Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 706 – 718

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

# Quantitative Trait Specific Differential Expression (qtDE)

Mrityunjay Sarkar[a], Aurpan Majumder[b,*]

[a]*Dept. of E.C.E, D.I.A.T.M, Durgapur 713212, INDIA*
[b]*Dept. of E.C.E, N.I.T Durgapur, Durgapur 713209, INDIA*

**Abstract**

Natural selection based on phenotypic traits may lead to genetic changes. To understand the relationship between phenotypic traits and gene expression profiles it is important to study the concordance between the two. In this work we have developed a simple procedure to find the differentially expressed (DE) genes across various tissues between phenotypes through linear correlation as well as non linear mutual information and polynomial regression between quantitative-trait and the gene expression profiles. Here we are making the use of mice gene expression data to find the differentially expressed genes between the male and female phenotypes exploring the dependency between the gene expression profiles of four tissues (brain, muscle, liver, adipose) and quantitative trait (weight). To prove the effectiveness of the method we have tested our results with a popular DE tool (DEGseq). In the results we have shown that mutual information based trait-specific DE genes are biologically more significant compared to the polynomial regression and linear correlative counterparts.


Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

*Keywords:* Quantitative trait; Differentially expressed genes; Correlation; MutualInfoAdjacency; T-statistics; Pathway

## 1. Introduction

Natural selection is the natural variation in the average reproductive success across phenotypes. Natural selection through genotypic variation (heritable variation) acts as a key factor towards evolution. So we can claim indirectly

* Corresponding author. Tel.: +91-343-2754388; fax: +91-343-2547375
*E-mail address:* aurpan.nitd@gmail.com


Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.131

that evolution via natural selection can drive ecological changes across phenotypes (sometime across different sex also)[1]. Some recent microarray studies puts in a great effort to find out the genes which are differentially expressed (DE) across different developmental stages through some specific patterns[2] . Till date different methods that have gained importance to find out the DE genes are single slide method[3], multiple slide method[4], Apo-AI and SR-BI method[3], and DEGseq[5] respectively.

All these methods have tried to find out DE genes at the genotype level (from the gene expression profile itself). In our work we have used the concept of differential dependencies between quantitative trait and gene expression profiles to stress more on natural selection.

However, to date some literatures[6] have given the concept of cluster based segregation of genes (by traits) through linear correlative association.

In this work we have conducted two kinds of trait based significance measures, one using linear correlation[7] and the other with non linear mutual information and polynomial regression. By using these measures we have found the DE genes from a trait based concept (henceforth represented as qtDE). In other words, we can call it as trait specific differential coexpression. In this context we have performed the statistical significance test (Student T-test)[8] to find out DE genes in the entire dataset. Next we have used a renowned DE tool (DEGseq) to find the DE genes in the same dataset and compared our qtDE results with those obtained from DEGseq.

In the results we have shown that our method proves to be more promising than DEGseq not only in terms of number of DE genes (the number of qtDE genes is higher than DEGseq) but also the extra DE genes found by qtDE are biologically enriched in terms of KEGG pathway analysis[9]. We have also observed that the number of DE genes participating in biologically enriched pathways as well as the number of such significant pathways is far higher in mutual information and polynomial regression based trait specific measures compared to the linear correlative counterpart.

The rest of the paper is as follows. In the following section we have discussed about the methodology. A detailed view of the problem and its implementation on mice data[10,11] is given in the results. A brief description on some important pathways found in the analysis has also been discussed. At the end we conclude with further simulations that need to be conducted in order to understand the differential ranking significance.

## 2. Methods

In this work we have implemented the trait specific concept using three different measures. They are correlation[12], mutual information[13] and polynomial regression[14].

### *2.1 Algorithm*

Suppose we have two kinds of phenotypes for a particular tissue .Here we have computed the trait based significance of each gene in order to determine the DE genes across these phenotypes. Let the sample trait be $T=(T_1,.....T_M)$ , where M is the total number of samples (values) in the quantitative trait. Again, X is assumed to be the gene expression matrix and each individual gene represented by $x_i$ .

The proposed algorithm given below finds out the trait specific DE genes (qtDE) using all the three measures stated above. In this algorithm variables ExV1 and ExV2 represent the gene expression matrices across phenotype1 and phenotype2 respectively (N represents the total number of genes). T1 and T2 represent the quantitative trait vectors. We have used mice weight as a quantitative trait. $\beta$ is the soft threshold parameter [15].

Step1 of the algorithm is dedicated to compute the gene significance values across both phenotypes. As mentioned previously we have explored three different ways to understand the gene significance from a differential prospective. First one computes linear gene significance by correlative measure, second one through non linear mutual information and the third one does this via non linear polynomial regression. Here *LinCor* function computes the correlation (linear gene significance) between the expression profile of each gene and quantitative trait; whereas *NLinMI* and *NLinPR* functions compute the same by non linear mutual information based uncertainty and polynomial regression based measures respectively. Gene significance values for phenotype1and phenotype2 are stored in GS1 and GS2 respectively.

**ALGORITHM** . Trait Specific Differential Gene Significance

**Input:** ExV1, ExV2, T1, T2, GN, $\beta \leftarrow 1$
**Output:** GS1, GS2, GS, TcdV, DE

```
Step1. s ← choose mode of computation
       for i in 1 : N do
          r1 ← rbind (ExV1[i,],T1)
          nr1 ← transpose (r1)
          r2 ← rbind (ExV2[i,],T2)
          nr2 ← transpose (r2)
        if s == 1 then
          GS1[i] ← LinCor (nr1, β )
          GS2[i] ← LinCor (nr2, β )
        else if s == 2 then
          GS1[i] ← NLinMI (nr1, β )
          GS2[i] ← NLinMI (nr2, β )
        else
          GS1[i] ← NLinPR (nr1, β )
          GS2[i] ← NLinPR (nr2, β )
        end if
       end for

Step2.  GS ← GS1 - GS2

Step3.  TcdV ← qt ((GS), set degree of freedom)

Step4.  ThV ← mean (TcdV)

Step5. for i in 1 : N do
          if TcdV[i] > ThV then
             s ← s+1
             DE[s] ← GN[i]
          end if
       end for

%%%%%%%% End of Main Program %%%%%%%%%
LinCor ← function (nr, β )
{
V ← cor(nr,set correlation method) ^ β
}

NLinMI ← function (nr, β )
{
V ← mutualInfoAdjacency (nr,discretize columns, set entropy estimation method, set the number of discretization bins) ^ β
}

NLinPR ← function(nr, β )
{
V ← adjacency.polyreg (nr,set the degree of polynomial, specify the method to symmetrise the pairwise model fitting index matrices) ^ β
}
```

Step2 computes the difference between gene significance values across the phenotypes. The corresponding result is stored in GS.

Step3 of the algorithm performs the T-statistics probability distribution of the gene significance difference values

obtained from step2 in order to compute the CDF values across each gene in the entire distribution. The result is stored in a variable TcdV.

Step4 computes the mean of the T-statistics CDF values obtained from step3 and outputs a threshold level in order to determine the DE genes.

In Step5 by comparing the T-statistics CDF (TcdV) values obtained from step3 with the threshold value from Step4 we come to know about genes differentially expressed between the phenotypes. If the TcdV value of a particular gene is greater than the threshold, it suggests a revealing difference of the significance values of the corresponding gene between phenotypes. Accordingly, this gene happens to be a DE gene.

## 3. Results

Implementation of the above mentioned algorithm is conducted on a publicly available dataset[10,11]. Details about the dataset, microarray analysis and data reduction by preprocessing has been given in[10].The dataset contains gene expression values of male and female mice across four types of tissues viz. brain, muscle, liver and adipose. Mice weight has been considered as a parameter of quantitative trait. Genes in all the four regions are same.

The sole intention of this work is to find genes that are differentially expressed in all these four regions across the male and female phenotypes.

Throughout this process we have taken the value of $\beta$ as 1.

The dataset contains 3600 genes in each of the four tissues. Mice weight (quantitative trait) has been given for both the male and female phenotypes. First of all, we segregate the mice-weight across male and female. Next according to mice –ID and strain we redistribute the mice-weight amongst the four tissues for male and female.

### *3.1. Linear method*

In this analysis we are following *LinCor* function given in the algorithm. This user defined function as mentioned in the algorithm invokes another function *cor* associated with the R package named *WGCNA* [16] to compute the linear gene significance by correlative measure. The function *cor* corresponds to the Pearson correlation operation being performed between each gene of a particular tissue and the redistributed mice weight of that tissue for both phenotypes. In our work, we have considered this to be the linear gene significance (*GS1* for male and *GS2* for female). Thereafter, we proceed through the remaining steps of the algorithm for the prediction of DE genes.

Following the above mentioned method we have found 837,856, 1132 and 579 qtDE genes in liver, adipose, muscle and brain respectively along with 213 common qtDE genes amongst these tissues. So, we can assume that with respect to weight, these genes are responsible for the evolution of two different sexes (male and female).

### *3.2. Non-linear method*

#### *3.2.1 Mutual information based approach*

The procedure to compute non linear mutual information based gene significance is quite similar to the linear one. As mentioned in the algorithm here we make the use of another user defined function *NlinMI* which computes the non linear gene significance by a symmetric uncertainty based mutual information adjacency measure. In this case it takes into consideration the R function mutualInfoAdjacency (as an entropy estimation method we are using maximum likelihood estimators with Miller-Madow bias correction) associated with the R package WGCNA[16] . This operation is similarly performed on each gene and the redistributed mice weight pair of the tissue under consideration for both phenotypes. The output reflects the idea of non linear gene significance (GS1 for male and GS2 for female).

Following the remaining sequence of operations we found that the qtDE genes in liver, adipose, muscle and brain are1479, 1236, 2503, and 1499 respectively and the common set of qtDE genes amongst these tissues equals 705.

Table 1. Significant pathways through linear correlative, non linear mutual informative and polynomial regression measure by the common qtDE genes across all tissues

| Nonlinear Method | | | | | | Linear Method | | |
|---|---|---|---|---|---|---|---|---|
| Mutual Information | | | Polynomial Regression | | | Correlation | | |
| Pathways | p-value | genes | Pathways | p-value | genes | Pathways | p-value | genes |
| Olfactory transduction | 6.6E-08 | 7 | Metabolic pathways | 1.8E-04 | 26 | Cell cycle | 1.1E-03 | 6 |
| Leukocyte transen-dothelial migration | 1.9E-05 | 14 | Nucleotide excision repair | 2.3E-04 | 4 | Chronic myeloid leukaemia | 2.8E-03 | 3 |
| Complement and coagulation cascades | 2.1E-03 | 8 | Glutathione metabolism | 4.6E-04 | 4 | Ether Lipid metabolism | 6.2E-03 | 2 |
| Ether Lipid metabolism | 3.4E-03 | 5 | Ether Lipid metabolism | 8.2E-04 | 3 | | | |
| Glycerolipid metabolism | 4.4E-03 | 6 | DNA replication | 9.7E-04 | 3 | | | |
| Glysero phospholipid metabolism | 9.6E-03 | 7 | Long Term Depression | 1.1E-03 | 4 | | | |
| | | | Mismatch repair | 3.1E-03 | 2 | | | |
| | | | Vascular smooth muscle contraction | 6.2E-03 | 4 | | | |

Table2. Significant pathways by the 9 common qtDE genes found between linear and nonlinear methods across all tissues

| Pathways | p-value | Gene names |
|---|---|---|
| Ether Lipid metabolism | 1.6E-03 | 2(***pla2g7,pld2***) |

### *3.2.2 Polynomial regression based approach*

Here, we proceed utilising the user defined function *NlinPR* which is there to compute non linear gene significance by polynomial regression. In this case to compute the measure *NlinPR* we perform the operation *adjacency.polyReg* (it calculates a network adjacency matrix by fitting polynomial regression models to pairs of variables) associated with the R package *WGCNA*[16] . Like the previous two cases this operation is also performed between each gene of a particular tissue and the redistributed mice weight of the same for both phenotypes. Thus we end up with another non linear gene significance measure (*GS1* for male and *GS2* for female).

Accordingly, in search of DE genes we have found 1395, 938, 1163, and 675 qtDE genes in liver, adipose, muscle and brain respectively with 364 common qtDE genes amongst these tissues.

Table 3.Significant pathways across different tissues through correlation, mutual information and polynomial regression based approach excluding the common genes

| Correlation | | | Mutual Information | | | Polynomial Regression | | |
|---|---|---|---|---|---|---|---|---|
| | | | ADIPOSE | | | | | |
| Pathways | p-value | genes | Pathways | p-value | genes | Pathways | p-value | genes |
| Olfactory transduction | 1.31E-06 | 12 | Olfactory transduction | 2.22E-11 | 9 | Olfactory transduction | 2.27E-17 | 12 |
| Cell cycle | 7.2E-05 | 24 | Leukocyte transendothelial migration | 2.82E-03 | 13 | Amoebiasis | 9.15E-04 | 8 |
| Cytokine-cytokine receptor interaction | 1.5E-03 | 2 | Leishmaniasis | 9.17E-03 | 8 | Cytokine-cytokine receptor interaction | 9.16E-04 | 9 |
| Fc gamma R-mediated phagocytosis | 1.8E-03 | 18 | Glutathione metabolism | 9.3E-03 | 7 | Steroid biosynthesis | 9.16E-04 | 7 |
| Chagas disease | 6.4E-03 | 17 | Cell adhesion molecules | 9.7E-03 | 7 | Arginine and proline metabolism | 8.8E-03 | 3 |
| | | | | | | Nitrogen metabolism | 8.8E-03 | 7 |
| | | | BRAIN | | | | | |
| Pathways | p-value | genes | Pathways | p-value | genes | Pathways | p-value | genes |
| Olfactory transduction | 2.29R-09 | 6 | Olfactory transduction | 2.41E-15 | 12 | Olfactory transduction | 1.2E-09 | 9 |
| Pyruvate metabolism | 7.9E-03 | 4 | Amoebiasis | 3.36E-05 | 28 | Metabolic pathways | 6.3E-05 | 34 |
| Metabolic pathways | 9.6E-03 | 3 | Leishmaniasis | 2.73E-05 | 14 | Maturity onset diabetes of the young | 3.6E-04 | 8 |
| Complement and coagulation cascades | 9.97E-03 | 4 | Focal adhesion | 2.89E-04 | 49 | Amino sugar and nucleotide sugar metabolism | 1.2E-03 | 10 |
| | | | Cytokine-cytokine receptor interaction | 3.23E-04 | 43 | Fc gamma R-mediated phagocytosis | 4.7E-03 | 13 |
| | | | | | | Insulin signaling pathway | 5.5E-03 | 8 |
| | | | | | | Focal adhesion | 7.9E-03 | 23 |
| | | | LIVER | | | | | |
| Pathways | p-value | genes | Pathways | p-value | genes | Pathways | p-value | genes |
| Olfactory transduction | 7.23E-07 | 8 | Leishmaniasis | 2.51E-04 | 12 | Cytokine-cytokine receptor interaction | 8.1E-05 | 31 |
| Cytokine-cytokine receptor interaction | 2.84E-05 | 32 | Focal adhesion | 4.05E-04 | 25 | Chagas disease | 2.4E-04 | 14 |
| Complement and coagulation cascades | 9.1E-03 | 13 | Amoebiasis | 1.6E-03 | 16 | Focal adhesion | 2.47E-04 | 27 |
| Hematopoietic cell lineage | 9.11E-03 | 14 | ECM- receptor interaction | 4.1E-03 | 12 | Hematopoietic cell lineage | 1.34E-03 | 13 |
| | | | Malaria | 4.1E-03 | 10 | Glutathione metabolism | 4.2E-03 | 10 |
| | | | | | | Arginine and proline metabolism | 4.8E-03 | 18 |

| | | | MUSCLE | | | | | |
|---|---|---|---|---|---|---|---|---|
| Pathways | p-value | genes | Pathways | p-value | genes | Pathways | p-value | genes |
| Olfactory transduction | 8.7E-11 | 4 | Olfactory transduction | 3.61E-21 | 19 | Olfactory transduction | 8.58E-12 | 8 |
| Focal adhesion | 3.7E-07 | 37 | Cytokine-cytokine receptor interaction | 7.4E-07 | 50 | Focal adhesion | 1.4E-05 | 12 |
| Metabolic pathways | 1.3E-03 | 48 | Focal adhesion | 8.9E-06 | 48 | | | |
| Nitrogen metabolism | 7.3E-03 | 8 | Amoebiasis | 7.6E-04 | 24 | Amoebiasis | 1.01E-04 | 23 |
| Type-II diabetes mellitus | 7.34E- 03 | 12 | Hematopoietic cell lineage | 6.3E-03 | 21 | Arginine and proline metabolism | 1.6E-03 | 13 |
| | | | | | | Cell cycle | 1.64E-03 | 23 |
| | | | | | | Chagas disease | 1.64E-03 | 20 |

In our experimentation KEGG Pathway[9] analysis of the genes reveals that the biological enrichment of the pathways by non linear methodologies are far better than the linear domain both in terms of p-value[17] as well as in the number of participating genes. Pathways having p-value at least 1E-03 and minimum 2 genes are considered to be significant.

Table 1 highlights the significant pathways by the common qtDE genes amongst brain, muscle, liver, and adipose individually for the linear (213 common DE genes) and non linear (705 common DE genes by mutual information and 364 by polynomial regression) processes.

Table 2 showing a significant biological pathway enrichment enlists those genes which are not only common among all the three methods but also amongst the four tissues. In this context we have obtained 9 common genes. A crucial pathway from these 9 genes is being depicted in this table.

The above analysis with common qtDE genes depicts the fact that *Ether Lipid Metabolism* happens to be the only notable pathway and that too with just 2 genes (*pla2g7*and *pld2*). In this connection it is noteworthy to mention that we do observe the biological enrichment of the pathways not having these 2 genes to be better than *Ether Lipid Metabolism*.

Table 4. Significant pathways through DEGseq excluding the 59 common genes

| Region | Pathways | p-value | Genes |
|---|---|---|---|
| | Leishmaniasis | 1.05E-05 | 9 |
| | Amoebiasis | 4.7E-04 | 11 |
| ADIPOSE | TGF-beta signalling pathway | 1.07E-03 | 9 |
| | Olfactory transduction | 5.46E-03 | 16 |
| | Jak-Stat signalling pathway | 5.8E-03 | 11 |
| | Fc gamma R-mediated phagocytosis | 8.8E-03 | 8 |
| | Tight junction | 7.43E-05 | 9 |
| | p53 signalling pathway | 1.5E-05 | 4 |
| BRAIN | Fc gamma R mediated phagocytosis | 1.6E-03 | 4 |
| | Glycerolipid metabolism | 1.6E-03 | 4 |

| | | | |
|---|---|---|---|
| LIVER | Maturity onset diabetes of the young | 6.03E-06 | 3 |
| | Galactose metabolism | 4.033E-05 | 5 |
| | Olfactory transduction | 7.43E-04 | 6 |
| | Cytokine-cytokine receptor interaction | 1.01E-03 | 9 |
| | Focal adhesion | 3.03E-03 | 7 |
| | TGF-beta signaling pathway | 4.25E-03 | 4 |
| MUSCLE | Galactose metabolism | 2.48E-04 | 3 |
| | Focal adhesion | 3.92E-04 | 7 |
| | Olfactory transduction | 8.4E-04 | 3 |
| | Glycosaminoglycan biosynthesis – keratan sulfate | 9.1E-04 | 3 |
| | Fc gamma R-mediated phagocytosis | 1.81E-03 | 4 |
| | Chagas disease | 1.64E-03 | 7 |

To gain a better insight we have shown in table 3 that excluding these 9 common genes the biological enrichment of pathways formed across the different tissues are better via non linear interactions. To be specific, the crucial pathways formed by the DE genes making use of mutual information based differential dependency are better than polynomial regression which further outperform the ones obtained via correlative measure.

As a next step to estimate the efficiency of our algorithm we have compared our results with the well established DE tool called DEGseq It is an R package to find out the differentially expressed genes from RNA-seq data used to provide gene expression measurement. In this package depending upon expression values of genes at different samples/time instants and by setting a particular p-value/ z-score/ q-value threshold, DE genes are computed. This comparison is performed to check the effectiveness of our method, though the procedures to find DE genes are different. While qtDE uses sample traits along with gene expression data to determine a DE gene, DEGseq makes the use of gene expression data only.

Table 5. Significant diseases formed by the common DE genes between our method (qtDE) and DEGseq along with the mutually exclusive sets of DE genes with respect to qtDE (Case1) and DEGseq (Case2) (the mutually exclusive sets of DE genes are given in bold)

| Pathways | Method & Organ | Case1 (Common DE + mutually exclusive qtDE) | | Case2 (Common DE + mutually exclusive DEGseq) | |
|---|---|---|---|---|---|
| Autoimmune thyroid disease | | p-value | Genes | p-value | Genes |
| | **Correlation** ADIPOSE | 2.01E-03 | 5(*Cd86, H2-DMa, H2-T10, H2-Ab1,* ***H2-DMb1***) | 1.3E-02 | 6(***H2-Aa***, ***H2-Q8***, *Cd86, H2-DMa, H2-T10, H2-Ab1*) |
| | **Correlation** MUSCLE | 3.43E-02 | 4(*H2-Ab1, Tnf, H2-Aa, H2-Eb1*) | 6.3E-03 | 8(*H2-Eb1,* ***Cd86***, *H2-Aa,* ***H2-Q8***, ***H2-DMa***, ***H2-DMb1***, *H2-Ab1, Tnf*) |
| | **Polynomial Regression** ADIPOSE | 2.3E-03 | 6(H2-*DMa, Tnf, H2-DMb1,Cd86, H2-Aa, H2-Eb1*) | 5.8E-03 | 8(*Ifng,* ***H2-T10***, *H2-DMa, Tnf, H2-DMb1, Cd86, H2-Aa,* ***H2-Eb1***) |

| | | | | | |
|---|---|---|---|---|---|
| Cardiac muscle contraction | **Polynomial Regression** MUSCLE | 1.06E-03 | 10(*Tpm3,Cacnb1, Myl3, Slc8a1, Cox6a2, Actc1, Cacna2d1, Tpm1, Cox7a1, Cox7a2*) | 7.9E-04 | 12(***Cox7b***, ***Myh7***, *Tpm3, Cacnb1, Myl3, Slc8a1, Cox6a2, Actc1, Cacna2d1, Tpm1, Cox7a1, Cox7a2*) |
| Dilated cardiomyopathy | **Polynomial Regression** MUSCLE | 1.7E-02 | 8(*Tpm3, Cacnb1, Myl3, Slc8a1, Actc1, Cacna2d1, Tpm1, Actb*) | 4.08e-03 | 11(***Itga8***, ***Myh7***, ***Tnf***, *Tpm3, Cacnb1, Myl3, Slc8a1, Actc1, Cacna2d1, Tpm1, Actb*) |
| Graft-versus-host disease | **Correlation** ADIPOSE | 2.01E-03 | 5(*Cd86, H2-DMa, H2-T10, H2-Ab1,* ***H2-DMb1***) | 3.3E-03 | 7(***H2-Aa***, ***H2-Q8***, ***Il1b***, *Cd86, H2-DMa, H2-T10, H2-Ab1*) |
| | **Correlation** MUSCLE | 1.4E-02 | 6(*H2-Aa, H2-Q8, H2-DMa, H2-DMb1, H2-Ab1, Tnf*) | 6.9E-03 | 8(***H2-Eb1***, ***Cd86***, *H2-Aa, H2-Q8, H2-DMa, H2-DMb1, H2-Ab1, Tnf*) |
| Drug metabolism cytochrome P450 | **Correlation** LIVER | 4.14E-03 | 6(*Cyp2c40,Gstm2, Mgst2, Cyp2d10, Ugt1a9, Mgst3*) | 1.7E-03 | 9(***Fmo3***,***Cyp2d22***, ***Gsta2***, *Cyp2c40, Gstm2, Mgst2, Cyp2d10, Ugt1a9, Mgst3*) |
| | **Mutual Information** BRAIN | 1.3E-02 | 3(*Cyp2b9,Cyp2c55, Gstm1*) | 1.3E-02 | 3(*Cyp2b9,Cyp2c55, Gstm1*) |
| | **Polynomial Regression** LIVER | 6.8E-03 | 9(Gsta2*, Cyp2d22, Cyp2c40, Gstm2, Mgst2, Cyp2d10, Mgst3, Cyp2c54, Fmo3*) | 2.1E-02 | 10(***Ugt1a9***, *Gsta2, Cyp2d22, Cyp2c40, Gstm2, Mgst2, Cyp2d10, Mgst3, Cyp2c54,Fmo3*) |
| Proteasome | **Polynomial Regression** MUSCLE | 6.3E-03 | 6(*Psma4, Ifng, Psmb3, Psmc6, Psmb9, Psma2*) | 2.3E-02 | 6(*Psma4, Ifng, Psmb3, Psmc6, Psmb9, Psma2*) |
| Viral myocarditis | **Correlation** ADIPOSE | 1.4E-02 | 5(*Cd86, H2-DMa, H2-T10, H2-Ab1,* ***H2-DMb1***) | 3.3E-03 | 9(***Rac2***, ***H2-Aa***, ***H2-Q8***, ***Itgal***, ***Casp3***, *Cd86, H2-DMa, H2-T10, H2-Ab1*) |
| | **Correlation** MUSCLE | 1.3E-02 | 8(*H2-Aa, H2-Q8, Myh7, H2-DMa, H2-DMb1, H2-Ab1, Fyn, Itgal*) | 5.1E-03 | 11(***Rac2***, ***H2-Eb1***, ***Cd86***, *H2-Aa, H2-Q8, Myh7, H2-DMa, H2-DMb1, H2-Ab1,Fyn, Itgal*) |
| | **Polynomial Regression** ADIPOSE | 2.1E-02 | 6(H2-*DMa, H2-DMb1, Cd86, H2-Aa, H2-Eb1, Myh2*) | 9.03E-03 | 9(***Casp3***, ***H2-T10***, ***Itgal***, *H2-DMa, H2-DMb1, Cd86, H2-Aa, H2-Eb1, Myh2*) |
| Metabolism of xenobiotic by cytochrome P450 | **Correlation** LIVER | 1.06E-02 | 5(*Cyp2c40, Gstm2, Mgst2, Ugt1a9, Mgst3*) | 3.4E-02 | 6(***Gsta2***, *Cyp2c40, Gstm2, Mgst2, Ugt1a9, Mgst3*) |
| | **Mutual Information** BRAIN | 1.4E-02 | 3(*Cyp2b9,Cyp2c55, Gstm1*) | 1.4E-02 | 3(*Cyp2b9,Cyp2c55, Gstm**1***) |

The DE genes found by DEGseq in adipose, brain, liver and muscle are 732, 373, 424 and 301 respectively. We have found 59 DE genes common across the four organs via DEGseq. In table 4 we have enlisted the significant pathways formed by the DE genes excluding the 59 common ones. While comparing the results obtained by our method and DEGseq we have noticed that in adipose between 856 qtDE genes (found by correlative measure) and 732 DE genes 584 genes are common. Similarly, between 938 qtDE genes (polynomial regression based measure)

and 732 DE genes 498 genes are common, but all the 732 DE genes become a subset of 1236 qtDE genes found by mutual information based measure. In brain all the 373 DE genes discovered by DEGseq are common in 579 (correlative), 1409 (mutual information), and 675 (polynomial regression) qtDE genes. In liver between 424 DE and 837 (by correlative measure) qtDE genes 379 genes are common. Again between 424 DE and 1395 (using polynomial regression) qtDE genes 392 genes are common, but all the 424 DE genes happens to be a subset of 1479 qtDE genes found by mutual information based approach. In muscle between 1132 qtDE (by correlative measure) and 301 DE genes 283 genes are common. Again between 301 DE and 1163 (by polynomial regression measure) qtDE genes 275 genes are common. Here also all the 301 DE genes come within the set of 2503 qtDE genes discovered by mutual information.

To indulge further we have taken the mutually exclusive DE genes present in qtDE and performed KEGG pathway analysis of the same. In this connection we have gone through some significant disease related KEGG pathways having well defined differential functionalities between different sexes of mice. We checked the enrichment of these pathways comprising of the mutually exclusive qtDE genes on one side and the same set of qtDE genes in addition to the disjoint set of DE genes acquired from DEGseq on the other (like in adipose only 584 genes are common between the correlative qtDE genes and DEGseq, so the remaining 732-584 = 148 genes are mutually exclusive in DEGseq).

Table 5 depicts the enrichment analysis of the aforementioned pathways. Comparing the results we do not observe any significant improvement in the p-value after adding the disjoint set of DE genes with respect to DEGseq. For certain cases there is null refinement (*Drug metabolism cytochrome P450* in case of mutual information based differential interaction in brain, *Proteasome* in case of polynomial regression based differential interaction in muscle and *Metabolism of xenobiotic by cytochrome P450* in case of mutual information based differential interaction in brain). Lastly in *Autoimmune thyroid disease* (earned from correlative based differential interaction in adipose) the results rather deteriorate after adding an extra gene from DEGseq. Thus we can claim a promising role of qtDE over DEGseq in the context of abnormal cell development.

## 4. Discussion

*Ether lipid metabolism* is the only significant pathway that we obtain from the common qtDE genes amongst the four tissues and between linear and non linear phenotypic trait based interactions. This can be observed from both the tables 1 and 2. Importance of this pathway in mice and other related primates has been discussed in[18] which shows the significant involvement of the same in tumor cell invasiveness, energy storage, signaling molecules and in cardiovascular disease.

From table 1 we do observe another important pathway i.e. *Olfactory transduction*. Tables 3 and 4 also highlight this pathway to be significant. The observations in this regard do confirm the significant contribution of the partially disjoint set of DE genes (i.e. not taking into consideration the 9 common DE genes) individually for the four tissues with respect to linear and non linear phenotype interactive approaches. The role of this pathway affecting different tissues of mice is discussed in[19] which clarifies the level of association with the development of obesity in both adipose and muscle tissues. Again[20] reveals the role of the pathway in connection with functioning of olfactory sensory neurons (OSN) in the septal tissue and in[19] as functioning of rodent olfactory epithelium on liver.

From tables 1 and 3 we can observe another expressive pathway i.e. *Leukocyte transendothelial migration*. It is found to be significant exclusively via the non linear mutual information based method in adipose tissue. It is involved in blood-brain barrier (BBB) which plays a critical role in central nervous system (CNS) homeostasis[20]. It also shows active involvement in junctional adhesion molecule[21] and pathogenesis of inflammation [22]. Tables 3 and 4 highlight *Amoebiasis* and *Focal adhesion* as significant pathways. Function of these pathways in mice has been given in[23, 24] . In this context from tables 1 and 3 another impressive pathway to be mentioned is *Complement and coagulation cascades*. Here we can detect the significant biological existence of the same in liver and brain. It plays a significant role in the pathogenesis of cardiovascular disease[25], Huntington disease[26], and in liver development[27]. Accordingly, we are able to interpret the contributory role of the quantitative trait specific approach using mutual information.

Other significant pathways observed from table 1 are *Glycerolipid metabolism* (also present in table 4) found to be involved in liver and aging cerebellum[28] as well as in neural oxidative metabolism[29] followed by *Glycerophospholipid metabolism* associated with energy storage and signaling molecules in mice/rat[30]. Additionally

pathways such as *Chronic myeloid leukemia* and *Nucleotide excision repair* taking part in liver, muscle and adipose tissue have been discussed in[31, 32].

Pathways generated by DE genes found exclusively significant through the non linear mutual information based trait specific method are given in second column of table 3. They are *Leishmaniasis* (present in all tissues under study), and *Cytokine-cytokine receptor interaction* (present in muscle and brain). Significance of these pathways across different tissues are discussed thoroughly in [33, 34]. From table 4 we can have an idea of equivalent significance of these pathways using DEGseq model.

First column of table 3 gives us enriched pathways formed by the exclusive DE genes through linear correlative method. These are *Cell cycle* (present in the adipose tissue), *Pyruvate metabolism* (present in the brain tissue), *Metabolic pathways* (present in brain tissue), and *Nitrogen metabolism* (present in muscle tissue). Significance of these pathways is discussed in [35, 36, 37, 38] .

Third column of table 3 exclusively shows some eloquent pathways by polynomial regression based method. The notable ones are *Chagas disease* (present in liver, can also be seen in table 4 but related to muscle), *Arginine and proline metabolism* (present in adipose, liver and muscle) and *Cytokine-cytokine receptor interaction* (present in liver and adipose). Essence of these pathways has been discussed in[39, 40]. In this context another crucial pathway present in table 3 as well as in table 4 is *Fc gamma R-mediated phagocytosis,* the essence of which has been discussed in[41].

Biological significance and differential functionalities of different disease related pathways enlisted in table 5 have been discussed in[42, 43, 44, 45, 46, 47].

## 5. Conclusion and Future work

In this work we have found the qtDE genes between phenotypes using the idea of quantitative trait. Till date existing methods have focused only on the gene expression value to find DE genes. In this aspect we have developed a novel procedure by incorporating quantitative trait along with gene expression value to compute qtDE genes. Here, we have extended the concept of gene significance[7] to find the qtDE genes.

Methodologically we have found the gene significance by linear correlative method as well as by non linear mutual information adjacency and polynomial regression methods. Then by applying student's T-distribution over the difference of gene significance values between two phenotypes and checking the T-statistics value with respect to some threshold we get our qtDE genes.

We have tested our algorithm on the gene expression data of mice. Our aim was to find genes differentially expressed across male and female mice in brain, muscle, liver and adipose tissues taking weight as a parameter for quantitative trait. In the results, we have shown that not only the number of DE genes across these four organs by non linear method are higher than the linear counterpart but also biological enrichment via pathway analysis of the DE genes is far more prominent through non linear trait specific interactions in terms of p-value as well as by the number of such genes participating in a pathway.

We can also claim that pathways having DE genes found by the non linear methods show a broader variety of important functionalities compared to those constituted of DE genes found by the linear correlative method.

We can extend this work over more organisms/tissues to crosscheck the properties established in this work. Next, in order to mathematically validate the biological significance/enrichment we can further rank[48] the DE genes found via linear and non linear methods, specifically to check whether the DE genes found to be biologically significant with these techniques do possess significant ranking too.

## References

1. Johnson MTJ, Vellend M, Stinchcombe JR. Evolution in plant populations as a driver of ecological changes in arthropod communities. *Phil.Trans. R. Soc.* 2009; **364**: p.1593-1605.
2. Khun E. From library screening to microarray technology: Strategies to determine gene expression profiles and to identify differentially regulated genes in plant. *Annals of Botany* 2000;**87**:139-55.
3. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002; **12**:111-39.
4. Gottardo R, Raftery AE, Yeung KY, Bumgarner RE. Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometric* 2006;**62**:10-8.

5. Zhang X, Wang X, Wang X, Wang L, Feng Z. DEGseq: an R package for identifying differentially expressed genes from RNA- seq data. *Bioinformatics* 2010; **26**:136-8.
6. Seo JH, Li Q, Fatima A, Eklund A, Szallasi Z, Polyak K, Richardson AL, Freedman ML. Deconvoluting complex tissues for expression quantitative trait locus-based analyses. *Philos Trans R Soc Lond B Biol Sci* 2013; **368**.
7. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology* 2008; **4**.
8. Available at http://www.mathworld.wolfram.com/ (last accessed on May 2014).
9. Kanehisa M, Goto S. KEGG: Kyoto encyclopaedia of genes and genomics. *Nucleic Acids Research.* 2000; **28**: p.27-30.
10. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. Weighted gene co-expression network analysis strategies applied to mouse weight. *Mamm Genome* 2007; **18**: 463-72.
11. Available at http:// www.genetics.ucla.edu/ labs/ horvath/CoexpressionNetwork/MouseWeight/ (Last accessed on February, 2014)
12. Fowler RL. Power and robustness in product-moment correlation. *Applied Psychological Measurement* 1987; **11**:419-28.
13. Paninski L. Estimation of entropy and mutual information, *Neural Computation* 2003; **15**:1191-253.
14. Chang YW, Hsieh CJ, Chang KW, Ringgaard M, Lin CJ. Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research* 2010; **11**: 1471-90.
15. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics* 2006; **2** :1182-92.
16. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**:559.
17. Gelman A. Commentary: P values and statistical practice. *Epidemiology* 2013; **24**: p. 69-72.
18. Zhang Y, Zou X, Ding Y, Wang H, Wu X, Liang B. Comparative genomics and functional study of lipid metabolic genes in caenorhabditis elegans. *BMC Genomics* 2013; **14**:164.
19. Choi Y, Hur CG, Park T. Induction of olfaction and cancer-related genes in mice fed a high-fat diet as assessed through the mode of action by network identification analysis. *PLoS One* 2013; **8**.
20. Oshimoto A, Wakabayashi Y, Garske A, Lopez R, Rolen S, Flowers M, Arevalo N, Restrepo D. Potential role of transient receptor potential channel M5 in sensing putative pheromones in mouse olfactory sensory neurons. *PLoS One* 2013; **8**.
21. Molinas A, Sicard G, Jakob I. Functional evidence of multidrug resistance transporters (MDR) in rodent olfactory epithelium. *PLoS One* 2012; **7**.
22. Valencia HA, Berdnikovs S, Cook-Mills JM. Mechanisms for vascular cell adhesion molecule-1 activation of ERK1/2 during leukocyte transendothelial migration. *PLoS One* 2011; **6**.
23. Sawangjaroen N, Sawangjaroen K, Poonpanang P. Effects of piper longum fruit, piper sarmentosum root and quercus infectoria nut gall on caecal amoebiasis in mice. *Journal of Ethnopharmacology* 2004; **91**:357-60.
24. Pandey AK, Somvanshi S, Singh VP. Focal adhesion kinase: An old protein with new roles. *Online Journal of Biological Sciences* 2012; **12**: 11-4.
25. Carter AM. Complement activation: An emerging player in the pathogenesis of cardiovascular disease. *Scientifica* 2012; **2012**: 402783.
26. Diamanti D, Mori E, Incarnato D, Malusa F, Fondelli C, Magnoni L, Pollio G. Whole gene expression profile in blood reveals multiple pathways deregulation in R6/2 mouse model. *Biomarker Research* 2013; **1**:28.
27. Thoren LA, Norgaard GA, Weischenfeldt J, Waage J, Jakobsen JS, Damgaard I, Bergstrom FC, Blom AM, Borup R, Bisgaard HC, Porse Bo T. UPF2 is a critical regulator of liver development, function and regeneration. *PLoS One* 2010; **5**.
28. Dwyer JR, Donkor J, Zhang P, Csaki LS, Vergnes L, Lee JM, Dewald J, Brindley DN, Atti E, Tetradis S, Yoshinaga Y, Jong PJD, Fong LG, Young SG, Reue K. Mouse lipin-1 and lipin-2 cooperate to maintain glycerolipid homeostasis in liver and aging cerebellum. *PNAS* 2012; **109**:2486-95.
29. Lee J, Wolfgang MJ. Metabolomic profiling reveals a role for CPT1c in neuronal oxidative metabolism. *BMC Biochemistry* 2012; **13**:23.
30. Hicks AM, DeLong CJ, Thomas MJ, Samuel M, Cui Z. Unique molecular signatures of glycerophospholipid species in different rat tissues analyzed by tandem mass spectrometry. *Biochimica et Biophysica Acta(BBA)-Molecular and Cell Biology of Lipids* 2006; **1761**: 1022-29.
31. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine* 2001; **344** : 1031-37.
32. Darwin KH, Nathan CF. Role for nucleotide excision repair in virulence of mycobacterium tuberculosis. *Infection and Immunity* 2005; **73**:4581-87.
33. Cruz I, Nieto J, Moreno J, Canavate C, Desjeux P, Alvar J. Leishmania/HIV co-infections in the second decade. *Indian J Med Res* 2006; **123**: 357- 88.
34. Patil A, Kumagai Y, Liang KC, Suzuki Y, Nakai K. Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response. *PLoS Computational Biology* 2013; **9**.
35. Blanchet E, Annicotte JS, Fajas L. Cell cycle regulators in the control of metabolism. *Cell Cycle* 2009; **8**: 4029-31.
36. Xu ZP, Wawrousek EF, Piatigorsky J. Transketolase haploinsufficiency reduces adipose tissue and female fertility in mice. *Molecular and Cellular Biology* 2002; **22**: 6142-47.
37. Jha MK, Jeon S, Suk K. Pyruvate dehydrogenase kinases in the nervous system: Their principal functions in neuronal-glial metabolic interaction and neuro-metabolic disorders. *Current Neuropharmacology* 2012; **10**: p.393-403.
38. Krappmann S, Braus GH. Nitrogen metabolism of aspergillus and its role in pathogenecity. *Medical Mycology Supplement* 2005;**43**:p.31-40.
39. Friere-de-Lima C, Pecanha LM, Dos Reis GA. Chronic experimental chagas disease: functional syngeneic T-B-cell cooperation in vitro in the absence of an exogenous stimulus. *Infection and Immunity* 1996; **64**: 2861-66.
40. Racke K, Warnken M. L-arginine metabolic pathways. *The Open Nitric Oxide Journal* 2010; **2**: p.9-19.
41. Menche J, Sharma A, Cho MH, Mayer RJ, Rennard SI, Celli B, Miller BE, Locantore N, Tal-Singer R, Ghosh S, Larminie C, Bradley G, Riley JH, Agusti A, Silverman EK, Barabasi A-L. A diVIsive Shuffling Approach (VIStA) for gene expression analysis to identify subtypes in chronic obstructive pulmonary disease. *BMC Systems Biology* 2014; **8** (Suppl 2):S8.

42. Ahmed SA, Penhale WJ, Talal N. Sex hormones, immune responses, and autoimmune diseases: Mechanisms of sex hormone action. *The American Journal of Pathology* 1985;**121**:531–51.
43. McKee LA, Chen H, Regan JA, Behunin SM, Walker JW, Walker JS, Konhilas JP. Sexually dimorphic myofilament function and cardiac troponin I phosphospecies distribution in hypertrophic cardiomyopathy mice. *Archives of Biochemistry and Biophysics* 2013;**535**: p.39-48.
44. Cvitic S, Longtine MS, Hackl H, Wagner K, Nelson MD, Desoye G, Hiden U. The human placental sexome differs between trophoblast epithelium and villous vessel endothelium. *PLOS One* 2013; **8**.
45. Jeong H. Altered drug metabolism during pregnancy: Hormonal regulation of drug-metabolizing enzymes. *Expert Opinion on Drug Metabolism & Toxicology* 2010; **6**: 689-99.
46. Li K., Xu W, Guo Q, Jiang Z, Wang P, Yue Y, Xiong S. Differential macrophage polarization in male and female BALB/c mice infected with coxsackievirus B3 defines susceptibility to viral myocarditis. *Circulation Research* 2009; **105**:353-64.
47. Moskalev A, Shaposhnikov M, Snezhkina A, Kogan V, Plyusnina E, Peregudova D, Melnikova N, Uroshlev L, Mylnikov S, Dmitriev A, Plusnin S, Fedichev P, Kudryavtseva A. Mining gene expression data for pollutants (Dioxin, Toluene, Formaldehyde) and low dose of gamma irradiation. *PLOS One* 2014; **9**.
48. Odibat O, Reddy CK . Ranking differential hubs in gene coexpression networks. *Journal of Bioinformatics and Computational Biology* 2012; **10**.