frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 06 December 2023
DOI 10.3389/fgene.2023.1286800

OPEN ACCESS

EDITED BY
Rosalba Giugno,
University of Verona, Italy

REVIEWED BY
Kaya Kuru,
University of Central Lancashire, United Kingdom
Xiaofeng Li,
Heilongjiang International University, China

*CORRESPONDENCE
Zuqi Li,
zuqi.li@kuleuven.be

†PRESENT ADDRESS
Diane Duroux,
ETH AI Center, ETH Zurich, Zurich, Switzerland

RECEIVED 31 August 2023
ACCEPTED 14 November 2023
PUBLISHED 06 December 2023

CITATION
Li Z, Melograna F, Hoskens H, Duroux D, Marazita ML, Walsh S, Weinberg SM, Shriver MD, Müller-Myhsok B, Claes P and Van Steen K (2023), netMUG: a novel network-guided multi-view clustering workflow for dissecting genetic and facial heterogeneity.
Front. Genet. 14:1286800.
doi: 10.3389/fgene.2023.1286800

netMUG: a novel network-guided multi-view clustering workflow for dissecting genetic and facial heterogeneity

Zuqi Li[1,2]*, Federico Melograna[1], Hanne Hoskens[1,2], Diane Duroux[3†], Mary L. Marazita[4,5], Susan Walsh[6], Seth M. Weinberg[4,5], Mark D. Shriver[7], Bertram Müller-Myhsok[8], Peter Claes[1,2,9,10] and Kristel Van Steen[1,3]

[1]BIO3 - Laboratory for Systems Medicine, Department of Human Genetics, KU Leuven, Leuven, Belgium, [2]Medical Imaging Research Center, University Hospitals Leuven, Leuven, Belgium, [3]BIO3 - Laboratory for Systems Genetics, GIGA-R Medical Genomics, University of Liège, Liège, Belgium, [4]Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, University of Pittsburgh, Pittsburgh, PA, United States, [5]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, United States, [6]Department of Biology, Indiana University Indianapolis, Indianapolis, IN, United States, [7]Department of Anthropology, Pennsylvania State University, State College, PA, United States, [8]Max Planck Institute of Psychiatry, Munich, Germany, [9]Department of Electrical Engineering, KU Leuven, Leuven, Belgium, [10]Murdoch Children's Research Institute, Melbourne, VIC, Australia

Introduction: Multi-view data offer advantages over single-view data for characterizing individuals, which is crucial in precision medicine toward personalized prevention, diagnosis, or treatment follow-up.

Methods: Here, we develop a network-guided multi-view clustering framework named netMUG to identify actionable subgroups of individuals. This pipeline first adopts sparse multiple canonical correlation analysis to select multi-view features possibly informed by extraneous data, which are then used to construct individual-specific networks (ISNs). Finally, the individual subtypes are automatically derived by hierarchical clustering on these network representations.

Results: We applied netMUG to a dataset containing genomic data and facial images to obtain BMI-informed multi-view strata and showed how it could be used for a refined obesity characterization. Benchmark analysis of netMUG on synthetic data with known strata of individuals indicated its superior performance compared with both baseline and benchmark methods for multi-view clustering. The clustering derived from netMUG achieved an adjusted Rand index of 1 with respect to the synthesized true labels. In addition, the real-data analysis revealed subgroups strongly linked to BMI and genetic and facial determinants of these subgroups.

Discussion: netMUG provides a powerful strategy, exploiting individual-specific networks to identify meaningful and actionable strata. Moreover, the implementation is easy to generalize to accommodate heterogeneous data sources or highlight data structures.

KEYWORDS

multi-view clustering, multi-modal data, obesity subtyping, facial images, genomics, personalized network, distinguishing genetics, social heterogeneity

1 Introduction

In machine learning, unsupervised clustering has been widely discussed and applied. It refers to a data-analysis problem where the true classification of individuals (or items) is unknown, and we derive clusters by exploiting between-individual similarity. Clustering serves as an important tool for population stratification, image segmentation, anomaly detection, etc. (Ghosal et al., 2020). Specifically, clustering in medicine helps subgroup patients with potential disease risks and characterize each subgroup with distinctive genetic information. Patient subgrouping or disease subtyping plays an essential role in precision medicine, given that traditional medicine tends to offer a one-size-fits-all solution over the entire population and often overlooks heterogeneity (Spycher et al., 2008; Saria and Goldenberg, 2015). While designing a drug customized for every patient may not be feasible, fine-scaled disease subtyping is tractable and can facilitate more personalized prevention, diagnosis, and treatment.

To better characterize a disease or phenotype, it is beneficial to turn to different sources, i.e., multi-view data, that are jointly more comprehensive and informative than single modalities (Gligorijević et al., 2016). For example, prior work has shown that multi-view clustering algorithms often outperform single-view ones (Abavisani and Patel, 2018; Chauvel et al., 2020; Wen et al., 2021). However, many multi-view clustering methods have difficulty finding the consensus between modalities or exploiting relationships within and between views. Canonical correlation analysis (CCA) provides a solution to obtain optimal linear transformations of every data type to maximize their correlation (Hotelling, 1936). Moreover, sparse CCA (sCCA) introduces a sparsity parameter for each view, which can enforce the canonical weights on most features to be zero (Witten et al., 2009). sCCA both reduces the feature dimensionality and removes noisy features. Therefore, we can use this method to select the most meaningful features from datasets as input for the subsequent clustering. Sparse multiple canonical correlation network analysis (SmCCNet) can take an extraneous variable to detect modules of multi-view features with maximal canonical correlation between the data views informed by a phenotype of interest (Shi et al., 2019).

Current multi-view sample clustering methods integrate data views in different ways, e.g., concatenating all features, mapping views to a shared low-dimensional space, and merging between-sample relationship matrices from every view. The approach iCluster+ was designed to predict cancer subtypes from various data types by constructing a common latent space from all views (Mo et al., 2013). The extension iClusterBayes adopts a Bayesian model in the latent space (Mo et al., 2018). PintMF imposes a sparsity penalty on matrix factorization to integrate multiple data types into a common space (Pierre-Jean et al., 2022). Other state-of-the-art methods focus on combining similarity matrices from every data view. For example, Spectrum is such a method with a self-tuning density-aware kernel and similarity network fusion (SNF) that transforms data views to sample networks and fuses them nonlinearly (Wang et al., 2014; John et al., 2019). MRGC learns a robust graph for each data view and unifies them afterward (Shi et al., 2023). However, all the above-mentioned methods either reform the feature space or compute the between-sample interactions from features, which becomes deficient with data containing much information in the between-feature interactions.

Feature interactions derived from large sample collections are not individual-specific, although this has been shown to highlight interwiring modules for risk prediction and corresponding sample subtyping. Some analysis flows that illustrate this are built on the weighted gene co-expression network analysis (WGCNA) algorithm (Langfelder and Horvath, 2008). Starting from global networks, namely, networks built on a collection of samples, each individual's perturbation to the global network can be used to derive a so-called individual-specific network or ISN (Kuijjer et al., 2019). Comparing ISNs thus implies comparing interaction profiles between individuals. In reality, different features contributing to individual wirings may have different origins, paving the way for system-level clustering of individuals in a precision medicine context.

In this work, we propose netMUG (**net**work-guided **MU**lti-view clusterin**G**), a novel multi-view clustering pipeline that clusters samples on the basis of these sample-specific feature interactions or wirings (Figure 1). In the presence of 2-view data, multi-view features are jointly selected via SmCCNet, based on canonical correlations and an additional extraneous variable. ISNs are constructed from the selected features, taking as a measure of edge strength the overall correlation between every pair of features and the extraneous data. The Euclidean distance metric representing dissimilarities between ISNs is fed into Ward's hierarchical clustering (Ward, 1963), using the R library Dynamic Tree Cut to automatically derive the number of clusters (Langfelder et al., 2008). After validating netMUG on synthetic data, we applied our workflow to a collection of participants of recent European ancestry with both genotypic information and 3D facial images available (White et al., 2021). The aim was to dissect and interpret between-individual heterogeneity, informed by BMI as extraneous data. The application showed the potential of netMUG to refine existing classifications for obesity and to identify novel gene-BMI associations not yet discovered by genome-wide association studies (GWAS).

Contributions of this paper:

- We developed a novel multi-view clustering method, netMUG, that combines feature selection, network construction, and downstream unsupervised learning in a single workflow.
- netMUG exploits the synergy between data views, as well as extraneous information, to assess heterogeneity between individuals and help in disease subtyping and stratified prevention.
- We simulated two data types whose features are cross-linked and whose samples have a complex clustering structure to validate the performance of netMUG.
- On a real-life dataset with genetic and facial information, netMUG informed by BMI as extraneous information highlighted known and novel characteristics of obesity.

2 Materials and methods

The code of netMUG is available on GitHub, along with its computational environment for reproducibility (https://github.com/

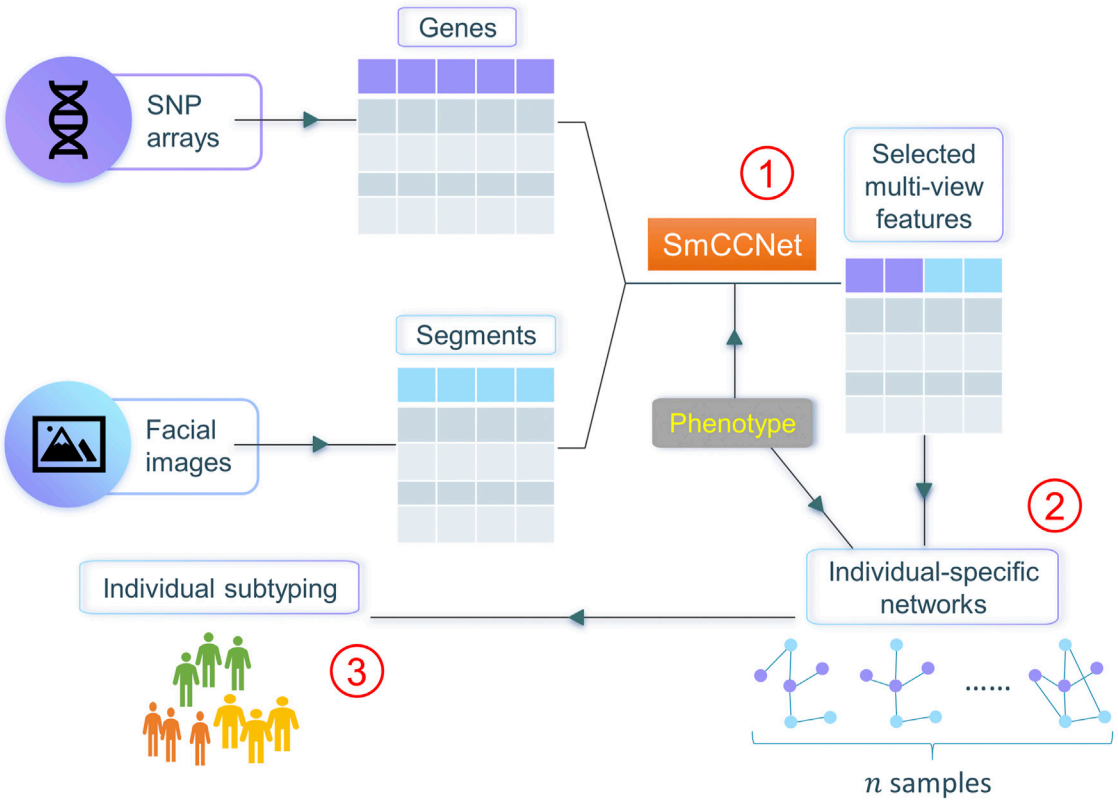

FIGURE 1
Workflow of netMUG (network-guided multi-view clustering). The pipeline consists of three parts: ① select phenotype-informed features from multiple data modalities via SmCCNet; ② build individual-specific networks based on the selected features and the phenotypic information; ③ subtype individuals via Ward's hierarchical clustering.

ZuqiLi/netMUG.git). The complete pipeline, simulation, and analyses were done in R (version 4.2.1). The data representation of the case study was computed in Python (version 3.9.7).

For the remainder of this paper, we denote the two data views as $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, respectively; the phenotypic variable is $Z \in \mathbb{R}^n$, where n is the number of individuals, p is the number of features in view 1, and q is the number of features in view 2.

2.1 SmCCNet

The SmCCNet pipeline consists of a sparse multiple canonical correlation analysis (SmCCA) and a module detection method (Shi et al., 2019). CCA and its variants are a set of multivariate statistical methods based on the cross-covariance matrix. The basic CCA aims to find a pair of vectors $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ that maximizes the correlation between Xa and Yb:

$$(\tilde{a}, \tilde{b}) = \underset{a,b}{\operatorname{argmax}}\, \operatorname{corr}(Xa, Yb) = \underset{a,b}{\operatorname{argmax}} \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}} \quad (1)$$

where Σ_{XY} denotes the cross-covariance matrix between X and Y, and Σ_{XX} and Σ_{YY} are the covariance matrix of X and Y, respectively. If we standardize both X and Y so that every feature has a mean of 0 and a standard deviation of 1, Eq. 1 can be reduced to

$$(\tilde{a}, \tilde{b}) = \underset{a,b}{\operatorname{argmax}} \frac{a^T X^T Y b}{\sqrt{a^T X^T X a}\sqrt{b^T Y^T Y b}} \quad (2)$$

If we further constrain the covariance matrices to be diagonal, Eq. 2 can be rewritten as

$$(\tilde{a}, \tilde{b}) = \underset{a,b}{\operatorname{argmax}}\, a^T X^T Y b, \text{s.t.} \|a\|^2 = \|b\|^2 = 1 \quad (3)$$

In case there are a lot of features with very little contribution to the canonical correlation, a sparse version of CCA (sCCA) is introduced, which applies the ℓ_1 regularization to the canonical weights a and b. Hence, the objective of sCCA is as follows:

$$(\tilde{a}, \tilde{b}) = \underset{a,b}{\operatorname{argmax}}\, a^T X^T Y b, \text{s.t.} \|a\|^2 = \|b\|^2 = 1, \|a\|^1 \le c_1, \|b\|^1 \le c_2 \quad (4)$$

where c_1 and c_2 are user-defined constants that regulate the sparsity, $c_1 \in [1, \sqrt{p}]$, and $c_2 \in [1, \sqrt{q}]$. Their values can be chosen via cross-validation.

CCA and sCCA are originally designed for only two data views; however, additional information may be available and it may be helpful to take into account the correlations between the existing views and the extra one. In the special case where the extra data type contains only a single feature, e.g. a phenotype of interest, this can be considered as finding the optimal canonical pair from the two existing views that also correlates with the phenotype. With the phenotypic variable denoted as $Z \in \mathbb{R}^n$, the objective of SmCCA becomes

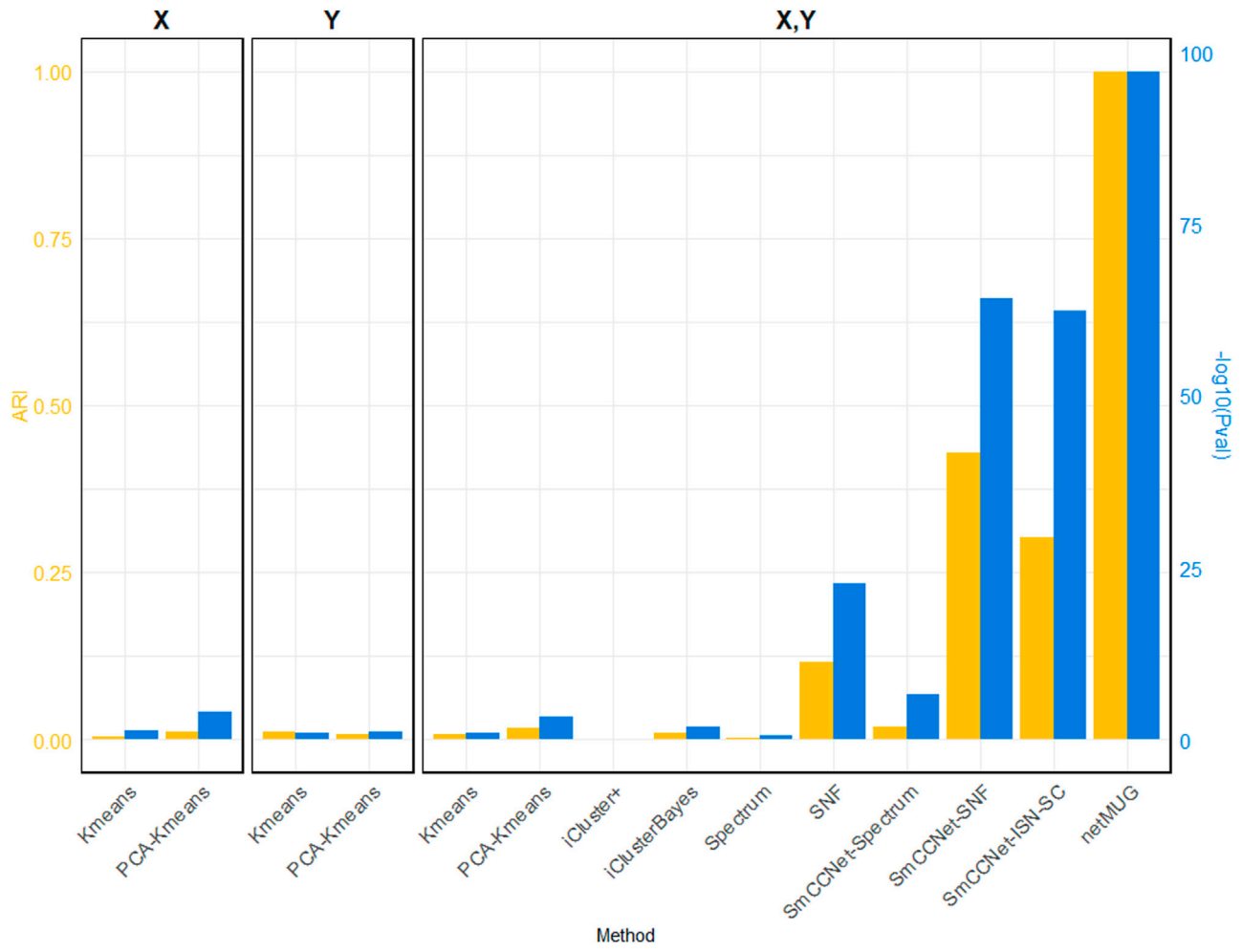

FIGURE 2
Performance of all methods on simulated data. The three barplot panels illustrate the ARI (on the left axis in yellow) and $-\log_{10}$*Pval* (on the right axis in blue) for the 4 single-view (X or Y) and 10 multi-view (both X and Y) models. Four out of six baseline models ("Kmeans" and "PCA-Kmeans") use a single view (X or Y). Two-view "Kmeans" and "PCA-Kmeans" models are based on the concatenation of X and Y. "SmCCNet-Spectrum" and "SmCCNet-SNF" are benchmark models with features selected by SmCCNet. "SmCCNet-ISN-SC" only differs from the proposed framework, netMUG, for the clustering method.

$$\left(\tilde{a},\tilde{b}\right) = \underset{a,b}{\operatorname{argmax}}\left(w_1 a^T X^T Y b + w_2 a^T X^T Z + w_3 b^T Y^T Z\right), \quad \text{s.t.} \|a\|^2 = \|b\|^2 = 1, \|a\|^1 \le c_1, \|b\|^1 \le c_2 \tag{5}$$

Coefficients w_1, w_2, and w_3 balance the three canonical correlations to account for the different correlation strengths among the multiple data types.

The pair of canonical weight vectors $(\tilde{a},\tilde{b})$ learned from the SmCCA are sparse indicators of how much every feature in X and Y contributes to the overall correlation among the two data modalities and the phenotype. If we concatenate $\tilde{a}$ and $\tilde{b}$ to form a new vector, $\tilde{ab}$, from which we can construct us a similarity matrix whose elements measure the relatedness between every two features:

$$S = \left|\tilde{ab} \otimes \tilde{ab}\right| \tag{6}$$

where $\otimes$ is the outer product operator and $|\cdot|$ takes the absolute value of every element in a matrix.

To make the canonical weights robust, SmCCNet integrates a feature subsampling scheme, which results in multiple pairs of $(\tilde{a},\tilde{b})$ with different subsets of features from X and Y. The similarity matrices from each pair are then averaged and divided by the maximum.

Hierarchical clustering with complete linkage is performed on the distance matrix $1 - \bar{S}$, where $\bar{S}$ denotes the averaged and rescaled similarity matrix. The dendrogram is subsequently cut at the user-specified height and modules with feature(s) from a single view are discarded to derive multi-view modules.

2.2 netMUG: workflow and algorithm

The input of netMUG is multi-view data with a phenotype (or more generally, extraneous variables), both describing the same set of samples. To reduce dimensionality and extract the most informative features, netMUG first incorporates SmCCNet for

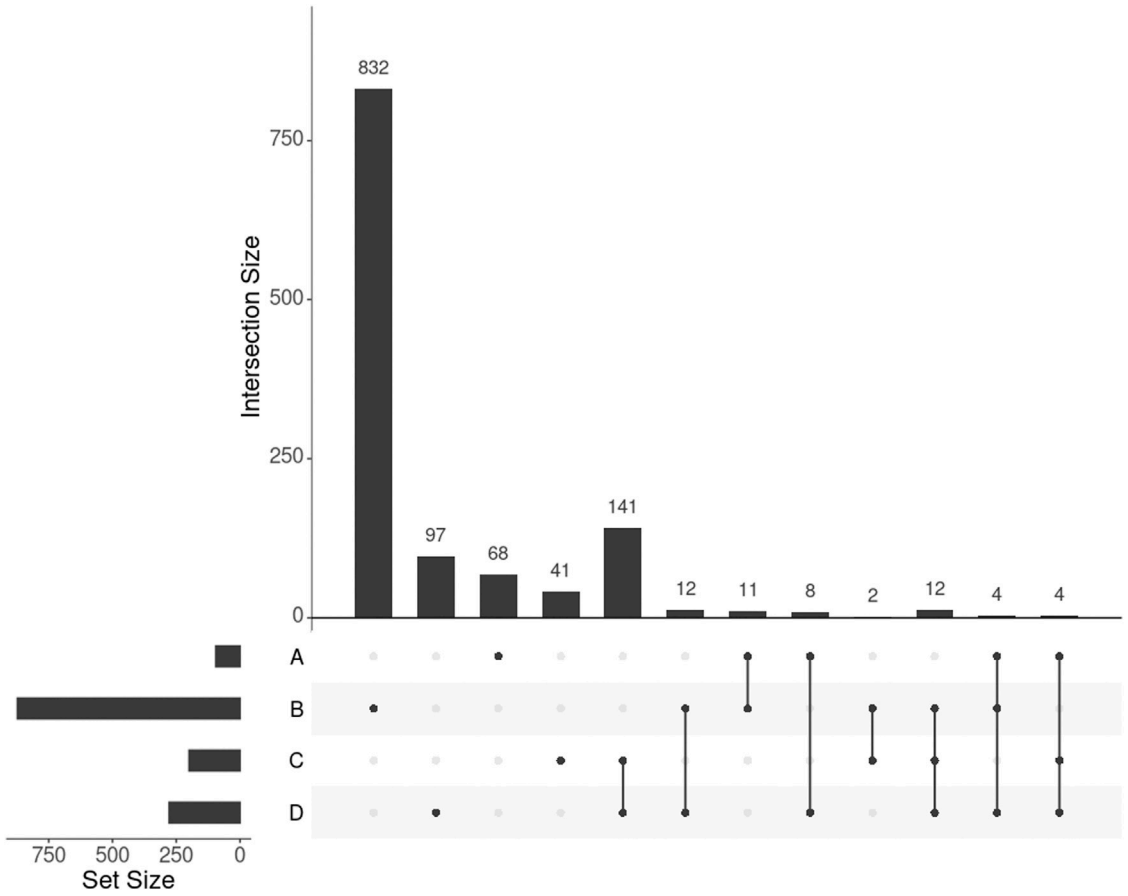

FIGURE 3
UpSet plot showing the intersections of four gene sets. A, B, C, and D represent the gene list found by GWAS, DisGeNET, uninformed SmCCNet, and standard SmCCNet, respectively. Uninformed SmCCNet maximizes the canonical correlation without extraneous information, whereas standard SmCCNet takes BMI into account to supervise CCA. Genes of the four sets are listed in the Supplementary Material.

feature selection, namely, features in the final modules detected by SmCCNet. We have found that in practice, it is difficult to assess and quantify the balancing weights on each pairwise correlation (see Eq. 5), so this property has been omitted in netMUG, i.e., $w_1 = w_2 = w_3 = 1$.

We then use the selected features to construct ISNs for each individual. These networks are characterized by individual-specific edges as well as individual-specific nodes. In particular, we first construct a global network $G^{(\alpha)} = (V, E^{(\alpha)})$ across all samples, where V denotes the set of selected features from both views and $E^{(\alpha)}$, the set of edge weights whose element $e_{ij}^{(\alpha)}$ is the sum of pairwise Pearson correlations between feature i, j, and the phenotype Z on all samples:

$$e_{ij}^{(\alpha)} = \text{corr}\left(V_i, V_j\right) + \text{corr}\left(V_i, Z\right) + \text{corr}\left(V_j, Z\right) \tag{7}$$

Second, we compute the leave-one-out network $G^{(\alpha-s)} = (V, E^{(\alpha-s)})$ whose edges $E^{(\alpha-s)}$ are derived from all samples but sample s, similarly to Eq. 7. Intuitively, the difference between the global and leave-one-out networks measures the perturbation on the network caused by sample s. Based on this, in our work, we define ISN by the absolute differential network:

$$e_{ij}^{(s)} = \left|e_{ij}^{(\alpha)} - e_{ij}^{(\alpha-s)}\right| \text{ for } \forall i, j \in \{1, 2, \cdots, r\} \tag{8}$$

where $e_{ij}^{(s)}$ is the edge value between node i and j in the network specific to individual s. Hence, $G^{(s)} = (V, E^{(s)})$, and r is the number of selected features. Eq. 8 is a variant of the original ISN construction method reported by Kuijjer et al. (2019) because deriving an individual-specific association is not essential when the final aim is to identify "distances" between individuals. A higher $e_{ij}^{(s)}$ indicates that individual s deviates from its population more than others concerning the joint correlation between features and phenotype. In addition, an ISN with patterns that are very different from those of another ISN means that their corresponding individuals may belong to different population subgroups.

The edge weights in an ISN can be seen as the coordinates of a data point in a k-dimensional Euclidean space, where k = number of edges. Accordingly, clusters can be achieved by hierarchical clustering on Euclidean distance and Ward's linkage. Because the Euclidean distance is not squared, we chose "ward.D2" as the method for the R function "hclust" to adopt the correct Ward's clustering criterion (Murtagh and Legendre, 2014). The clusters are obtained automatically via the R package "dynamicTreeCut", which iteratively cuts the hierarchical tree and determines the optimal

10.3389/fgene.2023.1286800

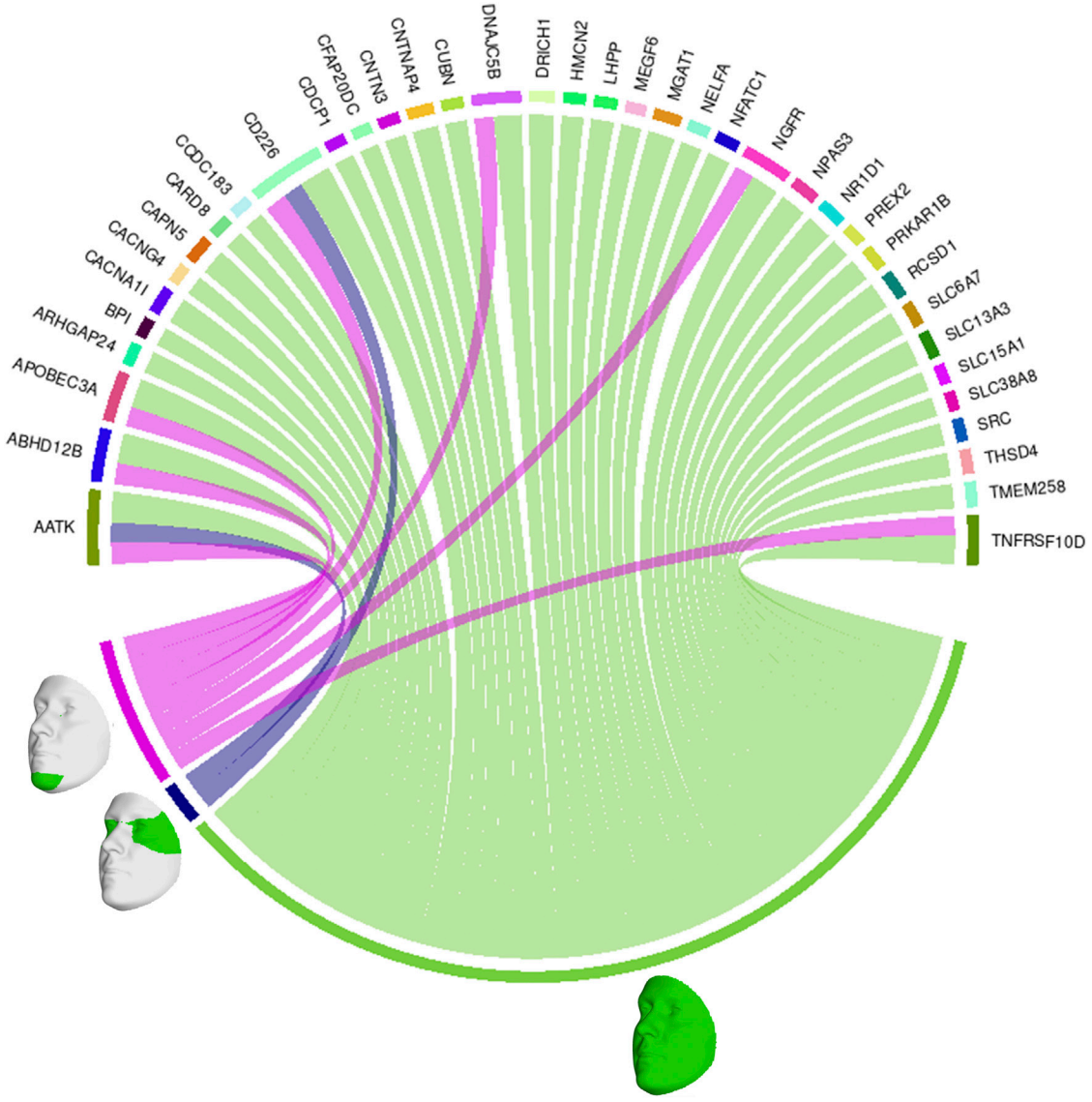

FIGURE 4
Genes and facial segments with the top 1% connections. Connections refer to the similarities between the canonical weights of genetic and facial PCs, as described in Methods. Only inter-modality relationships are considered. A thicker connection in the circus plot means a higher relatedness.

number of clusters based on their shape. We set its hyperparameter deepSplit = 1 to have fewer but larger clusters.

2.3 Simulation study

We synthesized 1,000 samples with two views, each of which contains 1,000 variables, simulating complex, cross-linked datasets, e.g., genetic and facial data. Samples are randomly distributed in three balanced clusters. Because real-life data often contain a large number of uninformative features, we generated 600 variables of each data view from standard normal distribution $N(0,1)$, representing the noise. Then, the remaining 400 features were correlated within and between the two views to different extents by linearly transforming 400 orthogonal vectors per view.

The first 200 orthogonal vectors in each view, X_{corr} and Y_{corr}, were sampled from a multivariate normal distribution with random covariance ranging from 0.5 to 0.8, i.e., $[X_{corr}\, Y_{corr}] \sim N(0, \Sigma)$, where $\Sigma = \begin{bmatrix} I & R \\ R & I \end{bmatrix}$, $I \in \mathbb{R}^{200\times200}$ is an identity matrix, and $R \in \mathbb{R}^{200\times200}$ is a diagonal matrix with the 200 covariances on the diagonal. The remaining 200 vectors in each view, X_{clust} and Y_{clust}, were generated similarly but with covariance ranging from 0.8 to 1. X_{clust} and Y_{clust} have three distinctive clusters of samples by multiplying 1, 2, and 3 by their values for the samples in every cluster. The clustering also gave rise to a phenotype that follows a different normal distribution in every cluster, i.e., $Z = [Z^{(1)}\, Z^{(2)}\, Z^{(3)}]^T$, where $Z^{(1)} \sim N(-1,1)$, $Z^{(2)} \sim N(0,1)$, and $Z^{(3)} \sim N(1,1)$. To make the correlation patterns more complex and representative of real-life data, we linearly

10.3389/fgene.2023.1286800

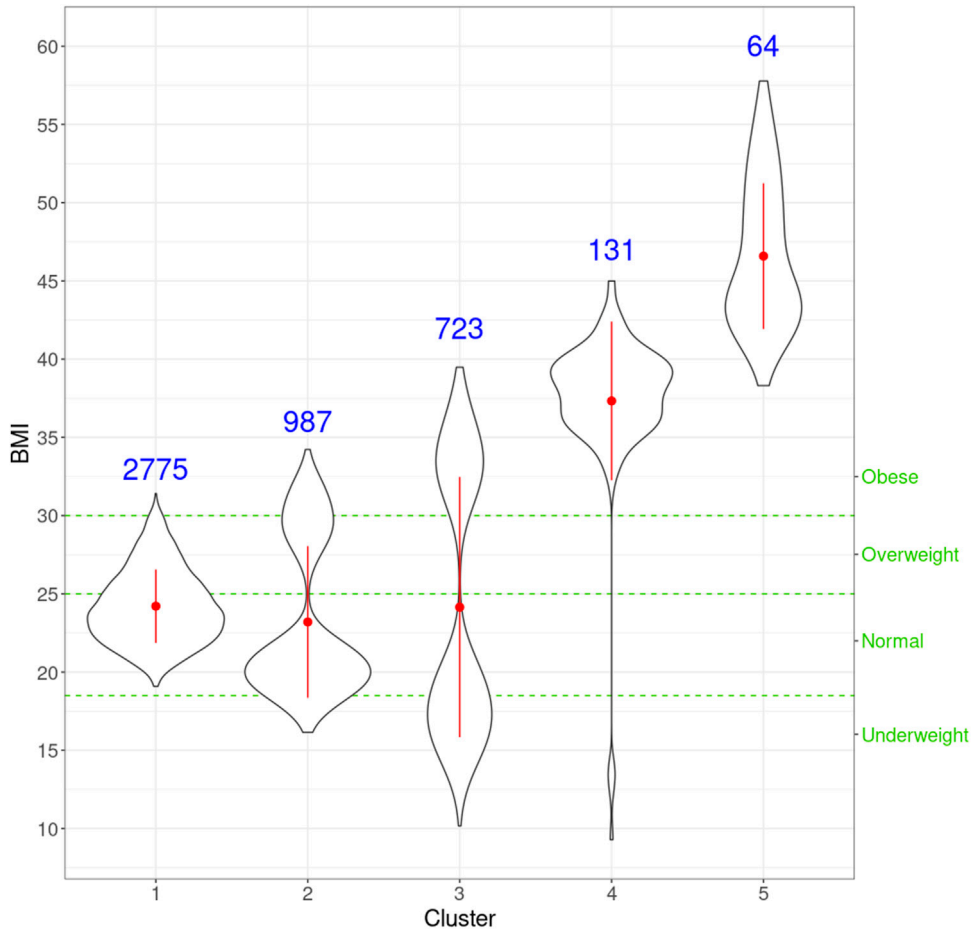

FIGURE 5
Violin plots for the distribution of BMI for individuals in every cluster. For each cluster, the red dot and vertical line indicate the mean and standard deviation, respectively, and the number in blue is the size of every cluster. The three green horizontal dashed lines represent the cut-offs of the four classic BMI categories shown on the right *Y*-axis.

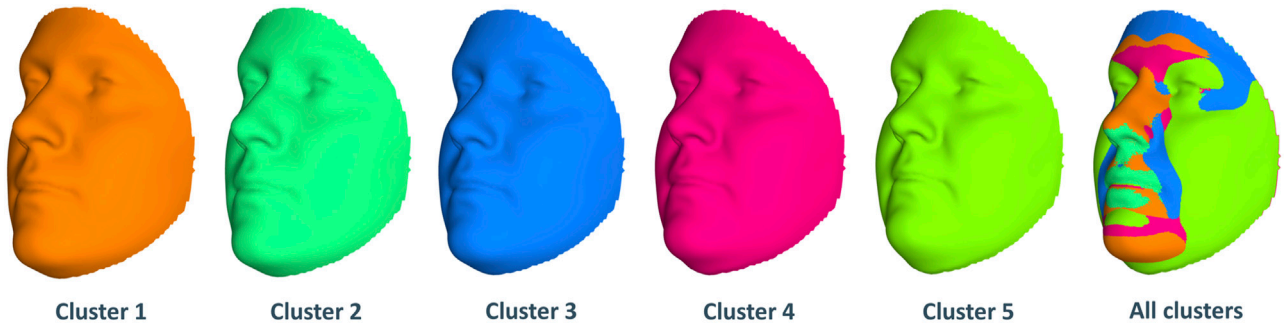

FIGURE 6
Mean facial shapes of every cluster and the superposition of them. The first five faces are the average of all individual faces in every cluster, while the last face was obtained by plotting the five mean faces on top of each other (colors of this superposition face correspond to the five mean faces).

transformed $[X_{corr}\ X_{clust}]$ and $[Y_{corr}\ Y_{clust}]$ by multiplying a square coefficient matrix whose values are random in the uniform range $[-3, 3]$.

We also investigated the execution time spent by each model. The bottleneck of netMUG is the SmCCNet step that computes a full SVD internally for each dataset, whose time complexity is

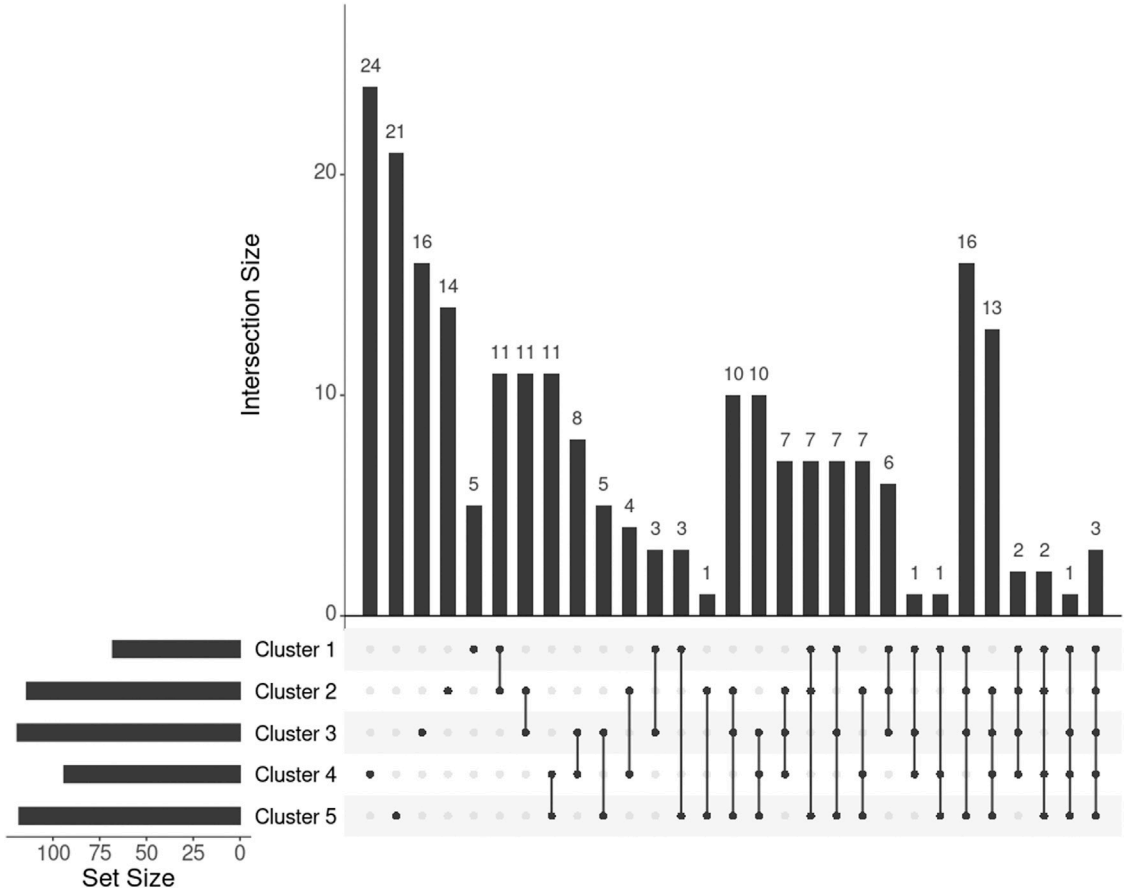

FIGURE 7
UpSet plot showing the intersections of genes extracted from the mean ISN of every cluster.

$\mathcal{O}(\min\ (mn^2, nm^2))$, where n and m are the two dimensions of the data matrix. To make the method applicable to large-scale datasets, e.g., the whole genomic data composed of millions of SNPs, we may replace the full SVD with truncated SVD in the future for much faster computation and less memory usage.

2.4 Case study

2.4.1 Data representation

Data pre-processing resulted in 265,277 SNPs and 7160 3D facial landmarks for all 4,680 individuals with European ancestry (detailed steps are described in the Supplementary Material). We subsequently reduced the dimensionality of both genomic and facial data via principal component analysis (PCA) (Pearson, 1901). Specifically, SNPs were first mapped to protein-coding genes if they fell within 2000 base pairs around a gene. Genes mapped with less than three SNPs were removed for insufficient information. Then, we performed PCA on the SNPs in every gene to obtain the principal components (PCs) explaining at least 90% of its variance to represent each gene by fewer but more informative features. Meanwhile, we hierarchically segmented 3D facial images into five levels and 63 segments via the method proposed by White *et al.* (2021). Finally, the optimal number of PCs representing every facial segment was determined via the simulation-based parallel analysis. As a result, we obtained 60,731 PCs for 9,077 genes and 1,163 PCs for 63 facial segments.

2.4.2 Group-level interpretation

To estimate the association of BMI with our genomic data, we first conducted a GWAS via PLINK to calculate the p-value of the Wald test (detailed steps are described in the Supplementary Material), which would identify whether an SNP is significantly associated with BMI. Only SNPs with FDR-adjusted p-values < 0.05 were kept and then mapped to genes, following the same mapping strategy as in the 'Data representation' step. Meanwhile, we also downloaded the list of BMI-relevant genes from the DisGeNET database with a decent filter (≥0.1) on the gene-disease association (GDA) score.

To further analyze the behavior of SmCCNet, we reran it with the same settings but without phenotypic information (the two sparsity parameters were re-determined via cross-validation). So far, four sets of genes have been obtained from the standard SmCCNet, GWAS, DisGeNET, and the uninformed SmCCNet. We investigated the overlap among them to see the agreement between SmCCNet and GWAS, the enrichment of DisGeNET genes in the gene set of SmCCNet (detailed steps are described in the Supplementary Material), and the difference in SmCCNet made by the presence of phenotype.

TABLE 1 Enriched Reactome pathways in every cluster. Pathways for cluster 5 are all in italics because no *p*-value was lower than 0.05 after multiple testing corrections (FDR). For cluster 5, the four pathways with raw *p*-values ≤ 0.01 are shown.

Cluster No.	Pathways
Cluster 1	Tryptophan catabolism
Cluster 2	TP53 regulates transcription of death receptors and ligands
	Defective GALNT12 causes CRCS1
	Defective GALNT3 causes HFTC
	NPAS4 regulates expression of target genes
	Transcriptional regulation by NPAS4
Cluster 3	Regulation of TP53 expression
	STAT3 nuclear events downstream of ALK signaling
	Signaling by ALK
	Regulation of TP53 expression and degradation
Cluster 4	Regulation of gene expression in early pancreatic precursor cells
	Regulation of gene expression in late-stage (branching morphogenesis) pancreatic bud precursor cells
	Rap1 signalling
	Defective B3GALTL causes PpS
	O-glycosylation of TSR domain-containing proteins
Cluster 5	*Glycogen storage disease type Ib (SLC37A4)*
	Interleukin-33 signaling
	RHOB GTPase cycle
	RHOC GTPase cycle

2.4.3 Cluster-level interpretation

We first evaluated the relationship between BMI and the clustering derived by netMUG. Namely, we conducted a Kruskal–Wallis test to determine whether there were statistically significant differences in BMI distributions among the clusters. Our obesity subtypes were also compared with the classic BMI categories, i.e., underweight (BMI <18.5), normal (18.5≤BMI <25), overweight (25≤BMI <30), and obese (BMI≥30) (WHO Global InfoBase team, 2005).

Further, we characterized every subgroup by facial and genetic information. By averaging all facial images of each subgroup, we represented it with a mean face shape. As for genetics, ISNs of every cluster were averaged, and a subnetwork was taken from the mean ISN based on the top 1% edge weights to only focus on the vital signals. Subsequently, we computed the largest connected component of every subnetwork and extracted all genes in this component. The overlap among clusters was analyzed, and we paid special attention to the cluster-specific genes to characterize each cluster. In particular, an enrichment analysis was done via the analysis tool of the Reactome website (Gillespie et al., 2022) to test which biological pathways are significantly over-represented in the gene list specific to every cluster.

2.4.4 ISN-level interpretation

Because our pipeline represents every individual by a network, namely, ISN, a fully-connected weighted graph, we need a lower-dimensional representation of the ISNs to examine their behavior better and visualize them. Therefore, the graph filtration curve method was applied, which computed a function or curve from each ISN, whose values were derived via graph evolution (O'Bray et al., 2021). More specifically, we set a series of thresholds on the edge weights and, at each threshold, constructed a subnetwork by taking only edges larger than that threshold. In such a manner, we got a series of subnetworks from smallest to largest for every ISN, and the largest subnetwork is the ISN itself. Then, graph property was calculated based on each subnetwork, and therefore, an ISN was converted to a function of graph property against the edge threshold. Here, in our project, we chose the mean node degree of the largest connected component (LCC) as the graph property because the subgraphs may not be fully connected anymore. The average degree is a simple yet powerful tool to measure graph density.

3 Results

netMUG was validated on a synthetic dataset and compared with both baseline and benchmark methods for multi-view clustering. We then applied it to real-life large-scale multi-view cohort data and characterized the resultant clusters by their representative faces and enriched pathways.

3.1 Validating netMUG on synthetic data

We simulated a scenario where a multi-view dataset contained complex feature patterns and many noisy features, representing real-

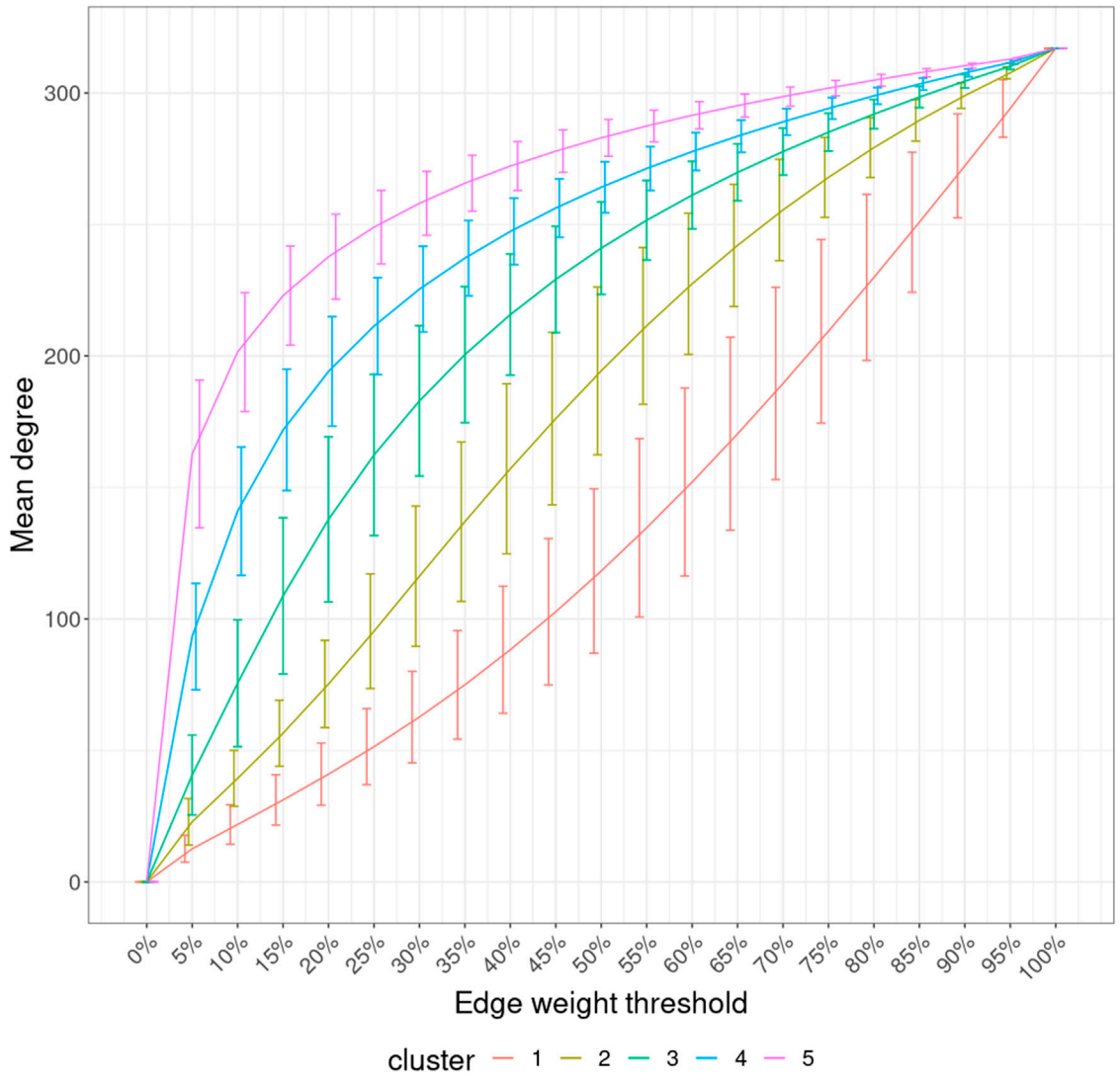

FIGURE 8
Filtration curves of ISNs grouped by the derived clustering. Every line linking the bottom-left and upper-right corners shows the mean filtration curve of each cluster. The vertical lines are the standard deviation of the function values at each threshold within each cluster.

life high-dimensional data, e.g., genomics and images. Two criteria were adopted to assess clustering performance, p-value from the Kruskal–Wallis test, and adjusted Rand index (ARI). The Kruskal–Wallis test was used as a non-parametric version of one-way ANOVA to test whether the distribution of an extraneous phenotype is similar across clusters because the simulated phenotype follows a multimodal distribution (Kruskal and Wallis, 1952). Therefore, a low p-value (< 0.05) indicates a significant association between the clustering and the phenotype. ARI was computed to show the similarity between the derived clustering and the phenotype (Rand, 1971): the higher, the better; an ARI of 1 means perfect fitting.

Various baseline and benchmark methods were considered in addition to netMUG, and their performances were compared. Baseline models included k-means on every single view, and both views concatenated, with or without principal component analysis (PCA) as a data dimensionality reduction strategy. We chose four benchmark models, iCluster+ (Mo et al., 2013), iClusterBayes (Mo et al., 2018), Spectrum (John et al., 2019), and SNF (Wang et al., 2014), and their codes are available in R (R Core Team, 2022). To show the effectiveness of SmCCNet as a feature selector, we also applied Spectrum and SNF on the features selected by SmCCNet, i.e., model "SmCCNet-Spectrum" and "SmCCNet-SNF'. Finally, as an alternative to our proposed framework, we replaced hierarchical clustering with spectral clustering as the last step of netMUG, leading to the model 'SmCCNet-ISN-SC'.

As shown in Figure 2 (exact values are listed in Supplementary Table S1), all baseline models performed poorly in terms of ARI. The p-values for models with PCA on single X and both views were lower than 0.05 but did not convey a big difference on the plot. Among the benchmarks, SNF outperformed iCluster+, iClusterBayes, and Spectrum. With features selected by SmCCNet, the performances of both Spectrum and SNF were substantially improved, implying the necessity of feature selection on high-dimensional and noisy

data. Integrating SmCCNet, ISN, and Dynamic Tree Cut, netMUG achieved the best p-value and ARI results. It retrieved the clustering of phenotype for the complex multi-view data we simulated.

The runtime of each method (Supplementary Table S1) shows that all baseline models spent less than half a minute and PCA brought an acceleration in speed for fewer features. iCluster+ and iClusterBayes spent much more time than Spectrum and SNF, without an improvement in performance. The SmCCNet feature selection step took 33 min on its own, which became the bottleneck of all the SmCCNet-based models ('SmCCNet-Spectrum', 'SmCCNet-SNF', 'SmCCNet-ISN-SC', and netMUG). This is mostly due to the subsampling scheme and the full SVD computation within SmCCNet.

3.2 Case study

To exemplify netMUG, we used a multi-view dataset of 4,680 individuals of recent European ancestry recruited from three independent studies in the US: 3D Facial Norms cohort (PITT), Pennsylvania State University (PSU), and Indiana University-Purdue University Indianapolis (IUPUI) (White et al., 2021). For each individual, facial images and genomic data were collected along with extra information, including age, sex, height, and weight. BMI was computed as BMI = weight(kg)/height$(m)^2$. Because both facial and genomic data are high-dimensional, we converted them separately into low-dimensional PC spaces before feeding them into a netMUG pipeline. More specifically, the genomic and facial data views had 60,731 and 1,163 PCs as features, respectively.

We interpreted netMUG analysis results at two levels: at the all-samples level, hereafter referred to as "group-level", and at the level of clustered individuals. Group-level interpretation refers to describing the multi-view features selected by SmCCNet. In addition, we assessed the overlap between SmCCNet-selected genes and genes found by a genome-wide association study (GWAS) or DisGeNET database. Cluster-level interpretations were made by evaluating the association between the final clustering and BMI and the characterization of every cluster in terms of facial or genetic characteristics. Finally, we applied graph filtration curves to represent and visualize ISNs in 2D space (O'Bray et al., 2021).

Two flavors of netMUG were implemented: one with SmCCNet informed by BMI as an extraneous variable, and one without such information.

3.2.1 Group-level interpretation

Informed by BMI. We chose the two sparsity parameters for SmCCNet, namely, c_1 and c_2 in Eq. 5, via 5-fold cross-validation, predicting the canonical correlation. The subsampling proportions were determined as 50% and 90% for genomic and facial data because of their substantial dimensional differences. We then computed the average canonical weights over 500 subsampling runs to derive the similarity matrix $\bar{S}$. Three modules with 316 PCs (see Methods "Data representation") were found by cutting the hierarchical tree at a height very close to 1 (0.999) and discarding modules with a single feature or features from a single view. The cutting threshold was determined following Shi WJ et al. (Shi et al., 2019). The selected features from the retained modules comprised 278 genes and 26 facial segments. All the subsequent analyses were performed on these features unless mentioned otherwise. One of the most well-known obesity-related genes, *FTO* (Fawcett and Barroso, 2010), was on the gene list. Another essential gene for obesity, *MC4R* (Loos et al., 2008) was not selected because it was filtered out for having less than three SNPs mapped. A recent epistasis analysis found two pairs of SNPs whose interactions were associated with BMI (*FTO–MC4R* and *RHBDD1 – MAPK1*) (D'Silva et al., 2022). SmCCNet did not detect *RHBDD1* or *MAPK1*, possibly caused by CCA's focus on inter-modal rather than intra-modal interactions.

GWAS detected 155 SNPs significantly associated with BMI mapped to 95 protein-coding genes (p-value < 0.05 adjusted by false discovery rate, or FDR). Out of these 95 genes, 16 (16.8%) were in common with the 278 genes selected by SmCCNet (Figure 3, set A and D), confirming the agreement between SmCCNet and GWAS but also explaining their differences in the involvement of facial information.

We found 1,014 genes from DisGeNET associated with BMI with a gene-disease association (GDA) score≥0.1 (Piñero et al., 2019), of which 873 were protein-coding genes. Of the 873 genes, 28 (3.2%) also appeared in the 278 genes selected by SmCCNet (Figure 3, set B and D). The hypergeometric test showed that the DisGeNET gene set was significantly enriched in the gene set from SmCCNet (p-value = 6.0×10^{-5}).

Uninformed by BMI. A total of 329 features were selected by the uninformed SmCCNet, resulting in 200 genes and 50 facial segments. Of the 200 genes, 157 (78.5%) were shared between the standard and the uninformed SmCCNet (Figure 3, set C and D). However, the uninformed SmCCNet found fewer GWAS genes (4) and DisGeNET genes (14) than informed SmCCNet (Figure 3, set A, B, and C). Furthermore, a lower percentage of BMI-related genes were selected by uninformed SmCCNet (2% in GWAS and 7% in DisGeNET) than informed SmCCNet (6% in GWAS and 10% in DisGeNET), highlighting the merits of supervised analysis.

We also looked at the top 1% of connections between genes and facial segments in the features selected by informed SmCCNet (Figure 4). It was clearly shown that the full face has high relatedness with all genes, indicating that those genes are primarily associated with facial morphology globally. On a local level, the selected genes are most strongly connected with the eyes (with the temporal area) and chin, in line with the fact that they are known facial signals of obesity. *AATK* and *CD226* are the genes with the most top connections. *AATK* plays a role in neuronal differentiation and has been known to be highly associated with BMI (Zhu et al., 2020). CD226 encodes a glycoprotein expressed on the surface of blood cells and is related to colorectal cancer (Storojeva et al., 2005), which could be a potential biomarker of obesity. Furthermore, *APOBEC3A*, *DNAJC5B*, and *NGFR* all affect body height and BMI (Kichaev et al., 2019).

3.2.2 Cluster-level interpretation

netMUG automatically detected five clusters with significantly different BMI distributions given the Kruskal–Wallis test (p-value = 7.4×10^{-150}) (Figure 5). The derived clustering is much more significantly associated with

BMI than the clustering uninformed by BMI (p-value = 1.8×10^{-17}). Clusters 4 and 5 are the two outstanding subgroups, and both fall into the classic category of 'obese.' (WHO Global InfoBase team, 2005). Nevertheless, as clearly shown in Figure 5, our clustering provides higher granularity on the obese subgroup, which is desired because it can lead to more precise identification and better characterization of obese people. Cluster 1 contains roughly half the people in the dataset, and it substantially follows the distribution of the whole population with most normal and normal-to-overweight individuals. Clusters 2 and 3 have similar bimodal distributions with different deviations. The two peaks of cluster 3 are further away from the center than cluster 2. Jointly looking at clusters 1, 2, and 3, they classify individuals by how far they are from the 'population normal' while treating underweight and overweight indifferently. This behavior suggests that underweight and overweight people deviate from the normal condition similarly in terms of multi-view interactions, which may be related to, e.g., the double burden of malnutrition. It has been shown that some crucial vitamins and minerals can affect both underweight and overweight individuals.

Next, we look at the genetic and facial characteristics of the identified clusters. The average facial shapes per cluster are depicted in Figure 6 and largely follow the profile of mean BMI across clusters (Figure 5). Again, clusters 4 and 5 stand out compared to clusters 1–3. Superposition of the average faces shows pronounced areas on the forehead or the chin for cluster 4 individuals, whereas cheek and eye areas are most responsible for cluster 5 differences in the rest of the samples. Cluster one to three faces have pronounced features around the nose and mouth.

For individuals within every cluster, we averaged their corresponding ISNs and binarized the mean ISNs, with the top 1% edge weights being 1 and the rest 0. Subsequently, we computed the largest connected component (LCC) from every binary network, resulting in LCCs of 86, 129, 144, 112, and 136 nodes (68, 114, 119, 94, and 118 genes) for the five clusters, respectively (Figure 7). The genes *MIGA1*, *CACNA1B*, and *SLC38A8* were common to all clusters. *MIGA1* regulates mitochondrial fusion and shows a strong relationship with BMI according to the GWAS Catalog (scoring 14.0) (Zhu et al., 2020). *CACNA1B* encodes a voltage-gated calcium channel subunit and GWAS Catalog records a strong association (scoring 12.2) between *CACNA1B* and acute myeloid leukemia (Lv et al., 2017), for which BMI is a known risk factor. *SLC38A8* has a high GWAS Catalog score (14.1) with adiponectin measurement (Spracklen et al., 2019), which directly affects insulin sensitivity and obesity, and a strong association with eye diseases, e.g., foveal hypoplasia 2 and anterior segment dysgenesis. The link from gene *SLC38A8* to both obesity and facial features may imply a novel relationship between obesity and facial morphology.

To further investigate the genes exclusively extracted from each subgroup, i.e., subtype-specific genes, cluster 4 has the most subtype-specific genes (24) and also has the highest proportion (25%) of all the genes in its LCC, followed by cluster 5 with 21 unique genes (17.8%). This observation is in line with the distinctive BMI distributions for these clusters. Meanwhile, there are only five genes specific to cluster 1 (7.4% of 68 genes in the LCC of cluster 1), suggesting that cluster 1 represents the "population normal."

One or more Reactome pathways were significantly enriched in genes that were specific to a cluster, except for cluster 5 (Table 1). The reason may be that genes enriching cluster 5 were also obtained in other clusters, so they were not considered specific to cluster 5. Another possibility is that those 21 genes exclusively in cluster 5 are too functionally diverse as obesity is involved in many different biological pathways.

3.2.3 ISN-level interpretation

With the mean node degree of the largest connected component being the function to describe a graph, the filtration curves of ISNs exhibited notable variation in their evolution trajectory, which implies the ability of ISNs to exploit the between-individual heterogeneity (Supplementary Figure S2). If we group the filtration curves by the netMUG-derived clustering, their values are significantly different at most edge thresholds (Figure 8). Clusters 4 and 5 have higher mean degree values overall than the rest along the graph evolution, indicating that their ISNs are depicted by more densely connected components. This result is in line with the association between BMI and every cluster (Figure 5) in the sense that people with severe obesity (clusters 4 and 5) show more prominent characteristics than normal or slightly overweight people (clusters 1, 2, and 3). The ISN-level results both mean that average degree is a good descriptor for ISNs and confirm that the clusters differ in terms of intrinsic graph properties, e.g., average degree.

4 Discussion

In this study, we introduced netMUG, a multi-view clustering strategy that links population-based networks to individual-specific networks, possibly informed by an extraneous variable. Synthetic data analysis showed promising outperformance of netMUG compared to baseline and benchmark multi-view clustering methods. An application to real-life cohort data and exploiting extraneous information via BMI revealed a refined classification of obese individuals and an increased understanding of genetic and facial segments linked to BMI-induced subgroups.

Because our workflow is highly modular, various extensions or adaptions can be implemented, at the discretion of the user. Here, we presented a basic version. Modifications can be made at several levels, which will be covered in the following paragraphs.

At the level of the data input, a single data view can be considered, single or multiple extraneous variables can be informative, and missing values can be inferred by imputation. Single-view netMUG replaces its basic SmCCNet implementation with one that targets a single dataset only. With multiple extraneous variables (for instance a symptom set of variables), weight optimization (in Eq. 5) is affected via prior knowledge or cross-validation. When adopting imputation strategies, it is important to account for within- and between-data relationships. We refer to (Song et al., 2020), who reviewed integrative imputation strategies for multi-omics datasets. Some of the proposed strategies may also apply to highly heterogeneous omics/non-omics mixed data. These include deep learning-inspired approaches. In addition, the data representation may

be impacted by different numbers of PCs or more complex component summaries. We considered diffusion kernel PCA (Walakira et al., 2022), which nonlinearly exploits the graphic structure in the data. However, it is computationally intensive and requires extra hyperparameter tuning, and therefore, was discarded in this study.

Furthermore, adaptions are possible at the level of computing relationships between sets of variables. For instance, structural equation models (SEMs) can be used to conduct CCA in the presence of missing data (Lu, 2019). It needs more work to see how sparsity and supervision are best introduced in SEMs as in sparse CCA, which was key to SmCCNet (Shi et al., 2019). For a review of sparse CCA extension to genomic data, we refer to Witten and Tibshirani (2009). Ideas on sparse supervised CCA and sparse multiple CCA (more than two data views) pave the way for extended applications of netMUG, for instance via the R function MultiCCA (Witten et al., 2009).

Once a filtered heterogeneous network is obtained, its nodes can serve as a template for constructing individual-specific edges. It is up to the user to define the most appropriate measure of association. Our analyses on real-life datasets showed a benefit to including the extraneous data in this measure (here in this work: BMI). The appropriateness of a measure of association between features is context-dependent. For instance, a microbial co-occurrence network has been utilized for building ISNs with microbiome data (Yousefi et al., 2023).

At the level of sample clustering, multiple graph clustering algorithms can be used. The diversity in algorithms is in part due to the many ways distance or similarity between graphs (here: ISNs) can be defined. Examples include edge difference distance or graph diffusion distance. These options and many more are discussed in a study by Kriege et al. (2020). The basic version of netMUG represents ISNs as points in a high-dimensional Euclidean space (axes refer to edges) and implements Ward's minimum variance hierarchical clustering method. The initial cluster distance is taken to be the squared Euclidean distance between points. As a potential replacement for Ward, a recently published linkage method, *k*-centroid, takes the samples around cluster centers into consideration and exhibits better performance than conventional methods (Dogan and Birant, 2022).

Several measures can be taken to further increase the robustness of netMUG. Our choice is to adopt the subsampling feature in SmCCNet, i.e., SmCCA is run multiple times, each time on a random subset of features. It relates to increasing the robustness of feature selection. Robustness may also be increased at the clustering level by fusing a variety of clustering algorithms or settings within a multiple clustering protocol (Zhang and Li, 2011). Deriving the optimal number of clusters via Dynamic Tree Cut settings or computationally more intensive multiscale bootstrap resampling may be included in such a protocol. Alternatively, the hierarchical clustering process and significance assessment are intertwined as in netANOVA (Duroux and Van Steen, 2022), with promising performance for ISNs.

We illustrated netMUG on cohort data, with BMI as extraneous information. Alternative applications of netMUG include disease subtyping and studies that explore the impact of confounders on clustering results. The latter studies can be carried out by comparing supervised (i.e., with the confounders or extraneous information) and unsupervised netMUG (i.e., without Z in Eq. 5). Our evaluation dataset consists of a large number of individuals and features and has been properly pre-processed by experts, contributing to the high clustering performance. A potential improvement could be achieved in the genotype imputation, which is essential for the quality and quantity of genetic data.

In conclusion, our proposed netMUG method exploits population-based and individual-specific network analyses to construct and interpret multi-view clusters. It takes advantage of SmCCNet for multi-view data integration, ISN for individual-specific network representation, and Ward's hierarchical clustering for cluster analysis. Clusters may or may not be supervised using extraneous data. The modular build-up of the workflow easily allows customization at several steps in the workflow, for instance, going beyond two data views. In the future, we will apply netMUG on other multi-view patient datasets for exploring disease subtyping and facilitating precision medicine, e.g., using RNA-Seq data and histopathology images for cancer subtyping.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Access to the 3D facial surface models in the 3D Facial Norms dataset requires proper institutional ethics approval and approval from the FaceBase data access committee. The PSU and IUPUI datasets were not collected with broad data sharing consent. This restriction is not because of any personal or commercial interests. Requests to access these datasets should be directed to the dbGaP controlled-access repository (http://www.ncbi.nlm.nih.gov/gap) and the FaceBase Consortium (https://www.facebase.org).

Ethics statement

Institutional review board (IRB) approval was obtained at each recruitment site. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian, for the publication of any potentially identifiable images or data included in this article.

Author contributions

ZL: Conceptualization, Formal Analysis, Methodology, Validation, Writing–original draft, Writing–review and editing. FM: Methodology, Writing–review and editing. HH: Methodology, Writing–review and editing. DD: Methodology, Writing–review and editing. MM: Resources, Writing–review and editing. SuW: Resources, Writing–review and editing. SeW: Resources, Writing–review and editing. MS: Resources, Writing–review and editing. BM-M: Supervision, Writing–review and editing. PC: Resources, Supervision, Writing–review and

editing. KVS: Conceptualization, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the European Union's Horizon 2020 research and innovation programme under the H2020 Marie Skłodowska-Curie grant agreement (No. 860895 to ZL, FM, and KVS; No. 813533 to DD and KVS). This study was supported by grants from the National Institute of Dental and Craniofacial Research at the National Institutes of Health (U01-DE020078 to SeW and MM; R01-DE016148 to MM and SeW; R01-DE027023 to SeW and PC).

Acknowledgments

Many thanks to the former PhD student Karlijne Indencleef at the Department of Electrical Engineering, ESAT/PSI, KU Leuven, who greatly aided the authors' understanding of how the US cohort data were pre-processed and how the imaging analysis was done.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1286800/full#supplementary-material

References

Abavisani, M., and Patel, V. M. (2018). Deep multimodal subspace clustering networks. *IEEE J. Sel. Top. Signal Process.* 12 (6), 1601–1614. doi:10.1109/JSTSP.2018.2875385

Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F., and Becker, J. (2020). Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings Bioinforma.* 21 (2), 541–552. doi:10.1093/bib/bbz015

D'Silva, S., Chakraborty, S., and Kahali, B. (2022). Concurrent outcomes from multiple approaches of epistasis analysis for human body mass index associated loci provide insights into obesity biology. *Sci. Rep.* 12 (1), 7306. doi:10.1038/s41598-022-11270-0

Dogan, A., and Birant, D. (2022). K-centroid link: a novel hierarchical clustering linkage method. *Appl. Intell.* 52 (5), 5537–5560. doi:10.1007/s10489-021-02624-8

Duroux, D., and Van Steen, K. (2022). 'netANOVA: novel graph clustering technique with significance assessment via hierarchical ANOVA'. Available at: https://www.biorxiv.org/content/10.1101/2022.06.28.497741v1.

Fawcett, K. A., and Barroso, I. (2010). The genetics of obesity: FTO leads the way. *Trends Genet.* 26 (6), 266–274. doi:10.1016/j.tig.2010.02.006

Ghosal, A., Nandy, A., Das, A. K., Goswami, S., and Panday, M. (2020). "A short review on different clustering techniques and their applications," in *Emerging technology in modelling and graphics*. Editors J. K. Mandal and D. Bhattacharya (Singapore: Springer Singapore), 69–83.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., et al. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 50, D687–D692. D1. doi:10.1093/nar/gkab1028

Gligorijević, V., Malod-Dognin, N., and Pržulj, N. (2016). Integrative methods for analyzing big data in precision medicine. *PROTEOMICS* 16 (5), 741–758. doi:10.1002/pmic.201500396

Hotelling, H. (1936). RELATIONS BETWEEN TWO SETS OF VARIATES. *Biometrika* 28 (3–4), 321–377. doi:10.1093/biomet/28.3-4.321

John, C. R., Watson, D., Barnes, M. R., Pitzalis, C., and Lewis, M. J. (2019). "Spectrum: fast density-aware spectral clustering for single and multi-omic data," in *Bioinformatics* Editor L. Cowen p btz704. doi:10.1093/bioinformatics/btz704

Kichaev, G., Bhatia, G., Loh, P. R., Gazal, S., Burch, K., Freund, M. K., et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* 104 (1), 65–75. doi:10.1016/j.ajhg.2018.11.008

Kriege, N. M., Johansson, F. D., and Morris, C. (2020). A survey on graph kernels. *Appl. Netw. Sci.* 5 (1), 6. doi:10.1007/s41109-019-0195-3

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47 (260), 583–621. doi:10.1080/01621459.1952.10483441

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. (2019). Estimating sample-specific regulatory networks. *iScience* 14, 226–240. doi:10.1016/j.isci.2019.03.021

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9 (1), 559. doi:10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24 (5), 719–720. doi:10.1093/bioinformatics/btm563

Loos, R. J. F., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., et al. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* 40 (6), 768–775. doi:10.1038/ng.140

Lu, Z. (2019). "Canonical correlation analysis with missing values: a structural equation modeling approach," in *Quantitative psychology*. Editor M. Wiberg (Cham: Springer International Publishing Springer Proceedings in Mathematics and Statistics), 243–254. doi:10.1007/978-3-030-01310-3_22

Lv, H., Zhang, M., Shang, Z., Li, J., Zhang, S., Lian, D., et al. (2017). Genome-wide haplotype association study identify the FGFR2 gene as a risk gene for Acute Myeloid Leukemia. *Oncotarget* 8 (5), 7891–7899. doi:10.18632/oncotarget.13631

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci.* 110 (11), 4245–4250. doi:10.1073/pnas.1208949110

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., and Hilsenbeck, S. G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* 19 (1), 71–86. doi:10.1093/biostatistics/kxx017

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* 31 (3), 274–295. doi:10.1007/s00357-014-9161-z

O'Bray, L., Rieck, B., and Borgwardt, K. (2021). "Filtration curves for graph representation," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event Singapore: ACM, Singapore, August 14 - 18, 2021, 1267–1275. doi:10.1145/3447548.3467442

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philosophical Mag. J. Sci.* 2 (11), 559–572. doi:10.1080/14786440109462720

Pierre-Jean, M., Mauger, F., Deleuze, J. F., and Le Floch, E. (2022). PIntMF: penalized integrative matrix factorization method for multi-omics data. *Bioinforma. Oxf. Engl.* 38 (4), 900–907. doi:10.1093/bioinformatics/btab786

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2019). *The DisGeNET knowledge platform for disease genomics: 2019 update*. Oxford, United Kingdom: Oxford University Press. gkz1021. doi:10.1093/nar/gkz1021

R Core Team (2022). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66 (336), 846–850. doi:10.1080/01621459.1971.10482356

Saria, S., and Goldenberg, A. (2015). Subtyping: what it is and its role in precision medicine. *IEEE Intell. Syst.* 30 (4), 70–75. doi:10.1109/MIS.2015.60

Shi, W. J., Zhuang, Y., Russell, P. H., Hobbs, B. D., Parker, M. M., Castaldi, P. J., et al. (2019). Unsupervised discovery of phenotype-specific multi-omics networks. ', *Bioinforma. Oxf. Engl.* 35 (21), 4336–4343. doi:10.1093/bioinformatics/btz226

Shi, X., Liang, C., and Wang, H. (2023). Multiview robust graph-based clustering for cancer subtype identification. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20 (1), 544–556. doi:10.1109/TCBB.2022.3143897

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., et al. (2020). A review of integrative imputation for multi-omics datasets. *Front. Genet.* 11, 570255. doi:10.3389/fgene.2020.570255

Spracklen, C. N., Karaderi, T., Yaghootkar, H., Schurmann, C., Fine, R. S., Kutalik, Z., et al. (2019). Exome-derived adiponectin-associated variants implicate obesity and lipid biology. *Am. J. Hum. Genet.* 105 (1), 15–28. doi:10.1016/j.ajhg.2019.05.002

Spycher, B. D., Silverman, M., Brooke, A. M., Minder, C. E., and Kuehni, C. E. (2008). Distinguishing phenotypes of childhood wheeze and cough using latent class analysis. *Eur. Respir. J.* 31 (5), 974–981. doi:10.1183/09031936.00153507

Storojeva, I., Boulay, J. L., Ballabeni, P., Buess, M., Terracciano, L., Laffer, U., et al. (2005). Prognostic and predictive relevance of DNAM-1, SOCS6 and CADH-7 genes on chromosome 18q in colorectal cancer. *Oncology* 68 (2–3), 246–255. doi:10.1159/000086781

Walakira, A., Ocira, J., Duroux, D., Fouladi, R., Moškon, M., Rozman, D., et al. (2022). Detecting gene–gene interactions from GWAS using diffusion kernel principal components. *BMC Bioinforma.* 23 (1), 57. doi:10.1186/s12859-022-04580-7

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11 (3), 333–337. doi:10.1038/nmeth.2810

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58 (301), 236–244. doi:10.1080/01621459.1963.10500845

Wen, Y., Song, X., Yan, B., Yang, X., Wu, L., Leng, D., et al. (2021). Multi-dimensional data integration algorithm based on random walk with restart. *BMC Bioinforma.* 22 (1), 97. doi:10.1186/s12859-021-04029-3

White, J. D., Indencleef, K., Naqvi, S., Eller, R. J., Hoskens, H., Roosenboom, J., et al. (2021). Insights into the genetic architecture of the human face. ', *Nat. Genet.* 53 (1), 45–53. doi:10.1038/s41588-020-00741-7

WHO Global InfoBase team (2005). *The SuRF Report 2. Surveillance of chronic disease Risk Factors: country-level data and comparable estimates*. Geneva: World Health Organization.

Witten, D. M., and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* 8 (1), Article28–27. doi:10.2202/1544-6115.1470

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostat. Oxf. Engl.* 10 (3), 515–534. doi:10.1093/biostatistics/kxp008

Yousefi, B., Melograna, F., Galazzo, G., van Best, N., Mommers, M., Penders, J., et al. (2023). Capturing the dynamics of microbial interactions through individual-specific networks. *Front. Microbiol.* 14, 1170391. doi:10.3389/fmicb.2023.1170391

Zhang, Y., and Li, T. (2011). "Consensus clustering+ meta clustering= multiple consensus clustering," in Twenty-Fourth International FLAIRS Conference, Palm Beach, Florida, USA, May 18-20, 2011.

Zhu, Z., Guo, Y., Shi, H., Liu, C. L., Panganiban, R. A., Chung, W., et al. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J. Allergy Clin. Immunol.* 145 (2), 537–549. doi:10.1016/j.jaci.2019.09.035