Multiset sparse partial least squares path modeling for high dimensional omics data analysis

Attila Csala*, Aeilko H. Zwinderman and Michel H. Hof

Abstract

Background: Recent technological developments have enabled the measurement of a plethora of biomolecular data from various omics domains, and research is ongoing on statistical methods to leverage these omics data to better model and understand biological pathways and genetic architectures of complex phenotypes. Current reviews report that the simultaneous analysis of multiple (i.e. three or more) high dimensional omics data sources is still challenging and suitable statistical methods are unavailable. Often mentioned challenges are the lack of accounting for the hierarchical structure between omics domains and the difficulty of interpretation of genomewide results. This study is motivated to address these challenges. We propose multiset sparse Partial Least Squares path modeling (msPLS), a generalized penalized form of Partial Least Squares path modeling, for the simultaneous modeling of biological pathways across multiple omics domains. msPLS simultaneously models the effect of multiple molecular markers, from multiple omics domains, on the variation of multiple phenotypic variables, while accounting for the relationships between data sources, and provides sparse results. The sparsity in the model helps to provide interpretable results from analyses of hundreds of thousands of biomolecular variables.

Results: With simulation studies, we quantified the ability of msPLS to discover associated variables among high dimensional data sources. Furthermore, we analysed high dimensional omics datasets to explore biological pathways associated with Marfan syndrome and with Chronic Lymphocytic Leukaemia. Additionally, we compared the results of msPLS to the results of Multi-Omics Factor Analysis (MOFA), which is an alternative method to analyse this type of data.

Conclusions: msPLS is a multiset multivariate method for the integrative analysis of multiple high dimensional omics data sources. It accounts for the relationship between multiple high dimensional data sources while it provides interpretable results through its sparse solutions. The biomarkers found by msPLS in the omics datasets can be interpreted in terms of biological pathways associated with the pathophysiology of Marfan syndrome and of Chronic Lymphocytic Leukaemia. Additionally, msPLS outperforms MOFA in terms of variation explained in the chronic lymphocytic leukaemia dataset while it identifies the two most important clinical markers for Chronic Lymphocytic Leukaemia.

Availability: http://uva.csala.me/mspls; https://github.com/acsala/2018_msPLS

Keywords: Multivariate analysis, High dimensional omics data, Partial least squares
Background

Technological developments have enabled the measurement and storage of a plethora of biomolecular data extracted from various omics domains, such as data from the genome, epigenome, proteome or metabolome. It has become common to measure hundreds of thousands of biomolecular variables. To explore biological pathways across multiple omics domains, which might be associated with phenotypic (e.g. disease) outcomes, a natural research direction is to simultaneously analyse these omics domains. Complex diseases, such as obesity, diabetes, and schizophrenia have genetic architectures that involve many biological pathways, since they are a result of interactions between genomic, epigenomic and environmental variables [1, 2]. Therefore, modeling biological pathways across multiple omics domains might help to better understand the underlying genetic architecture and biological processes of complex phenotypes, which in turn leads to improved diagnosis, prognosis and therapy [1].

There is ongoing research for suitable statistical methods that could help leverage the available omics data to better model and understand biological pathways and genetic architectures of complex phenotypes on the biomolecular level [3].

Some of the first statistical methods developed for the integrated (i.e. simultaneous) analysis of multiple high dimensional omics datasets are generalizations of well known multivariate methods; e.g. sparse Canonical Correlation Analysis (CCA) [4–8], sparse Redundancy Analysis (RDA) [9, 10], and Multi-Omics Factor Analysis (MOFA) [11]. Detailed reviews and discussions on multivariate methods for omics data analysis can be found in [3, 12–18]. Although there are various statistical methods available to analyse omics data, recent reports argue that the simultaneous analysis of multiple (i.e. three or more) omics data sources is still challenging and current statistical methods are suboptimal. Among the challenges are the lack of accounting for the hierarchical structure between omics domains (i.e. relationship between data sources) and the difficulty of interpretation of genomewide results [2, 3, 19, 20].

To address those challenges, we propose a multiset multivariate statistical method, called multiset sparse Partial Least Squares path modeling (msPLS). msPLS is the penalised extension of multi-block Partial Least Squares path modeling (PLS-PM). Given the situation where biomolecular variables from multiple omics domains are measured on the same patients with shared phenotypes of interest, msPLS models biological pathways by identifying biomarkers (i.e. biomolecular variables that are associated with the phenotypes of interest) in each omics domain. The omics domains are assumed to have a hierarchical structure between each other, and their relationship is modelled in terms of dependencies through explanatory and response domain pairs. The explanatory and response omics data source pairs can be determined through the hypothesised information transfer between data sources as follows [21]. In an asymmetric relationship, a response data source is dependent on an explanatory data source if the prevalent way of information transfer is from the explanatory to the response data source. In a symmetric relationship, there is a recursive information transfer between data sources, and both data sources are dependent on each other. In PLS-PM, latent variables (LVs) are used to model the relationships between explanatory and response manifest variables (MVs) [22, 23]. Similarly to PLS-PM, the LVs in msPLS are linear combinations of the MVs, and are estimated in an iterative regression framework [24]. The LVs are constructed so that the combination of the explanatory MVs account for the most variance either directly in the response MVs (in an asymmetric relationship), or in the LVs of the response MVs (in a symmetric relationship). In general, Partial Least Squares path modeling distinguishes between these two types of relationships between data sources (i.e. symmetric or asymmetric relationships) the same way as the two well known multivariate statistical methods Canonical Correlation Analysis [8] and Redundancy Analysis [10] do. In the “Methods” section, we describe msPLS’s direct correspondence with those two well known multivariate methods. We give a detailed description of msPLS in the “Methods” section.

To illustrate such an explanatory and response dependency structure, consider that we have biomolecular variables (i.e. genomewide epigenomic, transcriptomic and proteomic variables) measured in patients with Marfan syndrome. The goal of this analysis is to use msPLS to explore biological pathways associated with Marfan syndrome, through the simultaneous analysis of the data sources. For this setting, we assume that the proteomic variables are responses for both the epigenomic and transcriptomic variables. Thus the proteome data source has an asymmetric relationship with both the epigenome and the transcriptome data sources. Additionally, there is a symmetric relationship between the epigenome and the transcriptome data sources, assuming a recursive information transfer between the epigenome and transcriptome. These assumptions are based on the special biological sequential information transfers of the central dogma of molecular biology and its elaborated versions [25, 26]. Given the above relationship between omics domains, msPLS identifies the combination of epigenomic and transcriptomic biomarkers that explain the most variance in the proteomic variables, while the combination of the epigenomic and transcriptomic biomarkers have maximum possible correlation with each other. This example is elaborated in more detail in the “Results” section of this paper.
To provide interpretable results from analyses of hundreds of thousands of MVs is addressed through sparse variable selection. msPLS enforces sparse variable selection through penalization methods, such as through the Least Absolute Shrinkage and Selection Operator (LASSO), Ridge, and Elastic Net (ENet) penalization methods [27]. These penalization methods are introduced to PLS-PM by regularising the multivariate regression steps in the iterative regression framework. Introducing regularisation allows msPLS to deal with the characteristic high dimensionality of omics datasets, where the number of variables are much higher than the number of samples. In addition, regularisation improves the interpretability of the final model in the form of sparse variable selection. Once the final model is obtained, the identified biomarkers can be interpreted in terms of biological pathways that are associated with the interest of phenotypes. In the “Methods” section, we quantify msPLS’s ability to identify a handful of associated variables from multiple data sources among thousands of irrelevant variables.

The rest of the paper is structured as follows. In the next section, the results of the real data analyses are described, where msPLS was applied to genomewide biomolecular variables measured in Marfan patients in order to explain the variance in the phenotypic proteomic variables with the combination of biomarkers from the epigenome and transcriptome, while accounting for a hypothesised relationship in omics domains. Additionally, msPLS was applied to a second omics dataset containing data from patients with Chronic lymphocytic leukaemia, and its results were compared to the results of MOFA. We discuss these findings in the “Discussion” section. In the “Methods” section, we describe msPLS and its implementation in an iterative regression framework, along with a working example of the analysis of three related data sources. In addition, we describe how msPLS, and PLS in general, relate to two well known multivariate methods, CCA and RDA. Finally, we show the results from a simulation study that was performed to assess the ability of msPLS to deal with high dimensional data and its ability to extract explanatory MVs that explain the most variance in the response MVs and LVs.

Results
We applied msPLS to genomewide epigenomic, transcriptomic and proteomic data sources measured in Marfan patients [28]. In addition, we applied msPLS to genomic, epigenomic, transcriptomic, and drug response data sources measured in Chronic Lymphocytic Leukaemia (CLL) patients [29].

Marfan data
The goal of this analysis was to explore biological pathways associated with Marfan disease based on epigenomic, transcriptomic and proteomic data measured in 37 Marfan patients [30]. The 364,134 epigenomic methylation variables were obtained by Illumina Infinium HumanMethylation450 BeadChips from blood leukocytes, the 18,424 transcriptomic gene expression variables were obtained by Affymetrix Human Exon 1.0ST Arrays from skin biopsy, and the 47 proteomic cytokine variables were measured in blood plasma.

The model was constructed by extracting the combination of LVs from the epigenome and transcriptome that explain the most variance in the phenotypic proteome MVs (Fig. 1). We hypothesised a symmetric relationship between the epigenome and transcriptome and asymmetric relationships from the proteome to both the epigenome and the transcriptome, so that the proteomic variables were set as response MVs for both the epigenomic and transcriptomic MVs. We used Univariate Soft Thresholding (UST) penalisation with 10-fold cross validation (see “Methods” section) to find the penalisation parameter that optimised the sum of squared correlations between the combination of LVs from the epigenome and transcriptome with respect to the proteome variables (see Eq. (5) in Methods). The final model extracted 40 methylation markers and 52 gene expression markers that optimised the sum of squared correlation of the explanatory LVs of the epigenome and transcriptome with the MVs from the proteome (Fig. 2). The sum of squared correlations was 9.32. Through bootstrapping, we obtained a 95% confidence interval of [9.03, 9.56] and a p-value <0.01 after permutation (see “Methods” section). The best fitting model resulted in a set of LVs that captured 49% of variance in both the epigenome and transcriptome variables and 65% of variance in the proteome variables. The extracted biomarkers with their corresponding individual contribution towards the overall explained variance in the proteomic variables (i.e. illustrated by the methylation and gene expression weights) and the proteomic variables with their corresponding individual correlation strength with the combination of the explanatory LVs (i.e. illustrated by the cytokine weights) are listed in Table 1.

Subsequent set of LVs can be extracted by applying msPLS to the residual data of the epigenome and transcriptome data sources and to the original proteome data source (see “Methods” section). Doing so, we obtained a second set of LVs that explain a different portion of variance in the MVs than the first set of LVs (Fig. 3). After optimizing the model on the residual data, we obtained the second set of LVs that captured 67% of the remaining variance in both the epigenome and transcriptome variables and 91% of the remaining variance in the proteome variables. Thus the first two sets of LVs captured a total of 83% variance in the epigenome and transcriptome variables and a total of 97% variance in the proteome variables. The list of the second set of epigenomic and
Fig. 1 msPLS identified a combination of 40 epigenomic markers (denoted as $\xi_1$) and 52 transcriptomic markers (denoted as $\xi_2$) that explain the most variance in the proteome variables. The color scale represents the strength of $w$ regression weights.

Transcriptomic biomarkers and the proteomic variables with their corresponding weights can be found in Table 2.

A gene set enrichment analysis (available at https://reactome.org) was used to test the association of the resulting gene expression markers (see Table 1) with already known biological pathways. The gene set enrichment analysis identified 208 pathways (see Additional file 2). We ordered the pathways on their respective $p$-values from an over-representation analysis (see https://reactome.org). For the sake of interpretability, we assessed the pathways with $p$-values only lower than $5 \times 10^{-2}$. From the 208 pathways, 58 (28%) had a $p$-value $< 5 \times 10^{-2}$ (see Table 3). From these pathways, 44 (76%) can be associated with Marfan disease. From the 58 pathways there are 14 (24%) not known to be associated with Marfan disease, and from these 14 there are 12 pathways that can be associated with the Influenza Virus. This might suggest that Influenza as co-morbidity was present in the patients during data gathering.

Among the pathways that were identified, already known pathophysiological pathways associated with Marfan disease [31] were found, such as the "Extracellular matrix organization" ($p$-value $4.8 \times 10^{-3}$), the "Crosslinking of collagen fibrils" ($p$-value $1.2 \times 10^{-3}$), the "TGF-beta receptor signaling in EMT (epithelial to mesenchymal transition)" ($p$-value $3.92 \times 10^{-2}$), and the "Loss of Function of TGFBR2" ($p$-value $8.39 \times 10^{-3}$) pathway. The identified pathways can be further appraised in the context of known interactions of genes and genetic phenotypes. We queried the curated database of Online Mendelian Inheritance in Man (OMIM, available at https://www.omim.org). The OMIM query yielded 372 results (the full list can be found in Additional file 3). Among others, OMIM identified the TGF-beta, Collagen IV, Interleukin-6 loci. The identified pathways from these analysis suggest that some patients suffered from Marfan syndrome type 2, which is based on mutations in the TGFBR2 gene (associated pathway "Loss of Function of TGFBR2"). The mutation in the FBN1
Fig. 2 msPLS identified 40 methylation markers and 52 gene expression markers that optimised the sum of squared correlation of the explanatory LVs of the epigenome and transcriptome with the MVs from the proteome.

Chronic lymphocytic leukaemia data

We used msPLS for the simultaneous analysis of 69 genomic, 4248 epigenomic, 5000 transcriptomic and 310 drug response variables measured in 200 chronic lymphocytic leukaemia (CLL) patients. This data is publicly available through the Multi-Omics Factor Analysis (MOFA) R package [11]. We used MOFA to impute the missing variables as described in [11]. A detailed description of this dataset can be found in [29]. The goal of this analysis was to compare msPLS performance in terms of explained variance to the performance of MOFA, a state-of-art unsupervised statistical method for the integrative analysis of multiple omics data sources.

To construct the hierarchical structure between the data sources for the msPLS analysis, we hypothesised the following relationship structure between the data source pairs. We assumed symmetric relationships between the genomic, epigenomic and transcriptomic MVs, and the drug response variables were set as response to both the epigenomic and transcriptomic MVs. We used UST penalisation and we compared our results to the results of MOFA. MOFA's model selected 5 non-zero biomolecular variables in each LVs. To compare the results to the msPLS, we enforced the model to extract 5 genomic, 30 epigenomic and 30 transcriptomic MVs from the omics sources. Also, we extracted multiple set of LVs per data source, and compared the total captured variation of msPLS's LVs to MOFA's LVs. The final model of msPLS resulted in 3 set of LVs that together explained 92% of variance in the genomic variables, 97% of variance in the epigenomic variables, 98% of variance in the transcriptomic variables and 85% of variance in the drug response variables. In comparison, MOFA's first 10 LVs (i.e. referred to as factors in the MOFA model) together explained 23% of variance in the genomic variables, 24%
Table 1 The weights of the epigenomic, transcriptomic and proteomic variables extracted by msPLS from the Marfand data

| Site              | Methylation markers w | Gene expression markers w | Cytokine markers w |
|-------------------|------------------------|---------------------------|--------------------|
| cg02394578        | 0.93                   | AKAP4 0.65                | b NGF 0.43         |
| cg20322866        | 0.96                   | ANXA2P3 0.63              | CTACK 0.34         |
| cg00906428        | 0.91                   | ASMT_A 0.83               | GRO α 0.31         |
| cg05024291        | 0.93                   | ATHL1 0.73                | HGF 0.21           |
| cg05093318        | 0.95                   | B9D2 0.66                 | Hu Eotaxin 0.34    |
| cg18850112        | 0.97                   | C15orf52 0.76             | Hu GCSF 0.61       |
| cg07475117        | 0.95                   | C17orf54 0.44             | Hu GMCSF 0.43      |
| cg07588614        | 0.94                   | C1orf170 0.8              | Hu IFN γ 0.28      |
| cg10372701        | 0.94                   | C9orf98 0.4               | Hu IL 10 0.56     |
| cg01718788        | -0.9                   | CALHM1 0.84               | Hu IL 12 p70 0.75 |
| cg17061156        | 0.87                   | CHTF18 0.81               | Hu IL 13 0.51     |
| cg24000259        | 0.95                   | CLEC4C 0.39               | Hu IL 15 0.19     |
| cg25914270        | 0.92                   | COL4A6 0.69               | Hu IL 17 0.65     |
| cg14203970        | 0.91                   | Cxorf50B 0.84             | Hu IL 1b 0.48     |
| cg22859054        | 0.94                   | CYPIA2 0.64               | Hu IL 1a 0.58     |
| cg02972485        | 0.96                   | DKFZp779 0.85             | Hu IL 2 0.49      |
| cg22380215        | 0.88                   | FKRP 0.85                 | Hu IL 4 0.26      |
| cg09570676        | 0.96                   | FUZ 0.71                  | Hu IL 5 0.29      |
| cg14751679        | 0.93                   | GTF2IRD1 0.29             | Hu IL 6 0.04      |
| cg24477102        | 0.94                   | HIGD1B 0.53               | Hu IL 7 0.44      |
| cg24882220        | 0.93                   | ICAM5 0.85                | Hu IL 8 0.42      |
| cg25626012        | 0.95                   | JPH3 0.88                 | Hu IL 9 0.42      |
| cg26502549        | 0.94                   | LBP 0.53                  | Hu IP 10 0.72     |
| cg00015699        | 0.93                   | LOC64322 -0.44            | Hu MCP 1 0.04     |
| cg01281611        | 0.93                   | LRRCS0 0.66               | Hu MIP 1a 0.23    |
| cg07685390        | 0.95                   | MGC4294 0.73              | Hu MIP 1b 0.47    |
| cg26550872        | 0.97                   | MT1DP 0.43                |                |
| cg09236780        | 0.92                   | MT1L 0.51                 | Hu PDGF bb 0.29   |
| cg20618527        | 0.92                   | NDUFAF2 -0.47             | Hu RANTES 0.52    |
| cg01546046        | 0.89                   | OR11A1 -0.53              | Hu VEGF 0.59      |
| cg20999565        | 0.96                   | PEX11G 0.77               | IFNα 2a 0.15      |
| cg11152012        | 0.87                   | PIF1 0.88                 | Il 12 p40 0.28    |
| cg22958262        | 0.96                   | PIN4 0.65                 | Il 16 0.38        |
| cg26820811        | 0.9                    | POTEF -0.72               | Il 18 0.64        |
| cg06731730        | 0.95                   | PRIC285 0.83              | Il 1a 0.14        |
| cg08311998        | 0.94                   | PSPN 0.63                 | Il 2Rα 0.56       |
| cg19696891        | 0.92                   | PTGR1 -0.65               | Il 3 0.24         |
| cg27495444        | 0.93                   | RECB 0.72                 | LIF 0.57          |
| cg17840843        | 0.93                   | RINL 0.85                 | M-CSF 0.4          |
| cg05062854        | 0.95                   | ROM1 0.69                 | MCP 3 0.42        |
|                   |                        | SLC16A3 0.78              | MIF 0.27          |
|                   |                        | SLPI -0.63                | MIG 0.63          |
|                   |                        | SNORA39 0.81              | SCF 0.23          |
|                   |                        | SPATA2L 0.57              | SCGF b 0.42       |
|                   |                        | TMEM179 0.79              | SDF 1a 0.24       |
|                   |                        | TNFRSF6B 0.79             | TNF b 0.26        |
|                   |                        | TRBV12_3 0.46             | TRAIL 0.6         |
|                   |                        | UBQLN1 0.65               |                |
|                   |                        | UNQ6494 0.55              |                |
|                   |                        | ZNF215 -0.76              |                |
|                   |                        | ZNF574 0.74               |                |
|                   |                        | ZNF688 0.65               |                |
of variance in the epigenomic variables, 38% of variance in the transcriptomic variables and 40% of variance in the drug response variables (Table 4). We compared the correlations of Table 5 the selected MVs with their corresponding LVs (i.e., these correlations are referred to as loadings in the MOFA model) from msPLS’s and MOFA’s models. The biomarkers extracted with msPLS are listed with their corresponding loadings in Table 6.

We also compared the results of MOFA and msPLS in terms of clinical assessment of the outputs of both models (the full clinical assessment of MOFA’s results can be found in [11]). For this, we used the gene set enrichment analysis in MOFA’s environment. This query resulted in total more than 10,000 pathways, from which 241 pathways with p-values < 0.05 were identified with the gene sets obtained on the CLL data with MOFA, and 298 pathways with p-values < 0.05 were identified with the gene sets obtained on the CLL data with msPLS.

The first 1000 pathways (ordered by their corresponding p-values) for the gene sets from MOFA and msPLS can be found in Additional file 5 and 6. Out of these 1000 pathways, 811 (81%) were identified by both methods, and there are 158 (66% and 53%) overlapping pathways with p-values < 0.05 (see Additional file 7). Similarly to MOFA, msPLS extracted biomarkers from the genomic variables that can be associated with the pathophysiological pathways of CLL. After querying the gene sets from msPLS, the gene set enrichment analysis identified associations with biological pathways such as the “Transcriptional regulation of white adipocyte differentiation” (p-value $3.72 \times 10^{-4}$), the “Glycerophospholipid biosynthesis” (p-value $5.92 \times 10^{-5}$), and the “TP53 Regulates Metabolic Genes” (p-value $4.39 \times 10^{-4}$) pathway in the first LV. In the second LV, the pathways “Keratan sulfate/keratin metabolism” (p-value $5.16 \times 10^{-5}$), “Post NMDA receptor activation events” (p-value 1.15 $\times 10^{-4}$), and “Activation of NMDA receptor upon glutamate
| Methylation markers | Gene expression markers | Cytokine markers |
|---------------------|-------------------------|----------------|
| Site               | Genecode                | Markercode     | w    | w    | w    |
| cg23054189         | AGTR2                   | bNGF46         | 0.93 | -0.59| 0.57 |
| cg18347642         | C2orf43                 | CTACK72        | 0.93 | 0.87 | 0.52 |
| cg16489610         | CCDC112                 | GRo a 61       | 0.85 | 0.87 | 0.16 |
| cg27013696         | DKFZP434                | HGF62          | 0.91 | -0.58| -0.08|
| cg20457796         | GMCL1                   | Hu Eotaxin 43  | 0.93 | 0.88 | 0.12 |
| cg03181582         | LPO                     | Hu FGF basic 44| 0.91 | -0.77| 0.23 |
| cg10521851         | MAD2L1                  | Hu GCSF 57     | 0.9   | 0.74 | -0.57|
| cg19968840         | MGC4473                 | Hu GM CSF 34   | 0.92 | -0.81| -0.03|
| cg27648075         | NFAM1                   | Hu IFN g 21    | 0.92 | -0.7 | -0.03|
| cg22891500         | NMI                     | Hu Il 10 56    | 0.92 | 0.8  | 0.17 |
| cg05158197         | PF4                     | Hu Il 12 p70 75| 0.92 | -0.83| 0.43 |
| cg20119106         | PRDM14                  | Hu Il 13 51    | 0.93 | -0.71| 0.26 |
| cg02675353         | PSMA8                   | Hu Il 15 73    | 0.91 | -0.63| 0.67 |
| cg26991025         | RDM1                    | Hu Il 17 76    | 0.93 | 0.47 | -0.14|
| cg20643012         | RNF8                    | Hu Il 1b 39    | 0.92 | 0.69 | -0.03|
|                    | TTC30A                  | Hu Il 1r 25    | 0.92 | 0.8  | -0.03|
|                    | TTC30B                  | Tu Il 2 38     | 0.92 | 0.68 | 0.36 |
|                    | TTC4                    | Hu Il 4 52     | 0.92 | 0.81 | 0.67 |
|                    | UNQ8126                 | Hu Il 5 33     | 0.92 | -0.79| 0.23 |
|                    | ZNF677                   | Hu Il 6 19    | 0.92 | 0.06 | -0.04|
|                    |                         | Hu Il 7 74     | 0.92 | 0.81 | 0.44 |
|                    |                         | Hu Il 8 54     | 0.92 | 0.77 | 0.13 |
|                    |                         | Hu IP 10 48    | 0.92 | 0.69 | 0.73 |
|                    |                         | Hu MCP 1 MCAF 53| 0.92 | -0.69| 0.09 |
|                    |                         | Hu MIP 1a 55   | 0.92 | -0.83| -0.24|
|                    |                         | Hu MIP 1b 18.  | 0.92 | 0.69 | 0.26 |
|                    |                         | Hu PDGF bb 47  | 0.92 | 0.69 | 0.3  |
|                    |                         | Hu RANTES 37   | 0.92 | 0.69 | 0.84 |
|                    |                         | Hu VEGF 45     | 0.92 | 0.69 | 0.84 |
|                    |                         | IFN a 2.20     | 0.92 | -0.29| -0.09|
|                    |                         | Il 12 p40 28   | 0.92 | 0.35 | 0.35 |
|                    |                         | Il 16 27       | 0.92 | 0.18 | 0.18 |
|                    |                         | Il 18 42       | 0.92 | 0.18 | 0.18 |
|                    |                         | Il 1a 63       | 0.92 | 0.18 | 0.18 |
|                    |                         | Il 2 Ra 13     | 0.92 | 0.18 | 0.18 |
|                    |                         | LIF 29         | 0.92 | 0.18 | 0.18 |
|                    |                         | M CSF 67       | 0.92 | 0.18 | 0.18 |
|                    |                         | MCP 3 26       | 0.92 | 0.18 | 0.18 |
|                    |                         | MIF 35         | 0.92 | 0.18 | 0.18 |
|                    |                         | MIG 14         | 0.92 | 0.18 | 0.18 |
|                    |                         | SCF 65         | 0.92 | 0.18 | 0.18 |
|                    |                         | SCGF b 78      | 0.92 | 0.18 | 0.18 |
|                    |                         | SDF 1a 22      | 0.92 | 0.18 | 0.18 |
|                    |                         | TNF b 30       | 0.92 | 0.18 | 0.18 |
|                    |                         | TRAIL 66       | 0.92 | 0.18 | 0.18 |
Table 3 Over representation analysis results of the msPLS analysis on Marfan data

| Pathway name                                                                 | p-value          | Associated with Marfan disease through pathway                                                                 |
|------------------------------------------------------------------------------|------------------|---------------------------------------------------------------------------------------------------------------|
| Influenza Virus Induced Apoptosis                                            | $3.41 \times 10^{-5}$ | Not known*                                                                                                    |
| Non-integrin membrane-ECM interactions                                      | $2.92 \times 10^{-4}$ | Collagene formation [31]                                                                                     |
| Anchoring fibril formation                                                    | $4.73 \times 10^{-4}$ | Collagene formation [31]                                                                                     |
| ECM proteoglycans                                                            | $6.19 \times 10^{-4}$ | Extracellular matrix organization [31]                                                                         |
| Integrin cell surface interactions                                           | $7.90 \times 10^{-4}$ | Extracellular matrix organization [31]                                                                         |
| Transcriptional activation of mitochondrial biogenesis                       | $8.17 \times 10^{-4}$ | Possibly through reduced mitochondrial respiration [32]                                                       |
| Crosslinking of collagen fibrils                                             | $1.20 \times 10^{-3}$ | Collagene formation [31]                                                                                     |
| Laminin interactions                                                         | $1.98 \times 10^{-3}$ | Extracellular matrix organization [31]                                                                         |
| Mitochondrial biogenesis                                                     | $2.40 \times 10^{-3}$ | Possibly through reduced mitochondrial respiration [32]                                                       |
| NCAM1 interactions                                                           | $3.92 \times 10^{-3}$ | NCAM signaling for neurite out-growth [33]                                                                     |
| Collagen chain trimerization                                                 | $3.92 \times 10^{-3}$ | Collagene biosynthesis and modifying enzymes [31]                                                              |
| TGFBR2 MSI Frameshift Mutants in Cancer                                     | $4.20 \times 10^{-3}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| Extracellular matrix organization                                             | $4.82 \times 10^{-3}$ | Extracellular matrix organization [31]                                                                         |
| Host Interactions with Influenza Factors                                     | $5.02 \times 10^{-3}$ | Not known*                                                                                                    |
| Organelle biogenesis and maintenance                                         | $5.14 \times 10^{-3}$ | Possibly through reduced mitochondrial respiration [32]                                                       |
| Transfer of LPS from LBP carrier to CD14                                     | $6.30 \times 10^{-3}$ | Possibly through toll-like receptor-4 signaling [34]                                                          |
| Transport of HA trimmer, NA tetramer and M2 tetramer from the endoplasmic reticulum to the Golgi Apparatus | $6.30 \times 10^{-3}$ | Not known*                                                                                                    |
| Loss of Function of TGFBR2 in Cancer                                        | $8.39 \times 10^{-3}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| TGFBR1 LBD Mutants in Cancer                                                 | $8.39 \times 10^{-3}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| TGFBR2 Kinase Domain Mutants in Cancer                                       | $8.39 \times 10^{-3}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| Assembly of collagen fibrils and other multimeric structures                 | $8.81 \times 10^{-3}$ | Collagene formation [31]                                                                                     |
| Collagen degradation                                                         | $9.32 \times 10^{-3}$ | Degradation of the extracellular matrix [31]                                                                    |
| NCAM signaling for neurite out-growth                                        | $9.58 \times 10^{-3}$ | NCAM signaling for neurite out-growth [33]                                                                     |
| Interleukin-4 and Interleukin-13 signaling                                   | $9.78 \times 10^{-3}$ | Vascular inflammation through interleukins [35, 36]                                                            |
| Collagen biosynthesis and modifying enzymes                                  | $1.12 \times 10^{-2}$ | Collagene formation [31]                                                                                     |
| TGFBR1 KD Mutants in Cancer                                                  | $1.26 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| Loss of Function of TGFBR1 in Cancer                                         | $1.46 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| SMAD2/3 Phosphorylation Motif Mutants in Cancer                              | $1.46 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| Assembly of Viral Components at the Budding Site                            | $1.46 \times 10^{-2}$ | Not known*                                                                                                    |
| Loss of Function of SMAD2/3 in Cancer                                        | $1.67 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| RUNX3 regulates CDKN1A transcription                                         | $1.67 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [37]                                                                     |
| Signaling by TGF-beta Receptor Complex in Cancer                            | $1.88 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [31]                                                                     |
| Collagen formation                                                          | $2.02 \times 10^{-2}$ | Extracellular matrix organization [31]                                                                         |
| Transcriptional regulation of white adipocyte differentiation                | $2.17 \times 10^{-2}$ | Possibly by depleted or abnormal adipose tissue [38]                                                          |
| Aromatic amines can be N-hydroxylated by CYP1A2                              | $2.29 \times 10^{-2}$ | Not known                                                                                                    |
| Formation of annular gap junctions                                           | $2.29 \times 10^{-2}$ | Endothelial dysfunction [39]                                                                                   |
| Gap junction degradation                                                     | $2.50 \times 10^{-2}$ | Endothelial dysfunction [39]                                                                                   |
| Proton-coupled monocarboxylate transport                                    | $2.50 \times 10^{-2}$ | Not known                                                                                                    |
| RUNX3 regulates p14-ARF                                                      | $3.31 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [37]                                                                     |
| Fusion of the Influenza Virion to the Host Cell Endosome                     | $3.52 \times 10^{-2}$ | Not known*                                                                                                    |
| Packaging of Eight RNA Segments                                              | $3.52 \times 10^{-2}$ | Not known*                                                                                                    |
| Fusion and Uncoating of the Influenza Virion                                | $3.72 \times 10^{-2}$ | Not known*                                                                                                    |
| Uncoating of the Influenza Virion                                            | $3.72 \times 10^{-2}$ | Not known*                                                                                                    |
Table 3 Over representation analysis results of the msPLS analysis on Marfan data (Continued)

| Pathway name                                      | p-value          | Associated with Marfan disease through pathway                                      |
|--------------------------------------------------|------------------|---------------------------------------------------------------------------------------|
| Budding                                          | $3.72 \times 10^{-2}$ | Not known*                                                                            |
| Release                                          | $3.72 \times 10^{-2}$ | Not known*                                                                            |
| Biosynthesis of protectins                       | $3.72 \times 10^{-2}$ | Possibly by proresolving lipid mediators [40]                                          |
| Degradation of the extracellular matrix          | $3.87 \times 10^{-2}$ | Extracellular matrix organization [31]                                                |
| RHO GTPases Activate Formins                     | $3.92 \times 10^{-2}$ | Extracellular matrix organization [41]                                                |
| TGF-beta receptor signaling in EMT (epithelial to mesenchymal transition) | $3.92 \times 10^{-2}$ | Signaling by TGF-beta receptor complex [31]                                           |
| Cell-extracellular matrix interactions           | $3.92 \times 10^{-2}$ | Extracellular matrix organization [31]                                                |
| Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE) | $4.13 \times 10^{-2}$ | Arachidonic acid metabolism [42]                                                      |
| Entry of Influenza Virion into Host Cell via Endocytosis | $4.13 \times 10^{-2}$ | Not known*                                                                            |
| Virus Assembly and Release                       | $4.13 \times 10^{-2}$ | Not known*                                                                            |
| Biosynthesis of maresin-like SPMs                | $4.33 \times 10^{-2}$ | Possibly by proresolving lipid mediators [40]                                          |
| Biosynthesis of specialized proresolving mediators (SPMs) | $4.41 \times 10^{-2}$ | Possibly by proresolving lipid mediators [40]                                          |
| Cytokine Signaling in Immune system              | $4.49 \times 10^{-2}$ | Cytokine signaling [31]                                                              |
| Synthesis of epoxy (EET) and dihydroxyeicosatrienoic acids (DHET) | $4.73 \times 10^{-2}$ | Arachidonic acid metabolism [42]                                                      |
| Arachidonic acid metabolism                      | $4.76 \times 10^{-2}$ | Arachidonic acid metabolism [42]                                                      |

The pathway names and p-values are obtained from https://reactome.org. Not known associations marked with asterisk (*) are all biomolecular pathways associated with reactions to influenza virus.

**Discussion**

In this paper, we propose a penalised extension of multiset Partial Least Squares path modeling in response to recent reports pointing out the lack of appropriate statistical methods for the simultaneous analysis of multiple high dimensional omics data sources.

msPLS addresses two challenges of integrated high dimensional omics data analysis; namely, it accounts for the relationships between multiple data sources and it provides interpretable results from analyses of hundreds of thousands of biomolecular variables.

Firstly, msPLS accounts for the hierarchical relationship between multiple high dimensional data sources in terms of a explanatory-response dependency structure. It can model dependencies between data sources, such as a hypothesised sequential information transfer in biomolecular domains, through explanatory-response data source pairs. This relationship structure can be easily redefined prior to the analysis, based on most recent biological knowledge. When the relationship is set according to biological knowledge, the biologically relevant biomarkers are identified instead of the variables that explain the most variance in the (combination of) phenotypic variables. Secondly, msPLS provides interpretable results in the form of combinations of biomarkers that have the highest explanatory power for the variance in the phenotypic variables. The biomarkers are extracted along with their weights that indicate their strength of contribution to the overall explained variance. These biomarkers can

**binding and postsynaptic events** (p-value $2.03 \times 10^{-4}$) are among the identified ones. Finally, some of the pathways identified in the thirds LV are “**Downstream TCR signaling**” (p-value $7.27 \times 10^{-7}$), “**Translocation of ZAP-70 to Immunological synapse**” (p-value $1.52 \times 10^{-59}$), “**TCR signaling**” (p-value $3.14 \times 10^{-41}$), and “**Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell**” (p-value $8.68 \times 10^{-14}$). The two most important clinical markers for CLL, namely the immunoglobulin heavy chain gene (IGHV) and the trisomy of chromosome 12 (trisomy12) were extracted as the first and second LV, respectively (Table 6) [11]. Thus similarly to MOFA, the first two set of LVs from msPLS are aligned among IGHV and trisomy12 (the absolute loading of IGHV is 0.66 in the first LV and the absolute loading of trisomy12 is 0.65 in the second LV), and these can be seen as axis of disease heterogeneity. The samples can be clearly clustered based on their IGHV and trisomy 12 status (Fig. 5). Also, there were 140 pathways with p-values < 0.05 discovered by the gene sets from msPLS that are not overlapping with the pathways discovered by the gene sets from MOFA. Notable pathways that might signal new knowledge discovery are “**Regulation of TP53 Activity through Phosphorylation**” (p-value $1.93 \times 10^{-4}$), “**TP53 Regulates Transcription of Cell Death Genes**” (p-value $8.1 \times 10^{-4}$) [47], “**HDACs deacetylating histones**” (p-value $8.73 \times 10^{-25}$) [48], “**HS-GAG degradation**” (p-value $1.22 \times 10^{-4}$), “**HS-GAG biosynthesis**” (p-value $7.6 \times 10^{-4}$), and “**Heparan sulfate/heparin (HS-GAG) metabolism**” (p-value $5.66 \times 10^{-3}$) [49].
be further appraised in the context of known biological pathways, for example via gene set enrichment analysis.

Through simulation studies and analyses of omics datasets, we show that msPLS is able to find the combination of biomarkers with the highest explanatory power for the variance in the phenotypic variables, and it can capture a higher proportions of variance in data sources than MOFA, a state-of-art LV based method for multiset omics data analysis. True positive rates of msPLS are reported from the simulation studies (see “Methods” section) to quantify the ability of finding the combination of explanatory variables from the data sources that

---

*Fig. 4* The co-expression pattern of the resulting Marfan genes queried on their biological process based functions. The figure was produced with GeneMania (available at https://genemania.org)
### Table 4

The percentage variation in the chronic lymphocytic leukemia (CLL) data sources explained by the subsequent LVs of msPLS and MOFA

|                  | Genomic variables msPLS | MOFA | Epigenomic variables msPLS | MOFA | Transcriptomic variables msPLS | MOFA | Drug response variables msPLS | MOFA |
|------------------|-------------------------|------|---------------------------|------|-------------------------------|------|-------------------------------|------|
| LV 1             | 72%                     | 15%  | 92%                       | 17%  | 92%                           | 7.5% | 57%                           | 15%  |
| LV 2             | 18%                     | 8.2% | 4%                        | 0.5% | 5%                            | 4.7% | 21%                           | 3.5% |
| LV 3             | 2%                      | <0.1%| 1%                        | <0.1%| 1%                            | 1.4% | 7%                            | 11.2%|
| LV 4             | <0.1%                   | <0.1%| <0.1%                     | 9%   |                               | <0.1%|                               |      |
| LV 5             | <0.1%                   | <0.1%| <0.1%                     | 2.8% |                               | 6.1% |                               |      |
| LV 6             | <0.1%                   | <0.1%| <0.1%                     | 4.8% |                               | 3.4% |                               |      |
| LV 7             | <0.1%                   | <0.1%| <0.1%                     | 2.6% |                               | 0.9% |                               | 0.4% |
| LV 8             | <0.1%                   | <0.1%| <0.1%                     | 1.9% |                               | 1%   |                               |      |
| LV 9             | <0.1%                   | 2.6% | 0.9%                      | 0.5% |                               | 1%   |                               |      |
| LV 10            | <0.1%                   | <0.1%| <0.1%                     | 2.2% |                               |      |                               |      |
| Total            | 92%                     | 24%  | 97%                       | 24%  | 98%                           | 38%  | 85%                           | 41%  |

### Table 5

The weights of the genomic, epigenomic, and transcriptomic variables extracted by msPLS from CLL data sources

| Name             | w     | Genomic variables Site | w     | Epigenomic variables Site | w     | Transcriptomic variables Site | w     |
|------------------|-------|------------------------|-------|---------------------------|-------|-------------------------------|-------|
| del11q22.3       | 0.31  | cg00369076             | 0.036 |                           |       |                               |       |
| del17p13         | 0.16  | cg232449085            | 0.36  |                           |       |                               |       |
| BRAF             | 0.17  | cg12208353             | 0.36  |                           |       |                               |       |
| TP53             | 0.21  | cg04694619             | 0.037 |                           |       |                               |       |
| IGHV             | -0.66 | cg20782816             | 0.38  |                           |       |                               |       |
|                 |       | cg00832703             | 0.037 |                           |       |                               |       |
|                 |       | cg01399475             | -0.036|                           |       |                               |       |
|                 |       | cg21398469             | 0.037 |                           |       |                               |       |
|                 |       | cg11181763             | 0.036 |                           |       |                               |       |
|                 |       | cg01360627             | 0.036 |                           |       |                               |       |
|                 |       | cg09087901             | 0.036 |                           |       |                               |       |
|                 |       | cg04848693             | 0.037 |                           |       |                               |       |
|                 |       | cg12522599             | 0.038 |                           |       |                               |       |
|                 |       | cg11090458             | 0.037 |                           |       |                               |       |
|                 |       | cg00148025             | 0.038 |                           |       |                               |       |
|                 |       | cg12032915             | 0.036 |                           |       |                               |       |
|                 |       | cg07529149             | 0.039 |                           |       |                               |       |
|                 |       | cg23844018             | 0.037 |                           |       |                               |       |
|                 |       | cg05213414             | 0.037 |                           |       |                               |       |
|                 |       | cg01928411             | 0.037 |                           |       |                               |       |
|                 |       | cg07999978             | 0.036 |                           |       |                               |       |
|                 |       | cg03035162             | 0.036 |                           |       |                               |       |
|                 |       | cg03462096             | 0.039 |                           |       |                               |       |
|                 |       | cg08171667             | 0.036 |                           |       |                               |       |
|                 |       | cg26441291             | 0.038 |                           |       |                               |       |
|                 |       | cg21400896             | 0.037 |                           |       |                               |       |
|                 |       | cg15236196             | 0.036 |                           |       |                               |       |
|                 |       | cg21394039             | 0.038 |                           |       |                               |       |
|                 |       | cg04613057             | 0.036 |                           |       |                               |       |
|                 |       | cg08496123             | 0.036 |                           |       |                               |       |
Table 6 The loadings of the three subsequent LVs extracted by msPLS from the genomic variables of the CLL data set

|                  | 1st set of LVs |                                                                 | 2nd set of LVs |                                                                 | 3rd set of LVs |                                                                 |
|------------------|----------------|-----------------------------------------------------------------|----------------|-----------------------------------------------------------------|----------------|-----------------------------------------------------------------|
| Name             | loading        | Name                | loading        | Name                | loading        | Name                | loading        |
| del11q22.3       | 0.31           | del11q22.3          | -0.27          | NRAS                | 0.35           | COL6A5              | -0.34          |
| del17p13         | 0.16           | trisomy12           | 0.65           | FAM47A              | -0.35          | FAT4                | -0.39          |
| BRAF             | 0.17           | del13q14_any        | -0.37          | CREBBP              | 0.15           | PRPF8               | -0.52          |
| TP53             | 0.21           | del14q24.3          | 0.20           | CREBBP              | 0.15           | PRPF8               | -0.52          |
| IGHV             | -0.66          | CREBBP              | 0.15           |                    |                |                     |                |

explain the most variance in response variables. True positive rates range from 0.61 to 0.99, indicating that the precision of finding truly associated variables improves with increasing sample size. Similarly, true negative rates are reported to quantify msPLS’s ability to exclude irrelevant variables from the final model. True negative rates are above 0.99 for each simulation studies, indicating that the final model excludes irrelevant variables with high precision, regardless of sample size.

The analysis of a genomewide omics dataset of 364,134 epigenomic, 18,424 transcriptomic and 47 proteomic variables resulted in biological relevant pathways. msPLS identified a combination of 40 epigenomic biomarkers and 52 transcriptomic biomarkers that has the highest explanatory power for the variance in the phenotypic proteome variables. Despite the low sample size of 37, msPLS identified biomarkers that can be found in known biological pathways associated with the pathophysiology of Marfan disease. Similarly to other LV based multivariate methods, it is possible to extract subsequent LVs with msPLS in a way that they explain a different portion of variance in the data sources. These subsequent LVs are orthogonal.
to each other, thus the newly obtained biomarkers can be interpreted as biological pathways independent from the ones that were discovered in the previous set of LVs. Comparing the results of msPLS and MOFA on the analyses of the CLL dataset, we found that the three set of LVs from the msPLS model captured 92%, 97%, 98% and 85% of the variation in the genomic, epigenomic, transcriptomic and drug response data sources, respectively, while the first ten LVs of MOFA captured a total of 24%, 24%, 38% and 41% of variation in those same data sources, respectively. msPLS, similarly to MOFA, identified the two most important clinical markers for CLL in its first two LVs, and in the “Results” section we additionally report many highly associated and possible novel pathways found through gene enrichment analysis using the MOFAR package.

Note that the present framework of msPLS assumes linear relationships between data sources and that the omics data is measured on a single homogeneous population. As an interesting future direction to extend msPLS is to incorporate non-linear relations in the model or to extend the model such that it can identify different subgroups in the samples.

Conclusions

In summary, msPLS is an appropriate multiset multivariate method that can account for the relationships between high dimensional data sources while it provides interpretable results through its sparse solutions. In the “Methods” section we also describe the algorithm for msPLS and we provide an implementation of the algorithm in the open source R software, which is uploaded with the manuscript and available upon request from the authors. We provide open source code that facilitates the use of our msPLS method on new data with the aim to leverage more and more biomolecular data to model and better understand the genetic architectures and biological processes of complex phenotypes, and ultimately to transition the information synthesised from omics data analyses into medical knowledge to improve diagnosis, prognosis and therapy.

Methods

Multiset sparse Partial Least squares path modeling

Multiset sparse Partial Least Squares path modeling (msPLS) is a multivariate approach for modeling the relationship between Q related data sources \( (X_1, \ldots, X_q, \ldots, X_Q) \), with the help of latent variables (LVs). Each data source contains \( p_q \) number of manifest variables (MVs), measured on the same \( n \) samples (i.e. \( X_q \in \mathbb{R}^{n \times p_q} \)), each data source is assigned to its corresponding LV \( (\xi_1, \ldots, \xi_q, \ldots, \xi_Q) \). The LVs are linear combinations of their MVs \( \xi_q = X_q w_q \), where \( \xi_q \in \mathbb{R}^{n \times 1} \) and \( w_q \in \mathbb{R}^{p_q \times 1} \). The relationship between the data sources is encoded in a connectivity matrix, like in Partial Least Squares path modeling (PLS-PM), and modelled through a multiple regression model between the LVs:

\[
\xi_q = \sum_{m=1}^{M_q} \theta_{qm} \xi_{m \rightarrow q} + v_q, \tag{1}
\]

where \( \sum_{m=1}^{M_q} \xi_{m \rightarrow q} \) denotes the sum of \( M_q \) LVs that are explanatory for \( \xi_q \), \( \theta_{qm} \) is the coefficient capturing the effect of the \( m \)th \( \xi_{m \rightarrow q} \) on \( \xi_q \), and \( v_q \) is white noise, following the notation of [22, 24] for PLS-PM. A full description for the PLS-PM algorithm can be found in [24] (Algorithm 6). The weight vectors \( w_q \) are estimated as

\[
w_q = \left( X_q' X_q \right)^{-1} X_q' \xi_q, \tag{2}
\]

or as

\[
w_q = (1/n) X_q' \xi_q, \tag{3}
\]

depending on the mode of the regression. PLS-PM denotes Eq. (2) as Mode A and Eq. (3) as Mode B regression. For msPLS, Mode A (i.e. multiple univariate regression) is used for the weight vectors of MVs that do not have any response MVs, and Mode B (i.e. multivariate regression) is used for the weight vectors of MVs that do have response MVs. The descriptions of the objective functions of PLS-PM can be found in [22, 24] and the objective function for msPLS is given by Eq. (5) in the “General case” section.

In a high dimensional setting (i.e. \( p_q >> n \)), the covariance matrix of \( X_q \) in Eq. (2) is non-invertible. To solve this problem, we propose to replace Eq. (2) with Elastic Net (ENet) penalization. Replacing the ordinary least square estimator in Eq. (2) with ENet penalisation has two advantages; not only we overcome the multicollinearity issue encountered in a high dimensional setting, but ENet also enforces sparse variable selection, which ease the interpretability of the final model. Equation (2) then becomes

\[
w_q = \arg \min_{w_q} w_q' \left( \frac{X_q' X_q + \lambda_2 I}{1 + \lambda_2} \right) w_q - 2 \xi_q' X_q w_q + \lambda_1 w_q, \tag{4}
\]

where \( \lambda_1 \) denotes the LASSO penalty and \( \lambda_2 \) denotes the Ridge penalty parameters [27].

An example of msPLS with three data sources

Let us first examine an application of msPLS to three data sources. Given data sources \( X_1, X_2, \) and \( X_3 \) with \( p_1, p_2 \) and \( p_3 \) number of variables, measured on \( n \) samples (i.e. \( X_1 \in \mathbb{R}^{n \times p_1}, X_2 \in \mathbb{R}^{n \times p_2} \) and \( X_3 \in \mathbb{R}^{n \times p_3} \)), we consider the following relationships between the data sources: \( X_1 \) and \( X_2 \) have a symmetric relation (i.e. they are responses for each other). Furthermore, there are asymmetric relations between \( X_1 \) and \( X_3 \), and between \( X_2 \) and \( X_3 \), such that \( X_3 \)
is response for both $X_2$ and $X_1$ (Fig. 6). These relationships are encoded in a three dimensional connectivity matrix $C$ (i.e. $C \in \{0, 1\}^{3 \times 3}$), where the entry $c_{qj}$ is $1$ if data source $q$ is response for data source $q'$, and $0$ otherwise (where $q \neq q'$ and $c_{qq'}$ indicates the element from $q$th row and $q'$th column of matrix $C$). The objective of the analysis is then to simultaneously extract the MVs from $X_1$ and $X_2$ with the highest explanatory power for the variance in MVs of $X_3$.

**Three data sources msPLS algorithm** Given data sources $X_1$, $X_2$, and $X_3$, and $\Theta = C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$

1. **Preliminary steps**
   - (a) Center and scale $X_1$, $X_2$, and $X_3$
   - (b) Set $w_1^{(0)}$, $w_2^{(0)}$ and $w_3^{(0)}$ initial weight vectors to arbitrary vectors of $[1, 1, ..., 1]^T$ with length $p_1$, $p_2$ and $p_3$, respectively
   - (c) Define convergence criterion $CRT = 1$ and a small positive tolerance $\gamma = 10^{-6}$

2. **Iterative regression steps**
   While $CRT \geq \gamma$:
   a. **Estimate initial LVs**
      $$\xi_1 \propto X_1 w_1^{(0)}; \text{where } \propto \text{ indicates that } \xi_1 \text{ is normalised to unit variance}$$
      $$\xi_2 \propto X_2 w_2^{(0)}$$
      $$\xi_3 \propto X_3 w_3^{(0)}$$
   b. **Model the relationship between data sources**
      (i) Let vector $c_q$ be the $q$-th row of $C$ that indicates the data sources that are explanatory for data source $q$, i.e.
      $$c_1 = [0, 1, 0], c_2 = [1, 0, 0], c_3 = [1, 1, 0];$$
      indicating $X_1$ has one explanatory, $X_2$ has one explanatory and $X_3$ has two explanatory data sources
      If $\sum_{i=1}^{3} c_{qi} > 0$, i.e. if data source $q$ has any explanatory data sources:
      $$\Theta_{c_{qj}} = \left[ Z_{cq} Z_{cq} \right]^{-1} Z_{cq}^T c_q,$$
      where $Z$ is the matrix of column bind LVs, i.e. $Z = [\xi_1, \xi_2, \xi_3]$, and $Z_{cq}$ is the matrix of column bind explanatory LVs of data source $q$. Then $\theta_{c_{qj}}$ is calculated as follows:
      For $c_1$ we calculate
      $$\Theta_{c_{11}} = \theta_{11} = [\xi_1^T \xi_1]^{-1} [\xi_1, \xi_2] \xi_1,$$
      and the value of $\theta_{11}$ and $\theta_{31}$ remain $0$.
      For $c_2$ we calculate
      $$\Theta_{c_{22}} = \theta_{22} = [\xi_2^T \xi_2]^{-1} [\xi_2, \xi_3] \xi_2,$$
      and the value of $\theta_{22}$ and $\theta_{32}$ remain $0$, and for $c_3$ we calculate
      $$\Theta_{c_{33}} = \theta_{13} = [\xi_3^T \xi_3],$$
      where the entries $\theta_{13}$ and $\theta_{23}$ are obtained from the multiple regression step
      $$[\xi_1, \xi_2]^T [\xi_1, \xi_2]^{-1} [\xi_1, \xi_2] \xi_3,$$
      and the value of $\theta_{33}$ remains $0$.
      (ii) Let vector $c_{q'}$ be the $q'$-th column of $C$ that indicates the data sources that are response for data source $q'$, i.e.
      $$c_1 = [0, 1, 0], c_2 = [1, 0, 1], c_3 = [0, 0, 0];$$
      indicating $X^T$ has two responses, $X_2$ has two responses and $X_3$ has no response data sources
      If $\sum_{i=1}^{3} c_{iq} > 0$, i.e. if data source $q'$ has any response data sources:

---

**Fig. 6** The proposed relationship between three data sources $X_1$ and $X_2$ have a symmetric relation (i.e. they are responses for each other) and $X_3$ have asymmetric relation with both $X_1$ and $X_2$ (i.e. $X_3$ is response for both $X_1$ and $X_2$).


\[ \Theta_{c_q'q'} = \text{cor}(\xi_{q'}, \xi_{c_q}), \]

i.e., for \( c_1 \) we calculate

\[ \Theta_{c_11} = \frac{\theta_1}{\theta_{31}} = \text{cor}(\xi_1, \xi_2) \]

and for \( c_2 \) we calculate

\[ \Theta_{c_21} = \frac{\theta_2}{\theta_{32}} = \text{cor}(\xi_2, \xi_1) \]

After Steps (b-i) and (b-ii), the entries of \( \Theta \) are:

\[ \Theta = \begin{bmatrix} 0 & \text{cor}(\xi_1, \xi_2) & \theta_{13} \\ \text{cor}(\xi_1, \xi_3) & 0 & \theta_{23} \\ \text{cor}(\xi_2, \xi_3) & \theta_{13} & 0 \end{bmatrix}, \]

Notice that \( \theta_{21} \) and \( \theta_{12} \) in Step (b-i) are overwritten in Step (b-ii). This is because \( \xi_1 \) and \( \xi_2 \) are both responses to one another.

c. Re-estimate the latent variables

\[ [\xi_1, \xi_2, \xi_3] = [\xi_1, \xi_2, \xi_3] \Theta \]

d. Estimate the new \( \mathbf{w}_1^{(1)} \) weights

\[ \mathbf{w}_1^{(1)} = \text{argmin}_{\mathbf{w}_1^{(0)}} \left( \mathbf{X}_1 \mathbf{X}_1 + 2 \mathbf{I} \right) 
\]

\[ = \mathbf{X}_1^\top \mathbf{X}_1 \mathbf{w}_1^{(0)} + \lambda_1 \mathbf{w}_1^{(0)} \]

\[ \mathbf{w}_2^{(1)} = \text{argmin}_{\mathbf{w}_2^{(0)}} \left( \mathbf{X}_2 \mathbf{X}_2 + 2 \mathbf{I} \right) 
\]

\[ = \mathbf{X}_2^\top \mathbf{X}_2 \mathbf{w}_2^{(0)} + \lambda_2 \mathbf{w}_2^{(0)} \]

\[ \mathbf{w}_3^{(1)} = \left[ \left( \mathbf{X}_1^\top \mathbf{X}_1 \right)^{-1} \mathbf{X}_1 \mathbf{X}_2 \right] \]

e. Evaluate the convergence criteria and discard the old \( \mathbf{w}_0^{(0)} \) weights

\[ \text{CRT} = \sum_{q=1}^{Q} (\mathbf{w}_1^{(1)} - \mathbf{w}_1^{(0)})^2 \]

\[ \mathbf{w}_1^{(0)} = \mathbf{w}_1^{(1)}, \mathbf{w}_2^{(0)} = \mathbf{w}_2^{(1)} \]

and \( \mathbf{w}_3^{(0)} = \mathbf{w}_3^{(1)} \)

3. Upon convergence, return \( \mathbf{w}_1^{(0)}, \mathbf{w}_2^{(0)}, \) and \( \mathbf{w}_3^{(0)} \)

### General case

The general case for msPLS can be described as follows. Given \( Q \) related data sources \( X_1, ..., X_q, ..., X_Q \) with \( p_1, ..., p_q, ..., p_Q \) corresponding MVs, measured on \( n \) samples (i.e. \( X_1 \in \mathbb{R}^{n \times p_1}, ..., X_q \in \mathbb{R}^{n \times p_q}, ..., X_Q \in \mathbb{R}^{n \times p_Q} \)), and a \( Q \)-dimensional connectivity matrix \( \Theta \) (i.e. \( \Theta \in [0,1]^{Q \times Q} \)), where the entry \( \Theta_{q,q} \) is 1 if data source \( q \) is a response data source for data source \( q' \) and 0 otherwise. The goal of the analysis then is to optimise the following objective function (i) with respect to data source \( q' \):

\[
\text{OF} = \text{argmax} \begin{cases}
\sum_{q'=1}^{Q} \text{cor}(\xi_{q'}, \xi_{q})^2 & \text{if } \sum_{q'=1}^{Q} \xi_{q'} > 0 \\
\sum_{m=1}^{M_{q'}} \text{cor}(\xi_{m} \rightarrow \xi_{q'}, X_{q'}^{(i)})^2 & \text{otherwise}
\end{cases}
\]

(5)

where \( \xi_{q'} \) is the LV of \( X_{q'} \), \( c_{q'} \) indicates the \( q' \)th column of matrix \( \Theta \) (i.e. \( |c_{q'}| > 0 \) indicates that data source \( q' \) have at least one response data source), \( X_{q'}^{(i)} \) denotes the \( i \)th column of data source \( X_{q'} \) (i.e. the \( i \)th MV of \( X_{q'} \)), \( \sum_{q=1}^{Q} \xi_{q'} \rightarrow r \) denotes the sum of \( R_q' \) LVs that are response for \( \xi_{q'} \), and \( \sum_{m=1}^{M_{q'}} \xi_{m} \rightarrow q \) denotes the sum of \( M_{q'} \) LVs that are explanatory for \( \xi_{q'} \). In other words, if data source \( q' \) have at least one response data source, then the squared correlation between \( \xi_{q'} \) and the combination of its response LVs is maximised, and if data source \( q' \) does not have any response data sources, the correlation between the MVs of \( X_{q'} \) and the combination of the explanatory LVs for \( X_{q'} \) is maximised. The symmetric relationship between \( X_q \) and \( X_{q'} \) is indicated as \( c_{qq'} = c_{q'q} = 1 \), in which case the OF of their pairwise analysis is to maximise the correlation between their LVs \( \xi_q \) and \( \xi_{q'} \), corresponding to the characteristic objective function of Canonical Correlation Analysis (CCA) [8, 22, 50]. In an asymmetric relationship, the OF of a pairwise analysis is to maximise the sum of squared correlation between the explanatory LV \( \xi_q \) and the response MVs in \( X_{q'} \), corresponding to the characteristic objective function of Redundancy Analysis (RDA) [10, 22, 51]. This direct correspondence with CCA and RDA is described in Additional file 1 under the Modes of relationships between data sources section.

Next we describe the general algorithm for \( Q \) data sources.

### General msPLS algorithm

Given \( Q \) data sources \( X_1, ..., X_q, ..., X_Q \), and \( \Theta = \Theta \in [0,1]^{Q \times Q} \), where

\[
c_{q,q'} = \begin{cases} 
1 & \text{if } X_q \text{ response for } X_{q'} \\
0 & \text{otherwise}
\end{cases}
\]

1. **Preliminary steps**

   (a) Center and scale \( X_1, ..., X_q, ..., X_Q \)
   
   (b) Set \( w_0^{(0)} \) to arbitrary weight vectors \([1, 1, ..., 1]^\top\)

   (c) Define convergence criterion \( \text{CRT} = 1 \) and a small positive tolerance \( \gamma = 10^{-6} \)

2. **Iterative regression steps**

   While \( \text{CRT} \geq \gamma \):

   a. Estimate initial LVs

   \[ \xi_q \propto X_q w_0^{(0)} \]

   where \( q \) is the index from 1 to \( Q \) and \( \propto \) indicates that \( \xi_q \) is normalised to unit variance

   b. Model the relationship between data sources

   (i) Let vector \( c_{q} \) be the \( q \)-th row of \( C \) that indicates the data sources that are exploratory for data source \( q \)
If $\sum_{i=1}^{Q} c_{q_i} \geq 0$, i.e. if data source $q$ has any explanatory data sources:

$$\Theta_{c,q} = (Z_{c}^{T} Z_{c})^{-1} Z_{c}^{T} \xi_{q},$$

where $Z$ is the matrix of column bind LVs, i.e. $Z = [\xi_{1}, \xi_{2}, \xi_{3}]$, and $Z_{c}$ is the matrix of the column-bind explanatory LVs of data source $q$.

(ii) Let vector $c_{q}$ be the $q'$-th column of $C$ that indicates the data sources that are response for data source $q'$

If $\sum_{i=1}^{Q} c_{q_i} > 0$, i.e. if data source $q'$ has any responses:

$$\Theta_{c,q'} = \text{cor}(\xi_{q'}, \xi_{c,q'})$$

c. Re-estimate the LVs

$$\tilde{Z} = Z\Theta$$

d. Estimate the new $w_{q}^{(1)}$ weights

If $X_{q}$ doesn't have any response data sources:

$$w_{q}^{(1)} = \left[ [Z_{q}^{T} \tilde{Z}_{q}]^{-1} \right]_{q} X_{q}^{T} X_{q}$$

otherwise:

$$w_{q}^{(1)} = \arg\min_{w_{q}} w_{q}^{(0)}(X_{q}^{T} X_{q} + \lambda_{1} I)^{-1} X_{q}^{T} X_{q} w_{q}^{(0)} - 2 X_{q}^{T} X_{q} w_{q}^{(0)} + \lambda_{2} w_{q}^{(0)}$$

e. Evaluate the convergence criteria and discard the old $w_{q}^{(0)}$ weights and calculate OF from Eq. (5) with respect to each data sources

$$\text{CRT} = \sum (w_{q}^{(1)} - w_{q}^{(0)})^2$$

$$w_{q}^{(0)} = w_{q}^{(1)}$$

3. Upon convergence, return $w_{q}^{(0)}$

After the algorithm converges, the $w_{q}$ weights indicate the contribution of explanatory MVs from the $q$th data source towards the overall explained variance in the response MVs or LVs (see Additional file 1 under Modes of relationships between data sources section). Through the penalisation of the multivariate regression in Step (2-d), a small subset of explanatory MVs are extracted, namely those with the highest explanatory power for the variance in their response MVs or LVs. The extracted set of MVs can be further explored in terms of known biological pathways, for example through gene enrichment analysis.

Multiple latent variables per dataset

It is possible to extract multiple LVs per dataset in a way that they explain a different portion of variance in the MVs. The explained variance is based on the $R^2$ statistic obtained from the regression model from Step (2-d) in the general msPLS algorithm. The subsequent latent variables can be obtained by applying msPLS to the residual data sources, where the residuals data sources are calculated as

$$X_{q(\alpha)}^{\text{res}} = X_{q(\alpha)} - \hat{X}_{q(\alpha)}$$

$$= X_{q(\alpha)} - \xi_{q(\alpha)} [\xi_{q(\alpha)}^{T} \xi_{q(\alpha)}]^{-1} \xi_{q(\alpha)}^{T} X_{q(\alpha)}.$$
the confidence interval of the optimisation criterion was approximated through bootstrapping.

Data generation for simulation studies

For all simulation studies, we generated three data sources, $X_1$, $X_2$ and $X_3$, in such way that the relationship between data sources resembles the one we describe in “Chronic lymphocytic leukaemia data” section (Fig. 4).

All $X_q$s were assigned to $p_q$ number of MVs (i.e. $p_1 = p_2 = 1000$, $p_3 = 100$) from which $k_q$ variables were associated with their LVs and response MVs (i.e. $k_1 = k_2 = k_3 = 10$), and there were $j_q$ number of not associated MVs (i.e. $j_1 = 990$, $j_2 = 990$, $j_3 = 90$). The number of samples are denoted by $n$ samples (i.e. $X_q \in \mathbb{R}^{n \times p_q}$ with $k_q$ associated MVs and $j_q$ not associated MVs, $p_q = k_q + j_q$), and in the first three simulation studies $n$ varied from 1, 100, and 250.

$X_1$ and $X_2$ were generated from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma$, and their response MVs in $X_3$ was generated from LVs $\xi_1$ and $\xi_2$, as follows:

1. $\Sigma = I_{5000}$
2. Replace $\Sigma_{1001:1010;1:10} = \Sigma_{1:10,1001:10:10} = H$, where $H \in \mathbb{R}^{10 \times 10}$ distributed over $\mathcal{N}(0,0.05)$.
3. $D \sim \mathcal{N}(0, \Sigma)$ where $D \in \mathbb{R}^{n \times 2000}$
4. $X_1 = D_{1:n,1:1000}$ and $X_2 = D_{1:n,1001:2000}$

$\Sigma$ is a $p_1 + p_2$ dimensional identity matrix where elements $\Sigma_{1001:1010;1:10} = \Sigma_{1:10,1001:10:10}$ were replaced with $H$, where $H \in \mathbb{R}^{10 \times 10}$ was distributed over $\mathcal{N}(0,0.05)$. $D$ was sampled from the multivariate normal distribution with mean 0 and covariance matrix $\Sigma$, and $D$ was used to generate $X_1$ and $X_2$. Next, the weight vectors were generated;

5. $w_q = (w_q(1), w_q(2), ..., w_q(k_q), w_q(k_q+1), ..., w_q(p_q))$, $w_q(k_q) = w_{\text{associated}}, w_q(p_q) = 0$.

The associated $k$ MVs had higher regression weights with their LVs (with weights $w_{\text{associated}}^2 = 0.7$, $w_{\text{associated}}^2 = 0.6$, $w_{\text{associated}}^2 = 0.3$) than the not associated $j$ MVs (i.e. $w_q = (w_q(1), w_q(2), ..., w_q(k_q), w_q(k_q+1), ..., w_q(p_q))$, $w_q(k_q) = w_{\text{associated}}, w_q(p_q) = 0$). The LVs were generated as a linear combination of the MVs and weights,

6. $\xi_1 = X_1w_1$ and $\xi_2 = X_2w_2$.

$X_3$ was generated with from $\xi_1$ and $\xi_2$. The $k_3$ associated LVs were sampled from the normal distribution with mean $\theta_1 \xi_1 + \theta_2 \xi_2$ (where $\theta_2$ is the regression coefficient from Eq. (1), with $\theta_1 = 0.8$ and $\theta_2 = 0.7$) and standard deviation $\sqrt{1 - (w_3)^2}$. The $j_3$ not associated variables were sampled from the standard normal distribution;

7. $X_3 \in \mathbb{R}^{n \times 100}$
8. For $l = 1, ..., k_3$:

$$X_{3(l)} \text{ distributed } \mathcal{N}(\theta_1 \xi_1 + \theta_2 \xi_2 \sqrt{1 - (w_3)^2})$$

(9) For $i = k_3 + 1, ..., p_3$:

$$X_{3(l)} \text{ distributed } \mathcal{N}(0, 1)$$

In addition, we designed a fourth simulation study, where the size of the data resembled the size of the omics data sources, described in “Results” section (i.e. $p_1 = 360000$, $p_2 = 18000$, $p_3 = 47$, $k_1 = k_2 = 40$, $k_3 = 10$, and $n = 37$).

Simulation study results

We generated data as described in above with three different sample sizes, i.e. $n = 50$, $n = 100$, $n = 250$. To assess msPLS’s ability to identify the $k_q$ associated MVs from explanatory data sources $X_1$ and $X_2$, we used the true-positive rate (TPR) and true-negative rate (TNR) measures over 1000 simulations.

TPR measures the proportion of associated MVs included in the final model (i.e. those that are assigned to non-zero $w$ weights) to either the number of associated MVs that were generated, or to the total number of non-zero $w$ weights, whichever is smaller (i.e. $\text{TPR}_q = \sum_{kq}^{kq} I(w_q(kq) \neq 0)/\min(k_q, \sum_{l=1}^{pq} I(w_q(l) \neq 0))$). TPR ranges from 0.61 to 0.99 and increases with increasing sample size when the variable size held constant (Table 7).

TNR measures the proportion of not associated MVs excluded from the model to the number of not associated MVs that were generated (i.e. $\text{TNR}_q = \sum_{l=1}^{pq} I(w_q(l) = 0)/j_q$). TNR rates resulted in 0.99 and were not affected by the sample size (Table 7).

We assessed the statistical significance of the resulting models in respect to the optimised OFs through permutation, and the confidence intervals of the optimised OFs were constructed through bootstrapping (see the “Selecting the optimal penalisation parameters and assessing the statistical significance of the resulting model” section). All the three models obtained on the three different sample sizes with constant variable size were statistically significant, and the confidence interval of the optimised OFs shrank with increased sample size (Fig. 7). For $n = 50$, the optimised OF with respect to $X_3$ resulted in 114.21 (95% CI [85.02, 163.43], $p$-value <0.001), for $n = 100$ the optimised OF resulted in 117.12 (95% CI [71.58, 148.69], $p$-value <0.001), and for $n = 250$ the optimised OF resulted in 123.61 (95% CI [91.04, 130.94], $p$-value <0.001).

| $n$ | TPR | TNR |
|-----|-----|-----|
| 50  | 0.67 | 0.99 |
| 100 | 0.66 | 0.99 |
| 250 | 0.99 | 0.99 |
| 37  | 0.99 | 0.99 |
Fig. 7 The null distributions of the optimisation criteria (with respect to $X_3$) for the simulated data with different sample sizes ($n = 50, 100, 250$), obtained after 1000 permutations. The red bars indicate the optimisation criteria obtained applying msPLS to the original data with the optimal $\lambda_1$ parameters for UST. The red dots are the bootstrapped values, and the dashed red bars are the 95% confidence intervals.

Availability and Requirements
Project name: msPLS implementation
Project home page: http://uva.csala.me/mspls and https://github.com/acsala/2018_msPLS
Operating system(s): Platform independent
Programming language: R
Other requirements: additional R packages listed in the source code, freely available from the Comprehensive R Archive Network online
License: MIT
Any restrictions to use by non-academics: MIT license applies

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3286-3.

Additional file 1: Appendix. Supplementary materials for the proposed method and results in pdf file format.
Additional file 2: Supplementary materials for the gene set enrichment analysis results of the Marfan data, obtained from Reactome in comma separated file format.
Additional file 3: Supplementary OMIM Gene Map Retrieval results of the Marfan data, obtained from OMIM in xls file format.
Additional file 4: Supplementary co-expression pattern results of the Marfan data, obtained from Gene Mania in txt file format.

Additional file 5: Supplementary gene set enrichment analysis results of the Chronic lymphocytic leukaemia data analyzed by MOFA. Results are obtained by the MOFA R package and exported as txt file format.
Additional file 6: Supplementary gene set enrichment analysis results of the Chronic lymphocytic leukaemia data analyzed by msPLS. Results are obtained by the MOFA R package and exported as txt file format.
Additional file 7: Supplementary material on the overlapping gene set enrichment analysis results between msPLS and MOFA on the Chronic lymphocytic leukaemia data. Results are obtained by the MOFA R package and exported as txt file format.

Abbreviations
CCA: Canonical correlation analysis; CCL: Chronic lymphocytic leukaemia; CI: Confidence interval; ENeT: Elastic net; LVs: Latent variables; LASSO: Least absolute shrinkage and selection operator; MIVs: Manifest variables; MOFA: Multi-omics factor analysis; msPLS: Multiset sparse partial least squares path modeling; OF: Objective function; OMIM: Online mendelian inheritance in man; PLS-PM: Partial least squares path modeling; RDA: Redundancy analysis; TNR: True-negative rate; TPR: True-positive rate; UST: Univariate soft thresholding

Acknowledgements
This work was previously presented at the 40th Annual Conference of the International Society for Clinical Biostatistics. The authors would like to thank the anonymous reviewer and the editor for their valuable comments and suggestions.

Authors’ contributions
AHZ designed the idea and supervised the study process. ACS analysed the data, implemented the results and wrote the manuscript. AHZ and MHH revised the manuscript. All authors read and approved the final manuscript.
UC, Meyer N, Morita H, Olthausen J, O'Mahony L, Pezer M, Prati M, Rebane A, Rhyner C, Rinaldi A, Sokolowska M, Stanic B, Sugita K, Treis A, van de Veen W, Wanke K, Wawrzyniak J, Wawrzyniak P, Wirz OF, Zakuz JS, Akdis CA. Interleukins (from IL-1 to IL-38), interferons, transforming growth factor β, and TNF-α: Receptors, functions, and roles in diseases. J Allergy Clin Immun. 2016;138(4):984–1010. https://doi.org/10.1016/j.jaci.2016.06.033.

36. Ju X, Ijaz T, Sun H, LeJeune W, Vargas G, Shilagard T, Recinos A, Milewicz DM, Brasier AR, Tilton RG. IL-6 Regulates Extracellular Matrix Remodeling Associated With Aortic Dilation in a Fibrillin-1 Hypomorphic mgR/mgR Mouse Model of Severe Marfan Syndrome. J Am Heart Assoc. 2014;3(1):1–13. https://doi.org/10.1161/JAHA.113.000476.

37. Lenk GM, Tromp G, Weinheimer S, Gatalica Z, Berguer R, Kuivanen M. Whole genome expression profiling reveals a significant role for immune function in human abdominal aortic aneurysms. BMC Genomics. 2007;8(1):237. https://doi.org/10.1186/1471-2164-8-237.

38. Davis MR, Arner E, Dufy CRE, De Sousa PA, Dahlman I, Arner P, Summers KM. Expression of FBN1 during adipogenesis: Relevance to the lipodystrophy phenotype in Marfan syndrome and related conditions. Mol Genet Metab. 2016;119(1-2):174–85. https://doi.org/10.1016/j.mgme.2016.06.009.

39. Syypong H, Chung A, Yang H, van Breemen C. Dysfunction of endothelial and smooth muscle cells in small arteries of a mouse model of Marfan syndrome. British J Pharmacol. 2009;158(6):1597–608. https://doi.org/10.1111/j.1476-5381.2009.00439.x.

40. Rayner KJ. Cell Death in the Vessel Wall. Arterioscler Thromb Vasc Biol. 2017;37(7):75–81. https://doi.org/10.1161/ATVBAHA.117.309229.

41. Lukashov M. ECM signalling: orchestrating cell behaviour and misbehaviour. Trends Cell Biol. 1998;8(11):437–41. https://doi.org/10.1016/S0962-8924(98)01362-2.

42. Soto ME, Guarnier-Lans V, Herrera-Morales KY, Pérez-Torres I. Participation of Arachidonic Acid Metabolism in the Aortic Aneurysm Formation in Patients with Marfan Syndrome. Front Physiol. 2018;9(FEB):1–13. https://doi.org/10.3389/fphys.2018.00077.

43. Chung AW, Au Yeung K, Sandor GG, Judge DP, Dietz HC, Van Bremen C. Loss of elastic fiber integrity and reduction of vascular smooth muscle contraction resulting from the upregulated activities of matrix metalloproteinase-2 and-9 in the thoracic aortic aneurysm in marfan syndrome. Circ Res. 2007;101(5):512–22.

44. Neptune ER, Frischmeyer PA, Arking DE, Myers L, Bunton TE, Gayraud B, Ramirez F, Sakai LY, Dietz HC. Dysregulation of tgf-β activation contributes to pathogenesis in marfan syndrome. Nat Genet. 2003;33(3):407.

45. Bolar N, Van Laer L, Loey SL. Marfan syndrome: from gene to therapy. Curr Opin Pediatr. 2012;24(4):498–504.

46. Judge DP, Dietz HC. Marfan’s syndrome. Lancet. 2005;366(9501):1965–76.

47. Farooqui MZ, Valdez J, Martyr S, Aue G, Saba N, Niemann CU, Herman SE, Tian X, Marti G, Soto S, et al. Ibrutinib for previously untreated or relapsed/refractory chronic lymphocytic leukemia with t(14;18) aberrations: a phase 2, single-arm trial. Lancet Oncol. 2015;16(2):169–76.

48. Van Damme M, Crompton E, Meuleman N, Mineur P, Bron D, Lagneaux L, Stamatopoulos B. Hdad isoenzyme expression is deregulated in chronic lymphocytic leukemia b-cells and has a complex prognostic significance. Epigenetics. 2012;7(12):1403–12.

49. Sebestyen A, Kovalszky I, Mihalik R, Gallai M, Bocsi J, Laszlo E, Benedek S, Steeter L, Kopper L. Expression of syndecan-1 in human b cell chronic lymphocytic leukemia. Eur J Canc. 1997;33(13):2273–7.

50. Waaijenborg S, Zwinderman AH. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. BMC Bioinformatics. 2009;10(1):315. https://doi.org/10.1186/1471-2105-10-315.

51. Csala A, Voorbraak FPJM, Zwinderman AH, Hof MH. Sparse redundancy analysis of high-dimensional genetic and genomic data. Bioinformatics (Oxford, England). 2017;33(20):3228–34. https://doi.org/10.1093/bioinformatics/btx374.

**Publisher’s Note**
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.