Final Report

Paul Brunner, Marie Kleinert, Felipe Stünkel, Chloé Weiler

20/07/2022

Abstract

Abbreviations

GSEA	gene set enrichment analysis
GSVA	gene set variation analysis
LUAD	lung adenocarcinoma
PC	principal component
PCA	principal component analysis
UMAP	uniform manifold approximation and projection

Introduction

Methods

Overview of used packages

Table 2: **Tab. 1: Used packages in alphabetical order.**

Package Name	Application	**Reference**
babyplots	create interactive 3D visualizations	Trost (2022)
base	basic R functions	R Core Team (2022a)
bayesbio	calculate Jaccard coefficients	McKenzie (2016)
BiocParallel	novel implementations of functions for parallel evaluation	(Morgan et al., 2021)
biomaRt	access to genome databases	Durinck et al. (2009)
cinaR	combination of different packages	Karakaslar and Ucar (2022)
cluster	cluster analysis of data	Maechler et al. (2021)

Package Name	Application	**Reference**
ComplexHeatmap	arrange multiple heatmaps	(Gu et al., 2016)
edgeR	assess differential expression in gene expression profiles	Chen et al. (2016)
EnhancedVolcano	produce improved volcano plots	(Blighe et al., 2021)
enrichplot	visualization of gene set enrichment results (GSEA)	Yu (2022)
FactoMineR	perform principal component analysis (PCA)	Lê et al. (2008)
fgsea	Run GSEA on a pre-ranked list	Korotkevich et al. (2019)
ggplot2	visualization of results in dot plots, bar plots and box plots	Wickham (2016)
ggpubr	formatting of ggplot2-based graphs	Kassambara (2020)
grid	implements the primitive graphical functions that underlie the ggplot2 plotting system	R Core Team (2022b)
gridExtra	arrange multiple plots on a page	Auguie (2017)
GSVA	Run GSVA on a data set	(**GSVA?**)
gplots	plotting data	(Warnes et al., 2022a)
gtools	calculate foldchange, find NAs, logratio2foldchange	Warnes et al. (2022b)
knitr	creation of citations using write_bib	Xie (2014)
limma	"linear models for microarray data"	Ritchie et al. (2015)
msigdbr	provides the 'Molecular Signatures Database' (MSigDB) gene sets	Dolgalev (2022)
parallel	allows for parallel computation through multi core processing	R Core Team (2022c)
pheatmap	draw clustered heatmaps	Kolde (2019)
RColorBrewer	provides color schemes for maps	Neuwirth (2022)
scales	helps in visualization: r automatically determines breaks and labels for axes and legends	Wickham and Seidel (2022)
Seurat	visualize gene set enrichment results in dot plots	Satija et al. (2022)
tidyverse	collection of R packages, including ggplot2	Wickham et al. (2019)
uwot	performs dimensionality reduction and Uniform Manifold Approxiamtion and Projection (UMAP)	Melville (2021)

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a procedure used to perform linear dimension reduction. The goal is to reduce the dimension of a given data set whilst losing as little information as possible by retaining a maximum of the standardized data set's variation (Ringnér, 2008).

Principal components (PC) are a set of new orthogonal variables that are made up of a linear combination of the original variables. Principal components display the pattern of similarity of the observations and of the variables as points in maps (Abdi and Williams, 2010). By convention, the PCs are ordered in decreasing order according to the amount of variation they explain of the original data (Ringnér, 2008). It is important to note that all PCs are uncorrelated.

PCA is a useful tool for genome-wide expression studies and often serves as a first step before clustering or classification of the data. Dimension reduction is a necessary step for easy data exploration and visualization (Ringnér, 2008).

Uniform Manifold Approximation and Projection (UMAP)

Uniform manifold approximation and projection (UMAP) is a k-neighbour graph based algorithm that is used for nonlinear dimension reduction (Smets et al., 2019) (McInnes et al., 2018).

After data normalization, the Euclidean distances between points in a two-dimensional space of the graph are calculated and a local radius is determined (Vermeulen et al., 2021). In general the closer two points are to each other, the more similar they are. UMAP makes a density estimation to find the right local radius. This variable radius is smaller in high density regions of data points and larger in low density regions. In general, the density is higher when the k-nearest neighbour is close and vice versa. The number of k-nearest neighbours controls the number of neighbours whose local topology is preserved. Precisely, a large number of neighbours will ensure that more global structure is preserved whereas a smaller number of neighbours will ensure the preservation of more local structure (McInnes et al., 2018).

UMAP is a newer method tan PCA and it is generally believed to be easier to interpret and group data than by using PCA. Furthmore, UMAP has the advantage of not requiring linear data (Milošević et al., 2022).

Gene Set Enrichment Analysis (GSEA)

Gene set enrichment analysis (GSEA) is a computational method that is used to determine whether two gene expression states are significantly different from each other or not. In this project we compared gene expression profiles between healthy and tumorous tissue of LUAD(Subramanian et al., 2007).

Two data sets are compared and the genes are sorted from the most to the least differential expression between the data sets according to their p-values. This creates a ranked list L.

Referring to an *a priori* defined set of gene sets S, the goal is to locate for each pathway of S where its corresponding genes fall in L and find a discerning trend. If the genes of a given pathway are randomly distributed in L then the pathway is assumed to not significantly contribute to the particular tumor's phenotype. However if the genes are primarily clustered at the top or the bottom of L then a phenotypic significance of the given pathway can be assumed.

To determine the location of the genes, an enrichment score is calculated for each pathway. For this, a running-sum statistic is calculated as the list L is ran through. The running-sum is increased every time a gene belonging to the pathway in question is encountered and decreased otherwise. An enrichment score is thus calculated for each pathway. The enrichment score is defined as the maximum deviation from zero of the running-sum.

Lastly, adjustment for multiple hypothesis testing is performed by normalizing the enrichment score for each pathway to account for its size and a normalized enrichment score is obtained.

GSEA is a useful tool for interpretation of gene expression data.

(Subramanian et al., 2005)

Gene Set Variation Analysis (GSVA)

Gene set variation analysis (GSVA) is an unsupervised method to estimate pathway activities based on gene expression data. Contrarily to the aforementioned GSEA, GSVA does not rely on phenotypic characterisation of the data sets into two categories but rather quantifies enrichment in a sample-wise manner which makes GSVA the better choice to perform on the tcga_exp data set.

GSVA estimates a cumulative distribution for each gene over all samples. The gene expression values are then converted according to these estimated cumulative distributions into scaled values. Based on these new values, the genes are ranked in each sample. Next, the genes are classified into two distributions and a Komogorow-Smirnow statistic is calculated to judge how similar the distributions are to each other and to obtain an ES.

The GSVA corresponds to either the maximum deviation between both running sums or the GSVA score can be defined as the difference of the maximum deviations in the positive and in the negative direction. A highly positive or negative GSVA score indicates that the studied gene set is positively or negatively enriched compared to the genes not in the gene set, respectively. If the GSVA score for a given gene set is close to zero, then the gene set is probably not differentially expressed compared to the genes not in the gene set.

(Hänzelmann et al., 2013)

Results

Conclusion

References

Abdi, H., and Williams, L.J. (2010). Principal component analysis. WIREs Computational Statistics *2*, 433–459.

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics.

Blighe, K., Rana, S., and Lewis, M. (2021). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

Chen, Y., Lun, A.A.T., and Smyth, G.K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-seq experiments using rsubread and the edgeR quasi-likelihood pipeline. F1000Research *5*, 1438.

Dolgalev, I. (2022). Msigdbr: MSigDB gene sets for multiple organisms in a tidy data format.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. Nature Protocols *4*, 1184–1191.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 1–15.

Karakaslar, O., and Ucar, D. (2022). cinaR: A computational pipeline for bulk 'ATAC-seq' profiles.

Kassambara, A. (2020). Ggpubr: 'ggplot2' based publication ready plots.

Kolde, R. (2019). Pheatmap: Pretty heatmaps.

Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. bioRxiv.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. Journal of Statistical Software *25*, 1–18.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). Cluster: Cluster analysis basics and extensions.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction (arXiv).

McKenzie, A. (2016). Bayesbio: Miscellaneous functions for bioinformatics and bayesian statistics.

Melville, J. (2021). Uwot: The uniform manifold approximation and projection (UMAP) method for dimensionality reduction.

Milošević, D., Medeiros, A.S., Stojković Piperac, M., Cvijanović, D., Soininen, J., Milosavljević, A., and Predić, B. (2022). The application of uniform manifold approximation and projection (UMAP) for unconstrained ordination and classification of biological indicators in aquatic ecology. Science of The Total Environment *815*, 152365.

Morgan, M., Wang, J., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2021). BiocParallel: Bioconductor facilities for parallel evaluation.
Neuwirth, E. (2022). RColorBrewer: ColorBrewer palettes.
R Core Team (2022a). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
R Core Team (2022b). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
R Core Team (2022c). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
Ringnér, M. (2008). What is principal component analysis? Nature Biotechnology *26*, 303–304.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research *43*, e47.
Satija, R., Butler, A., Hoffman, P., and Stuart, T. (2022). SeuratObject: Data structures for single cell data.
Smets, T., Verbeeck, N., Claesen, M., Asperger, A., Griffioen, G., Tousseyn, T., Waelput, W., Waelkens, E., and De Moor, B. (2019). Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. Analytical Chemistry *91*, 5706–5714.
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., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550.
Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J.P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics *23*, 3251–3253.
Trost, N. (2022). Babyplots: Easy, fast, interactive 3D visualizations for data exploration and presentation.
Vermeulen, M., Smith, K., Eremin, K., Rayner, G., and Walton, M. (2021). Application of uniform manifold approximation and projection (UMAP) in spectral imaging of artworks. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy *252*, 119547.
Warnes, G.R., Bolker, B., and Lumley, T. (2022b). Gtools: Various r programming tools.
Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2022a). Gplots: Various r programming tools for plotting data.
Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (Springer-Verlag New York).
Wickham, H., and Seidel, D. (2022). Scales: Scale functions for visualization.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. Journal of Open Source Software *4*, 1686.
Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In Implementing Reproducible Computational Research, V. Stodden, F. Leisch, and R.D. Peng, eds. (Chapman; Hall/CRC),.
Yu, G. (2022). Enrichplot: Visualization of functional enrichment result.