Supplementary material of the article “Multiple hot-deck imputation for network inference from RNA sequencing data”

A. Imbert et al.

October 26, 2017

Contents

1 Hyperparameters of the method 1

2 Material description 2
  2.1 GTEx data description ......................................................... 2
  2.2 DiOGenes data description ................................................. 2

3 Implementation details 3

4 Supplementary figures and tables 3

5 Supplementary results 6
  5.1 GTEx ................................................................. 6
  5.1.1 Choice of $\sigma$ .................................................... 6
  5.1.2 Different rates of missing individuals ............................. 7
  5.2 DiOGenes ............................................................. 9
  5.2.1 Choice of $\sigma$ .................................................... 9
  5.2.2 Results for CID2 (20% missing individuals) .................... 10
  5.2.3 Different rates of missing individuals ............................. 11
  5.2.4 Impact of the similarity chosen to create the pool of donors 13

6 DiOGenes: gene modules for CID1 and CID2 14
  6.1 Module 2 (CID1) and module 5 (CID2) ............................. 15
  6.2 Other modules .......................................................... 18
  6.3 Module description and consistency with previous findings 28

7 DiOGenes: GO term enrichment analysis 28

8 Lists of figures and tables and references 30

1 Hyperparameters of the method

The imputation method does not require to tune many hyperparameters: only a parameter for the definition of the pool of donors $\sigma$ and the number $M$ of replicates are required. The combination step of multiple imputation also requires to fix the reliability threshold $r_0$. Sound choices for these parameters are discussed in this section.

The method with an affinity score requires to choose a threshold $\sigma$. We propose to choose one which makes a good trade-off between similarity within the pool of donors (low average variance within the pools of donors) and variety (large enough number of donors in every pool). To do so, the average variance intra-$D(i)$ (for all $i = n_1 + 1, \ldots, n$) is first calculated:

$$V_{\text{intra}}(\sigma) = \frac{\sum_{i \in \alpha(n)} (x_i - x_{i'})^2}{d_i}$$

in which $d_i$ is the number of donors for the individual $i$ (the number of donors in $D_i$). Small values of $V_{\text{intra}}(\sigma)$ ensure a good homogeneity within all pools of donors but $V_{\text{intra}}(\sigma)$ is always increasing when $\sigma$ increases. Hence, we propose to plot $V_{\text{intra}}(\sigma)$ versus $\sigma$ in order to leverage the choice of $\sigma$ which has to fulfill similarity...
and variety in pools of donors, as mentioned above. A standard “elbow rule” is finally applied to choose a relevant value for \( \sigma \) (see Supplemental material).

The number of imputed datasets, \( M \), should be large enough in order to limit the influence of some specific individuals in the imputation and in order to well estimate the variability induced by the imputation on the results of the analysis (here, network inference). In our case, the value \( M = 100 \) has been chosen and was proved satisfactory in practice.

Finally, \( r_0 \) is chosen in accordance to the distribution of \( r(e) \) among all pairs of genes, \( e \). An ideal case is the one in which a large amounts of these values are very small (i.e., most of the pairs of genes, \( e \), are in the set of predicted edges 1-5 times within the \( M = 100 \) inferred networks) and only a small fraction of them are close to the maximum (100%). In practical applications, such behavior has been experienced and the values \( r_0 = 90 - 100\% \) all produced interesting results (see below for detailed results on the influence of \( r_0 \) on the performance).

2 Material description

2.1 GTEx data description

Genotype-Tissue Expression (GTEx) project collects and analyzes multiple human tissues from 544 donors (Lonsdale et al., 2013). RNA-seq expression data were acquired from many donors, across 53 distinct tissues and are available at http://gtexportal.org. We confined our analysis to two tissues which were found to have similar patterns of gene expression in (Melé et al., 2015): lung and thyroid. Lung expression dataset was used as primary dataset, \( X_0 \), and thyroid expression dataset was used as the auxiliary dataset, \( Y \). The complete dataset contained 221 samples common to lung and thyroid expression datasets (available on GTEx portal). RNA-seq expression data were normalized with edgeR (TMM normalization) and network inference was performed using the most variable genes (\( p = 100 \) genes for \( X_0 \) and \( q = 50 \) for \( Y \)). 36 genes are common to both datasets.

2.2 DiOGenes data description

DiOGenes is a pan-European research project that is a longitudinal (2-phases: low-calorie diet, weight maintenance diet) dietary intervention, undertaken in 8 European countries (Larsen et al., 2010). During the first phase, obese individuals were on a low-calorie diet during 8 weeks with the objective of more than 8% weight loss. Many biological measures have been performed before the beginning of the diet (CID1) and at its end (CID2). In particular, transcriptomic measures have been obtained from biopsies of adipose tissues with several techniques, including RT-qPCR and RNA-seq. In the present article, we tackle the issue of RNA-seq data imputation for a better understanding of the impact of calorie-restriction on the relationships between genes, as shown by graphical models.

RNA-seq expression have been measured for 433 individuals in CID1 and 307 individuals in CID2 with only 189 individuals in common between the two time steps. The RNA-seq expression dataset is available in Gene Expression Omnibus through the accession number GSEXXXX. In the present article, the measure of expression of a selected set of 317 genes were used. Genes were selected for regulation according to obesity, weight changes, metabolic adaptation and fat cell types based on literature and own published (Viguerie et al. (2012); Barquissau et al. (2016)) and unpublished datasets. RNA-seq data were normalized with edgeR (upper-quartile normalization).

RT-qPCR data (auxiliary dataset \( Y \)) have been measured for 413 individuals in CID1 and for 428 individuals in CID2. They contain the measures of the expression of 272 genes already included in the previous selected set of 317 genes. The RT-qPCR expression dataset is available in Gene Expression Omnibus through the accession number GSE60946. A Venn diagram of the number of common samples between the different datasets is provided in Figure 2. RT-qPCR data were normalized using the reference gene \( 2^{-\Delta C_t} \) (Viguerie et al. (2012); Livak and Schmittgen (2001)).

For the evaluation of the method, CID1 and CID2 datasets were processed independently, using only the 189 samples common to RNA-seq datasets at CID1 and CID2. For \( Y \) (RT-qPCR), samples are restricted to samples which are common with the previous 189 samples. Thus, 166 samples are used for CID1 and 172 for CID2.

For the application of the method on the whole DiOGenes dataset, networks at CID1 and CID2 were inferred from, respectively 433 and 307 observed individuals. Auxiliary datasets contained 413 individuals for

\(^1\)GEO repository, http://www.ncbi.nlm.nih.gov/geo/
\(^2\)Note from authors: available soon.
CID1 and 428 for CID2. CID1 auxiliary dataset is made of 313 individuals already observed in RNA-seq at CID1 and 100 individuals not observed in RNA-seq at CID1. CID2 auxiliary dataset is made of 276 individuals already observed in RNA-seq at CID2 and of 152 individuals not observed in RNA-seq at CID2 (see Figure 4 of supplementary material).

The difference between evaluation and application on DiOGenes dataset is illustrated in Figure 3 of supplementary material.

3 Implementation details

All analyses were conducted using R (R Core Team, 2016). Multiple hot-deck imputation for unit non-response was adapted from the single hot-deck standard imputation in the package hot.deck. The network inference method was implemented using the packages PoiClaClu (for transforming the count variables in order to take their overdispersion into account) and glmnet (Friedman et al., 2010) (to fit the Poisson generalized linear models).

The multiple PCA, MIPCA was performed using the package missMDA (Josse and Husson, 2016). Finally, networks were analyzed and compared using the package igraph (Csardi and Nepusz, 2006), which was also used to perform node clustering (with an approximation of the optimization for the modularity, using the function cluster_spinglass). Finally, GO term enrichment analyses (Section 7 of this Supplementary material) were performed with the packages biomaRt and topGO, querying the biomart database http://www.biomart.org/.

4 Supplementary figures and tables

![Figure 1: Pattern of missing values in RNA-seq dataset (X) and in auxiliary dataset (Y).](image)

![Figure 2: DiOGenes: Venn diagram (number of common samples) CID1 vs CID2 and RNA-seq vs RT-qPCR.](image)
RNAseq datasets

**CID1**
- n = 433

**CID2**
- n = 307

**common samples** (n = 189) used for the validation (see Figure 4 of Supplementary material and Sections 3.2 and 4.1 of the article) with:
- Y for CID1
  - n = 166
- Y for CID2
  - n = 172

RT-qPCR datasets

**CID1**
- n = 433

**CID2**
- n = 307

**all samples** used for network inference (“Application”, Sections 3.3 and 4.2 of the article) with:
- Y for CID1
  - n = 413
- Y for CID2
  - n = 428

RT-qPCR datasets

Figure 3: Flowchart for DiOGenes datasets. Left part illustrates the validation of the method based on this dataset (Sections 3.2 and 4.1 of the article) and right part illustrates the application of the method of the full dataset (Sections 3.3 and 4.2 of the article) to obtain two networks: one at CID1 and the other at CID2. See Figure 2 for further details about the number of samples in each case.

Figure 4: Overview of the evaluation process.
Figure 5: GTEx: Distribution of the appearance of an edge in the $M = 100$ networks for datasets imputed by \textit{hd-MI}, 20% missing individuals.

Figure 6: DiOGenes, CID1: Distribution of the appearance of an edge in the $M = 100$ networks for datasets imputed by \textit{hd-MI}, 20% missing individuals.
Table 1: Global properties of inferred networks, 20% missing individuals for GTEx. The number of edges, the density, the transitivity, the diameter and the size of the largest component are given for every network (reference, missing, mean, MIPCA and hd-MI). For missing and mean, the chosen network is the one corresponding to the value of $\lambda$ selected by StARS. For MIPCA and hd-MI, the chosen network is the one obtained from a threshold equal to $r_0 = 0.9$.

Table 2: Global properties of inferred networks, 20% missing individuals for DiOGenes at CID1. The number of edges, the density, the transitivity, the diameter and the size of the largest component are given for every network (reference, missing, mean, MIPCA and hd-MI). For missing and mean, the chosen network is the one corresponding to the value of $\lambda$ selected by StARS. For MIPCA and hd-MI, the chosen network is the one obtained from a threshold equal to $r_0 = 0.9$.

Table 3: GTEx: Number of gene modules and NMI (compared to modules obtained reference).

Table 4: DiOGenes: Number of gene modules and NMI (compared to modules obtained for reference and for the network obtained at CID2 with the same method), CID1, 20% missing individuals.

5 Supplementary results

This section shows additional results related to Section 4 of the article. This section is organized as follow: in Section 5.1, the results on GTEx datasets are extended. More precisely, the choice of $\sigma$ is discussed and the results for the other rates of missing individuals (10%, 30% and 40%) are presented. In Section 5.2, the results on the evaluation of the method on DiOGenes datasets are extended. The results for CID2 data and for various other rates of missing individuals are presented.

5.1 GTEx

5.1.1 Choice of $\sigma$

$V_{\text{intra}}$ was computed for various $\sigma$. The evolution of $V_{\text{intra}}$ versus $\sigma$ is displayed in Figure 7. The elbow rules gives $\sigma = 2$. 


5.1.2 Different rates of missing individuals

**PR curves** PR curves with 10%, 30% and 40% missing individuals are given in Figure 8.
Gene modules The number of gene modules obtained for GTEx data for the different rates of missing individuals are given in Table 5 and the NMI values in Table 6.

The number of gene modules for reference network for GTEx is 7.

Table 5: GTEx: Number of gene modules according to the rate of missing individuals, as compared to reference.

<table>
<thead>
<tr>
<th>Number of gene modules</th>
<th>10% missing</th>
<th>30% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>mean</td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>MIPCA</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>hd-MI</td>
<td>8</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 6: GTEx: Module NMI according to the rate of missing individuals, as compared to modules of reference.

<table>
<thead>
<tr>
<th>NMI</th>
<th>10% missing</th>
<th>30% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>0.657</td>
<td>0.681</td>
<td>0.615</td>
</tr>
<tr>
<td>mean</td>
<td>0.634</td>
<td>0.619</td>
<td>0.535</td>
</tr>
<tr>
<td>MIPCA</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>hd-MI</td>
<td>0.687</td>
<td>0.609</td>
<td>0.594</td>
</tr>
</tbody>
</table>

5.2 DiOGenes

5.2.1 Choice of \( \sigma \)

The evolution of \( V_{\text{intra}} \) versus \( \sigma \) is displayed in Figure 9. The elbow rule gives \( \sigma = 3 \).

In addition, the impact of the scales of the different variables on this result has been assessed. The affinity score has been computed on data previously scaled to unit variance and the evolution of \( V_{\text{intra}} \) versus \( \sigma \) is displayed in Figure 10. The elbow rule also gives \( \sigma = 3 \).
Moreover, the impact of the choice of the same affinity score on data scaled to unit variance on the final inference is assessed in Section 5.2.4 of this Supplementary material.

5.2.2 Results for CID2 (20% missing individuals)

Choice of $\sigma$ for hd-MI with an affinity score  

The evolution of $V_{\text{intra}}$ versus $\sigma$ is displayed in Figure 9. The elbow rules gives $\sigma = 3$.

Global properties  

The global properties for the inferred networks for CID2 are summarized in Table 7.
Table 7: DiOGenes: Global properties, CID2, 20% missing individuals.

<table>
<thead>
<tr>
<th>graph</th>
<th></th>
<th>edges</th>
<th>density</th>
<th>transitivity</th>
<th>diameter</th>
<th>size of largest component</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference</td>
<td></td>
<td>1486</td>
<td>0.0297</td>
<td>0.102</td>
<td>6</td>
<td>301</td>
</tr>
<tr>
<td>missing</td>
<td></td>
<td>1493</td>
<td>0.0298</td>
<td>0.099</td>
<td>7</td>
<td>301</td>
</tr>
<tr>
<td>mean</td>
<td></td>
<td>1497</td>
<td>0.0299</td>
<td>0.096</td>
<td>6</td>
<td>301</td>
</tr>
<tr>
<td>MIPCA</td>
<td></td>
<td>220</td>
<td>0.0044</td>
<td>0.046</td>
<td>17</td>
<td>145</td>
</tr>
<tr>
<td>hd-MI</td>
<td></td>
<td>1050</td>
<td>0.0210</td>
<td>0.076</td>
<td>7</td>
<td>299</td>
</tr>
</tbody>
</table>

**Precision/Recall curves** PR curves for CID2 are displayed in Figure 12.

![PR curves](image)

Figure 12: DiOGenes: PR curves, CID2, 20% missing individuals.

**Gene modules** The number of gene modules and NMI values for the different imputation methods for CID2, are given in Table 8.

Table 8: DiOGenes: Number of gene modules and NMI, CID2, 20% missing individuals. NMI are obtained by comparing gene modules with those of reference.

<table>
<thead>
<tr>
<th>graph</th>
<th>complete</th>
<th>missing</th>
<th>mean</th>
<th>MIPCA</th>
<th>hd-MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>♯ gene modules</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>NMI (CID2)</td>
<td>0.565</td>
<td>0.589</td>
<td>0.324</td>
<td>0.488</td>
<td></td>
</tr>
<tr>
<td>NMI with CID1</td>
<td>0.423</td>
<td>0.427</td>
<td>0.439</td>
<td>0.277</td>
<td>0.392</td>
</tr>
</tbody>
</table>

5.2.3 Different rates of missing individuals

**PR curves** PR curves for for CID2 are displayed in Figure 13 for varying rates of missing individuals (10%, 30% and 40%).
Figure 13: DiOGenes: PR curves for CID2 at different rates of missing individuals.

**Gene modules**  
- **For CID1**

The numbers of gene modules for different rates of missing individuals for CID1 are given in Table 9 and the value of NMI in Table 10.

The number of gene modules for **reference** at CID1 is 7.

Table 9: DiOGenes: Number of gene modules according to the rate of missing individuals, CID1.

<table>
<thead>
<tr>
<th>Number of gene module</th>
<th>10% missing</th>
<th>30% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>mean</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>MIPCA</td>
<td>14</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>hd-MI</td>
<td>8</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>
Table 10: DiOGenes: NMI according to the rate of missing individuals, CID1.

<table>
<thead>
<tr>
<th></th>
<th>10% missing</th>
<th>30% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>0.466</td>
<td>0.442</td>
<td>0.506</td>
</tr>
<tr>
<td>mean</td>
<td>0.478</td>
<td>0.488</td>
<td>0.472</td>
</tr>
<tr>
<td>MIPCA</td>
<td>0.397</td>
<td>0.332</td>
<td>0.683</td>
</tr>
<tr>
<td>hd-MI</td>
<td>0.463</td>
<td>0.375</td>
<td>0.391</td>
</tr>
</tbody>
</table>

For CID2

The numbers of gene modules for different rates of missing individuals for CID2 are given in Table 11 and the value of NMI in Table 12.

The number of gene modules for reference at CID2 is 9.

Table 11: DiOGenes: number of gene modules according to the rate of missing individuals, CID2.

<table>
<thead>
<tr>
<th>Number of gene module</th>
<th>10% missing</th>
<th>30% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>mean</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>MIPCA</td>
<td>15</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>hd-MI</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 12: DiOGenes: NMI according to the rate of missing individuals, CID2.

<table>
<thead>
<tr>
<th></th>
<th>10% missing</th>
<th>30% missing</th>
<th>40% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>0.597</td>
<td>0.622</td>
<td>0.62</td>
</tr>
<tr>
<td>mean</td>
<td>0.62</td>
<td>0.641</td>
<td>0.63</td>
</tr>
<tr>
<td>MIPCA</td>
<td>0.454</td>
<td>0.379</td>
<td>0.259</td>
</tr>
<tr>
<td>hd-MI</td>
<td>0.507</td>
<td>0.491</td>
<td>0.511</td>
</tr>
</tbody>
</table>

5.2.4 Impact of the similarity chosen to create the pool of donors

In this section, we evaluate the impact of the choice of the similarity used to create the pool of donors. Different solutions are compared:

- the affinity score described in the article computed on raw data or on data for which all variables have been scaled to unit variance (as in Section 5.2.1 of this supplementary material). These scores are respectively denoted by “affinity score (raw)” and “affinity score (scaled)” in this section;

- a $k$-nearest neighbor ($k$-NN) approach, in which the pool of donors is composed of the first $k = 5$ nearest neighbor for $Y$ of the missing individual in $X$. The choice of $k = 5$ is a standard one for such an approach and to leverage the impact of correlation between variables or of different scales, the $k$-NN have been obtained not only with the standard Euclidean distance between individuals but also with the Mahalanobis distance. The approaches are refered as “$k$-NN (Euclidean)” and “$k$-NN (Mahalanobis)” in this section;

- in addition, to compare the method with a well-known integration method, we used a CCA-based $k$-NN approach for the imputation. CCA (Hotelling, 1936) is a well-known method designed so as to find the study the correlations between two datasets and it has been used as a basis for $k$-NN imputation in the R package yaImpute (Crookston and Finley, 2008). However, the method does not directly apply to our datasets in which the number of samples is smaller than the number of variables: in this case, CCA is an ill-posed problem and has to be performed with some penalization or regularization method. Most widely used approaches include sparse CCA (Witten et al., 2009) and ridge CCA (Vinod, 1976).

The methods based on the affinity scores have been implemented as previously described in Section 3 of this supplementary material. Mahalanobis distance has been obtained using the package biotools, which relies on the pseudo-inverse of the covariance matrix for badly conditionned case such as our. For CCA-based $k$-NN, we have used the ridge version of CCA implemented in the R package mixOmics to implement this approach.

The comparison of the different methods to create the pool of donors is assessed through the PR curves obtained for CID1 dataset and with 20% of missing individuals. These curves are displayed in Figure 14 and
show that all methods provide very similar performances. This indicates that all approaches provide reasonable imputations for the missing individuals, with no clear advantage for any of them.

Figure 14: PR curves obtained with different methods for creating a pool of donors (affinity score on raw data or data scaled to unit variance; k-NN pools of donors with respect to Euclidean distance, Mahalanobis distance or based on CCA).

6 DiOGenes: gene modules for CID1 and CID2

The different gene modules obtained for CID1 and CID2 (except module 1 which is in the article, Figure 3) with the application on DiOGenes (Sections 3.3 and 4.2 of the article) are displayed below. All have been displayed with the R package igraph. Node sizes are proportional to their degrees and node colors represent betweenness (the darker the color, the higher the betweenness). Edges are colored according to the sign of correlation: blue for negative correlations and pink for positive correlations.
6.1 Module 2 (CID1) and module 5 (CID2)

The two following modules contain common features discussed in Section 6.3 (below).
Figure 16: DiOGenes: Module 2, CID1
Figure 17: DiOGenes: Module 5, CID2
6.2 Other modules

Figure 18: DiOGenes: Module 3, CID1
Figure 19: DiOGenes: Module 4, CID1
Figure 20: DiOGenes: Module 5, CID1
Figure 21: DiOGenes: Module 6, CID1
Figure 22: DiOGenes: Module 7, CID1
Figure 23: DiOGenes: Module 8, CID1
Figure 24: DiOGenes: Module 2, CID2
Figure 25: DiOGenes: Module 3, CID2
Figure 26: DiOGenes: Module 4, CID2
Figure 27: DiOGenes: Module 6, CID2
6.3 Module description and consistency with previous findings

At CID1, module 2 (Figure 16) showed a link between \(ABHD5\) and \(PLIN5\), two major regulators of adipose tissue lipase (Granneman et al., 2011). This module also contains \(G0S2\) and \(PLIN1\) which encode alternate lipid droplet surface-associated proteins (Tansey et al., 2001). Module 3 (Figure 18) contained only positively correlated genes. Among them, a link between \(AZGP1\) and \(GPD1L\), both of which being a metabolic syndrome signature previously described in adipose tissue (Viguerie et al., 2012; Montastier et al., 2015). In module 4 (Figure 19), \(PPARGC1A\), encoding the transcription coactivator PGC-1\(\alpha\), a main adipocyte browning inducer (Besseiche et al., 2015), showed only negative correlations. Module 7 (Figure 22) contained several correlated immune genes (C1q and MS4A family genes, notably).

As common features between CID1 and CID2, module 2 at CID1 (Figure 16) and module 5 at CID2 (Figure 17) showed persistent link between \(FADS1\) and \(FADS2\) which are located at same cluster on chromosome 11 and encode family members of enzymes involved in synthesis of unsaturated fatty acids. Some features were common to CID1 and CID2. Modules 4 at CID1 and CID2 (Figures 19 and 26 respectively) showed positive relationship between \(ALOX5\) and \(ALOX5AP\) that encode enzymes, which initialize the biosynthesis of leukotrienes from arachidonic acid.

7 DiOGenes: GO term enrichment analysis

Tables 13 and 14 provide the results of GO term enrichment tests for (respectively) CID1 and CID2 network modules. The last column contains the result of between cluster comparison of the level of the correlation between gene expression and clinical variables (HOMA-IR, body fat). These comparisons were performed with Kruskal-Wallis test with Nemenyi post-hoc tests. Reported significant results correspond to a higher absolute value of correlation with a threshold of 5% (but all \(p\)-values of significant results were found smaller than \(10^{-3}\)).
At CID1, absolute value of correlation between gene expression and HOMA-IR was found significantly higher in modules 3, 4 and 7 than in module 5 and 8. Absolute correlation with fat mass was higher in modules 2, 3 and 5 compared to genes in module 4. At CID1, correlations between gene expression and HOMA-IR were mostly positive in modules 4 and 7 and mostly negative in module 3. Correlations between gene expression and body fat were mostly positive in modules 2, 3 and 5. At CID2, absolute value of correlation between gene expression and HOMA-IR was found higher in module 2 than in module 7 and absolute value of correlation between gene expression and body fat was found higher in module 2 than in modules 4 and 7. At CID2, correlations between gene expression and HOMA-IR were mostly negative in module 2 and correlations between gene expression and body fat were mostly positive in module 2.

Table 13: GO enrichment analysis for the modules of CID1 network.

<table>
<thead>
<tr>
<th>Module</th>
<th>Top GO biological process</th>
<th>Percent. of called genes</th>
<th>Gene expression correlation to bioclinical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>regulation of leukocyte differentiation</td>
<td>58.33%</td>
<td>ns</td>
</tr>
<tr>
<td>2</td>
<td>no significant enrichment</td>
<td>-</td>
<td>body fat</td>
</tr>
<tr>
<td>3</td>
<td>regulation of cellular response to stress</td>
<td>45.45%</td>
<td>HOMA-IR, body fat</td>
</tr>
<tr>
<td>4</td>
<td>acute inflammatory response</td>
<td>55.56%</td>
<td>HOMA-IR</td>
</tr>
<tr>
<td>5</td>
<td>positive regulation of hormone secretion</td>
<td>60.00%</td>
<td>body fat</td>
</tr>
<tr>
<td>6</td>
<td>regulation of transcription</td>
<td>40.00%</td>
<td>ns</td>
</tr>
<tr>
<td>7</td>
<td>cofactor biosynthetic process</td>
<td>28.578%</td>
<td>HOMA-IR</td>
</tr>
<tr>
<td>8</td>
<td>skeletal system development</td>
<td>35.00%</td>
<td>ns</td>
</tr>
</tbody>
</table>

Table 14: GO enrichment analysis for the modules of CID2 network.

<table>
<thead>
<tr>
<th>Module</th>
<th>Top GO biological process</th>
<th>Percent. of called genes</th>
<th>Gene expression correlation to bioclinical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no significant enrichment</td>
<td>-</td>
<td>ns</td>
</tr>
<tr>
<td>2</td>
<td>lipoprotein metabolic process</td>
<td>25%</td>
<td>body fat, HOMA-IR</td>
</tr>
<tr>
<td>3</td>
<td>regulation of gene expression</td>
<td>15.38%</td>
<td>ns</td>
</tr>
<tr>
<td>4</td>
<td>no significant enrichment</td>
<td>-</td>
<td>ns</td>
</tr>
<tr>
<td>5</td>
<td>response to acid chemical</td>
<td>50%</td>
<td>ns</td>
</tr>
<tr>
<td>6</td>
<td>cellular modified amino acid metabolic</td>
<td>71.43%</td>
<td>ns</td>
</tr>
<tr>
<td>7</td>
<td>regulation of cell cycle</td>
<td>50%</td>
<td>ns</td>
</tr>
</tbody>
</table>
8 Lists of figures and tables and references

List of Figures

1 Pattern of missing values in RNA-seq dataset (X) and in auxiliary dataset (Y). 3
2 DiOGenes: Venn diagram (number of common samples) CID1 vs CID2 and RNA-seq vs RT-qPCR. 3
3 Flowchart for DiOGenes datasets. Left part illustrates the validation of the method based on this dataset (Sections 3.2 and 4.1 of the article) and right part illustrates the application of the method of the full dataset (Sections 3.3 and 4.2 of the article) to obtain two networks: one at CID1 and the other at CID2. See Figure 2 for further details about the number of samples in each case. 4
4 Overview of the evaluation process. 4
5 GTEx: Distribution of the appearance of an edge in the M = 100 networks for datasets imputed by hd-MI, 20% missing individuals. 5
6 DiOGenes, CID1: Distribution of the appearance of an edge in the M = 100 networks for datasets imputed by hd-MI, 20% missing individuals. 5
7 GTEx: Evolution of $V_{\text{intra}}$ versus $\sigma$. 7
8 GTEx: PR curves for different rates of missing individuals.
   (a) 10% missing individuals. 8
   (b) 30% missing individuals. 8
   (c) 40% missing individuals. 8
9 DiOGenes: Evolution of $V_{\text{intra}}$ versus $\sigma$, CID1, 20% missing individuals. 9
10 DiOGenes: Evolution of $V_{\text{intra}}$ versus $\sigma$, CID1, 20% missing individuals, for dataset scaled to unit variance. 10
11 DiOGenes: Evolution of $V_{\text{intra}}$ versus $\sigma$, CID2, 20% missing individuals. 10
12 DiOGenes: PR curves, CID2, 20% missing individuals. 11
13 DiOGenes: PR curves for CID2 at different rates of missing individuals.
   (a) CID2, 10% missing individuals. 12
   (b) CID2, 30% missing individuals. 12
   (c) CID2, 40% missing individuals. 12
14 PR curves obtained with different methods for creating a pool of donors (affinity score on raw data or data scaled to unit variance; k-NN pools of donors with respect to Euclidean distance, Mahalanobis distance or based on CCA). 14
15 Legend for the clusters (Figures 16-17 below and Figure 3 of the main article). 15
16 DiOGenes: Module 2, CID1. 16
17 DiOGenes: Module 5, CID2. 17
18 DiOGenes: Module 3, CID1. 18
19 DiOGenes: Module 4, CID1. 19
20 DiOGenes: Module 5, CID1. 20
21 DiOGenes: Module 6, CID1. 21
22 DiOGenes: Module 7, CID1. 22
23 DiOGenes: Module 8, CID1. 23
24 DiOGenes: Module 2, CID2. 24
25 DiOGenes: Module 3, CID2. 25
26 DiOGenes: Module 4, CID2. 26
27 DiOGenes: Module 6, CID2. 27
28 DiOGenes: Module 7, CID2. 28

List of Tables

1 Global properties of inferred networks, 20% missing individuals for GTEx. The number of edges, the density, the transitivity, the diameter and the size of the largest component are given for every network (reference, missing, mean, MIPCA and hd-MI). For missing and mean, the chosen network is the one corresponding to the value of $\lambda$ selected by StARS. For MIPCA and hd-MI, the chosen network is the one obtained from a threshold equal to $r_0 = 0.9$. 6
Global properties of inferred networks, 20% missing individuals for DiOGenes at CID1. The number of edges, the density, the transitivity, the diameter and the size of the largest component are given for every network (reference, missing, mean, MIPCA and hd-MI). For missing and mean, the chosen network is the one corresponding to the value of $\Delta$ selected by StARS. For MIPCA and hd-MI, the chosen network is the one obtained from a threshold equal to $r_\lambda = 0.9$.

GTEx: Number of gene modules and NMI (compared to modules obtained for reference).

DiOGenes: Number of gene modules and NMI (compared to modules obtained for reference and for the network obtained at CID2 with the same method), CID1, 20% missing individuals.

GTEx: Number of gene modules according to the rate of missing individuals, as compared to reference.

DiOGenes: Module NMI according to the rate of missing individuals, as compared to modules of reference.

DiOGenes: Global properties, CID2, 20% missing individuals.

DiOGenes: Number of gene modules and NMI, CID2, 20% missing individuals. NMI are obtained by comparing gene modules with those of reference.

DiOGenes: Number of gene modules according to the rate of missing individuals, CID1.

DiOGenes: NMI according to the rate of missing individuals, CID1.

DiOGenes: Number of gene modules according to the rate of missing individuals, CID2.

DiOGenes: NMI according to the rate of missing individuals, CID2.

GO enrichment analysis for the modules of CID1 network.

GO enrichment analysis for the modules of CID2 network.

References


