Multiple hot-deck imputation for network inference from RNA sequencing data

Alyssa Imbert, Armand Valsesia, Caroline Le Gall, Claudia Armenise, Gregory Lefebvre, Pierre-Antoine Gourraud, Nathalie Viguerie, Nathalie Vialaneix

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Original Papers

Multiple hot-deck imputation for network inference from RNA sequencing data

Alyssa Imbert [1,*], Armand Valsesia [2], Caroline Le Gall [3], Claudia Armenise [4], Gregory Lefebvre [2], Pierre-Antoine Gourraud [3], Nathalie Viguerie [5] and Nathalie Villa-Vialaneix [1]

[1] MIAT, Université de Toulouse, INRA, F-31326 Castanet-Tolosan, France
[2] Nestlé Institute of Health Sciences, CH-1015, Lausanne, Switzerland
[3] Methodomics, F-31200 Toulouse, France
[4] Quartzbio, CH-1202 Geneva, Switzerland
[5] UMR1048, Obesity Research Laboratory, Institute of Metabolic and Cardiovascular Diseases (I2MC), Inserm, F-31024 Toulouse, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Network inference provides a global view of the relations existing between gene expression in a given transcriptomic experiment (often only for a restricted list of chosen genes). However, it is still a challenging problem: even if the cost of sequencing techniques has decreased over the last years, the number of samples in a given experiment is still (very) small compared to the number of genes.
Results: We propose a method to increase the reliability of the inference when RNA-seq expression data have been measured together with an auxiliary dataset that can provide external information on gene expression similarity between samples. Our statistical approach, **hd-MI**, is based on imputation for samples without available RNA-seq data that are considered as missing data but are observed on the secondary dataset. **hd-MI** can improve the reliability of the inference for missing rates up to 30% and provides more stable networks with a smaller number of false positive edges. On a biological point of view, **hd-MI** was also found relevant to infer networks from RNA-seq data acquired in adipose tissue during a nutritional intervention in obese individuals. In these networks, novel links between genes were highlighted, as well as an improved comparability between the two steps of the nutritional intervention.
Availability: Software and sample data are available as an R package, **RNAseqNet**, that can be downloaded from the Comprehensive R Archive Network (CRAN).
Contact: Alyssa Imbert - alyssa.imbert@inra.fr

1 Introduction

In the last decades, biology and medicine have been profoundly renovated by the access to a large amount of molecular information, at various 'omics levels. Among high-throughput sequencing techniques, RNA-seq measures the expression of several thousands of genes for a given tissue. The large amount of generated data has created a need for multiple bioinformatics and statistical post-processing of the raw experimental data. In particular, a great deal of attention has been granted to the search of various types of relations between the genes (co-expression or regulation) (Zhang and Mallick (2013); Montastier *et al.* (2015), among others): better understanding those relations gives an insight on the global functioning of the cell in given environments and is a key point to reveal signaling pathways and to identify target genes for a given biological problem. Moreover, network visualization facilitates global analysis of the datasets.

RNA-seq expression data are count data and are thus discrete so standard GGM models usually used for network inferrence and that are based on Gaussianity assumption are not suited to such data. Recent works have considered using a generalized linear model (GLM) based on the

Poisson distribution (log-linear graphical model Allen and Liu (2012) or hierarchical Poisson log-normal model Gallopin *et al.* (2013)). In such methods, edge selection is handled by a L_1 penalty as in the continuous case. However, network inference is still a difficult issue as explained in Verzelen (2012), since the number of available samples (n) is generally much lower than the number of parameters to estimate (that scales as p^2, where p is the number of genes). Second, network inference can be very sensitive to missing individuals in key genes (Picheny *et al.*, 2014) or to the presence of "influential" individuals. Having a large number of observations is thus a key point for ensuring reliable results in statistical analyses of RNA-seq data Liu *et al.* (2014).

In this paper, we propose a method to increase the reliability of the inference when RNA-seq expression data have been measured together with other biological data closely related to the phenomenon under study. One typical case for such studies is the one in which RNA-seq measures have been performed on the same individuals simultaneously to another expression experiment using another (less costly) technique (*e.g.*, RT-qPCR). Another case is the one in which two RNA-seq experiments have been performed on two different tissues on the same individuals. It is relevant to try to use the external information brought by the auxiliary dataset in order to improve the quality of inference: when the acquisition cost is lower or when the ability to collect samples is easier, the number of samples acquired on the auxiliary dataset can be much larger. It thus provides additional information on relations between individuals and on biological variability.

We have designed an approach based on imputation in which individuals which are not observed in RNA-seq dataset but are observed on the secondary dataset are considered as missing data. It is called **hd-MI** and is presented in Section 2). A wide variety of methods falls under the general heading of imputation (Enders, 2010; Little and Rubin, 2002) but most of them impute missing values from different variables independently, based on the (non missing) values of the other variables for the same individual. Here, two additional issues are to be faced: first, entire individuals are considered as missing (*unit non-responses* case) means that entire rows are missing in $\tilde{X}$ and second, for network inference, the correlation structure between variables must be preserved during the imputation and the standard methods described above do not fulfill this need.

hd-MI is based on hot-deck and addresses both of these issues. In addition, the part of the uncertainty in the final result that comes from the imputation process is assessed using the general framework of *multiple imputation* (Rubin, 1987; Schafer, 1999; Rubin, 2012). This method has the additional benefit of providing more stable results and only assumes that the secondary dataset provides useful information about the resemblance between individuals and not about the network inference itself. The approach is assessed on two real RNA-seq datasets: one coming from a study on human tissue gene expression and the other coming from a longitudinal study on adipose tissue: the datasets and the methodology used to evaluate the method are presented in Section 3. Results are given and discussed in Section 4.

2 Methods

2.1 Background and notations

In the sequel, $\mathbf{X}$ will denote the RNA-seq expression dataset with n_1 rows (individuals) and p columns (genes). x_{ij} is the count of gene j, $j \in \{1, ..., p\}$, for individual i. Additionally, an auxiliary dataset is available, which will be denoted by $\mathbf{Y}$. $\mathbf{Y}$ is supposed to have q columns and $n > n_1$ rows, including rows corresponding to the same n_1 individuals already observed in $\mathbf{X}$. Without loss of generality, the common individuals between $\mathbf{X}$ and $\mathbf{Y}$ are supposed to correspond to the first n_1 rows of $\mathbf{Y}$. y_{ij} will denote the observation of variable j, $j \in \{1, ..., q\}$, for individual i. As already stated in the introduction, the above problem can be viewed as a missing value problem in the matrix $[\tilde{\mathbf{X}}, \mathbf{Y}]$ with dimensions $n \times (p+q)$ for which row number i is $\tilde{x}_i = \begin{cases} x_i & \text{if } i \leq n_1 \\ \text{missing} & \text{otherwise} \end{cases}$. Such a framework is called "unit non-responses", because missing values correspond to the absence of a complete individual. (see Supplementary Figure 1).

In this paper, missing individuals $i \in \{n_1 + 1, ..., n\}$ in the RNA-seq dataset are supposed to be Missing Completely At Random. This is a standard assumption if individuals have not been chosen according to a specific feature within $\{1, \ldots, n\}$ but because of a random choice or of technical constrains such as failed experiments or lack of tissue or to cost constrains if individual data are expensive or difficult to acquire.

2.2 hd-MI

"hot-deck" imputation is often used to impute non-response in surveys Andridge and Little (2010). It is based on the concept of "donors": if a respondent, i, called "recipient", has a missing value, $\tilde{x}_{ij}$, a set of similar individuals (donors) are pulled from $\{i' : i' \neq i \text{ st } \tilde{x}_{i'j} \text{ is not missing}\}$. This set of donors usually depends on the respondent itself. It is called *donor pool* and is denoted by $\mathcal{D}(i)$. One of the donors is finally randomly selected within $\mathcal{D}(i)$ to provide its value $\tilde{x}_{i'j}$ for imputing $\tilde{x}_{ij}$. Hot-deck imputation generally preserves the univariate distributions of the data and does not attenuate the variability of the filled-in data to the same extent as other imputation methods (Enders, 2010).

However, in basic hot-deck imputation, the correlation structure between variables is still modified during the imputation because the imputation of the different variables for an individual i are performed independently. To address this issue in the case of unit non-response problems, Voillet *et al.* (2016) proposed to impute simultaneously all variables $(\tilde{x}_{ij})_{j=1,\ldots,p}$ by the values coming from a single donor $i' \in \mathcal{D}(i)$ for a 'omic data integration problem.

Our approach, **hd-MI** (see Figure 1) is closely related to the method described in Voillet *et al.* (2016). It is adapted to the case of network inference with an auxiliary dataset. A multiple hot-deck imputation is performed, in which the imputation step is performed as described below:

1. firstly, for all missing individuals in $\tilde{\mathbf{X}}$, $i = n_1 + 1, \ldots, n$, the pool of donors $\mathcal{D}(i)$ is created and contains all individuals $i' \leq n_1$ which are "similar" to i. To estimate the similarity between individuals, the auxiliary dataset $\mathbf{Y}$ is used and different similarities can be calculated between individuals based on this dataset. Among them, we propose to use an affinity score, as in Cranmer and Gill (2012). This affinity score is computed for all individuals j by $s(i, i') = \frac{1}{q}\sum_{j=1}^{q} \mathbb{I}_{\{|y_{ij} - y_{i'j}| < \sigma\}}$, in which σ is a fixed threshold. The pool of donors is then obtained as $\mathcal{D}(i) = \{i' : s(i, i') = \max_{l=1,\ldots,n_1} s(i, i')\}$. This score is the average number of observed variables for which the individuals i and i' are "close";
2. in a second step, an individual, i' is picked at random in $\mathcal{D}(i)$ and the entire row i of $\tilde{\mathbf{X}}$ is imputed with row i' of $\tilde{\mathbf{X}}$. This step is repeated for all $i = n_1 + 1, \ldots, n$ to produce a complete case dataset $\mathbf{X}^*$.

In the framework of multiple imputation, the whole procedure is repeated M times independently to obtain M complete case datasets $\mathbf{X}^{*,m}$. The second step of the analysis consists in inferring the network for all these complete case datasets, using the model proposed in Allen and Liu (2012) (LLGM). The M networks are finally combined by studying the number of times an edge is predicted among the M networks: $r(e) = \frac{\text{number of times the edge } e \text{ is predicted}}{M}$. A reliability threshold, r_0 is finally chosen and the final network is composed of the edges e such

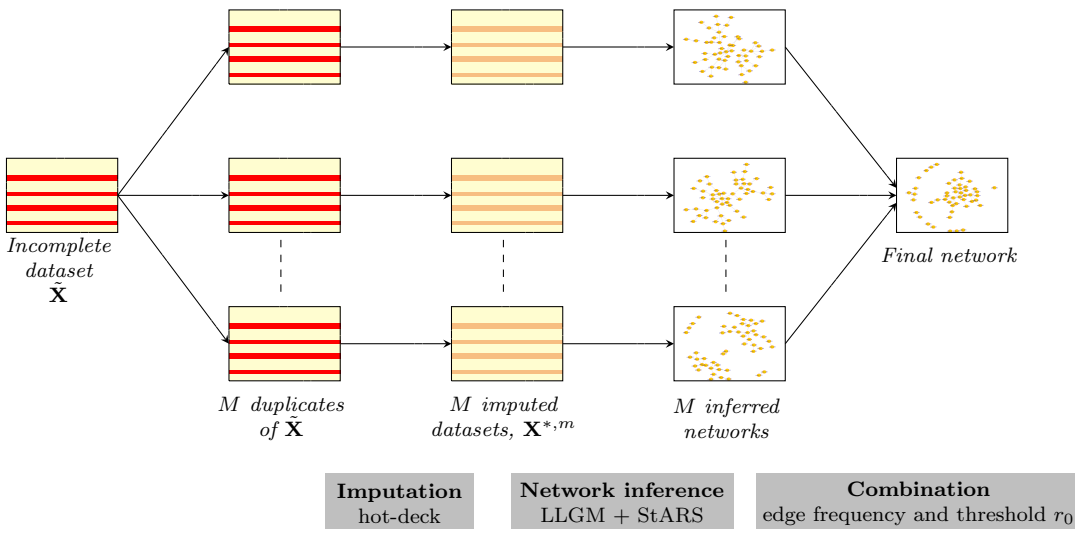


Fig. 1: Overview of **hd-MI**. The original dataset ($\tilde{\mathbf{X}}$, left) is duplicated M times (second column). For every duplicate, each missing row is imputed by hot-deck (third column, $\mathbf{X}^{*,m}$). A network is inferred from each imputed dataset (fourth column), with LLGM (StARS is used to choose the regularization parameter, λ, in the method). Finally the networks are combined into a single network using a threshold r_0 for edge frequency among the M networks (fifth column).

that $r(e) \geq r_0$. This approach is similar to the stability criterion described in Meinshausen and Bühlmann (2006).

The uncertainty of the imputation is thus handled in a way that is similar to standard approaches for improving the quality of network inference (Allouche *et al.*, 2013; Ballouz *et al.*, 2015), which use averaged weights or averaged ranks between multiple networks coming from different bootstrap resampling or from independent experiments.

Finally **hd-MI** does not require to tune many hyperparameters: only a parameter for the definition of the pool of donors (σ) and the number M of repeats are required. The combination step of multiple imputation also requires to fix the reliability threshold r_0 and network inference requires to tune a regularization parameter. Sound choices for these parameters are discussed in Supplementary Section 1.

3 Implementation and evaluation

3.1 Data description

To evaluate the performance of **hd-MI**, two real datasets were used coming from 2 distinct projects:

- GTEx (Lonsdale *et al.*, 2013), in which RNA-seq expression were acquired on several human tissues. We confined our analysis to two tissues: lung and thyroid. Lung expression dataset was used as primary dataset, $\mathbf{X}_0$, and thyroid expression dataset was used as the auxiliary dataset, $\mathbf{Y}$;
- DiOGenes (Larsen *et al.*, 2010), in which RNA-seq expression were acquired simultaneously to the measure of gene expression with another technique (RT-qPCR), on the same human tissue (adipose tissue) at two different time steps (CID1 and CID2) of a dietary intervention (before and after a 8 week low calorie diet). For $n = 189$ individuals, RNA-seq expression data were acquired at both CID1 and CID2 but for other individuals, either CID1 or CID2 data only were acquired (see Supplementary Figure 2 for a Venn Diagramm of data acquisition).

Further details on datasets are provided in Supplementary Section 2. The relevance of the method was assessed in two different types of analyses:

1. firstly, common samples between the primary and auxiliary datasets were kept to form the complete case dataset. The selected datasets do not contain missing individual. Artificially removing individuals from the primary dataset, these datasets are used as a reference to evaluate **hd-MI** and to compare it with other methods (Section 3.2);
2. secondly, all individuals from the DiOGenes project (CID1 and CID2, both common samples and CID specific samples) were used to infer one network for each CID (Section 3.3). These networks were further investigated and evaluated by comparison to previously inferred networks, obtained from different datasets.

The difference between evaluation and application on DiOGenes dataset is illustrated in Supplementary Figure 3.

3.2 Evaluation and comparison with existing methods

For each case, only common samples between two RNA-seq datasets ($n = 221$, for GTEx) or between two expression datasets (RNA-seq and RT-qPCR) and two time steps ($n = 189$ for DiOGenes) were kept to form the complete case dataset, used as a reference. $p = 100$ and $q = 50$ genes were selected for being the most variable in GTEx and $p = 317$ genes were selected for biological reasons for DiOGenes. The selected datasets do not contain missing individual and will be denoted by $\mathbf{X}_0$ in the sequel.

Datasets with missing individuals were then generated by randomly removing some samples in $\mathbf{X}_0$. More precisely, starting from a complete dataset $\mathbf{X}_0$, a given percentage f of rows were randomly removed (with $f \in \{10\%, 20\%, 30\%, 40\%\}$) to produce a dataset $\tilde{\mathbf{X}}$ with missing values. The corresponding complete case dataset is again denoted by $\mathbf{X}$.

Imputation was then performed using **hd-MI** and two alternative methods: 1/ a simple imputation method, *i.e.*, imputation by the mean and 2/ a state-of-the-art imputation method, Multiple Imputation by PCA (MIPCA) (Josse *et al.*, 2011). This led us to infer networks from:

- the full dataset $\mathbf{X}_0$: the obtained network is referred as **reference** in the result section. It is used as a gold standard for our comparison;
- datasets with varying rates of missing individuals with no imputation (hence the networks were inferred from the complete case dataset $\mathbf{X}$). These networks are referred as **missing** (possibly followed by a missing rate) in the result section. They are used as the worse case scenario;
- datasets with values imputed by the mean. These networks are referred as **mean** (possibly followed by a missing rate) in the result section;
- M datasets with values imputed by MIPCA, in which the dataset $[\tilde{\mathbf{X}}, \mathbf{Y}]$ is used as an input. For DiOGenes, $\mathbf{Y}$ (RT-qPCR expression) were scaled before affinity computation. Negative imputations (if any) were replaced by 0. These networks are referred as **MIPCA** (possibly followed by a missing rate) in the result section;
- M datasets with values imputed by our approach (using either the affinity score or the k-NN approach to define the donor sets). These networks are referred as **hd-MI**, possibly followed by a missing rate and, for the second case, by a number indicating the value of k.

For all the datasets described above, networks were inferred as follow:

- for the cases in which a single inference is performed (all datasets except the ones in which multiple imputation is used), the model described in Section 2.2 is performed with a full regularization path for the regularization parameter of the sparse penalty, λ. For each dataset we thus obtained a network for every value of λ in the regularization path. We also computed the StARS criterion to obtain the value of λ related to the most stable network along the path;
- for the cases of multiple imputation, a single network was inferred from every imputed dataset $\mathbf{X}^{*,m}$, using the regularization parameter λ selected by the StARS criterion. This led us to obtain M networks which were combined with varying reliability rates r_0: a final network was obtained for every value of r_0.

The detailed evaluation process is illustrated in Supplementary Figure 4. The results obtained with these different methods were assessed through global comparison of the network structures and through more local comparisons that were generally computed using **reference** as a gold-standard. In addition, for the DiOGenes datasets, 20 replicates of the whole evaluation procedure were obtained, so as to assess the stability of our findings.

3.3 Illustration of the interest on complete DiOGenes data

Finally, two networks (one for CID1 and one for CID2) were inferred using all individuals from RNA-seq datasets for the inference and RT-qPCR datasets as the auxiliary dataset. Supplementary Figure 3 provides the flow chart of the DiOGenes dataset processing, distinguishing the evaluation (Section 3.2) from the application for network inference (current section).

All experiments have been performed using R, version 3.2.2 (R Core Team, 2016). Details about which packages have been used are given in Supplementary Section 3. The R package **RNAseqNet** provides functions to perform **hd-MI** and network inference using the model of Allen and Liu (2012). Facilities for choosing hyperparameters are also provided.

4 Results and discussion

4.1 Evaluation and comparison with existing methods

The distribution of the appearance of an edge in the M ($M = 100$) inferred networks from datasets imputed by **hd-MI** is provided in Supplementary Figures 5 and 6, respectively for GTEx and DiOGenes datasets. A large proportion of edges are present in less than 10% of networks: these edges are not stable and show the sensitivity of network inference to some individuals. However, a small proportion of edges are very stable and inferred in more than 90% of the M imputed datasets. When evaluating the goodness of networks, the value $r_0 = 0.9$ was then chosen. In addition, Supplementary Tables 1 and 2 provide the global characteristics of the inferred networks and show that **hd-MI** network is in line with **reference** with respect to these measures, even if the number of inferred edges is slightly less than for networks **reference**, **missing** and **mean**.

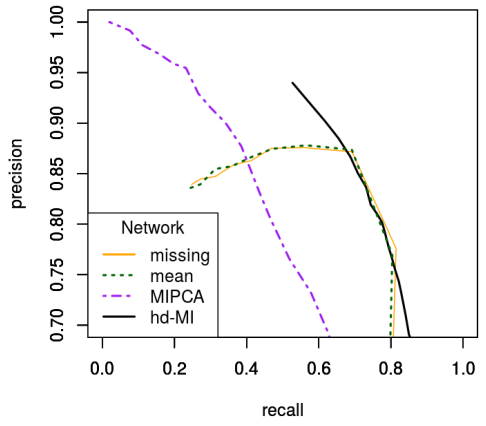
Precision / Recall (PR) curves, as compared to network **reference**, are displayed in Figure 2. PR curves for **mean** and **missing** are obtained for varying values of λ and PR curves for **MIPCA** and **hd-MI** are obtained for varying values of r_0. Top figures show the results obtained for DiOGenes CID1 (left) and GTEx (right) for 20% of missing individuals and bottom figures show the effect of varying the rate of missing individuals for DiOGenes CID1. **missing** and **mean** have similar curves, which shows that naive imputation method as **mean** does not perform better, in this framework, than simply using complete case datasets. On the contrary, **hd-MI** has the best recall for the highest precision rates. Since real biological networks are known to be sparse, the first few edges are the most important to recover: highest precisions have to be favored over recall for these applications. In the best case (DiOGenes), the precision is much larger than for the naive approaches. In the worst case (GTEx), naive methods do only slightly worse than **hd-MI** but with no indication on the reliability associated to each edge and on its sensitivity to missing values, contrary to **hd-MI**.

MIPCA shows poor performance: for all precision rates, **MIPCA** has the worse precision/recall curve for both GTEx and DiOGenes. **hd-MI** is much better adapted to the case of network inference than other state-of-the-art approach, such as **MIPCA**. The reason is twofold: firstly, as already stated before, by imputed missing values in different variables independently from each other, **MIPCA** does not preserve the correlation structure between variables. Secondly, as shown on this real data problem, **MIPCA** does not constrain the range of imputed values to be the same than the range of observed values. Sometimes, irrelevant values (*e.g.*, here negative values) are imputed and can strongly affect the results. Naive approaches do not have these drawbacks. Their performances can thus be more similar to those of our method.

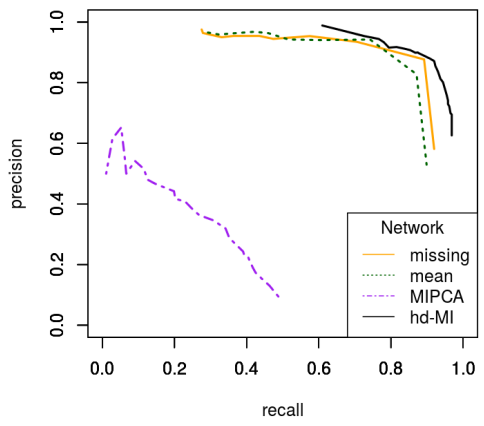
For the other rates of missing individuals, the results remain very similar even though the global performances of all methods are deteriorated by an increasing rate of missing individuals (as expected) and if the differences between methods tend to slightly decrease when the rate of missing individuals increases.

To assess the stability of our results, the whole simulation procedure was repeated 20 times for the DiOGenes dataset, at CID1 with 20% of missing individuals. Results show that only **hd-MI** is consistently able to reach a good recall for the highest precision rates: **missing** and **mean** manage to reach the precision rate of 85% for 18 curves over 20 and never manage to reach a 90% of precision rates. **MIPCA** always reaches the targeted precision rate but with very poor recalls. Statistics (*i.e.*, minimum, maximum and mean) of the recall for these two target precision rates are given in Table 1. They confirm this conclusion by exhibiting a much lower variability of the results obtained for **hd-MI** and a better recall (in average), as compared to the other methods.

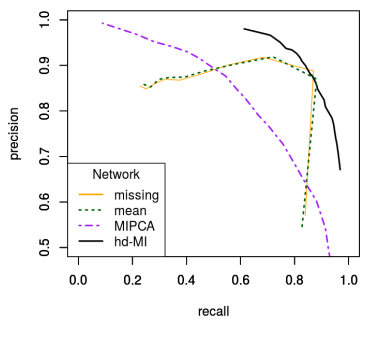
Supplementary Section 5 provides additional results that are similar to the ones presented in the present section. Choices of σ for **hd-MI** are illustrated in Supplementary 5.1.1 and 5.2.1., respectively for GTEx and for DiOGenes, CID1 (20% of missing individuals). Supplementary material also contains results obtained for CID2 of DiOGenes (Supplementary Section 5.2.2), results obtained with different rates of missing individuals (for both datasets in Supplementary Sections 5.1.2 and 5.2.3).



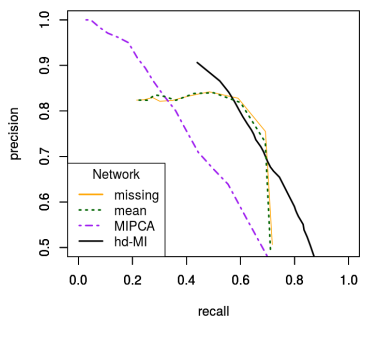
(a) DiOGenes - 20%



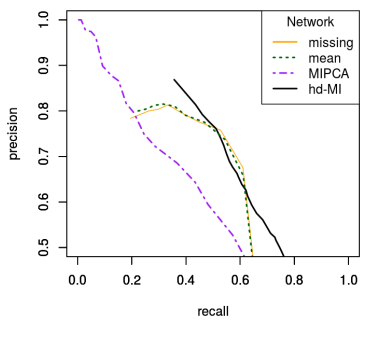
(b) GTEx - 20%



(c) DiOGenes - 10%



(d) DiOGenes - 30%



(e) DiOGenes - 40%

Fig. 2: **PR curves**, for every method and both datasets. Top figures show the results obtained for DiOGenes CID1 (left) and GTEx (right) for 20% of missing individuals and bottom figures show the effect of varying the rate of missing individuals for DiOGenes CID1. **hd-MI** provides an improved recall at highest precision rates, especially for DiOGenes and the smallest rates of missing individuals.

Table 1. **Recall statistics for precision rates of 85% (left) and 90% (right)**. For 85% precision rate, the recalls of the two cases where **missing** and **mean** did not reach the targeted precision rate were replaced by the recall for the highest precision.

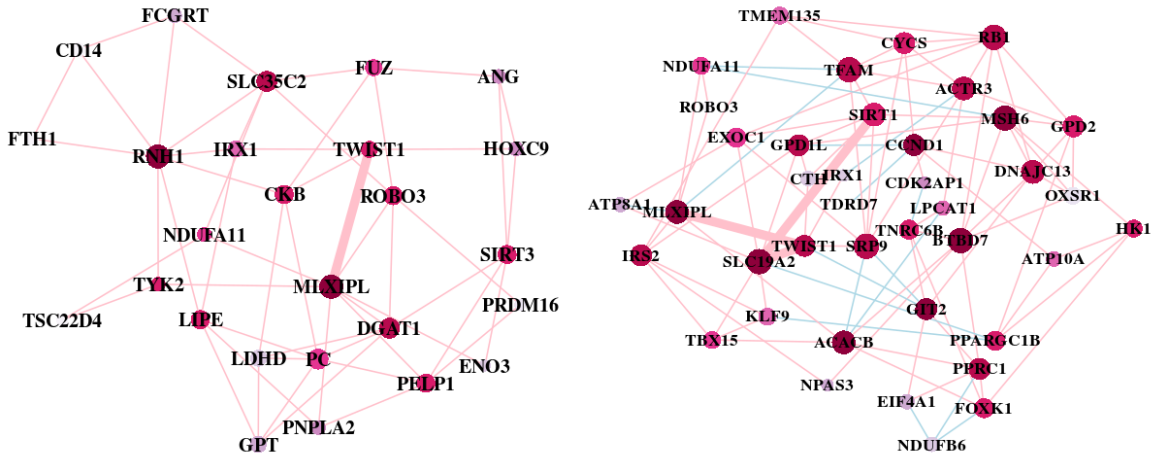
method	min	mean	max
missing	0.352	0.649	0.746
mean	0.487	0.641	0.733
MIPCA	0.324	0.355	0.397
hd-MI	0.580	0.658	0.729

method	min	mean	max
MIPCA	0.227	0.277	0.310
hd-MI	0.545	0.593	0.655

In addition, the impact of different methods to create a pool of donors have been tested. More precisely, we have compared our affinity based approach with an affinity performed on scaled data and with different types of k-NN imputations, respectively performed with Euclidean distance, Mahalanobis distance (so as to avoid the effect of different scales and strong correlations between variables) and a method similar to the one described in (Crookston and Finley, 2008) based on ridge regularized CCA (Vinod, 1976). The results prove that all these imputation methods perform very similarly with no visible changes in the inferred network (see Supplementary Section 5.2.4).

As explained in (de Smet and Marchal, 2010; Villa-Vialaneix *et al.*, 2013), gene networks are more relevant to identify groups of related genes (gene modules) than to study pairwise relationships between genes. To evaluate the preservation of gene modules, node clustering was performed by maximizing the modularity quality criterion (Newman and Girvan, 2004) in **reference** and in all inferred networks. To avoid an irrelevant number of clusters, clustering was performed only on the largest connected component of the graph. The resemblance between the module structure in **reference** and in the other inferred networks was assessed using the normalized mutual information (NMI, Danon *et al.* (2005)). NMI is a quality criterion ranging from 0 to 1, with a maximum equal to 1 when the two sets of clusters (modules) are identical. NMI was computed restricted to the genes in the intersection of the two largest connected components of the two networks (the **reference** network and the inferred network under study).

The number of gene modules and NMI values for network clustering are given in Supplementary Tables 3 and 4 for GTEx and DiOGenes CID1, respectively. Usually, gene modules are better recovered by **hd-MI** for GTEx dataset than for all the other methods for all missingness rates but this



(a) CID1 (b) CID2

Fig. 3: Module 1 for (resp.) CID1 (a) and CID2 (b) as obtained after clustering nodes in the two networks obtained with **hd-MI**. These modules show direct links between *TWIST1* and *MLX1PL* (a and b) and a novel link of *TWIST1* to *SIRT1* via *SLC19A2* at CID2 (b).

is not the case for DiOGenes. For the latter case, this is explained by the fact that **hd-MI** has closer performances to **missing** and **mean** than in GTEx and the resemblance with the modules of **reference** is artificially favored by the network selection. Indeed, final network selection is performed by StARS in **missing**, **mean** and **reference** and by r_0 thresholding in **hd-MI**: this corresponds to two different precision levels in PR curves. However, even in this case, modules in **hd-MI** network are rather similar to **reference**, illustrating the good preservation of network structure.

4.2 Analysis of networks inferred for complete DiOGenes data

We applied **hd-MI** to adipose tissue gene expression data obtained using RNA-seq during a dietary intervention. Gene modules were extracted using modularity optimization, as described above. Eight and 7 gene modules were found at CID1 and CID2, respectively (Supplementary Section 6). Because calorie restriction is known to promote body fat loss and alleviate insulin resistance, correlations between gene expression within each module with fat mass or the insulin resistance index, HOMA-IR, were computed and GO term enrichment analysis was performed for each module (Supplementary Section 7).

Some features were common to CID1 and CID2. For example, *MLX1PL* and *TWIST1* showed persistent link between module 1 at CID1 and at CID2 (Figure 3).

As a hallmark of the effects of calorie restriction, CID1 and CID2 shown differential signatures. The human *TWIST1* gene encodes a transcription factor abundantly expressed in adipocytes from lean individuals that is positively correlated to insulin sensitivity and is potential regulator of adipose tissue remodeling (Pettersson *et al.*, 2011). At CID1, *TWIST1* was connected to *MLX1PL* and *PNPLA2*. At CID2, it was connected to *MLX1PL* and to *SLC19A2*, which was linked to *SIRT1*. The *SLC19A2* gene encodes hTHTR-1, a transporter of thiamine, that plays an essential role in glycolysis. *SLC19A2* is one of the 41 strongest candidate gene regions associated to positive natural selection that are involved in nutrient metabolism (Sabeti *et al.*, 2007). The *SIRT1* gene encodes a deacetylase that regulates various metabolic pathways (Cao *et al.*, 2016). Calorie restriction is known to promote histone deacetylase expression. *MLX1PL* encodes the transcription factor ChREBP, whose activity is induced by glucose (Filhoulaud *et al.*, 2013). Together with *TWIST1* and *SIRT1* after calorie restriction it may function as glucose sensor and insulin sensitizer.

Inferred networks were found coherent with previous findings on these genes, which indicates that **hd-MI** data imputation does not induce a distortion in the relationship between gene expression (as described in Supplementary Section 6). The DiOGenes adipose tissue RNA samples have been previously analyzed using RT-qPCR (Viguerie *et al.*, 2012; Montastier *et al.*, 2015) and recently using RNA-seq (Armenise *et al.*, 2017). Network analyses were performed using RT-qPCR data on men and women (Viguerie *et al.*, 2012) or a subset of women (Montastier *et al.*, 2015). Several features were common to both studies and were also present in our networks. In particular, a module containing the same group of correlated genes which encode enzymes involved in lipogenesis including *FADS1*, *FADS2* and *AACS* was found in all CID1 networks (either inferred on women and on men). More interestingly, the links between *FADS1* and *AACS* and between *FADS2* and *AACS* persisted at CID2 only for our inferred network (Supplementary Figure 17). This persistence might be a positive effect of our imputation method, that provides a more comparable basis for CID1 and CID2 because it imputes missing individuals in one of the two CID.

By contrast to our previous network analyses that used a priori selected genes, the present study, by using RNA-seq data, revealed novel features. Especially, a persistent link was found between two transcription factors involved in insulin sensitivity, *TWIST1* and *MLX1PL* (this latter gene was not available in the previous analyses) and connections to novel genes, such as *SIRT1* and *SLC19A2*, appeared after calorie restriction.

5 Conclusion

We have designed a method to improve network inference from RNA-seq data from additional information about gene expression similarity between individuals. The method **hd-MI** is based on multiple hot-deck imputation and preserves the correlation structure between variables in a unit non-response framework (using the hot-deck approach) while estimating the uncertainty linked to the imputation (with a multiple imputation approach).

hd-MI shows a better precision for edge detection than complete case or naive imputation methods, with the additional advantage to provide information about the reliability of the edge and its sensitivity to missing

individuals. **hd-MI** has been used in a real world application related to the impact of a low calorie diet on adipose tissue expression. It succeeded in providing relevant networks that were similar to previously inferred networks, based on different dataset and different subsets of individuals. It also predicted the persistence of the links between *AACS*, *FADS1* and *FADS2* at CID2 and enlightened adipose tissue *SLC19A2* as novel partner in glucose homeostasis, besides *TWIST1* and *MLX1PL*. Its precise role as transporter or undiscovered function is still to be investigated.

Acknowledgements

The authors thank Mélina Gallopin for sharing personal R scripts. We also thank the three anonymous reviewers for valuable comments and suggestions which helped to improve the quality of the paper.

Funding

AI is a PhD fellow supported by the région Occitanie and Methodomics `http://www.methodomics.com`. PAG is founder of Methodomics. CLG is employed by Methodomics. CA is employed by Quartz Bio SA. GL, AV are employed by Nestlé Institute of Health Sciences. This work was supported by Inserm, Paul Sabatier University, Agence Nationale de la Recherche (ANR-12-BSV1-0025Obelip), Région Midi-Pyrénées (OBELIP and ILIP projects), Commission of the European Communities (FP6-513946 DiOGenes) and Innovative Medicines Initiative Joint Undertaking (grant agreement n°31 115372). The funders had no role in the study design, analyses, results interpretation and decision to publish.

References

Allen, G. and Liu, Z. (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Allouche, D. *et al.* (2013). A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context. In A. de la Fuente, editor, *Verification of Methods for Gene Network Inference from Systems Genetics Data*. Springer.

Andridge, R. and Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, **78**(1), 40–64.

Armenise, C. *et al.* (2017). Transcriptome profiling from adipose tissue during a low-calorie diet reveals predictors of weight and glycemic outcomes in obese, nondiabetic subjects. *The American Journal of Clinical Nutrition*, **106**(3), 736–746.

Ballouz, S., Verleyen, W., and Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, **31**(13), 2123–2130.

Cao, Y. *et al.* (2016). SIRT1 and insulin resistance. *Journal of Diabetes and its Complications*, **30**(1), 178–183.

Cranmer, S. and Gill, J. (2012). We have to be discrete about this: a non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, **43**, 425–449.

Crookston, N. and Finley, A. (2008). yaImpute: an R package for kNN imputation. *Journal of Statistical Software*, **23**, 10.

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics*, **2005**, P09008.

de Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, **8**, 717–729.

Enders, C. (2010). *Applied Missing Data Analysis*. Guilford Press.

Filhoulaud, G., Guilmean, S., Dentin, R., Girard, J., and Postic, C. (2013). Novel insights into ChREBP regulation and function. *Trends in Endocrinology and Metabolism*, **24**(5), 257–268.

Gallopin, M., Rau, A., and Jaffrézic, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS ONE*, **8**(10).

Josse, J., Pagès, J., and Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, **5**(3), 231–246.

Larsen, T. *et al.* (2010). The diet, obesity and genes (diogenes) dietary study in eight European countries - A comprehensive design for long-term intervention. *Obesity Reviews*, **11**(1), 76–91.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley.

Liu, Y., Zhou, J., and White, K. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**(3), 301–304.

Lonsdale, J. *et al.* (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics*, **45**, 580–585.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistic*, **34**(3), 1436–1462.

Montastier, E. *et al.* (2015). System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance. *PLoS Computational Biology*, **11**(1), e1004047. First co-author.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review, E*, **69**, 026113.

Pettersson, A. *et al.* (2011). Twist1 in human white adipose tissue and obesity. *The Journal of Clinical Endocrinology and Metabolism*, **96**(1), 133–41.

Picheny, V., Vandel, J., Vignes, M., and Villa-Vialaneix, N. (2014). Reconstruction quality of a biological network when its constituting elements are partially observed. In *AI & Statistics*, number L014, Reykjavik, Iceland.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. (1987). *Multlipe Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. (2012). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**(434), 473–489.

Sabeti, P. *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**(7164), 913–918.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**(1), 3–15.

Verzelen, N. (2012). Minimax risks for sparse regressions: ultra-high-dimensional phenomenons. *Electronic Journal of Statistics*, **6**, 38–90.

Viguerie, N. *et al.* (2012). Determinants of human adipose tissue gene expression: impact of diet, sex, metabolic status and cis genetic regulation. *PLoS Genetics*, **8**(9), e1002959.

Villa-Vialaneix, N. *et al.* (2013). The structure of a gene co-expression network reveals biological functions underlying eQTLs. *PLoS ONE*, **8**(4), e60045.

Vinod, H. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, **4**(2), 147–166.

Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., and Gonzáles, I. (2016). Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, **17**(402). Forthcoming.

Zhang, L. and Mallick, B. (2013). Inferring gene networks from discrete expression data. *Biostatistics*, **14**(4), 708–722.