RESEARCH

Identifying significantly enriched gene sets with NMF-derived metagenes and their difference vectors

Charlie Lonergan*
and Haixuan Yang

Keywords: Central nervous system (CNS); embryonal tumours; gene-expression (GE) data; non-negative matrix factorisation (NMF); gene set enrichment analysis (GSEA)

Abstract

NMF is a class of unsupervised machine learning algorithm that is used in many clustering applications, and has gained popularity as a method to reduce the dimensionality of gene expression data, particularly in the field of tumour classification. NMF algorithms decompose large, non-negative matrices such as gene-expression microarray data into the product of two new matrices: $\mathbf{A} \approx \mathbf{WH}$. The columns of $\mathbf{W}$ are interpreted as 'metagenes', or patterns of gene expression, whereas each column of $\mathbf{H}$ represents the metagene expression pattern of the samples. This paper investigates the ability to infer meaningful biological insights, not only from each metagene as in previous studies, but also from the vector differences between these metagenes ($\Delta\mathbf{W}$). To achieve this, each metagene and difference vector are treated as gene lists in a given pathway enrichment analysis procedure.

This study shows that the use of these difference vectors improved the discovery of significant cancer-promoting pathways in 3 subclasses of CNS embryonal tumour, where the use of metagenes alone did not expose these associations, such as the upregulation of the JAK-STAT signalling pathway in gliomas, downregulation of synaptic vesicle cycle pathways in medulloblastomas, and upregulation of heterocycle catabolism in rhabdoid CNS tumours. The use of difference vectors as metagenes therefore shows great potential in aiding the discovery of enriched pathways, particularly when the number of underlying subtypes is unknown, or when these subtypes are phenotypically similar in expression.

1 Introduction

With the availability of large volumes of experimental omics data, such as microarray gene expression data, comes the need to extract meaningful biological information from these data through various computational techniques. Microarray data contains gene expression values from thousands of genes, so dimensionality reduction has become a key mission in understanding complex biological systems like tumours. Dimensionality reduction techniques such as principle component analysis (PCA), self-organising maps (SOM), hierarchical clustering (HC), and their derivatives are widely used in the field of data mining[1].

NMF[2, 3] has shown significant advantages over these techniques in the context of microarray data analysis, including:

i ease of interpreting NMF "metagenes" vs. PCA "eigengenes"[4],
ii the stability of NMF over SOM clustering,
iii NMF's robustness over large inputs, and
iv NMF's ability to discover hierarchical structure, without enforcing it as in HC[5].

NMF has already been used in previous studies to not only discover the underlying structures of such microarray data, but also determine the biological relevance of these structures through various enrichment analyses. Wilson *et al.*(2012)[6] used this approach to reveal many new interactions between genes of the *Arabidopsis thaliana* plant, while Northcott

*Correspondence: haixuan.yang@nuigalway.ie
School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway,
Full list of author information is available at the end of the article

et al. (2011)[7] identified crucial differences in the presentation of Shh-medulloblastomas in infants and adults. More recently, Wang *et al.* (2016)[8] used enrichment analysis to benchmark their own NMF algorithm against previous methods. What is common to all of these approaches is the use of NMF-derived "metagenes" - overall patterns of gene expression - as gene lists when conducting enrichment analysis. This paper explores whether the differences between these metagenes can provide more insight than their direct analysis.

1.1 Background to NMF

NMF was first introduced in 1994 by Paatero & Tapper[2], under the name 'Positive Matrix Factorisation' (PMF), as an alternative to Factor Analysis (FA) and PCA in the context of environmetrics. PMF was preferable to FA in this context due to the positivity of its solutions, since FA required rotational tools in order to interpret ambiguous negative results when analysing environmental data[9]. This made interpreting its results much more intuitive than was previously possible. However, PMF was not yet suitable for many other applications due to issues with the algorithm's convergence and its lack of generality in other fields[10].

Lee & Seung (1999)[3] expanded on this work by developing an iterative update algorithm for NMF, demonstrating its power as a parts-based representation algorithm by applying it to facial images and comparing the representations to those from PCA and vector quantisation (VQ). While NMF decomposed the facial images into their constituent parts - eyes, noses, mouths, etc., the PCA and VQ results were much more holistic representations, and therefore harder to interpret. This popularised NMF in many other fields, and since Lee & Seung's work the algorithm has branched into many subdivisions, such as Constrained NMF (CNMF), Structured NMF (SNMF) and Generalized NMF (GNMF)[10]. Numerous variations of these subdivisions can be used depending on the desired application, such as Sparse NMF (sNMF)[11], Non-Smooth NMF (nsNMF)[12], and Multi-layer NMF (MLNMF)[13]. For a comprehensive review of NMF methods, consult "Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation" by Zdunek *et al.*(2009)[14].

1.2 Description of the NMF Algorithm

Given a matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{N \times M}$ and a rank $r \in \mathbb{N}$, NMF approximates $\mathbf{A} \approx \mathbf{WH} : \mathbf{W} \in \mathbb{R}_{\geq 0}^{N \times r}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times M}$, where N is the number of features and M is the number of objects in the dataset $\mathbf{A}$. The coefficient matrix $\mathbf{H}$ defines cluster membership such that $h_{kj} > h_{ij}\ \forall i \neq k$ implies the j-th object in $\mathbf{A}$ belongs to the k-th cluster, whereas the basis matrix $\mathbf{W}$ defines each cluster in terms of the features of $\mathbf{A}$. NMF randomly generates matrices $\mathbf{W}$ and $\mathbf{H}$ before iteratively updating these matrices to minimize a loss function D that measures the quality of the approximation. This is summarised by the following optimisation problem:

$$\min_{\mathbf{W},\mathbf{H} \geq 0} [D(\mathbf{A}, \mathbf{WH}) + R(\mathbf{W}, \mathbf{H})] \tag{1}$$

There is a choice of loss functions available for D, which are usually based on either the Euclidean distance between $\mathbf{A}$ and $\mathbf{WH}$ (2) or the Kullback-Leibler (K-L) divergence (3)[15]:

$$||\mathbf{A} - \mathbf{WH}||^2 = \sum_{ij} (a_{ij} - (\mathbf{WH})_{ij})^2 \tag{2}$$

$$D(\mathbf{A}||\mathbf{WH}) = \sum_{ij} (a_{ij} log \frac{a_{ij}}{(\mathbf{WH})_{ij}}) \tag{3}$$

Euclidean distance is typically used in applications where the noise is expected to be Gaussian distributed, whereas K-L divergence is associated with Poisson-distributed noise[16]. From equation 1, $R(\mathbf{W}, \mathbf{H})$ is a regularisation function that can enforce optional properties on $\mathbf{W}$ and $\mathbf{H}$, including smoothness or sparseness[12, 14, 17].

In 2001, Lee & Seung formally proved that equations 2 & 3 converge to a stationary point under the multiplicative update rules 4 & 5 respectively[15]:

$$\begin{aligned} h_{\alpha\mu} &\leftarrow h_{\alpha\mu} \frac{(\mathbf{W}^T\mathbf{A})_{\alpha\mu}}{(\mathbf{W}^T\mathbf{WH})_{\alpha\mu}}; \\ w_{i\alpha} &\leftarrow w_{i\alpha} \frac{(\mathbf{AH}^T)_{i\alpha}}{(\mathbf{WHH}^T)_{i\alpha}} \end{aligned} \tag{4}$$

$$\begin{aligned} h_{\alpha\mu} &\leftarrow h_{\alpha\mu} \frac{\sum_i w_{i\alpha} a_{i\mu} / (\mathbf{WH})_{i\mu}}{\sum_k w_{k\alpha}}; \\ w_{i\alpha} &\leftarrow w_{i\alpha} \frac{\sum_\mu h_{\alpha\mu} a_{i\mu} / (\mathbf{WH})_{i\mu}}{\sum_v w_{\alpha v}} \end{aligned} \tag{5}$$

Following this development, NMF became a widely adopted dimensionality reduction technique suitable to many applications.

1.3 NMF Rank Estimation

The factorisation rank r defines how many basis vectors are present in $\mathbf{W}$ and $\mathbf{H}$, and represents the number of clusters present in the underlying data. r is therefore a crucial parameter in any NMF method, especially in class discovery applications, so there are a variety of methods available to estimate it. The most common approach in clustering studies is to use the *cophenetic correlation coefficient* $\rho_r(\bar{\mathbf{C}})$, introduced by Brunet *et al.*[5], to measure clustering stability over a range of values of r.

To evaluate ρ_r, an NMF algorithm is first run on the input matrix $\mathbf{A}$ multiple times (typically 20–30 runs is sufficient[18]). When the algorithm clusters the data well, it is expected that these clusters do not change very much over each of the runs. Cluster assignment is described in each run by the $M \times M$ *connectivity matrix* $\mathbf{C}$[19], where:

$$c_{ij} := \begin{cases} 1 & \text{if objects } i \text{ and } j \text{ cluster together;} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The *consensus matrix* $\bar{\mathbf{C}}$ defines the average connectivity over all runs, where $\bar{c}_{ij} \in [0,1]$ relates to the probability that objects i and j share the same cluster. If a clustering is very stable, it is expected that elements $\bar{c}_{ij}$ tend towards 0 or 1; perfect clustering across all runs yields a consensus matrix with $\bar{c}_{ij} \in 0, 1 \; \forall i, j$. In order to calculate ρ_r, average linkage HC is used to rearrange the rows and columns of $\bar{\mathbf{C}}$. ρ_r is then calculated as the Pearson correlation coefficient between $\mathbf{I} - \bar{\mathbf{C}}$ and the distance matrix between objects induced by this HC rearrangement, giving $\rho_r = 1$ for perfect clustering consensus. Brunet *et al.* observed that high values of ρ_r correspond to NMF decompositions where r represents an accurate number of clusters, and suggested choosing the smallest value of r for which ρ_r begins to decrease.

Hutchins *et al.* (2008)[18] proposed that the optimal value of r is chosen where the plotted residual sum of squares (RSS) between $\mathbf{A}$ and $\mathbf{WH}$ shows an inflection point, whereas Frigyesi *et al.* (2008) recommended choosing the smallest value of r where the decrease in the RSS is lower than the decrease of the RSS obtained from random data[20]. However, often such inflection points in the RSS curve can be difficult to identify, especially when dealing with low rank estimates, while Frigyesi *et al.*'s method can be challenging to visualise at times. Kim & Park (2007)[21] prescribed measuring the *dispersion* δ of each consensus matrix $\bar{\mathbf{C}}$ (Equation 7), and choosing the rank r that maximises δ over a given range.

$$\delta(\bar{\mathbf{C}}) := \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} 4(\bar{c}_{ij} - \frac{1}{2})^2 \in [0,1] \tag{7}$$

In practice it is most helpful to use multiple methods when estimating the rank of an NMF decomposition, however it should be mentioned that the *cophenetic correlation* and *dispersion* methods are often the simplest to visualise. The resulting cluster assignments should also be inspected to confirm that the estimated rank is appropriate, especially when the data's underlying structure is known *a priori*. This can be easily achieved by visualising $\mathbf{W}$, $\mathbf{H}$, and $\bar{\mathbf{C}}$ as heatmaps[22].

1.4 NMF Decomposition Quality Measures

Once $\mathbf{A}$ has been factored, it is important to estimate the quality of this factorisation. NMF's algorithmic performance is often measured via the computational time taken to complete a factorisation or the residual error between $\mathbf{A}$ and $\mathbf{WH}$. Other performance measures include the *sparseness*, *purity*, and *entropy* of a NMF decomposition.

The *sparseness*[17] of a n-dimensional vector $\mathbf{x}$ is defined by Equation 8, and describes the proportion of $x_i \in \mathbf{x}$ which are non-zero elements, such that *sparseness*$(\mathbf{x}) \to 1$ iff $\mathbf{x}$ contains a single non-zero element, and *sparseness*$(x) \to 0$ iff $x_i = c \neq 0 \; \forall i$ for some constant c.

$$sparseness(\mathbf{x}) := \frac{\sqrt{n} - \frac{\sum_i |x_i|}{\sqrt{\sum_i x_i^2}}}{\sqrt{n} - 1} \tag{8}$$

Sparseness can be enforced on the columns of $\mathbf{W}$ or rows of $\mathbf{H}$ by constraining Equation 1 such that $sparseness(\mathbf{W}_i) = S_W \; \forall i$, or $sparseness(\mathbf{H}_i) = S_h \; \forall i$, depending on the desired application. This can improve one's ability to interpret the basis or coefficient matrix ($\mathbf{W}$ or $\mathbf{H}$). For example, it is often desirable to enforce low sparseness on the columns of $\mathbf{W}$ to cluster the underlying data into distinct, low-dimensional subtypes.

Pascual-Montano *et al.* (2006)[12] used the notion of sparseness to compare their nsNMF algorithm to other methods, by evaluating how the *explained variance*, a function of the RSS between $\mathbf{A}$ and $\mathbf{WH}$, changes with sparseness. They assert that a NMF algorithm that maintains high explained variance, meaning an accurate approximation of $\mathbf{A} \approx \mathbf{WH}$, over a wide range of sparseness is desirable, and demonstrated that nsNMF

achieves high explained variance while preserving high sparseness in both **W** and **H**. This was an improvement over previous models, since enforcing sparseness on one of the factors tended to force smoothness ("non-sparseness") on the other in order to preserve accurate approximation, due to a trade off resulting from the multiplicative update rules. The researchers achieved this by multiplying both factors by a *smoothing matrix* $\mathbf{S} \in \mathbb{R}^{r \times r}$ at each update step of the algorithm:

$$\mathbf{S} := (1-\theta)\mathbf{I} + \frac{\theta}{r}\mathbf{1}\mathbf{1}^T \;:\; \theta \in [0,1] \tag{9}$$

Kim & Park (2007)[21] defined the concepts of *purity* and *entropy* as measures of NMF's clustering performance in applications where the data's true clustering is known *a priori*. Let n be the number of objects in a dataset, l the number of true clusters present, and k the number of clusters generated by a given NMF algorithm. The *purity* of a decomposition is given by

$$purity := \frac{1}{n}\sum_{q=1}^{k} max_{1 \le j \le l}(n_q^j), \tag{10}$$

where n_q^j is the number of objects in cluster q that belong to the true cluster j. The purity of a decomposition is bounded by 0 and 1, and increases with accurate clustering performance.

The *entropy* of a decomposition is defined as

$$entropy := -\frac{1}{n \log_2 l}\sum_{q=1}^{k}\sum_{j=1}^{l} n_q^j \log_2 \frac{n_q^j}{n_q}, \tag{11}$$

where n_q gives the size of cluster q. Entropy is also bounded between 0 and 1, but decreases with clustering accuracy. Kim & Park demonstrated how their sNMF algorithm performed with higher purity and lower entropy than other algorithms over a range of sparseness values.

1.5 Background to Pathway Enrichment Analysis

GSEA is a widely practised enrichment analysis technique for extracting meaningful biological pathways from gene expression data that was first introduced by Mootha *et al.* (2003)[23] and was later improved by Subramanian *et al.* (2005)[24]. Given a predefined gene set S and an ordered list of L genes associated with a phenotypic class, the objective of GSEA is to determine whether the genes in S are randomly distributed throughout L, or if they appear at its extremities with statistical significance. There are many options available for the selection of S, including genes in a known biological pathway, genes targeted by transcription factors or miRNA, or known oncogenic gene sets, depending on the desired application. Such gene sets can be accessed through online pathway databases such as Reactome[25], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[26], and Gene Ontology (GO)[27]. What is important is that the gene list L has been ordered in a statistically valid manner and the selection of S is biologically meaningful. An outline of the Subramanian GSEA method is as follows.

1 Measure the difference between expression values in two phenotypes, *e.g.* case vs. control, and calculate an association statistic between the two classes for each gene. There are many association statistics to consider, such as the t-test[28], Z-score[29], or ANOVA[30]. Sort the genes by the desired statistic to give the ranked gene list L.

2 Calculate an enrichment score (ES) to measure the degree to which S is differentially expressed in L. GSEA traverses L and increases a running-sum gene set statistic when a gene in S is found, and decreases for genes not found in S. This statistic is often based on a weighted Kolmogorov-Smirnov (WKS) test, and the ES of S is defined as the maximum deviation of this statistic from zero.

3 Evaluate the statistical significance of the ES against its null distribution. This is achieved by permuting the phenotypic class labels and recalculating ES for S, and the derived significance level is calculated relative to this distribution. It is possible to derive a null distribution by permuting each gene, however this removes any genetic correlations present in the data and is not recommended[31].

4 Calculate the normalised enrichment score (NES). When evaluating multiple gene sets it is necessary to account for multiple hypothesis testing. The NES is calculated by normalising each ES to account for the number of genes in S, and the false discovery rate (FDR) of each NES is derived by comparing the tails of its observed and null distributions, typically via the Benjamini-Hochberg method[32, 31]. Each pathway's "leading edge" can also give valuable insight into the particular genes responsible for high NES scores. This is defined as the collection of genes that contribute towards the maximum deviation before the score returns to zero.

One advantage of this methodology over its predecessor is that the WKS test statistic is weighted according

to each gene's association with the given phenotype. In the Mootha method this running-sum statistic was equally weighted with each increment, meaning genes in S could be assigned high ES when found in the middle of L. This was not an ideal solution, since the most important gene sets of interest should show the greatest difference in gene expression for most applications. This improvement is demonstrated as follows.

Let N be the total number of genes and M the number of samples observed, and let $L = \{g_i : i \in [1, N]\}$ be the list of N genes ordered by phenotypic association statistics $\{r_i : i \in [1, N]\}$. The *enrichment score* of gene set S is defined as $ES(S) := \max_{g_i \in L}(P_+(S,i) - P_-(S,i))$, where P_+ and P_- denote the proportion of genes $g_i \in S$ before position i and the proportion of genes $g_i \notin S$ before position i, respectively, for all positions $i \in [1 : N]$.

In Mootha's GSEA method, P_+ and P_- are defined by Equation 12, where N_R denotes the number of genes $g_i \in S$ before position i.

$$\begin{aligned} P_+(S,i) &= \sum_{g_i \in S, j \leq i} \frac{1}{N_R}; \\ P_-(S,i) &= \sum_{g_i \notin S, j \leq i} \frac{1}{N - N_R} \end{aligned} \tag{12}$$

Subramanian *et al.* used a weighting factor $p \in [0, 1]$ to dampen $ES(S)$ for genes that appear in the middle of L (Equation 13). When $p = 1$, P_+ and P_- weight the genes by their corresponding association statistics r_j directly, and when $p = 0$ the $|r_j|^p$ terms reduce to 1, recovering Equation 12.

$$\begin{aligned} P_+(S,i) &= \sum_{g_i \in S, j \leq i} \frac{|r_j|^p}{N_R}; \\ P_-(S,i) &= \sum_{g_i \notin S, j \leq i} \frac{1}{N - N_R} : \\ N_R &= \sum_{g_i \in S} |r_j|^p \end{aligned} \tag{13}$$

While this demonstrates an improvement on the previous method, GSEA has since seen multiple revisions with many variants available for choosing association statistics, gene set statistics, ES significance estimation and FDR discovery methods. In a comprehensive survey of many of these variants, Ackermann & Strimmer (2009)[33] found that the choice of association statistic does not impact enrichment results as significantly as that of the gene set statistic. Of these, the Wilcoxon rank sum (WRS)[34], *maxmean*[35], and χ^2 statistics are often used as alternatives to WKS. Efron & Tibshirani (2007)[35] demonstrated that their *maxmean* statistic outperforms WKS analytically as well as in simulated datasets, however Ackermann & Strimmer showed that WRS was better than *maxmean* in some cases, particularly for gene sets with high levels of intra-group correlation.

Efron & Tibshirani also developed "restandardisation" as an alternative approach when estimating ES significance, which randomises genes as well as permuting the samples before scaling the *maxmean* statistic, and recently, Korotkevich *et al.* (2021)[36] developed a Monte-Carlo sampling-based method that can estimate ES significance with much greater sensitivity than the original approach. While it is not yet clear whether this Monte-Carlo approach is appropriate for wide ranges of gene sets, Ackermann & Strimmer stress the importance of considering one's null hypothesis when evaluating ES significance, as this has a considerable impact on the recommended method. The null hypotheses used are generally of two varieties, "competitive" and "complete" null hypotheses (Q_1 & Q_2). Q_1 states that "the genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes", whereas Q_2 states "the gene set does not contain any genes whose expression levels are associated with the phenotype of interest.[37]" Q_1 is suited to permuting genes, while sample permutation methods are more appropriate for Q_2. Finally, when correcting for multiple hypothesis testing, the Benjamini-Hochberg[32] method is generally preferred[31] over Bonferroni correction[38] since Bonferroni correction can result in many false negatives. Benjamini-Hochberg assumes p-values are uniformly distributed however, which may not be appropriate for some gene set enrichment studies due to intra-group correlation between different gene sets.

It is beyond the scope of this paper to assess best practice when choosing these parameters, but some recent methods have been developed to address some of these challenges. Pathway Analysis with Down-weighting of Overlapping Genes (PADOG)[39] gives more weight to genes that appear in few pathways when assessing their significance, with the logic that a gene being present in one pathway only is evidence that the pathway is more enriched in a given phenotype, thus reducing the number of false positive pathways found. Recently, Network-Based Gene Set Enrichment Analysis (NGSEA) as developed by Han *et al.* (2019)[40] and assesses the ES of pathways by considering the expression values of each gene's neighbours in a pathway

network, in addition to the gene's individual expression value. Not only did NGSEA outperform GSEA in identifying relevant phenotypes, it also overcomes a limitation of GSEA in that it identifies genes that are regulated in both directions, whereas GSEA is is inherently one-directional. NGSEA is similar in concept to several topology-based enrichment analysis tools such as Pathway Express[41], CePaGSA[42], and PathNet[43], in that each of these methods consider the impact of differentially expressed genes on pathways as *networks*, though they differ in how they measure this impact. Pathway Express measures how the expression of a gene propagates through the topology of a pathway, whereas CePaGSA weights each pathway's statistic in terms of a variety of network centrality measures. Pathnet considers the impact of differentially expressed genes on *all* pathways to capture the connectivity of a pathway network.

While it is clear that there are many enrichment analysis tools available that possess advantages over GSEA, it is still a widely used tool that is often used as to benchmark other methods' performance. This paper explores the potential for incorporating NMF-derived metagenes and their difference vectors into virtually any enrichment analysis technique, so for this purpose GSEA is as good a candidate as any other. As Nguyen *et al.* (2019)[44] remark in their review of pathway analysis techniques, "no method is perfect".

2 Motivation

Since NMF can cluster large volumes of gene expression data into their biological subtypes, this makes it an intuitive candidate for deriving phenotypically relevant gene lists, especially when these subtypes are not known *a priori*.

For a given matrix of gene expression values $\mathbf{A} \in \mathbb{R}_{\geq 0}^{N \times M}$, corresponding to the expression values of N genes in M samples, a rank r NMF decomposition factors $\mathbf{A}$ into two positive matrices, $\mathbf{A} \approx \mathbf{WH}$, where r is the number of classes/clusters present in the dataset (*e.g.* tumour classifications). $\mathbf{W} \in \mathbb{R}_{\geq 0}^{N \times r}$ defines r metagenes (gene expression patterns), such that element w_{ij} is the coefficient of gene i in metagene j. $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times M}$ defines the metagene expression patterns of each sample such that element h_{ij} represents the expression coefficient of metagene i in sample j[5].

Once a gene expression matrix has undergone a NMF decomposition of sufficient quality, the challenge becomes one of extracting biological meaning from this decomposition. Since each column of $\mathbf{W}$ represents a pattern of gene expression for each class, these columns can be sorted by coefficient value into ordered gene lists. These gene list can then undergo various enrichment analyses to elucidate the biological meaning of each metagene, since the coefficient values of each metagene are associated with their corresponding phenotype. Previous studies[8, 6, 7] have used this approach with some success to gain biological insights from NMF-derived metagenes. However, what has not yet been attempted is the derivation of *difference vectors* from these metagenes and using these as gene lists in pathway analyses, which is the focus of this paper.

Suppose metagene $\mathbf{W}_M$ represents the gene expression pattern associated with medulloblastomas, and metagene $\mathbf{W}_G$ is the gene expression pattern associated with gliomas. Both metagenes represent a class of central nervous system (CNS) embryona, and as such are likely to share many genes in the same biological pathways, so the expression coefficients of these genes may be similar in both metagenes. Therefore, pathway analysis of these metagenes may not reveal many nuances in the expression of distinct, yet similar tumour classes. However, the (sorted) gene list $\Delta\mathbf{W}_{MG} := \mathbf{W}_M - \mathbf{W}_G$ would rank those genes with similar expression coefficients towards the middle of the gene list, and those with differing expression values towards its extremities, thus exposing them in any subsequent pathway analyses.

It should be noted that one particular benefit of the non-negative property of $\mathbf{W}$ is it ensures that genes with low expression values are not pushed to the extremities of $\Delta\mathbf{W}_{ij}$ by this negation. Suppose for example that for some $n \in [1, N]$, $i, j \in [1, r]$, there exist small expression coefficients $w_{i,n} \in \mathbf{W}_i$, $w_{j,n} \in \mathbf{W}_j$: $w_{i,n} = -w_{j,n}$. Then $\delta w_{ij,n} = w_{i,n} - w_{j,n} = 2w_{i,n}$, which would artificially inflate the value of these coefficients in the subsequent analysis and could lead to spurious results. This non-negative property also means that subsequent pathway analyses allude to a dual interpretation; the upregulated genes in gene list $\Delta\mathbf{W}_{ij}$ are downregulated in gene list $\Delta\mathbf{W}_{ji}$, and vice versa. Therefore, if gene list $\Delta\mathbf{W}_{ij}$ demonstrates a NES score of N in a given pathway, then gene list $\Delta\mathbf{W}_{ji}$ would in theory produce an NES score of $-N$ since the running sum statistics increase and decrease by the same magnitude, assuming that the absolute gene-level statistics in the leading edge of $\Delta\mathbf{W}_{ij}$ are equal to those in the trailing edge of $\Delta\mathbf{W}_{ji}$.

3 Methods and Materials

3.1 Data

Sample CNS gene expression microarray data was taken from Pomeroy *et al.*(2002)[45], which were

scanned using Affymetrix Hu6800 assays capturing 7129 probe readings. The primitive neuroectodermal tumour (PNET) samples were excluded as in Brunet *et al.*(2004)[5], and the remaining data from 34 samples (10 classic medulloblastomas, 10 malignant gliomas, 10 rhabdoids and 4 normals) underwent background correction and normalisation. 15 rows were excluded that are not associated with unique, anti-sense strand detecting probe sets ("at" suffixes), such as probes in mixed or identical probe-sets ("xt" and "st" suffixes respectively - see [46]). Negative values were then set to zero, due to the non-negativity constraint of NMF, and any rows with all zero entries were removed (246 rows in total). This dataset shall be referred to as "Dataset 1". While setting some values to zero may lose some information, negative expression level suggests a given gene is not important to our analysis.

An alternative additive method was attempted to achieve this non-negativity constraint, where each column of A was offset by its minimum value, setting the minimum value of $\mathbf{A}$ to zero. This method did not produce a rank $r = 4$ estimation of the data, nor did the subsequent NMF decomposition sufficiently cluster the samples into their known phenotypic classifications, and so this dataset ("Dataset 1-A") was ignored (see Figures 1, 6, 7, Table 1).

Dataset 1 is compared to a second dataset, "Dataset 2", from the same experiment that underwent a tumour purification procedure. A K-L divergence based NMF was used to perform the tumour purification step, enforcing a constraint that the columns of H sum to 1 at each iteration of NMF algorithm. This ensures that the columns of $\mathbf{W}$ correspond to specific gene expression signatures for each cell type [47]. Since it is known from previous studies that the samples are of 4 tumour classes, it was assumed that the data has a rank $r = 5$ decomposition to include an impure gene expression profile that corresponds to the component with the most ambiguous cluster assignment. This impure component was removed from the model by subtracting $w_{ji}h_{ik}$ from $\mathbf{A}$ $\forall j \in [1, 7129], k \in [1, 34]$, where h_{ik} is associated with the *lowest* clustering frequency for each sample. As with Dataset 1, 15 rows associated with mixed or identical probe-sets were removed.

3.2 NMF Analysis

Rank r estimation of each dataset was performed in line with Brunet *et al.*(2004)[5] to ensure the samples are clustered accurately. nsNMF[12] was performed over 50 iterations on each dataset for ranks $r \in [2, 8]$, in order to generate more sparse metagenes than the basic NMF algorithm. Here, the cophenetic correlation coefficient (ρ) curve spiked at rank $r = 4$ before decreasing, which was expected from the authors' results and is consistent with the four biological classifications of the samples. The dispersion (δ) was also measured across each rank and decreased after $r = 4$ in both datasets, indicating that $r = 4$ is the optimal choice of rank for these data, and would be chosen if the underlying subtypes were unknown. Comparing Datasets 1 & 2, both ρ and δ were higher in Dataset 2 than in Dataset 1 overall, which suggests the tumour purification step increases clustering consistency over each run (see Figure 1).

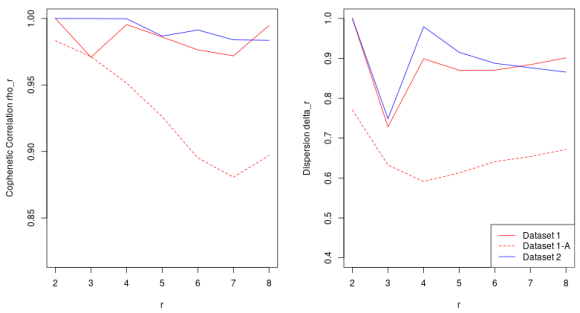

Figure 1 Rank estimation survey. 50 iterations were performed on datasets 1 & 2 for ranks $r \in [2 : 8]$. Cophenetic coefficient ρ_r and dispersion δ_r curves indicate highest cluster reproducibility at rank $r = 4$.

nsNMF was then performed on each dataset at rank $r = 4$ over 50 iterations to produce the final $\mathbf{A} \approx \mathbf{WH}$ decompositions. Comparing each decomposition (see Table 1), Dataset 2 displays lower purity and higher entropy than Dataset 1, while the average *sparseness(***W***)* of Dataset 2 was marginally higher than in Dataset 1, with equal average *sparseness(***H***)*. However, in the context of the purity and entropy scores, this marginal result can be considered negligible. These results indicate the nsNMF algorithm clustered the samples more accurately in Dataset 1, while the decomposition of Dataset 2 showed greater consistency.

dataset	sparseness(W)	sparseness(H)	purity	entropy
Dataset 1-A	0.06	0.35	0.68	0.46
Dataset 1	0.70	0.88	0.91	0.19
Dataset 2	0.72	0.88	0.88	0.26

Table 1 nsNMF decomposition quality measures for each dataset at rank $r = 4$ over 50 iterations.

Examining the consensus and coefficient matrices, nsNMF clustered the samples into their known phe-

notypic classes with considerable accuracy. In Dataset 1, nsNMF misclassified one medulloblastoma sample as normal, one rhabdoid as a medulloblastoma and one glioma as a rhabdoid (Figures 2 & 3). In Dataset 2 nsNMF performed similarly, but misclassified an additional medulloblastoma sample as being in the rhabdoid cluster $\mathbf{H}_R$ (Figures 4 & 5), which is reflected in Dataset 2's poorer purity and entropy measures. This misclassification can lead to some mixed results after enrichment analysis, but these effects should be overcome by the majority of samples being clustered appropriately.

One limitation of many NMF algorithms is that the resulting decomposition is not always unique. This can result in different clustering solutions, leading to varying interpretations and biological implications. In order to enforce unique decompositions on each dataset before performing GSEA, $\mathbf{W}$ was normalised via the post-processing method described in Yang & Seoighe (2016)[48], by dividing each column of $\mathbf{W}$ by its maximum. The difference matrix $\Delta\mathbf{W}$ was generated by subtracting each column of $\mathbf{W}$ from one another, resulting in 6 columns of $\Delta\mathbf{W}$. There are 12 total permutations of the 4 metagenes, however 6 of these are redundant for this analysis due to the dualistic interpretation of each gene set as discussed. The resulting columns of $\mathbf{W}$ (metagenes) and their difference vectors, $\Delta\mathbf{W}$, were then sorted by coefficient value and passed through GSEA as ranked gene lists for both datasets 1 & 2.

nsNMF was implemented with the R/CRAN package *NMF*[22], using the nonsmooth 'nsNMF' method to minimise the K-L divergence loss function over 50 iterations. The 'seed' parameter was kept as 'random' when initialising basis and coefficient matrices. It is possible that enforcing sparseness only on $\mathbf{W}$ through sNMF[11] could improve the accuracy of enrichment results for the derived gene lists, however this is not necessary when demonstrating the differences between the enrichment results of $\mathbf{W}$ and $\Delta\mathbf{W}$.

3.3 Gene Set Enrichment Analysis

Each gene list was compared against both KEGG[26] and GO:BP[27] gene sets. ES was calculated using the WKS statistic over 1000 permutations with a weighting factor of $p = 1$, corresponding to a "complete" null hypothesis (Q_2) as discussed above. For computational reasons, genes with an absolute expression coefficient value $|w| < 0.01$ were excluded from each gene list, leaving only the highest and lowest ranked genes. The p-value boundary was set to 1e−20 to allow for more accurate ES p-value estimation, using an integrated Monte-Carlo method outlined in Korotkevich *et al.* (2021)[36]. NES were calculated for each gene set as the mean ES as in Subramanian *et al.* (2005)[24], and multiple testing correction was performed via the Benjamini-Hochberg method[32]. While this correction method does not account for intra-group correlation as in other network-/topology-based methods, the differences between the results from $\mathbf{W}$ and $\Delta\mathbf{W}$ can still be observed. These results were then filtered for adjusted significance levels of $p_{adj} < 0.01$, and the top 5 pathways were collated for each gene list - a common approach for discovering the most important enriched genes (see Tables). Complete results, including the leading edge genes of each gene list, are available at `https://github.com/c-lonergan/nmf_gsea` as .Rdata files, in addition to the code used and a Docker[49] container with the R packages used in the analysis.

GSEA was performed using the R/Bioconductor package *clusterProfiler* v.4.1.1[50], and enrichment maps were plotted with the R/Bioconductor package *enrichplot* v.1.13.1[51].

4 Results

Considering the top 5 pathways for each gene list (Tables 5 - 8), Datasets 1 & 2 display shared enrichment in 4 KEGG pathways and 20 GO:BP pathways, but also identify some pathways uniquely. Table 2, displays these shared pathways as subsets of Tables 5 - 8 without gene-set level statistics, since these were calculated and adjusted independently of each other and cannot be pooled in a meaningful way. Dataset 1 shows enrichment in 8 KEGG pathways and 4 GO:BP pathways that are not identified by Dataset 2 (Table 3), while Dataset 2 shows unique enrichment in 4 KEGG pathways and 7 GO:BP pathways (Table 4). Of the KEGG pathways shared by both datasets, hsa03010 (Ribosome) and hsa05171 (COVID-19) showed significant enrichment across most metagenes and difference vectors, including the normalised metagene $\mathbf{W}_N$. While hsa03010 has been shown to be associated with other pathways involved in cancers such as melanoma, glioma, leukemia, and others[52], hsa03010 and hsa05171 share many overlapping genes, and since the underlying data was sampled many years before the emergence of COVID-19, these pathways will be ignored.

Of particular interest are hsa05012 (Parkinson's disease) and hsa05020 (Prion disease) which are share genes with pathways involved in the formation of brain tumours. Shared top 5 GO:BP pathways are mostly involved in normal cellular activity like translation, protein transport, RNA degredation and gene regulation. Some notable pathways enriched in both datasets

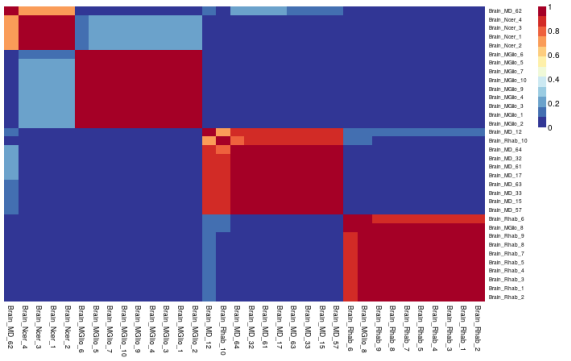

Figure 2 Consensus matrix of Dataset 1 at rank $r = 4$ indicates high clustering consistency across 50 runs.

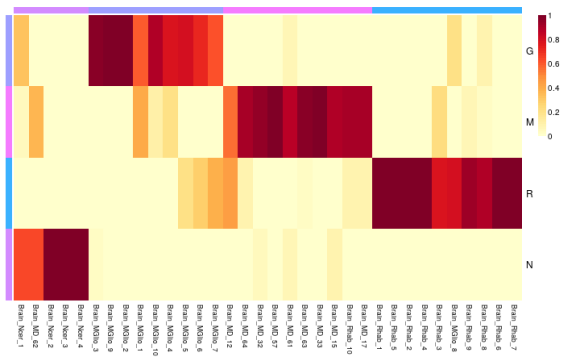

Figure 3 Coefficient matrix of Dataset 1 at rank $r = 4$ indicates sparse phenotypic association across 50 runs.

include various intracellular and cell-cell adhesion pathways (GO:0022610, GO:0033365, GO:0072594, GO:0090150, GO:0098609, GO:0098742) and viral gene expression (GO:0016032 GO:0019083).

Dataset 1 identifies several known pathways involved in CNS cancers such as hsa05169 (Epstein-Barr virus infection), hsa04721 (synaptic vesicle cycle) and hsa05022 (neurodegeneration), while Dataset 2 identifies hsa00010 (glycolysis / gluconeogenesis), hsa04630 (JAK-STAT signaling), hsa05014 (amyotrophic lateral sclerosis), hsa05016 (Huntington's Disease), and GO:0044270 (heterocycle catabolism).

ID	Description
hsa03010	Ribosome
hsa05012	Parkinson disease
hsa05020	Prion disease
hsa05171	Coronavirus disease - COVID-19
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
GO:0000956	nuclear-transcribed mRNA catabolic process
GO:0006401	RNA catabolic process
GO:0006402	mRNA catabolic process
GO:0006412	translation
GO:0006612	protein targeting to membrane
GO:0006614	SRP-dependent cotranslational protein targeting to membrane
GO:0006886	intracellular protein transport
GO:0007155	cell adhesion
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules
GO:0010629	negative regulation of gene expression
GO:0016032	viral process
GO:0019083	viral transcription
GO:0022610	biological adhesion
GO:0033365	protein localization to organelle
GO:0044403	biological process involved in symbiotic interaction
GO:0072594	establishment of protein localization to organelle
GO:0090150	establishment of protein localization to membrane
GO:0098609	cell-cell adhesion
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules

Table 2 Enriched pathways found in both Datasets 1 & 2

4.1 Normal Metagene: $\mathbf{W}_N$

$\mathbf{W}_N$ shows enrichment in most of the pathways one would expect from normal biological function, as can be seen from Figures 8 and 9.

ID	Description
hsa04141	Protein processing in endoplasmic reticulum
hsa04650	Natural killer cell mediated cytotoxicity
hsa04721	Synaptic vesicle cycle
hsa04940	Type I diabetes mellitus
hsa05022	Pathways of neurodegeneration - multiple diseases
hsa05169	Epstein-Barr virus infection
hsa05332	Graft-versus-host disease
hsa05416	Viral myocarditis
GO:0015031	protein transport
GO:0019080	viral gene expression
GO:0043043	peptide biosynthetic process
GO:0051649	establishment of localization in cell

Table 3 Enriched pathways found in Dataset 1 only.

ID	Description
hsa00010	Glycolysis / Gluconeogenesis
hsa04630	JAK-STAT signaling pathway
hsa05014	Amyotrophic lateral sclerosis
hsa05016	Huntington disease
GO:0006413	translational initiation
GO:0006605	protein targeting
GO:0006613	cotranslational protein targeting to membrane
GO:0044270	cellular nitrogen compound catabolic process
GO:0046700	heterocycle catabolic process
GO:0072599	establishment of protein localization to endoplasmic reticulum
GO:0072657	protein localization to membrane

Table 4 Enriched pathways found in Dataset 2 only.

4.2 Medulloblastoma: Metagene $\mathbf{W}_M$

When considered alone, $\mathbf{W}_M$ did not demonstrate any meaningful enrichment in any oncogenic pathways (see Tables 5 - 8). However, by examining the difference vector $\Delta\mathbf{W}_{MR}$, metagene $\mathbf{W}_M$ demonstrates positive NES when compared to metagene $\mathbf{W}_R$ in viral gene expression pathways (GO:0019080, GO:0019083), and negative NES in the synaptic vesicle cycle pathway hsa04721. hsa04721 has been shown to display down-regulation in medulloblastoma patients[53], while viral gene expression of the JCV polyomavirus has been linked with early onset medulloblastomas[54].

4.3 Glioma Metagene: $\mathbf{W}_G$

While $\mathbf{W}_G$ identified no upregulated KEGG pathways in either dataset, the difference vector $\Delta\mathbf{W}_{GN}$ identified the JAK-STAT signalling pathway hsa04630 as being positively enriched in Dataset 2, which has been found to predict poor survival when upregulated in glioma patients[55]. JAK-STAT is known to interact with various cellular adhesion molecules and affects other forms of cancer progression.

In Dataset 1, $\Delta\mathbf{W}_{GN}$ identified upregulation of hsa05169 (Epstein-Barr virus infection), hsa05332 (Graft-versus-host disease), and hsa05416 (Viral myocarditis). Epstein-Barr is a neurotropic virus that is strongly associated with oncogenesis in other cancers, including gliomas[56]. Figures 10 and 11 illustrate the genes shared between the JAK-STAT pathway and various GO:BP pathways responsible for cellular adhesion, as well as an association between genes in Epstein-Barr virus infection and other viral pathways and how these pathways are upregulated in $\mathbf{W}_G$.

4.4 Rhabdoid Metagene: $\mathbf{W}_R$

$\mathbf{W}_R$ identified hsa0494 (type I diabetes mellitus) in Dataset 1 (Table 5), and hsa05012 (Parkinson's disease), hsa00010 (glycolysis/gluconeogenesis) and hsa05020 (Prion disease) as significantly enriched in Dataset 2 (Table 7). Postoperative diabetes insipidus has previously been associated with rhabdoid tumours[57] and in some cases rhabdoid tumours have been found to present as diabetes insipidus[58]. In addition to this, the difference vector $\Delta\mathbf{W}_{GR}$ shows negative enrichment in the hsa05022 neurodegeneration pathway (Table 5) and negative NES in the hsa05016 (Huntington's disease) pathway (Table 7), indicating metagene $\mathbf{W}_R$ shows positive enrichment in these pathways when compared to metagene $\mathbf{W}_G$. Figures 12 and 13 illustrate how these pathways share many of the same genes as in hsa05012 and hsa05020, though it is not clear from this analysis how these pathways contribute to rhabdoid CNS tumours despite being associated with other neurodegenerative diseases. However, Figure 12 shows that these pathways share genes with hsa05208 (chemical carcinogenesis - reactive oxygen species), which is often responsible for oxidative stress induced transcription of oncogenic genes.

$\Delta\mathbf{W}_{RN}$ also identified GO:0046700 (heterocycle catabolism) as being upregulated in metagene $\mathbf{W}_R$ in Dataset 2. Nitrogen-based heterocycles have been shown to be effective components in anti-cancer drugs[59, 60] that are effective against CNS tumours, so upregulating pathways that catabolise these heterocycles could be evidence of heterocyclic resistance in the $\mathbf{W}_R$ metagene. Upregulation of nitrogen compound catabolism is also shown in Figure 13, but this did not appear in the top 5 pathways listed.

5 Discussion

The use of NMF-derived difference vectors demonstrates improved performance in discovering enriched pathways associated with distinct phenotypes, when compared to the use of metagenes as gene lists in a gene set enrichment analysis. These difference vectors exposed significant pathways involved in the development of each class of CNS tumour, where the use of metagenes alone did not rank these pathways as highly. This approach allows one to conduct enrichment analysis with a dual interpretation, gaining insights into the differences between phenotypically similar sample classes, and would therefore be a valuable addition to virtually any enrichment analysis procedure.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

C.L. designed the study, conducted the NMF and GSE analyses and wrote the manuscript. H.Y. prepared the underlying datasets and provided conceptual advice.

Acknowledgements

Many thanks go to the academic staff of the School of Mathematics, Statistics and Applied Mathematics, NUI Galway who were involved in the M.Sc. Genomics programme, for all their support and patience.

References

1. Song, M., Yang, H., Siadat, S.H., Pechenizkiy, M.: A comparative study of dimensionality reduction techniques to enhance trace clustering performances. Expert Systems with Applications **40**(9), 3722–3737 (2013). doi:10.1016/j.eswa.2012.12.078
2. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994). doi:10.1002/env.3170050203
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999). doi:10.1038/44565
4. Liu, W., Yuan, K., Ye, D.: Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. Journal of Biomedical Informatics **41**(4), 602–606 (2008). doi:10.1016/j.jbi.2007.12.003
5. Brunet, J.-P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences **101**(12), 4164–4169 (2004). doi:10.1073/pnas.0308531101
6. Wilson, L. Tyler J.and Lai, Ban, Y., Ge, S.X.: Identification of metagenes and their interactions through large-scale analysis of arabidopsis gene expression data. BMC Genomics **13**(1), 237 (2012). doi:10.1186/1471-2164-13-237
7. Northcott, P.A., Hielscher, T., Dubuc, A., Mack, S., Shih, D., Remke, M., Al-Halabi, H., Albrecht, S., Jabado, N., Eberhart, C.G., Grajkowska, W., Weiss, W.A., Clifford, S.C., Bouffet, E., Rutka, J.T., Korshunov, A., Pfister, S., Taylor, M.D.: Pediatric and adult sonic hedgehog medulloblastomas are clinically and molecularly distinct. Acta Neuropathologica **122**(2), 231–240 (2011). doi:10.1007/s00401-011-0846-7
8. Wang, D., Liu, J.-X., Gao, Y.-L., Yu, J., Zheng, C.-H., Xu, Y.: An nmf-l2,1-norm constraint method for characteristic gene selection. PLOS ONE **11**(7), 1–12 (2016). doi:10.1371/journal.pone.0158494

9. Paatero, P., Hopke, P.K.: Rotational tools for factor analytic models. Journal of Chemometrics **23**(2), 91–100 (2009). doi:10.1002/cem.1197
10. Wang, Y.-X., Zhang, Y.-J.: Nonnegative matrix factorization: A comprehensive review. IEEE Transactions on Knowledge and Data Engineering **25**(6), 1336–1353 (2013). doi:10.1109/TKDE.2012.51
11. Hoyer, P.O.: Non-negative sparse coding. In: Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 557–565 (2002). doi:10.1109/NNSP.2002.1030067
12. Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D., Pascual-Marqui, R.D.: Nonsmooth nonnegative matrix factorization (nsnmf). IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(3), 403–415 (2006). doi:10.1109/TPAMI.2006.60
13. Cichocki, A., Zdunek, R.: Multilayer nonnegative matrix factorization. electronics letters 42, 947-948. Electronics Letters **42**, 947–948 (2006). doi:10.1049/el:20060983
14. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.-I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. Wiley, Chichester (2009). doi:10.1002/9780470747278
15. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. Adv. Neural Inform. Process. Syst. **13** (2001)
16. Cichocki, A., Zdunek, R., Amari, S.-i.: New algorithms for non-negative matrix factorization in applications to blind source separation, vol. 5, p. (2006). doi:10.1109/ICASSP.2006.1661352
17. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004). doi:10.5555/1005332.1044709
18. Hutchins, L.N., Murphy, S.M., Singh, P., Graber, J.H.: Position-dependent motif characterization using non-negative matrix factorization. Bioinformatics **24**(23), 2684–2690 (2008). doi:10.1093/bioinformatics/btn526
19. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning **52**(1), 91–118 (2003). doi:10.1023/A:1023949509487
20. Frigyesi, A., Höglund, M.: Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes. Cancer Informatics **6**, 606 (2008). doi:10.4137/CIN.S606. PMID: 19259414
21. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007). doi:10.1093/bioinformatics/btm134
22. Gaujoux, R., Seoighe, C.: A flexible r package for nonnegative matrix factorization. BMC Bioinformatics **11**(1), 367 (2010). doi:10.1186/1471-2105-11-367
23. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C.: Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics **34**(3), 267–273 (2003). doi:10.1038/ng1180
24. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005). doi:10.1073/pnas.0506580102
25. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., D'Eustachio, P.: The Reactome pathway knowledgebase. Nucleic Acids Research **42**(D1), 472–477 (2013). doi:10.1093/nar/gkt1102
26. Kanehisa, M., Goto, S.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**(1), 27–30 (2000). doi:10.1093/nar/28.1.27
27. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (2000). doi:10.1038/75556
28. Jing, S., Michael G., W.: Gene set enrichment analysis (gsea) for interpreting gene expression profiles. Current Bioinformatics **2**(2), 133–137 (2007). doi:10.2174/157489307780618231
29. Kim, S.-Y., Volsky, D.J.: Page: Parametric analysis of gene set enrichment. BMC Bioinformatics **6**(1), 144 (2005). doi:10.1186/1471-2105-6-144
30. Al-Shahrour, F., Díaz-Uriarte, R., Dopazo, J.: Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics **21**(13), 2988–2993 (2005). doi:10.1093/bioinformatics/bti457
31. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., DeLisi, C.: Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in Bioinformatics **13**(3), 281–291 (2011). doi:10.1093/bib/bbr049
32. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
33. Ackermann, M., Strimmer, K.: A general modular framework for gene set enrichment analysis. BMC bioinformatics **10**, 47–47 (2009). doi:10.1186/1471-2105-10-47
34. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945). doi:10.2307/3001968
35. Efron, B., Tibshirani, R.: On testing the significance of sets of genes. The Annals of Applied Statistics **1**(1), 107–129 (2007). doi:10.1214/07-AOAS101
36. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., Sergushichev, A.: Fast gene set enrichment analysis. bioRxiv (2021). doi:10.1101/060012
37. Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., Park, P.J.: Discovering statistically significant pathways in expression profiling studies. Proceedings of the National Academy of Sciences **102**(38), 13544–13549 (2005). doi:10.1073/pnas.0506577102
38. Shaffer, J.P.: Multiple hypothesis testing. Annual review of psychology **46**(1), 561–584 (1995)
39. Tarca, A.L., Draghici, S., Bhatti, G., Romero, R.: Down-weighting overlapping genes improves gene set analysis. BMC bioinformatics **13**, 136–136 (2012). doi:10.1186/1471-2105-13-136
40. Han, H., Lee, S., Lee, I.: Ngsea: Network-based gene set enrichment analysis for interpreting gene expression phenotypes with functional gene sets. Molecules and cells **42**(8), 579–588 (2019). doi:10.14348/molcells.2019.0065
41. Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., Romero, R.: A systems biology approach for pathway level analysis. Genome research **17**(10), 1537–1545 (2007). doi:10.1101/gr.6202607
42. Gu, Z., Liu, J., Cao, K., Zhang, J., Wang, J.: Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC systems biology **6**, 56–56 (2012). doi:10.1186/1752-0509-6-56
43. Dutta, B., Wallqvist, A., Reifman, J.: Pathnet: a tool for pathway analysis using topological information. Source code for biology and medicine **7**(1), 10–10 (2012). doi:10.1186/1751-0473-7-10
44. Nguyen, T.-M., Shafi, A., Nguyen, T., Draghici, S.: Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biology **20**(1), 203 (2019). doi:10.1186/s13059-019-1790-4
45. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature (2002). doi:10.1038/415436a
46. Affymetrix Support Page. http://www.affymetrix.com/support/help/faqs/mouse_430/faq_8.affx

47. Gaujoux, R., Seoighe, C.: Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. Infection, Genetics and Evolution **12**(5), 913–921 (2012). doi:10.1016/j.meegid.2011.08.014
48. Yang, H., Seoighe, C.: Impact of the choice of normalization method on molecular cancer class discovery using nonnegative matrix factorization. PLoS ONE (2016). doi:10.1371/journal.pone.0164880
49. Merkel, D.: Docker: lightweight linux containers for consistent development and deployment. Linux journal **2014**(239), 2 (2014)
50. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G.: clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 100141 (2021). doi:10.1016/j.xinn.2021.100141
51. Yu, G.: Enrichplot: Visualization of Functional Enrichment Result. (2021). R package version 1.13.1. https://yulab-smu.top/biomedical-knowledge-mining-book/
52. Li, G., Pan, W., Yang, X., Miao, J.: Gene co-expression network and function modules in three types of glioma. Mol Med Rep **11**(4), 3055–3063 (2015). doi:10.3892/mmr.2014.3014
53. Liu, Z., Zhang, R., Sun, Z., Yao, J., Yao, P., Chen, X., Wang, X., Gao, M., Wan, J., Du, Y., Zhao, S.: Identification of hub genes and small-molecule compounds in medulloblastoma by integrated bioinformatic analyses. PeerJ **8**, 8670 (2020). doi:10.7717/peerj.8670
54. Del Valle, L., Gordon, J., Enam, S., Delbue, S., Croul, S., Abraham, S., Radhakrishnan, S., Assimakopoulou, M., Katsetos, C.D., Khalili, K.: Expression of Human Neurotropic Polyomavirus JCV Late Gene Product Agnoprotein in Human Medulloblastoma. JNCI: Journal of the National Cancer Institute **94**(4), 267–273 (2002). doi:10.1093/jnci/94.4.267
55. Tu, Y., Zhong, Y., Fu, J., Cao, Y., Fu, G., Tian, X., Wang, B.: Activation of jak/stat signal pathway predicts poor prognosis of patients with gliomas. Medical Oncology **28**(1) (2011). doi:10.1007/s12032-010-9435-1
56. Akhtar, S., Vranic, S., Cyprian, F.S., Al Moustafa, A.-E.: Epstein-barr virus in gliomas: Cause, association, or artifact? Frontiers in oncology **8**, 123–123 (2018). doi:10.3389/fonc.2018.00123
57. Das, J.M., Abraham, M., Nandeesh, B.N., Nair, S.N.: Pediatric suprasellar atypical teratoid rhabdoid tumor arising from the third ventricle: A rare tumor at a very rare location. Asian journal of neurosurgery **13**(3), 873–876 (2018)
58. Huq, S., Mahalakshmi, H., Ebru, S.: Adult atypical sellar teratoid tumor presenting as diabetes insipidus. Endocrine Practice **24** (2018)
59. Damanpreet, K.L., Rajwinder, K., Rashmi, A., Balraj, S., Sandeep, A.: Nitrogen-containing heterocycles as anticancer agents: An overview. Anti-Cancer Agents in Medicinal Chemistry **20**(18), 2150–2168 (2020). doi:10.2174/1871520620666200705214917
60. Sherer, C., Snape, T.J.: Heterocyclic scaffolds as promising anticancer agents against tumours of the central nervous system: Exploring the scope of indole and carbazole derivatives. European Journal of Medicinal Chemistry **97**, 552–560 (2015). doi:10.1016/j.ejmech.2014.11.007

6 Figures

7 Tables

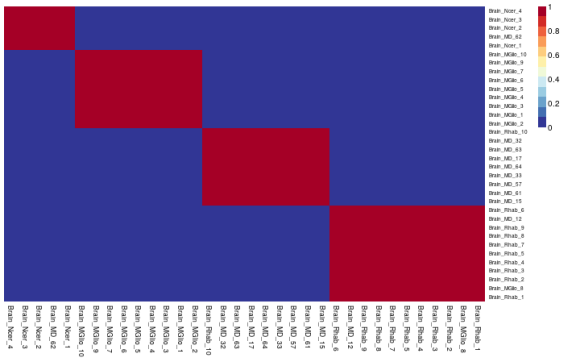

Figure 4 Consensus matrix of Dataset 2 at rank $r = 4$ indicates consistent sample clustering across 50 runs.

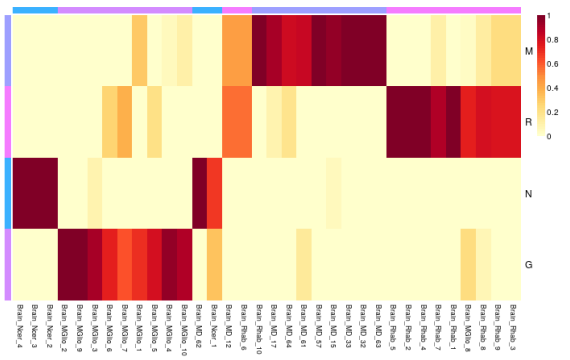

Figure 5 Coefficient matrix of Dataset 2 at rank $r = 4$ indicates sparse phenotypic association across 50 runs.

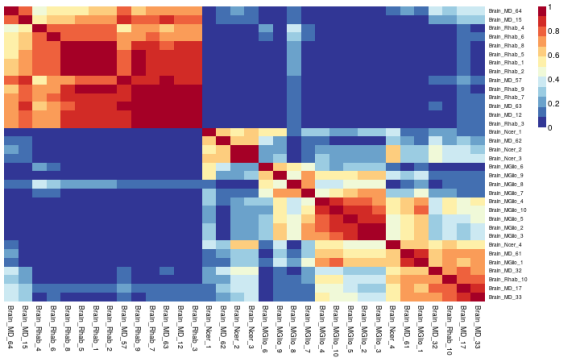

Figure 6 Consensus matrix of Dataset 1-A at rank $r = 4$ indicates inconsistent clustering across 50 runs.

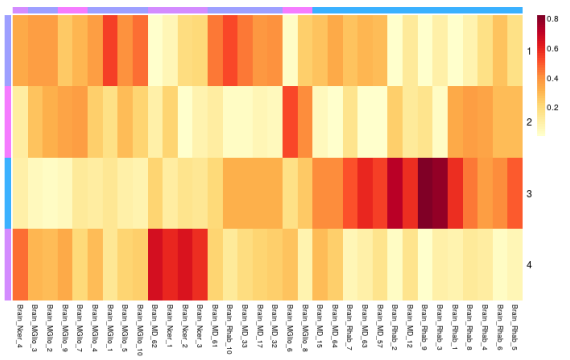

Figure 7 Coefficient matrix of Dataset 1-A at rank $r = 4$ indicates non-sparse phenotypic association across 50 runs.

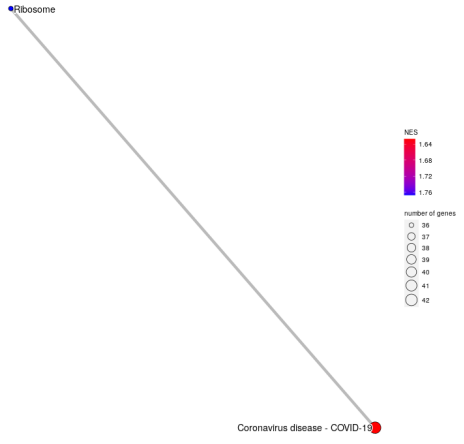

Figure 8 Enrichment map of KEGG pathways enriched in normal metagene for Dataset 1.

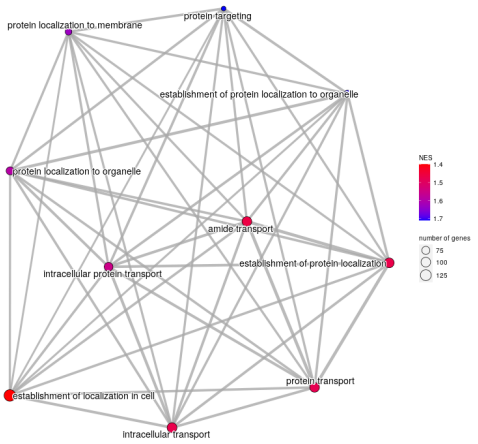

Figure 9 Enrichment map of GO:BP pathways enriched in normal metagene for Dataset 1.

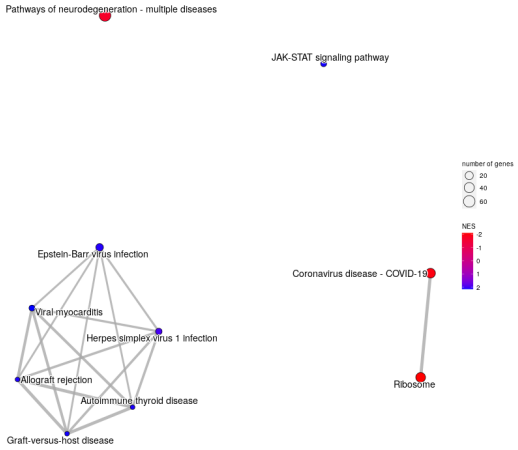

Figure 10 Enrichment map of KEGG pathways enriched in difference vector of metagenes G and N for Dataset 1.

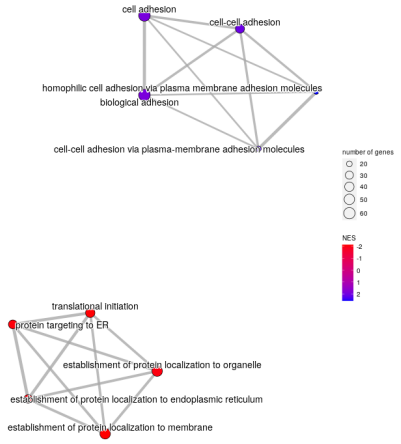

Figure 11 Enrichment map of GO:BP pathways enriched in difference vector of metagenes G and N for Dataset 1.

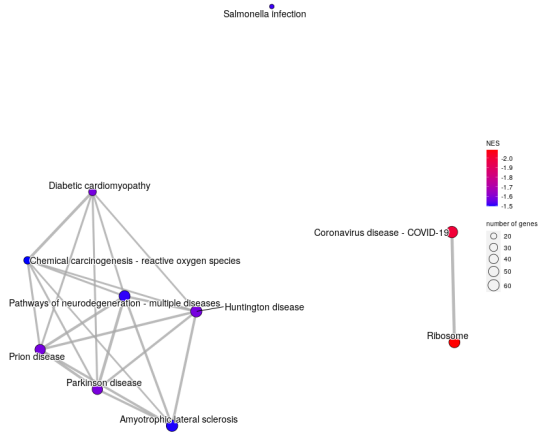

Figure 12 Enrichment map of KEGG pathways enriched in difference vector of metagenes G and R for Dataset 2. Negative NES indicates association with metagene R.

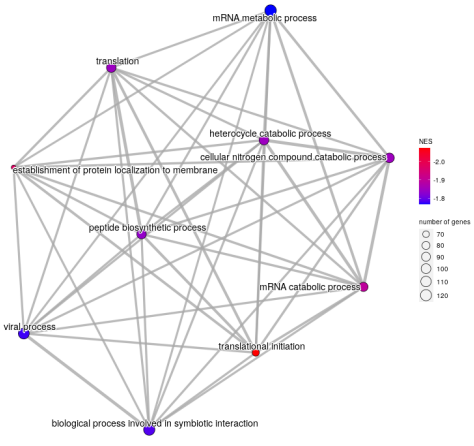

Figure 13 Enrichment map of GO:BP pathways enriched in difference vector of metagenes G and R for Dataset 2. Negative NES indicates association with metagene R.

Gene List	ID	Description	NES	p_{adj}	q
N	hsa03010	Ribosome	1.77	1.88e-05	1.88e-05
N	hsa05171	Coronavirus disease - COVID-19	1.63	1.04e-04	1.04e-04
M	hsa03010	Ribosome	1.75	1.00e-18	1.00e-18
M	hsa05171	Coronavirus disease - COVID-19	1.74	1.00e-18	1.00e-18
R	hsa03010	Ribosome	1.92	1.18e-18	1.08e-18
R	hsa05171	Coronavirus disease - COVID-19	1.85	1.18e-18	1.08e-18
R	hsa04940	Type I diabetes mellitus	1.60	4.99e-02	4.59e-02
GM	hsa03010	Ribosome	-2.22	1.09e-18	1.07e-18
GM	hsa05171	Coronavirus disease - COVID-19	-2.20	1.09e-18	1.07e-18
GR	hsa03010	Ribosome	-2.20	1.30e-18	1.25e-18
GR	hsa05171	Coronavirus disease - COVID-19	-2.16	1.30e-18	1.25e-18
GR	hsa05022	Pathways of neurodegeneration - multiple diseases	-1.46	1.58e-02	1.51e-02
GR	hsa05012	Parkinson disease	-1.50	1.73e-02	1.66e-02
GR	hsa05020	Prion disease	-1.49	2.69e-02	2.58e-02
GN	hsa03010	Ribosome	-2.08	1.55e-04	1.42e-04
GN	hsa05171	Coronavirus disease - COVID-19	-1.96	4.36e-04	3.99e-04
GN	hsa05169	Epstein-Barr virus infection	2.08	5.10e-03	4.67e-03
GN	hsa05332	Graft-versus-host disease	2.12	5.10e-03	4.67e-03
GN	hsa05416	Viral myocarditis	2.13	5.10e-03	4.67e-03
MR	hsa03010	Ribosome	2.46	1.49e-08	1.19e-08
MR	hsa05171	Coronavirus disease - COVID-19	2.39	1.70e-08	1.36e-08
MR	hsa04940	Type I diabetes mellitus	-1.88	1.23e-03	9.88e-04
MR	hsa04650	Natural killer cell mediated cytotoxicity	-1.90	3.03e-03	2.42e-03
MR	hsa04141	Protein processing in endoplasmic reticulum	-1.78	3.81e-03	3.05e-03
MN	hsa03010	Ribosome	2.42	1.23e-18	1.22e-18
MN	hsa05171	Coronavirus disease - COVID-19	2.43	1.23e-18	1.22e-18
MN	hsa04721	Synaptic vesicle cycle	-1.90	2.36e-02	2.33e-02
RN	hsa05171	Coronavirus disease - COVID-19	2.20	2.72e-18	2.39e-18
RN	hsa03010	Ribosome	2.27	3.65e-18	3.21e-18

Table 5 Top 5 KEGG pathways per metagene/difference vector, identified by GSEA for Dataset 1.

Gene List	ID	Description	NES	p_{adj}	q
N	GO:0072594	establishment of protein localization to organelle	1.71	1.37e-08	1.29e-08
N	GO:0033365	protein localization to organelle	1.59	1.05e-07	9.91e-08
N	GO:0006886	intracellular protein transport	1.55	1.58e-07	1.49e-07
N	GO:0051649	establishment of localization in cell	1.40	1.58e-07	1.49e-07
N	GO:0015031	protein transport	1.48	1.58e-07	1.49e-07
G	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	2.14	1.20e-17	1.17e-17
G	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	2.10	1.20e-17	1.17e-17
G	GO:0098609	cell-cell adhesion	1.57	2.99e-05	2.90e-05
G	GO:0022610	biological adhesion	1.38	1.93e-02	1.87e-02
G	GO:0007155	cell adhesion	1.38	1.93e-02	1.87e-02
M	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.72	5.98e-19	5.98e-19
M	GO:0000956	nuclear-transcribed mRNA catabolic process	1.71	5.98e-19	5.98e-19
M	GO:0006401	RNA catabolic process	1.63	5.98e-19	5.98e-19
M	GO:0006402	mRNA catabolic process	1.63	5.98e-19	5.98e-19
M	GO:0006412	translation	1.61	5.98e-19	5.98e-19
R	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.88	1.08e-18	1.01e-18
R	GO:0000956	nuclear-transcribed mRNA catabolic process	1.85	1.08e-18	1.01e-18
R	GO:0006401	RNA catabolic process	1.64	1.08e-18	1.01e-18
R	GO:0006402	mRNA catabolic process	1.65	1.08e-18	1.01e-18
R	GO:0006412	translation	1.66	1.08e-18	1.01e-18
GM	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	-2.17	5.22e-19	4.92e-19
GM	GO:0000956	nuclear-transcribed mRNA catabolic process	-2.19	5.22e-19	4.92e-19
GM	GO:0006401	RNA catabolic process	-2.11	5.22e-19	4.92e-19
GM	GO:0006402	mRNA catabolic process	-2.11	5.22e-19	4.92e-19
GM	GO:0006412	translation	-2.12	5.22e-19	4.92e-19
GR	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	-2.18	1.14e-18	1.07e-18
GR	GO:0000956	nuclear-transcribed mRNA catabolic process	-2.17	1.14e-18	1.07e-18
GR	GO:0006401	RNA catabolic process	-1.95	1.14e-18	1.07e-18
GR	GO:0006402	mRNA catabolic process	-1.96	1.14e-18	1.07e-18
GR	GO:0006412	translation	-1.91	1.14e-18	1.07e-18
GN	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	2.58	3.07e-08	2.69e-08
GN	GO:0090150	establishment of protein localization to membrane	-2.13	2.77e-06	2.42e-06
GN	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	2.34	1.57e-05	1.38e-05
GN	GO:0072594	establishment of protein localization to organelle	-2.02	1.76e-05	1.55e-05
GN	GO:0022610	biological adhesion	1.90	2.04e-05	1.79e-05
MR	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	2.63	1.44e-08	1.13e-08
MR	GO:0019080	viral gene expression	2.50	1.44e-08	1.13e-08
MR	GO:0006612	protein targeting to membrane	2.54	1.44e-08	1.13e-08
MR	GO:0000956	nuclear-transcribed mRNA catabolic process	2.53	1.44e-08	1.13e-08
MR	GO:0019083	viral transcription	2.50	1.44e-08	1.13e-08
MN	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.41	7.38e-19	6.93e-19
MN	GO:0000956	nuclear-transcribed mRNA catabolic process	2.42	7.38e-19	6.93e-19
MN	GO:0006401	RNA catabolic process	2.36	7.38e-19	6.93e-19
MN	GO:0006402	mRNA catabolic process	2.37	7.38e-19	6.93e-19
MN	GO:0006412	translation	2.38	7.38e-19	6.93e-19
RN	GO:0000956	nuclear-transcribed mRNA catabolic process	2.23	5.74e-18	5.08e-18
RN	GO:0010629	negative regulation of gene expression	1.90	5.74e-18	5.08e-18
RN	GO:0016032	viral process	1.96	5.74e-18	5.08e-18
RN	GO:0043043	peptide biosynthetic process	2.04	5.74e-18	5.08e-18
RN	GO:0044403	biological process involved in symbiotic interaction	1.97	5.74e-18	5.08e-18

Table 6 Top 5 GO:BP pathways per metagene/difference vector, identified by GSEA for Dataset 1.

Gene List	ID	Description	NES	p_{adj}	q
N	hsa03010	Ribosome	1.79	1.04e-05	1.04e-05
N	hsa05171	Coronavirus disease - COVID-19	1.63	8.46e-05	8.46e-05
M	hsa03010	Ribosome	1.59	4.95e-19	4.95e-19
M	hsa05171	Coronavirus disease - COVID-19	1.59	4.95e-19	4.95e-19
R	hsa03010	Ribosome	1.89	1.17e-18	1.05e-18
R	hsa05171	Coronavirus disease - COVID-19	1.80	1.17e-18	1.05e-18
R	hsa05012	Parkinson disease	1.35	1.85e-02	1.65e-02
R	hsa00010	Glycolysis / Gluconeogenesis	1.58	1.85e-02	1.65e-02
R	hsa05020	Prion disease	1.34	1.85e-02	1.65e-02
GM	hsa03010	Ribosome	-2.42	1.09e-18	1.09e-18
GM	hsa05171	Coronavirus disease - COVID-19	-2.43	1.09e-18	1.09e-18
GR	hsa03010	Ribosome	-2.09	1.70e-16	1.57e-16
GR	hsa05171	Coronavirus disease - COVID-19	-2.00	2.39e-15	2.20e-15
GR	hsa05012	Parkinson disease	-1.58	1.56e-03	1.44e-03
GR	hsa05016	Huntington disease	-1.58	1.56e-03	1.44e-03
GR	hsa05020	Prion disease	-1.57	1.56e-03	1.44e-03
GN	hsa03010	Ribosome	-2.18	3.22e-05	2.96e-05
GN	hsa05171	Coronavirus disease - COVID-19	-2.05	1.87e-04	1.71e-04
GN	hsa04630	JAK-STAT signaling pathway	2.08	1.18e-02	1.08e-02
MR	hsa03010	Ribosome	3.04	1.23e-18	9.58e-19
MR	hsa05171	Coronavirus disease - COVID-19	2.83	1.23e-18	9.58e-19
MR	hsa05014	Amyotrophic lateral sclerosis	-1.70	1.09e-05	8.52e-06
MR	hsa05020	Prion disease	-1.73	3.53e-05	2.75e-05
MR	hsa05016	Huntington disease	-1.64	3.32e-04	2.59e-04
MN	hsa03010	Ribosome	2.70	1.21e-18	1.20e-18
MN	hsa05171	Coronavirus disease - COVID-19	2.80	1.21e-18	1.20e-18
RN	hsa03010	Ribosome	2.00	1.01e-08	8.81e-09
RN	hsa05171	Coronavirus disease - COVID-19	1.93	1.01e-08	8.81e-09
RN	hsa05014	Amyotrophic lateral sclerosis	1.63	2.41e-03	2.10e-03
RN	hsa05012	Parkinson disease	1.62	4.34e-03	3.79e-03
RN	hsa05020	Prion disease	1.59	7.09e-03	6.19e-03

Table 7 Top 5 KEGG pathways per metagene/difference vector, identified by GSEA for Dataset 2.

Gene List	ID	Description	NES	p_{adj}	q
N	GO:0072594	establishment of protein localization to organelle	1.73	1.27e-07	1.18e-07
N	GO:0072657	protein localization to membrane	1.66	4.61e-07	4.29e-07
N	GO:0033365	protein localization to organelle	1.60	4.61e-07	4.29e-07
N	GO:0006886	intracellular protein transport	1.55	6.64e-07	6.18e-07
N	GO:0090150	establishment of protein localization to membrane	1.74	7.49e-07	6.97e-07
G	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	2.14	1.97e-17	1.91e-17
G	GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	2.11	1.97e-17	1.91e-17
G	GO:0098609	cell-cell adhesion	1.58	1.44e-05	1.40e-05
G	GO:0022610	biological adhesion	1.39	9.45e-03	9.18e-03
G	GO:0007155	cell adhesion	1.39	1.20e-02	1.16e-02
M	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.60	2.83e-19	2.73e-19
M	GO:0000956	nuclear-transcribed mRNA catabolic process	1.59	2.83e-19	2.73e-19
M	GO:0006401	RNA catabolic process	1.56	2.83e-19	2.73e-19
M	GO:0006402	mRNA catabolic process	1.56	2.83e-19	2.73e-19
M	GO:0006412	translation	1.53	2.83e-19	2.73e-19
R	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.85	1.85e-18	1.63e-18
R	GO:0000956	nuclear-transcribed mRNA catabolic process	1.81	1.85e-18	1.63e-18
R	GO:0006413	translational initiation	1.83	1.85e-18	1.63e-18
R	GO:0006605	protein targeting	1.71	1.85e-18	1.63e-18
R	GO:0006612	protein targeting to membrane	1.84	1.85e-18	1.63e-18
GM	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	-2.46	5.16e-19	4.70e-19
GM	GO:0000956	nuclear-transcribed mRNA catabolic process	-2.47	5.16e-19	4.70e-19
GM	GO:0006401	RNA catabolic process	-2.45	5.16e-19	4.70e-19
GM	GO:0006402	mRNA catabolic process	-2.46	5.16e-19	4.70e-19
GM	GO:0006412	translation	-2.39	5.16e-19	4.70e-19
GR	GO:0006413	translational initiation	-2.07	1.04e-17	9.27e-18
GR	GO:0016032	viral process	-1.78	1.04e-17	9.27e-18
GR	GO:0044403	biological process involved in symbiotic interaction	-1.79	1.04e-17	9.27e-18
GR	GO:0090150	establishment of protein localization to membrane	-1.97	4.83e-17	4.31e-17
GR	GO:0044270	cellular nitrogen compound catabolic process	-1.85	5.64e-17	5.04e-17
GN	GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	2.52	9.68e-09	8.80e-09
GN	GO:0090150	establishment of protein localization to membrane	-2.23	2.09e-06	1.90e-06
GN	GO:0007155	cell adhesion	1.96	2.96e-06	2.69e-06
GN	GO:0072594	establishment of protein localization to organelle	-2.08	2.96e-06	2.69e-06
GN	GO:0006413	translational initiation	-2.22	3.05e-06	2.77e-06
MR	GO:0006613	cotranslational protein targeting to membrane	3.12	6.04e-18	4.31e-18
MR	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	3.12	6.04e-18	4.31e-18
MR	GO:0019083	viral transcription	3.01	6.04e-18	4.31e-18
MR	GO:0072599	establishment of protein localization to endoplasmic reticulum	3.13	6.04e-18	4.31e-18
MR	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3.11	6.04e-18	4.31e-18
MN	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.75	6.76e-19	6.10e-19
MN	GO:0000956	nuclear-transcribed mRNA catabolic process	2.77	6.76e-19	6.10e-19
MN	GO:0006401	RNA catabolic process	2.72	6.76e-19	6.10e-19
MN	GO:0006402	mRNA catabolic process	2.73	6.76e-19	6.10e-19
MN	GO:0006412	translation	2.72	6.76e-19	6.10e-19
RN	GO:0044403	biological process involved in symbiotic interaction	1.84	3.33e-17	2.80e-17
RN	GO:0016032	viral process	1.82	1.06e-15	8.93e-16
RN	GO:0010629	negative regulation of gene expression	1.79	2.54e-14	2.13e-14
RN	GO:0006412	translation	1.88	2.28e-12	1.92e-12
RN	GO:0046700	heterocycle catabolic process	1.88	2.28e-12	1.92e-12

Table 8 Top 5 GO:BP pathways per metagene/difference vector, identified by GSEA for Dataset 2.