insight offered into the lives of individuals from whom the data originates.

In this context, the concept of informed consent is especially relevant. While, the analysis of patient data has immense potential in advancing healthcare for millions, it is also possible for the increasingly intrusive nature of the analysis to cause harm in the personal lives of individuals. Thus, it is imperative that the consequences of sharing medical data are transparent and well understood.

## 4.4 Acknowledgement

In conclusion, I want to express my gratitude towards my supervisor Lukas Käll for involving me in this ground-breaking project. He has been a major source of inspiration for me and I would not have found my footing in statistical biology without his guidance. I also want to thank my colleagues in the statistical biotechnology group for their support and encouragement. I am especially grateful to Richard Tjörnhammar who is a post-doctoral associate in the group. His advice, collaboration and support was crucial in helping me overcome the challenges I faced during this project. I would also like to acknowledge Juleen Zierath et. al. for providing the data without which this project would not have been possible.

# Bibliography

[1] Purvesh Khatri, Marina Sirota, and Atul J. Butte. "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges". In: *PLOS Computational Biology* 8.2 (Feb. 2012), pp. 1–10. DOI: 10.1371/journal.pcbi.1002375.

[2] Saikat Chowdhury and Ram Rup Sarkar. "Comparison of human cell signaling pathway databases - Evolution, drawbacks and challenges". In: *Database* 2015.January (2015), pp. 1–25. ISSN: 17580463. DOI: 10.1093/database/bau126.

[3] A. Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. arXiv: NIHMS150003.

[4] Ariel E. Bayá et al. "Gene Set Enrichment Analysis Using Non-parametric Scores". In: *Advances in Bioinformatics and Computational Biology*. Ed. by Marie-France Sagot and Maria Emilia M. T. Walter. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 12–21. ISBN: 978-3-540-73731-5.

[5] Jui-Hung Hung et al. "Gene set enrichment analysis: performance evaluation and usage guidelines". In: *Briefings in Bioinformatics* 13.3 (2012), pp. 281–291. DOI: 10.1093/bib/bbr049.

[6] Rachel Cavill et al. "Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells". In: *PLoS Computational Biology* 7.3 (2011). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1001113.

[7] Atanas Kamburov et al. "Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA". In: *Bioinformatics* 27.20 (2011), pp. 2917–2918. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr499.

[8] Charles J. Vaske et al. "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM". In: *Bioinformatics* 26.12 (2010), pp. 237–245. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq182.

[9] Tiinamaija Tuomi et al. "The many faces of diabetes: A disease with increasing heterogeneity". In: *The Lancet* 383.9922 (2014), pp. 1084–1094. ISSN: 1474547X. DOI: 10.1016/S0140-6736(13)62219-9.

[10] Harry S. Glauber, Naphtali Rishe, and Eddy Karnieli. "Introduction to Personalized Medicine in Diabetes Mellitus". In: *Rambam Maimonides Medical Journal* 5.1 (2014), e0002. ISSN: 20769172. DOI: 10.5041/RMMJ.10136.

[11] "Confirmatory Factor Analysis". In: *Methods of Multivariate Analysis*. Wiley-Blackwell, 2012. Chap. 14, pp. 479–500. ISBN: 9781118391686. DOI: 10.1002/9781118391686.ch14.

[12] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[13] Christopher M. Bishop. "12.2.4 Factor Analysis". In: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006, pp. 583–586. ISBN: 0387310738.

[14] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[15] John D. Storey and Robert Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445. ISSN: 0027-8424. DOI: 10.1073/pnas.1530509100.

[16] Jason K. Kim. *Hyperinsulinemic-euglycemic clamp to assess insulin sensitivity in vivo*. Vol. 560. 2009, pp. 221–238. ISBN: 9781934115152. DOI: 10.1007/978-1-59745-448-3_15.

[17] P Paatero and U Tapper. “Analysis of different modes of factor analysis as least squares fit problems”. In: *Chemom. Intell. Lab. Syst.* 18 (1993), pp. 183–194. DOI: 10.1016/0169-7439(93)80055-M.

[18] Kentaro Hayashi, Peter M. Bentler, and Ke-Hai Yuan. “On the Likelihood Ratio Test for the Number of Factors in Exploratory Factor Analysis”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 14.3 (2007), pp. 505–526. DOI: 10.1080/10705510701301891.

[19] Ahmad Barghash and Taner Arslan. “Robust Detection of Outlier Samples and Genes in Expression Datasets”. In: *Journal of Proteomics & Bioinformatics* 09.02 (2016), pp. 38–48. ISSN: 0974276X. DOI: 10.4172/jpb.1000387.

[20] David Croft et al. “The Reactome pathway knowledgebase”. In: *Nucleic acids* 42.D1 (2014), pp. D472–D477. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1102.

[21] Antonio Fabregat et al. “The Reactome Pathway Knowledgebase”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D649–D655. ISSN: 13624962. DOI: 10.1093/nar/gkx1132. arXiv: NIHMS150003.

[22] L. Monnier et al. “Continuous glucose monitoring in patients with type 2diabetes: Why? When? Whom?” In: *Diabetes and Metabolism* 33.4 (2007), pp. 247–252. ISSN: 12623636. DOI: 10.1016/j.diabet.2006.11.007.

[23] J Schlichtkrull, O Munck, and M Jersild. “M-Value, an Index for Blood Sugar Control in Diabetics”. In: *Ugeskrift for laeger* 126 (1964), pp. 815–20. ISSN: 0041-5782. URL: http://www.ncbi.nlm.nih.gov/pubmed/14161920.

[24] F. John Service. “Glucose variability”. In: *Diabetes* 62.5 (2013), pp. 1398–1404. ISSN: 00121797. DOI: 10.2337/db12-1396.

[25] World Health Organisation. *Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus*. Vol. 93. 3. 2011, pp. 299–309. DOI: 10.1016/j.diabres.2011.03.012.

[26] A. G. Jones and A. T. Hattersley. “The clinical utility of C-peptide measurement in the care of patients with diabetes”. In: *Diabetic Medicine* 30.7 (2013), pp. 803–817. ISSN: 07423071. DOI: 10.1111/dme.12159.

[27] D R Matthews et al. “Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man.” In: *Diabetologia* 28.7 (1985), pp. 412–419. ISSN: 0012-186X. DOI: 10.1007/BF00280883. arXiv: 3899825.

[28] Kohei Okita et al. “Homeostasis model assessment of insulin resistance for evaluating insulin sensitivity in patients with type 2 diabetes on insulin therapy”. In: *Endocrine Journal* 60.3 (2013), pp. 283–290. ISSN: 0918-8959. DOI: 10.1507/endocrj.EJ12-0320.

[29] Katherine T. Tonks et al. “Skeletal muscle and plasma lipidomic signatures of insulin resistance and overweight/obesity in humans”. In: *Obesity* 24.4 (2016), pp. 908–916. ISSN: 1930739X. DOI: 10.1002/oby.21448.

[30] Roxana Valdés-Ramos et al. “Vitamins and type 2 diabetes mellitus.” In: *Endocrine, metabolic & immune disorders drug targets* 15.1 (2015), pp. 54–63. ISSN: 2212-3873. DOI: EMIDDT-EPUB-63314[pii].

[31] Masaru Maebashi et al. “Therapeutic Evaluation of the Effect of Biotin on Hyperglycemia in Patients with Non-Insulin Dependent Diabetes Mellitus”. In: *Journal of Clinical Biochemistry and Nutrition* 14.3 (1993), pp. 211–218. ISSN: 0912-0009. DOI: 10.3164/jcbn.14.211.

[32] Annika Mehlem et al. “PGC-1$\alpha$ Coordinates Mitochondrial Respiratory Capacity and Muscular Fatty Acid Uptake via Regulation of VEGF-B”. In: *Diabetes* 65.4 (2016), pp. 861–873. ISSN: 0012-1797. DOI: 10.2337/db15-1231.

[33] Gary Lewis et al. “Disordered fat storage and mobilization in the pathogenesis of insulin resistance and type 2 diabetes.” In: *Endocrine reviews* 23.2 (2002), p. 201. ISSN: 0163-769X. DOI: 10.1210/edrv.23.2.0461.

[34] Ralph A. DeFronzo and Devjit Tripathy. “Skeletal muscle insulin resistance is the primary defect in type 2 diabetes.” In: *Diabetes care* 32 Suppl 2 (2009). ISSN: 19355548. DOI: 10.2337/dc09-S302.

[35] Esther Phielix and Marco Mensink. “Type 2 Diabetes Mellitus and Skeletal Muscle Metabolic Function”. In: *Physiology and Behavior* 94.2 (2008), pp. 252–258. ISSN: 00319384. DOI: 10.1016/j.physbeh.2008.01.020.

[36] Lori Laffel. “Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes”. In: *Diabetes/Metabolism Research and Reviews* 15.6 (1999), pp. 412–426. ISSN: 1520-7552. DOI: 10.1002/(SICI)1520-7560(199911/12)15:6<412::AID-DMRR72>3.0.CO;2-8.

[37] Reza Norouzirad, Pedro González-Muniesa, and Asghar Ghasemi. “Hypoxia in Obesity and Diabetes: Potential Therapeutic Effects of Hyperoxia and Nitrate”. In: *Oxidative Medicine and Cellular Longevity* 2017 (2017). ISSN: 19420994. DOI: 10.1155/2017/5350267.

[38] Huiyun Liang and Walter F. Ward. “PGC-1$\alpha$: a key regulator of energy metabolism”. In: *Advances in Physiology Education* 30.4 (2006). PMID: 17108241, pp. 145–151. DOI: 10.1152/advan.00052.2006.

[39] Mun Chun Chan and Zolt Arany. “The Many roles of PGC-1a in Muscle - Recent Developments”. In: *Metabolism* 63.4 (Apr. 2014). 24559845[pmid], pp. 441–451. ISSN: 0026-0495. DOI: 10.1016/j.metabol.2014.01.006.

[40] Xue Zhao et al. “The Relationship between Branched-Chain Amino Acid Related Metabolomic Signature and Insulin Resistance: A Systematic Review”. In: *Journal of Diabetes Research* 2016 (2016). ISSN: 23146753. DOI: 10.1155/2016/2794591.

[41] Christopher B. Newgard. “Interplay between lipids and branched-chain amino acids in development of insulin resistance”. In: *Cell Metabolism* 15.5 (2012), pp. 606–614. ISSN: 15504131. DOI: 10.1016/j.cmet.2012.01.024.

[42] Greet Pison et al. “Robust factor analysis”. In: *Journal of Multivariate Analysis* 84.1 (2003), pp. 145–172. ISSN: 0047-259X. DOI: https://doi.org/10.1016/S0047-259X(02)00007-6.

# A. Appendix

| Pathway description | measured/actual | | Integrated analysis | | Transcriptomic analysis | | Metabolomic analysis | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Amino acid transport across the plasma membrane | 28/34 | 2/34 | 3.96E-15 | 5.62E-12 | 3.85E-15 | 5.62E-12 | 6.42E-01 | 7.26E-01 |
| VEGF ligand-receptor interactions | 7/8 | 0/0 | 1.76E-14 | 8.34E-12 | 1.76E-14 | 8.57E-12 | - | - |
| VEGF binds to VEGFR leading to receptor dimerization | 7/8 | 0/0 | 1.76E-14 | 8.34E-12 | 1.76E-14 | 8.57E-12 | - | - |
| Lysine catabolism | 9/13 | 0/32 | 1.39E-12 | 4.92E-10 | 1.39E-12 | 5.06E-10 | - | - |
| Utilization of Ketone Bodies | 5/5 | 0/10 | 6.57E-12 | 1.87E-09 | 6.57E-12 | 1.92E-09 | - | - |
| Regulation of gene expression by Hypoxia-inducible Factor | 8/11 | 0/1 | 1.06E-11 | 2.50E-09 | 1.06E-11 | 2.57E-09 | - | - |
| Activation of PPARGC1A (PGC-1alpha) by phosphorylation | 9/10 | 0/3 | 1.69E-11 | 3.43E-09 | 1.69E-11 | 3.52E-09 | - | - |
| Amino acid synthesis and interconversion (transamination) | 28/34 | 1/47 | 6.16E-11 | 1.01E-08 | 6.17E-11 | 1.04E-08 | - | - |
| Defects in vitamin and cofactor metabolism | 20/23 | 0/20 | 6.40E-11 | 1.01E-08 | 6.40E-11 | 1.04E-08 | - | - |
| Defects in biotin (Btn) metabolism | 7/9 | 0/4 | 8.54E-11 | 1.21E-08 | 8.54E-11 | 1.25E-08 | - | - |
| Defective HLCS causes multiple carboxylase deficiency | 6/8 | 0/3 | 1.10E-10 | 1.42E-08 | 1.10E-10 | 1.46E-08 | - | - |
| Methylation | 12/14 | 0/38 | 2.00E-10 | 2.36E-08 | 2.00E-10 | 2.43E-08 | - | - |
| Synthesis of Prostaglandins (PG) and Thromboxanes (TX) | 16/16 | 0/54 | 2.33E-10 | 2.55E-08 | 2.33E-10 | 2.62E-08 | - | - |
| Pyruvate metabolism | 26/31 | 0/28 | 3.89E-10 | 3.22E-08 | 3.89E-10 | 3.31E-08 | - | - |
| Apoptotic factor-mediated response | 7/7 | 0/2 | 3.93E-10 | 3.22E-08 | 3.93E-10 | 3.31E-08 | - | - |
| Activation of caspases through apoptosome-mediated cleavage | 5/5 | 0/2 | 3.97E-10 | 3.22E-08 | 3.97E-10 | 3.31E-08 | - | - |
| Cytochrome c-mediated apoptotic response | 5/5 | 0/2 | 3.97E-10 | 3.22E-08 | 3.97E-10 | 3.31E-08 | - | - |
| Release of apoptotic factors from the mitochondria | 4/4 | 0/0 | 4.18E-10 | 3.22E-08 | 4.18E-10 | 3.31E-08 | - | - |
| GRB2 events in EGFR signaling | 6/8 | 0/2 | 4.31E-10 | 3.22E-08 | 4.31E-10 | 3.31E-08 | - | - |
| Formation of apoptosome | 3/3 | 0/2 | 4.87E-10 | 3.46E-08 | 4.87E-10 | 3.56E-08 | - | - |
| Biotin transport and metabolism | 10/12 | 0/8 | 1.29E-09 | 8.73E-08 | 1.29E-09 | 8.97E-08 | - | - |
| ERBB2 Regulates Cell Motility | 14/15 | 0/1 | 1.79E-09 | 1.16E-07 | 1.79E-09 | 1.19E-07 | - | - |
| Glyoxylate metabolism and glycine degradation | 26/31 | 0/43 | 2.13E-09 | 1.31E-07 | 2.13E-09 | 1.35E-07 | - | - |
| Ubiquinol biosynthesis | 7/8 | 0/31 | 3.14E-09 | 1.86E-07 | 3.14E-09 | 1.91E-07 | - | - |
| Glutathione synthesis and recycling | 11/12 | 1/13 | 4.46E-09 | 2.46E-07 | 6.61E-09 | 3.33E-07 | - | - |
| Pyruvate metabolism and Citric Acid (TCA) cycle | 48/56 | 0/45 | 4.51E-09 | 2.46E-07 | 4.51E-09 | 2.64E-07 | - | - |
| Propionyl-CoA catabolism | 5/5 | 0/9 | 4.84E-09 | 2.55E-07 | 4.84E-09 | 2.72E-07 | - | - |
| Regulation of pyruvate dehydrogenase (PDH) complex | 14/16 | 0/15 | 5.42E-09 | 2.75E-07 | 5.42E-09 | 2.93E-07 | - | - |
| Branched-chain amino acid catabolism | 21/23 | 2/47 | 6.66E-09 | 3.26E-07 | 6.55E-09 | 3.33E-07 | 1.88E-01 | 4.31E-01 |
| Mitochondrial biogenesis | 77/95 | 2/14 | 8.58E-09 | 4.06E-07 | 8.61E-09 | 4.19E-07 | 3.41E-01 | 6.57E-01 |
| Defects in cobalamin (B12) metabolism | 13/14 | 0/17 | 1.00E-08 | 4.58E-07 | 1.00E-08 | 4.71E-07 | - | - |
| EGFR interacts with phospholipase C-gamma | 2/3 | 0/2 | 1.19E-08 | 4.95E-07 | 1.19E-08 | 5.09E-07 | - | - |
| Signaling by Overexpressed Wild-Type EGFR in Cancer | 2/2 | 0/4 | 1.19E-08 | 4.95E-07 | 1.19E-08 | 5.09E-07 | - | - |
| Inhibition of Signaling by Overexpressed EGFR | 2/2 | 0/4 | 1.19E-08 | 4.95E-07 | 1.19E-08 | 5.09E-07 | - | - |
| Metabolism of cofactors | 18/19 | 0/62 | 1.90E-08 | 7.70E-07 | 1.90E-08 | 7.92E-07 | - | - |
| Citric acid cycle (TCA cycle) | 20/23 | 0/28 | 2.14E-08 | 8.43E-07 | 2.14E-08 | 8.67E-07 | - | - |
| Defective OPLAH causes 5-oxoprolinase deficiency (OPLAHD) | 1/1 | 1/3 | 2.34E-08 | 8.98E-07 | | | - | - |
| Synthesis of PI | 5/5 | 0/9 | 2.40E-08 | 8.98E-07 | 2.40E-08 | 9.48E-07 | - | - |
| ERBB2 Activates PTK6 Signaling | 11/13 | 0/2 | 2.62E-08 | 9.53E-07 | 2.62E-08 | 1.01E-06 | - | - |
| Sialic acid metabolism | 27/41 | 0/27 | 3.07E-08 | 1.09E-06 | 3.07E-08 | 1.15E-06 | - | - |
| Cholesterol biosynthesis | 23/25 | 0/46 | 4.30E-08 | 1.49E-06 | 4.30E-08 | 1.57E-06 | - | - |
| Organelle biogenesis and maintenance | 251/315 | 2/21 | 5.18E-08 | 1.75E-06 | 5.19E-08 | 1.85E-06 | 3.41E-01 | 6.57E-01 |
| Cholesterol biosynthesis via desmosterol | 4/4 | 0/10 | 6.15E-08 | 1.98E-06 | 6.15E-08 | 2.09E-06 | - | - |
| Cholesterol biosynthesis via lathosterol | 4/4 | 0/10 | 6.15E-08 | 1.98E-06 | 6.15E-08 | 2.09E-06 | - | - |
| Cristae formation | 25/31 | 0/0 | 9.75E-08 | 3.08E-06 | 9.75E-08 | 3.23E-06 | - | - |
| PLCG1 events in ERBB2 signaling | 3/4 | 0/2 | 1.01E-07 | 3.12E-06 | 1.01E-07 | 3.28E-06 | - | - |
| PI3K events in ERBB2 signaling | 14/16 | 0/4 | 1.04E-07 | 3.14E-06 | 1.04E-07 | 3.30E-06 | - | - |
| Sulfur amino acid metabolism | 17/26 | 0/83 | 1.58E-07 | 4.68E-06 | 1.58E-07 | 4.91E-06 | - | - |
| Hydrolysis of LPE | 2/2 | 0/9 | 1.71E-07 | 4.94E-06 | 1.71E-07 | 5.19E-06 | - | - |
| TP53 Regulates Metabolic Genes | 76/90 | 0/25 | 2.10E-07 | 5.97E-06 | 2.10E-07 | 6.26E-06 | - | - |
| CaMK IV-mediated phosphorylation of CREB | 2/3 | 0/7 | 2.26E-07 | 6.30E-06 | 2.26E-07 | 6.61E-06 | - | - |

**Table A.1:** Fifty most significantly differentially regulated pathways with respect to M-index.

| Pathway description | measured/actual | | Integrated analysis | | Transcriptomic analysis | | Metabolomic analysis | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Defective HLCS causes multiple carboxylase deficiency | 6/8 | 0/3 | 5.25E-07 | 4.19E-04 | 5.25E-07 | 4.22E-04 | - | - |
| Defects in biotin (Btn) metabolism | 7/9 | 0/4 | 6.35E-07 | 4.19E-04 | 6.35E-07 | 4.22E-04 | - | - |
| Defects in vitamin and cofactor metabolism | 20/23 | 0/20 | 1.44E-06 | 4.94E-04 | 1.44E-06 | 4.98E-04 | - | - |
| Propionyl-CoA catabolism | 5/5 | 0/9 | 1.50E-06 | 4.94E-04 | 1.50E-06 | 4.98E-04 | - | - |
| Methylation | 12/14 | 0/38 | 2.46E-06 | 6.50E-04 | 2.46E-06 | 6.55E-04 | - | - |
| Biotin transport and metabolism | 10/12 | 0/8 | 6.01E-06 | 1.32E-03 | 6.01E-06 | 1.33E-03 | - | - |
| Branched-chain amino acid catabolism | 21/23 | 2/47 | 8.67E-06 | 1.63E-03 | 8.64E-06 | 1.64E-03 | 9.98E-01 | 9.98E-01 |
| Signaling by Overexpressed Wild-Type EGFR in Cancer | 2/2 | 0/4 | 1.50E-05 | 1.98E-03 | 1.50E-05 | 2.00E-03 | - | - |
| Inhibition of Signaling by Overexpressed EGFR | 2/2 | 0/4 | 1.50E-05 | 1.98E-03 | 1.50E-05 | 2.00E-03 | - | - |
| EGFR interacts with phospholipase C-gamma | 2/3 | 0/2 | 1.50E-05 | 1.98E-03 | 1.50E-05 | 2.00E-03 | - | - |
| Utilization of Ketone Bodies | 5/5 | 0/10 | 2.40E-05 | 2.29E-03 | 2.40E-05 | 2.31E-03 | - | - |
| Apoptotic factor-mediated response | 7/7 | 0/2 | 2.44E-05 | 2.29E-03 | 2.44E-05 | 2.31E-03 | - | - |
| Activation of caspases through apoptosome-mediated cleavage | 5/5 | 0/2 | 2.45E-05 | 2.29E-03 | 2.45E-05 | 2.31E-03 | - | - |
| Cytochrome c-mediated apoptotic response | 5/5 | 0/2 | 2.45E-05 | 2.29E-03 | 2.45E-05 | 2.31E-03 | - | - |
| Release of apoptotic factors from the mitochondria | 4/4 | 0/ | 2.61E-05 | 2.29E-03 | 2.61E-05 | 2.31E-03 | - | - |
| Glyoxylate metabolism and glycine degradation | 26/31 | 0/43 | 2.78E-05 | 2.29E-03 | 2.78E-05 | 2.31E-03 | - | - |
| tRNA processing in the mitochondrion | 4/7 | 0/6 | 3.14E-05 | 2.44E-03 | 3.14E-05 | 2.45E-03 | - | - |
| Amino acid transport across the plasma membrane | 28/34 | 2/34 | 3.79E-05 | 2.61E-03 | 4.46E-05 | 2.63E-03 | 1.63E-01 | 3.36E-01 |
| SeMet incorporation into proteins | 10/11 | 0/6 | 4.44E-05 | 2.61E-03 | 4.44E-05 | 2.63E-03 | - | - |
| Synthesis of Prostaglandins (PG) and Thromboxanes (TX) | 16/16 | 0/54 | 4.45E-05 | 2.61E-03 | 4.45E-05 | 2.63E-03 | - | - |
| The proton buffering model | 5/5 | 0/6 | 4.54E-05 | 2.61E-03 | 4.54E-05 | 2.63E-03 | - | - |
| The fatty acid cycling model | 5/5 | 0/3 | 4.54E-05 | 2.61E-03 | 4.54E-05 | 2.63E-03 | - | - |
| Mitochondrial Uncoupling Proteins | 5/5 | 0/8 | 4.54E-05 | 2.61E-03 | 4.54E-05 | 2.63E-03 | - | - |
| Formation of apoptosome | 3/3 | 0/2 | 5.57E-05 | 3.07E-03 | 5.57E-05 | 3.09E-03 | - | - |
| Defects in cobalamin (B12) metabolism | 13/14 | 0/17 | 6.50E-05 | 3.25E-03 | 6.50E-05 | 3.27E-03 | - | - |

**Table A.2:** Twenty five most significantly differentially regulated pathways with respect to Glycated haemoglobin.

| Pathway description | measured/actual | | Integrated analysis | | Transcriptomic analysis | | Metabolomic analysis | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Defective HLCS causes multiple carboxylase deficiency | 6/8 | 0/3 | 5.25E-07 | 4.19E-04 | 5.25E-07 | 4.22E-04 | - | - |
| Amino acid transport across the plasma membrane | 28/34 | 2/34 | 1.57E-07 | 1.07E-04 | 1.55E-07 | 1.07E-04 | 9.22E-01 | 9.57E-01 |
| VEGF binds to VEGFR leading to receptor dimerization | 7/8 | 0/ | 1.94E-07 | 1.07E-04 | 1.94E-07 | 1.07E-04 | - | - |
| VEGF ligand-receptor interactions | 7/8 | 0/ | 1.94E-07 | 1.07E-04 | 1.94E-07 | 1.07E-04 | - | - |
| Regulation of gene expression by Hypoxia-inducible Factor | 8/11 | 0/1 | 2.53E-07 | 1.07E-04 | 2.53E-07 | 1.07E-04 | - | - |
| Misspliced LRP5 mutants have enhanced beta-catenin-dependent signaling | 6/6 | 0/ | 5.94E-07 | 2.01E-04 | 5.94E-07 | 2.01E-04 | - | - |
| Hydrolysis of LPE | 2/2 | 0/9 | 7.26E-07 | 2.04E-04 | 7.26E-07 | 2.04E-04 | - | - |
| Lysine catabolism | 9/13 | 0/32 | 9.80E-07 | 2.36E-04 | 9.80E-07 | 2.36E-04 | - | - |
| Branched-chain amino acid catabolism | 21/23 | 2/47 | 1.39E-06 | 2.94E-04 | 1.37E-06 | 2.89E-04 | 3.35E-02 | 7.38E-02 |
| Defects in biotin (Btn) metabolism | 7/9 | 0/4 | 1.69E-06 | 3.17E-04 | 1.69E-06 | 3.17E-04 | - | - |
| Defective HLCS causes multiple carboxylase deficiency | 6/8 | 0/3 | 2.37E-06 | 3.95E-04 | 2.37E-06 | 3.95E-04 | - | - |
| Utilization of Ketone Bodies | 5/5 | 0/10 | 2.79E-06 | 3.95E-04 | 2.79E-06 | 3.95E-04 | - | - |
| Activation of PPARGC1A (PGC-1alpha) by phosphorylation | 9/10 | 0/3 | 2.81E-06 | 3.95E-04 | 2.81E-06 | 3.95E-04 | - | - |
| Activation of RAC1 | 10/13 | 0/2 | 7.75E-06 | 1.01E-03 | 7.75E-06 | 1.01E-03 | - | - |
| Defective OPLAH causes 5-oxoprolinase deficiency (OPLAHD) | 1/1 | 1/3 | 8.54E-06 | 1.03E-03 | - | - | - | - |
| EGFR interacts with phospholipase C-gamma | 2/3 | 0/2 | 1.22E-05 | 1.22E-03 | 1.22E-05 | 1.29E-03 | - | - |
| Signaling by Overexpressed Wild-Type EGFR in Cancer | 2/2 | 0/4 | 1.22E-05 | 1.22E-03 | 1.22E-05 | 1.29E-03 | - | - |
| Inhibition of Signaling by Overexpressed EGFR | 2/2 | 0/4 | 1.22E-05 | 1.22E-03 | 1.22E-05 | 1.29E-03 | - | - |
| ERBB2 Regulates Cell Motility | 14/15 | 0/1 | 1.65E-05 | 1.55E-03 | 1.65E-05 | 1.64E-03 | - | - |
| Defects in vitamin and cofactor metabolism | 20/23 | 0/20 | 1.88E-05 | 1.67E-03 | 1.88E-05 | 1.77E-03 | - | - |
| Metabolism of folate and pterines | 15/17 | 0/26 | 2.29E-05 | 1.94E-03 | 2.29E-05 | 2.04E-03 | - | - |
| Amino acid synthesis and interconversion (transamination) | 28/34 | 1/47 | 2.94E-05 | 2.36E-03 | 2.96E-05 | 2.50E-03 | - | - |
| Abasic sugar-phosphate removal via the single-nucleotide replacement pathway | 2/2 | 0/1 | 3.73E-05 | 2.75E-03 | 3.73E-05 | 2.87E-03 | - | - |
| Biotin transport and metabolism | 10/12 | 0/8 | 3.74E-05 | 2.75E-03 | 3.74E-05 | 2.87E-03 | - | - |
| Biosynthesis of DPAn-3-derived maresins | 1/3 | 1/12 | 4.84E-05 | 3.26E-03 | - | - | - | - |

**Table A.3:** Twenty five most significantly differentially regulated pathways with respect to Homoeostatic model assessment-Insulin resistance.

| Pathway description | measured/actual | | Integrated analysis | | Transcriptomic analysis | | Metabolomic analysis | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Amino acid transport across the plasma membrane | 28/34 | 2/34 | 7.34E-09 | 1.43E-05 | 7.49E-09 | 1.45E-05 | 8.73E-01 | 9.06E-01 |
| VEGF ligand-receptor interactions | 7/8 | 0/0 | 4.18E-08 | 2.71E-05 | 4.18E-08 | 2.70E-05 | - | - |
| VEGF binds to VEGFR leading to receptor dimerization | 7/8 | 0/0 | 4.18E-08 | 2.71E-05 | 4.18E-08 | 2.70E-05 | - | - |
| Regulation of gene expression by Hypoxia-inducible Factor | 8/11 | 0/1 | 6.76E-08 | 3.29E-05 | 6.76E-08 | 3.28E-05 | - | - |
| Lysine catabolism | 9/13 | 0/32 | 6.74E-07 | 2.62E-04 | 6.74E-07 | 2.61E-04 | - | - |
| Misspliced LRP5 mutants have enhanced beta-catenin-dependent signaling | 6/6 | 0/0 | 1.16E-06 | 3.75E-04 | 1.16E-06 | 3.74E-04 | - | - |
| Defects in biotin (Btn) metabolism | 7/9 | 0/4 | 1.42E-06 | 3.95E-04 | 1.42E-06 | 3.93E-04 | - | - |
| Defective HLCS causes multiple carboxylase deficiency | 6/8 | 0/3 | 2.22E-06 | 5.39E-04 | 2.22E-06 | 5.37E-04 | - | - |
| Activation of RAC1 | 10/13 | 0/2 | 3.17E-06 | 6.63E-04 | 3.17E-06 | 6.62E-04 | - | - |
| Activation of PPARGC1A (PGC-1alpha) by phosphorylation | 9/10 | 0/3 | 3.46E-06 | 6.63E-04 | 3.46E-06 | 6.62E-04 | - | - |
| Hydrolysis of LPE | 2/2 | 0/9 | 3.86E-06 | 6.63E-04 | 3.86E-06 | 6.62E-04 | - | - |
| Calcineurin activates NFAT | 9/9 | 0/8 | 4.10E-06 | 6.63E-04 | 4.10E-06 | 6.62E-04 | - | - |
| Amino acid synthesis and interconversion (transamination) | 28/34 | 1/47 | 4.57E-06 | 6.83E-04 | 4.63E-06 | 6.90E-04 | - | - |
| CREB phosphorylation | 6/8 | 0/2 | 5.58E-06 | 7.64E-04 | 5.58E-06 | 7.62E-04 | - | - |
| Utilization of Ketone Bodies | 5/5 | 0/10 | 5.90E-06 | 7.64E-04 | 5.90E-06 | 7.62E-04 | - | - |
| Branched-chain amino acid catabolism | 21/23 | 2/47 | 6.45E-06 | 7.84E-04 | 6.36E-06 | 7.71E-04 | 7.65E-02 | 1.45E-01 |
| Sialic acid metabolism | 27/41 | 0/27 | 8.54E-06 | 9.67E-04 | 8.54E-06 | 9.74E-04 | - | - |
| Glutathione synthesis and recycling | 11/12 | 1/13 | 9.28E-06 | 9.67E-04 | 1.15E-05 | 1.17E-03 | - | - |
| CaMK IV-mediated phosphorylation of CREB | 2/3 | 0/7 | 9.45E-06 | 9.67E-04 | 9.45E-06 | 1.02E-03 | - | - |
| Pyruvate metabolism | 26/31 | 0/28 | 1.64E-05 | 1.42E-03 | 1.64E-05 | 1.42E-03 | - | - |
| Biotin transport and metabolism | 10/12 | 0/8 | 1.66E-05 | 1.42E-03 | 1.66E-05 | 1.42E-03 | - | - |
| ERBB2 Regulates Cell Motility | 14/15 | 0/1 | 1.70E-05 | 1.42E-03 | 1.70E-05 | 1.42E-03 | - | - |
| ERBB2 Activates PTK6 Signaling | 11/13 | 0/2 | 1.79E-05 | 1.42E-03 | 1.79E-05 | 1.42E-03 | - | - |
| EGFR interacts with phospholipase C-gamma | 2/3 | 0/2 | 1.98E-05 | 1.42E-03 | 1.98E-05 | 1.42E-03 | - | - |
| Signaling by Overexpressed Wild-Type EGFR in Cancer | 2/2 | 0/4 | 1.98E-05 | 1.42E-03 | 1.98E-05 | 1.42E-03 | - | - |

**Table A.4:** Twenty five most significantly differentially regulated pathways with respect to serum C-peptide level.

| Pathway description | measured/actual | | Integrated analysis | | Transcriptomic analysis | | Metabolomic analysis | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Metabolites | p value | q value | p value | q value | p value | q value |
| Hydrolysis of LPE | 2/2 | 0/9 | 1.35E-07 | 2.41E-04 | 1.35E-07 | 2.40E-04 | - | - |
| Defective HLCS causes multiple carboxylase deficiency | 6/8 | 0/3 | 3.20E-06 | 2.48E-03 | 3.20E-06 | 2.47E-03 | - | - |
| Defects in biotin (Btn) metabolism | 7/9 | 0/4 | 4.18E-06 | 2.48E-03 | 4.18E-06 | 2.47E-03 | - | - |
| Synthesis of Prostaglandins (PG) and Thromboxanes (TX) | 16/16 | 0/54 | 9.03E-06 | 2.75E-03 | 9.03E-06 | 2.73E-03 | - | - |
| EGFR interacts with phospholipase C-gamma | 2/3 | 0/2 | 1.08E-05 | 2.75E-03 | 1.08E-05 | 2.73E-03 | - | - |
| Signaling by Overexpressed Wild-Type EGFR in Cancer | 2/2 | 0/4 | 1.08E-05 | 2.75E-03 | 1.08E-05 | 2.73E-03 | - | - |
| Inhibition of Signaling by Overexpressed EGFR | 2/2 | 0/4 | 1.08E-05 | 2.75E-03 | 1.08E-05 | 2.73E-03 | - | - |
| Lysine catabolism | 9/13 | 0/32 | 1.70E-05 | 3.45E-03 | 1.70E-05 | 3.42E-03 | - | - |
| Branched-chain amino acid catabolism | 21/23 | 2/47 | 1.74E-05 | 3.45E-03 | 1.74E-05 | 3.42E-03 | 8.43E-01 | 9.09E-01 |
| Propionyl-CoA catabolism | 5/5 | 0/9 | 2.08E-05 | 3.70E-03 | 2.08E-05 | 3.68E-03 | - | - |
| Classical antibody-mediated complement activation | 40/95 | 0/2 | 3.50E-05 | 5.43E-03 | 3.50E-05 | 5.40E-03 | - | - |
| Biotin transport and metabolism | 10/12 | 0/8 | 3.66E-05 | 5.43E-03 | 3.66E-05 | 5.40E-03 | - | - |
| Creation of C4 and C2 activators | 47/103 | 0/8 | 5.28E-05 | 6.99E-03 | 5.28E-05 | 6.95E-03 | - | - |
| Defects in vitamin and cofactor metabolism | 20/23 | 0/20 | 5.50E-05 | 6.99E-03 | 5.50E-05 | 6.95E-03 | - | - |
| tRNA processing in the mitochondrion | 4/7 | 0/6 | 5.93E-05 | 7.00E-03 | 5.93E-05 | 6.97E-03 | - | - |
| FCGR activation | 45/101 | 0/4 | 6.30E-05 | 7.00E-03 | 6.30E-05 | 6.97E-03 | - | - |
| Role of phospholipids in phagocytosis | 55/114 | 0/18 | 7.71E-05 | 8.07E-03 | 7.71E-05 | 8.03E-03 | - | - |
| Trafficking of myristoylated proteins to the cilium | 5/5 | 0/2 | 8.96E-05 | 8.86E-03 | 8.96E-05 | 8.81E-03 | - | - |
| FCERI mediated Ca+2 mobilization | 63/116 | 0/17 | 1.03E-04 | 9.61E-03 | 1.03E-04 | 9.56E-03 | - | - |
| GRB2 events in EGFR signaling | 6/8 | 0/2 | 1.55E-04 | 1.38E-02 | 1.55E-04 | 1.37E-02 | - | - |
| Glyoxylate metabolism and glycine degradation | 26/31 | 0/43 | 1.66E-04 | 1.40E-02 | 1.66E-04 | 1.40E-02 | - | - |
| Utilization of Ketone Bodies | 5/5 | 0/10 | 1.88E-04 | 1.52E-02 | 1.88E-04 | 1.51E-02 | - | - |
| Role of LAT2/NTAL/LAB on calcium mobilization | 47/101 | 0/8 | 2.09E-04 | 1.62E-02 | 2.09E-04 | 1.61E-02 | - | - |
| FCERI mediated MAPK activation | 62/118 | 0/7 | 2.53E-04 | 1.88E-02 | 2.53E-04 | 1.87E-02 | - | - |
| Metabolism of folate and pterines | 15/17 | 0/26 | 2.83E-04 | 1.99E-02 | 2.83E-04 | 2.00E-02 | - | - |

**Table A.5:** Twenty five most significantly differentially regulated pathways with respect to fasting plasma glucose level.

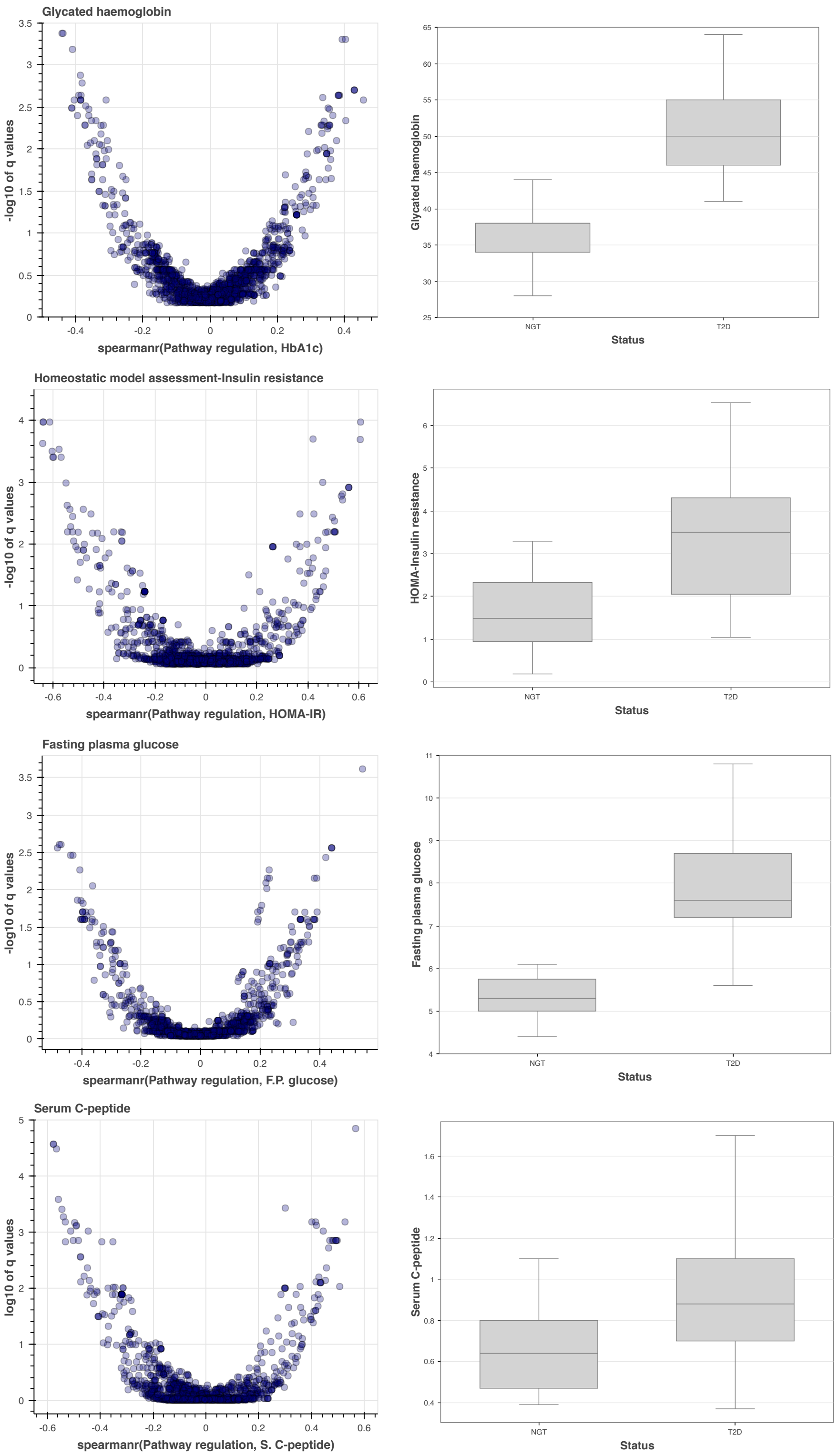

**Figure A.1:** The plots on the left show the relationship between significance of differential regulation of a pathway with the extent of differential regulation, for clinical variables in the analysis. The plots on the right show the the relationship between clinical variables and type II diabetes status. The box and whiskers are quartiles.

| Clinical variable | Unit | Average ± SEM | |
|---|---|---|---|
| | | T2D group | NGT group |
| Patient count | - | 25 | 24 |
| Age | *years* | 62 ± 1.5 | 58 ± 2.2 |
| Weight | *kg* | 88 ± 1.7 | 83.5 ± 2.2 |
| BMI | $kg/m^2$ | 27.3 ± 0.5 | 26.8 ± 0.4 |
| Body Fat | % | 24 ± 1 | 22.8 ± 1 |
| M-index | *µmol/(kg*min)* | 15.3 ± 1.9 | 34.2 ± 3.1 |
| HbA1c | *mmol/mol* | 51 ± 1.2 | 37 ± 0.7 |
| HOMA-IR | - | 3.6 ± 0.4 | 1.8 ± 0.3 |
| F.P. glucose | *mmol/L* | 7.8 ± 0.3 | 5.4 ± 0.1 |
| S. C-peptide | *nmol/L* | 0.89 ± 0.1 | 0.67 ± 0.05 |

**Table A.6:** Summary of clinical data for T2D and NGT groups

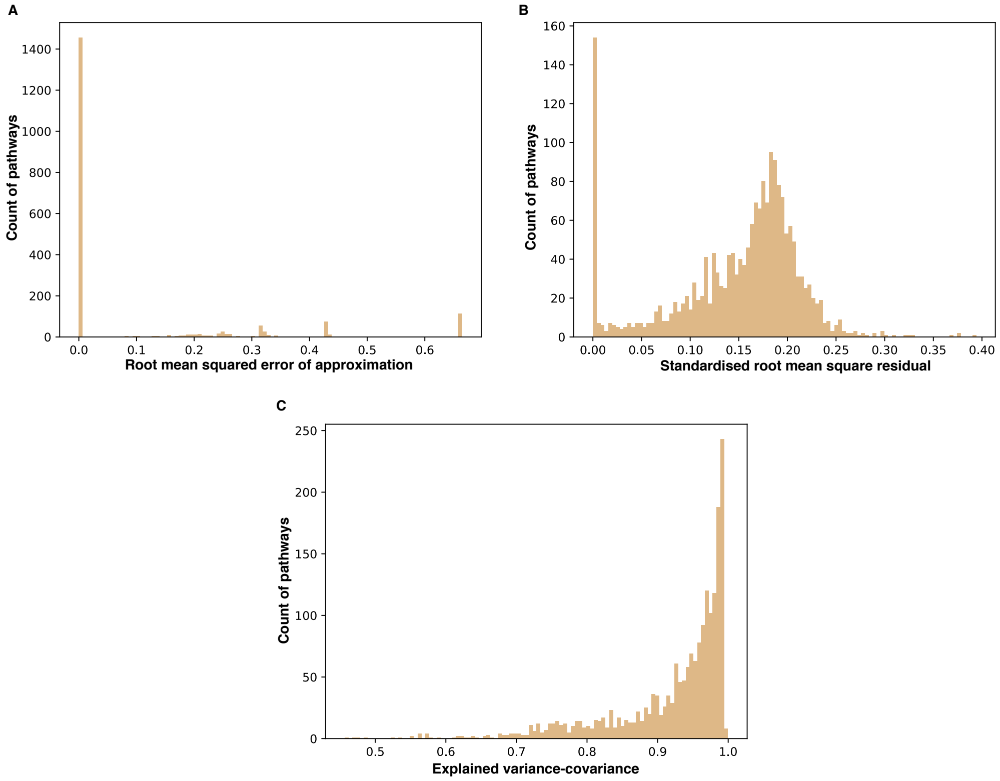

**Figure A.2:** Goodness of fit statistics were calculated for each pathway model. **A.** Root mean squared error of approximation is based upon the chi-square statistic. Values are in the range(0,1) with lower values implying a better fit. **B.** Standardised root mean squared residual is based on the sum of squares of difference between the sample co-variance matrix and the modelled covariance matrix. Values are in the range(0,1) with lower values implying a better fit. **C.** Explained variance-covariance is the ratio of the root sum of squares of the modelled covariance matrix with the sample covariance matrix. Values are in the range(0,1) with higher values implying a more explanatory model. For formulaic details consult “Confirmatory Factor Analysis” [11]

# B. Computing environment

**System: Kernel:** 4.4.0-17134-Microsoft x86_64 (64 bit)
**Distro:** Ubuntu 16.04 xenial
**CPU:** Hexa core Intel Core i7-8700K (-HT-MCP-) cache: 256 KB
clock speeds: max: 3696 MHz 1: 3696 MHz 2: 3696 MHz 3: 3696 MHz 4: 3696 MHz 5: 3696 MHz 6: 3696 MHz 7: 3696 MHz 8: 3696 MHz 9: 3696 MHz 10: 3696 MHz 11: 3696 MHz 12: 3696 MHz
**Partition:** ID-1: / size: 238G used: 174G (74%) fs: lxfs dev: N/A
**Info:** Processes: 5 Uptime: 0 min Memory: 5675.1/16330.9MB
**Client:** Shell (bash)
**Package list:** Python 3.6.5 was installed using Anaconda 5.1. The full package list can be found at,
https://github.com/aron0093/integrated_pathway_analysis/blob/master/package_list.txt
**Random state:** 0 for all such options.