hdmax2, an R package to perform high dimension mediation analysis

Florence Pittion, Basile Jumentier, Aurélie Nakamura, Johanna Lepeule, Olivier Francois, Magali Richard

hdmax2, an R package to perform high dimension mediation analysis

Florence Pittion[1], Basile Jumentier[2], Aurélie Nakamura[3], Johanna Lepeule[3], Olivier Francois [#, 1], and Magali Richard [#, 1]

DOI not yet assigned

Abstract

Mediation analysis plays a crucial role in epidemiology, unraveling the intricate pathways through which exposures exert influence on health outcomes. Recent advances in high-throughput sequencing techniques have generated growing interest in applying mediation analysis to explore the causal relationships between patient environmental exposures, molecular features (such as omics data) and various health outcomes. Mediation analysis handling high-dimensional mediators raise a number of statistical challenges. Despite the emergence of numerous methods designed to tackle these challenges, the majority are limited to continuous outcomes. Furthermore, these advanced statistical approaches have yet to find widespread adoption among epidemiologists and health data scientists in their day-to-day practices. To address this gap, we introduce an R package specifically tailored for high-dimensional mediation analysis using the max-squared method (HDMAX2). This tool aims to mitigate these obstacles by providing a practical solution for researchers and practitioners eager to explore intricate causal relationships in health data involving complex molecular features. Here we improve the HDMAX2 method, and expand its capabilities to accommodate multiple exposures and non-continuous variables. This improvement enables its application to a diverse array of mediation analysis scenarios, mirroring the complexity often encountered in healthcare data. To enhance accessibility for users with varying expertise, we release an R package called `hdmax2`. This package allows users to estimate the indirect effects of mediators, calculate the overall indirect effect of mediators, and facilitates the execution of high-dimensional mediation analysis.

Keywords: High-dimension; Mediation; Multivariate analysis; Confounding effect; Causal analysis

[1]Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France, [2]Research Center of the Sainte-Justine University Hospital, University of Montreal, Montreal, Quebec, Canada , [3]Université Grenoble-Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Development and Respiratory Health, Institute for Advanced Biosciences, Grenoble, France, [#]Equal contribution

Correspondence
olivier.francois@univ-grenoble-alpes.fr, magali.richard@univ-grenoble-alpes.fr

Introduction

When a statistical association is observed between an external exposure (X) and an individual outcome (Y), one or more intermediate variables (M) (such as gene expression or epigenetic changes) may mediate this effect. Identifying and assessing the mediating role of these variables in the effect of X on Y is crucial for deciphering underlying causal mechanisms in epidemiological and clinical research. This process, known as mediation analysis, involves studying mediator variables to define the causal structure between X and Y. The mediated effect, termed the indirect effect, is equal to the effect of X on Y mediated through M, to distinguish from the direct effect of X on Y unexplained by M (Richiardi et al., 2013). Historically, mediation analysis has predominantly focused on single mediation (i.e. one intermediary variable) (Baron and Kenny, 1986), running separate statistical tests for the effects of X on M and M on Y, followed by estimation of the indirect effect (Imai et al., 2010; Sobel, 1982). However, in the realm of high-dimensional molecular data (M, e.g., omic data), extending mediation analysis to high dimensions poses challenges, including correction for multiple testing, controlling the false discovery rate (FDR), addressing reverse causation, adjusting for confounding effects, considering interactions among mediators, and integrating multimodal data types (Blum et al., 2020; Zeng et al., 2021). Currently, there remains no consensus on the optimal combination of models and methodologies for high-dimensional mediation analysis. With the increasingly prevalent use of next-generation sequencing technologies, there is now an urgent need to develop high-performing methods in high dimensionality and make them accessible. While recent methods has partially addressed these needs (Dai et al., 2022; Djordjilović et al., 2022; Sampson et al., 2018; Zhang et al., 2016), particularly in controlling the type I error in high dimension, they do not account for unmeasured confounding factors and do not allow for the consideration of multiple exposures.

In this article, we introduce an R package called `hdmax2`. The HDMAX2 method was originally proposed by Jumentier et al., 2023. The fundamental concept behind HDMAX2 methods is to use a latent factor mixed regression model for estimating unobserved latent factors while conducting high-dimensional association analysis. HDMAX2 also implements a novel procedure known as the max-squared test to assess the satistical significance of potential mediators (see Materials and Methods section). Finally, HDMAX2 enables the calculation of an overall indirect effect from a single model that includes all mediators simultaneously. This approach therefore takes into account correlations between mediators. A case study investigating the impact of maternal smoking on birth weight identified epigenetic regions mediating the indirect effect of this exposure (Jumentier et al., 2023). In this article, we introduce an enhanced version of the HDMAX2 method, expanding its functionalities and making the method accessible to practitioners through a packaged format. The `hdmax2` program has been enriched with numerous features, including the ability to manage various types of variables in the exposure (continuous, binary, categorical, and multivariate), as well as the capability to incorporate binary outcomes. This package enables users to (i) investigate associations between the variables X, M, and Y, (ii) compute the mediated effect for each potential mediator, (iii) assess the overall indirect effect for the total model, and (iv) visualize these results. A graphical model of the package is presented in Fig 1. The package is open-source and accessible on our GitHub page at https://github.com/bcm-uga/hdmax2.

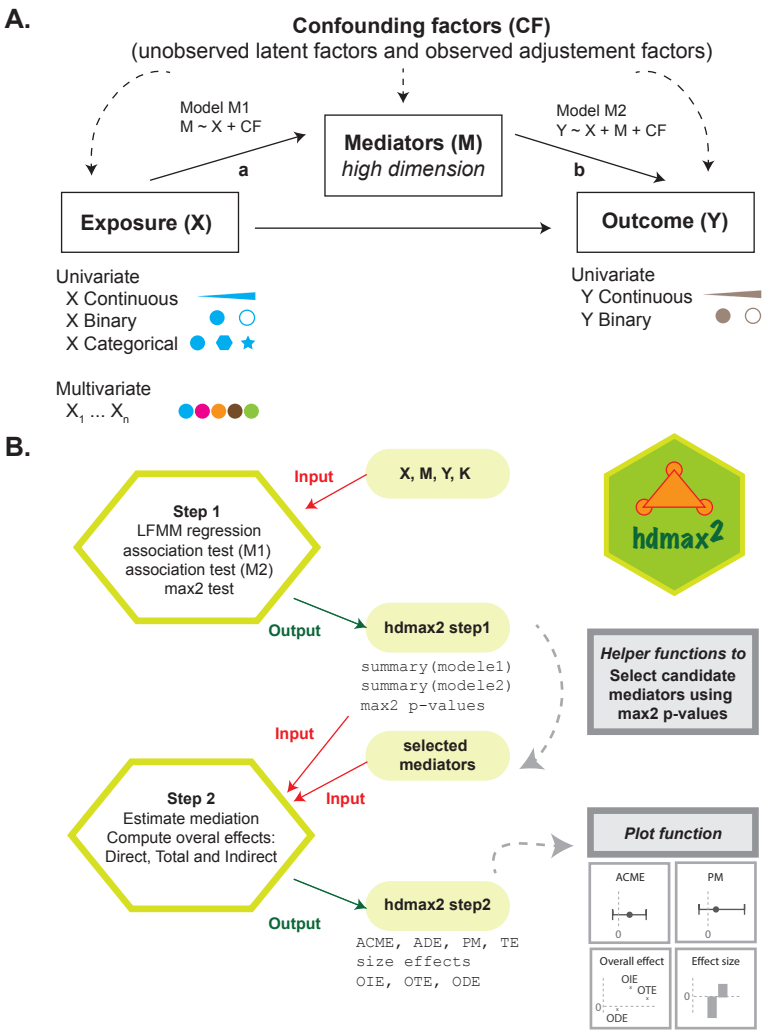

Figure 1 – Graphical model of the `hdmax2` R package. A. Acyclic graph of HDMAX2 method for conducting high-dimension mediation analysis. Exposure variables (X) can be univariate of multivariate, the colors represent different types of variable, and the symbols schematically represent the modalities of the same variable. Intermediary variables (M) are continuous variables. The outcome variable (Y) can be binary or continuous. Confounding factors represent observed adjustment factors and unobserved latent factors estimated by LFMM regression. The values a and b represent the effect sizes for each regression of the mediation model. **B.** Core structure of the `hdmax2` package. Step 1: First, latent factors are estimated by LFMM multivariate regression. Then, the association of X and M and of M and Y are statistically tested in simple regression models and the significance values obtained are combined, for each potential mediator. Step 2: the indirect effects of the mediator variables are estimated, along with various causal measures of interest: ACME (Average Causal Mediated Effect), ADE (Average Direct Effect), PM (Proportion Mediated), TE (Total Effect), size effects, OIE (Overall Indirect Effect), OTE (Overall Total Effect), ODE (Overall Direct Effect).

Materials and methods

The package comprises a set of core functions along with a visualization function. Its usage is guided through several vignettes and test datasets. The statistical methodologies embedded within the package are described below, with the initial application case documented by Jumentier et al., 2023.

Input data. The `hdmax2` package is designed to accept one or several exposure variables, X, which can be continuous, binary, or categorical. The user must provide an exposure *data.frame* having at least one column as input. We define N as the number of samples within the tested cohort. Therefore, each exposure and outcome variable consists of N measurements. In the `R` language, categorical variables are encoded as factor objects. The function `as.factor()` can be used to encode categorical variables. The functions `levels()` and `ordered()` can be used to define the order of the modalities of categorical variables. By convention, `hdmax2` uses the first modality as a reference to calculate the effects associated with the other modalities of the variable, as encoded by default in `lm()` function in `R`. Continuous intermediary variables, denoted as M, are represented as a matrix encompassing potential mediators, such as methylome or transcriptome molecular features. The matrix M should be entered as a separated input, without missing values. The intermediary variable matrix M is of dimension $J \times N$, with J the total number of intermediate variables. The outcome variable, denoted as Y, corresponds to a vector of dimension N, which supports both continuous and binary formats. Continuous and binary variables must be encoded in numeric format. Optional covariates, Z, can be included as observed adjustment factors in the model. The package `hdmax2` also takes as input the number of latent factors to be estimated, K.

Step 1: Identification of potential mediators. The function `hdmax2::run_AS()` evaluates the association between exposure variables, intermediary variables and the outcome variable using a latent factor mixed model (LFMM (Caye et al., 2019)) to estimate K unobserved latent factors U. First this function tests the significance of association between the exposure variables and a the potential mediator variables. Then it tests association between the potential mediator variables and the outcome variable. Finally it evaluates the significance of the indirect effects.

For univariate (continuous, binary and categorical) exposure, a significance value $P_{1,j}$ is computed by the `hdmax2` program for the test of a null effect size for exposure variable X on intermediary variable M_j, for each j (with j ranging from 1 to J, the total number of intermediary variables). For multivariate exposure, the `hdmax2` program applies partial regression models, and returns a single P-value, $P_{1,j}$, for the test of a null effect size of the full exposure variable X on the intermediary variable M_j, for each j (see Fig 1.A, regression 1). Then, the `hdmax2` program returns a significance value $P_{2,j}$ corresponding to the association of each intermediary variable M_j with the outcome variable Y (see Fig 1.A, regression 2).

Finally `hdmax2` identifies potential mediators by combining the significance values P_1 and P_2 to compute a P-value for each intermediary variable using the max-squared ($\max^2$) test (equation (1)). This test rejects the null hypothesis of no effect of exposure on potential mediators or no effect of potential mediators on the outcome.

$$P = \max(P_1, P_2)^2. \quad (1)$$

The `hdmax2::run_AS()` function returns an object of class *hdmax2_step1*, including the following attributes:

- a vector of J mediation P-values, resulting of the max-squared test,

- U, score matrix of size $N \times K$, corresponding to the K unobserved latent factors calculated from an LFMM regression (model M1),

- the input variables of the model: exposure variables X, the outcome variable Y, and adjustment factors Z (when applicable).

Step 2: Estimation of indirect effects. The function `hdmax2::estimate_effect()` takes as input an object *hdmax2_step1* and a list of potential mediators M^S to be analyzed in subsequent steps. The subset M^S is defined by the user based on the output of *hdmax2_step1*. We provide a series of helper functions to guide the user in the selection of M^S. These functions include a False Discovery Rate (FDR) control approach, or the possibility to aggregate mediator regions according to their location on the genome. Aggregating mediator regions allows, for instance, the inference of differentially methylated regions (DMRs) when working with DNA methylation data. Illustrations of these approaches can be found in Jumentier et al., 2023 and in the `hdmax2` vignettes.

For each univariate exposure variable and the subset of mediators M^S, the `hdmax2::estimate_effect()` function computes several estimates to evaluate the indirect effects in the path between exposure variables and the outcome variable. Initially, this function assesses each mediator variable M_j^S individually and computes causal measures of interest such as (i) the Average Causal Mediated Effect (ACME, corresponding to the indirect effect) and (ii) the Proportion Mediated (PM). The ACME differs from the Average Direct Effect (ADE), which represents the unmediated effect, and from the Total Effect (TE) which is equal to the sum of direct and indirect effect. PM corresponds to the proportion of the total effect that is mediated by the mediator (ratio of the indirect effect to the total effect). ACME and PM are computed by the `mediation::mediate()` function of the package `mediation`, that automatically detects the type of statistical model used in the mediation analysis (Tingley et al., 2014). The function `mediation::mediate()` calculates uncertainty estimates by a quasi-Bayesian Monte Carlo approach described in (Imai et al., 2010). In addition, it estimates the intermediary effect sizes a_j and b_j and their standard deviations (see Fig 1A). Eventually, `hdmax2` calculates an Overall Indirect Effect (OIE) from a single model that includes all mediators M^S simultaneously. The OIE corresponds to the sum of the indirect effect associated with all mediators. The confidence interval (CI) of the OIE is estimated by a bootstrap approach. Along with the OIE, `hdmax2` estimates the Overall Total Effect (OTE) corresponding to the effect of exposure variables on the outcome variable, as well as the Overall Direct Effect (ODE) corresponding to the effect of exposure variables on the outcome variable when the mediators M^S are included in the model. For categorical variables, all estimates (ACME, TE, OIE, OTE, ODE and size effects) are calculated relative to a reference corresponding to the first modality of the variable. In the case of a multivariate exposure, each variable is treated independently, the other variables being included in the covariable matrix of the mediation model.

The `hdmax2::estimate_effect()` function returns an object of class *hdmax2_step2*, including the following attributes:

- ACME (average causal mediated effect), ADE (average direct effect), PM (proportion mediated) and TE (total effect), for each mediator,

- OIE (overall indirect effect), OTE (overall total effect) and ODE (overall direct effect),

- summaries of regression models adjusted during the mediation analysis.

The function `hdmax2::plot()` takes as input an object *hdmax2_step2* and enables graphical visualization of mediated effects. This function returns an ACME forest plot, a PM forest plot, a plot of the overall effects and a plot of the indirect effect sizes *a* and *b*.

Data collection and preprocessing. The breast cancer dataset was collected from the TCGA-BRCA public repository. DNA methylation data underwent filtering to remove probes containing NA values, resulting in the retention of approximately 20,000 CpG sites. Gene expressions were normalized using standard `DESeq2` parameters (Love et al., 2014) and pseudo-log transformed. The multiple sclerosis dataset was collected from the GEO public repository (accession number : GSE137143 (Kim et al., 2021)). RNA-seq data underwent normalization using standard `DESeq2` parameters (Love et al., 2014) and were filtered to retain only coding genes with detectable expression (>0 in at least one sample). Filtered data were subsequently transformed using pseudo-logarithmic transformation.

Bioinformatic analysis. Gene set enrichment analysis were performed using the `fgsea` (Korotkevich et al., 2021) and `msigdbr` (Dolgalev, 2022) `R` packages, using the defaults parameters. Gene ranks correspond to -log10(max-squared pvalues).

Results

Description of the package vignette. We propose several vignettes to explain the use of the package based on simulated data: an analysis case with a univariate exposure, an analysis case with a multivariate exposure and a vignette to illustrate the usage of helper functions.

First use case : HER2 and breast cancer. In this example study, we employed mediation analysis to assess the potential causal role of DNA methylation in the pathway linking HER2 status of Breast Cancer to a survival prognostic factor, namely the risk score. Our investigation utilized data from the TCGA-BRCA repository. The risk score was derived from a six-gene expression signature, as described by (Yin et al., 2024), and is inversely correlated with patient survival. HER2 positive status contributes to the molecular subtyping of breast cancer, which includes 'estrogen and progesterone receptor-positive, HER2 negative' (Luminal A), 'hormone positive and HER2 negative' (Luminal B), as well as 'HER2 positive' and 'triple-negative breast cancer (TNBC)' subtypes (Orrantia-Borunda et al., 2022). The computation of the risk score was based on RNAseq data, following the methodology outlined by (Yin et al., 2024). We categorized patients based on HER2 status, dividing them into HER2-positive (n=176) and HER2-negative (n=752) groups, while filtering out equivocal (n=22) and indeterminate (n=8) cases. Our analytical model included X as a binary exposure variable representing HER2 expression (0 = negative, 1 = positive), M encompassing 20,000 methylation probes, and Y denoting the continuous risk score.

In our analysis, after adjusting for the confounding effect of age, we found that the total effect of HER2-positive status resulted in a 0.30 higher risk score (t-test, p=0.007, sd = 0.11). For the initial step of the HDMAX2 approach, we opted to use $K = 2$ latent factors in the association study. Subsequently, we identified the top 10 potential mediators with the lowest max-squared P-values. We then estimated the individual indirect effect of each mediator by computing the Average Causal Mediation Effect (ACME) as depicted in Fig 2A, along with the proportion mediated shown in Fig 2B. Following this, we calculated the Overall Indirect Effect and its corresponding Overall Direct Effect, as illustrated in Fig 2C. In our analysis, we observed a negative direct effect of HER2 on the risk score, suggesting that HER2 expression in breast tumors has a protective effect on survival. However, the indirect effect mediated by the top 10 CpG probes corresponded to a 0.61 increase in the risk score (standard deviation = 0.07). This indicates that the mediated

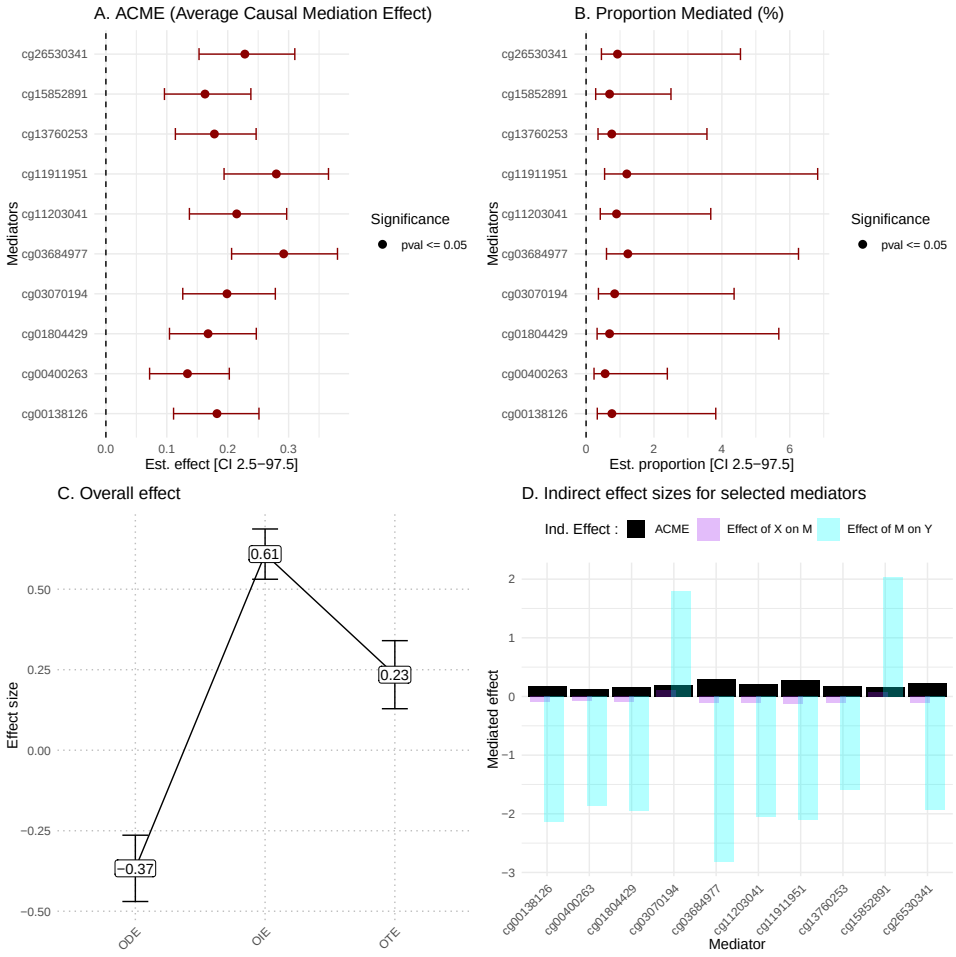

Figure 2 – **Summary of the breast cancer use case. A.** Estimates of indirect effect (ACME) and **B.** proportions of mediated effect (PM) for the top 10 mediators. The effect estimate is represented by a dot and its 95% CI by the bar. Symbols correspond to the significance cut off of 5% (square for P-value ≥ 0.05, circle P-value < 0.05). Colors correspond to the sign of the effect (green for estimated effect ≤ 0 , red for estimated effect > 0). **C.** Effect sizes of Overall Direct Effect (ODE), Overall Indirect Effect (OIE) and Overall Total Effect (OTE). Error bars correspond to standard deviation (ODE and OTE) or confidence interval (OIE). **D.** Indirect effect sizes for the selected mediators. Black corresponds to the ACME, violet to the effect of exposure X on mediator M in the model $X \sim M$, and blue corresponds to the effect of mediator M on outcome Y in the model $Y \sim M + X$.

effect is detrimental to patient survival, resulting in the observed total effect of 0.30 on the risk score. Interestingly, in our analysis, we noted that for 8 out of the 10 identified mediators, the positive Average Causal Mediation Effect (ACME) resulted from a double negative effect: firstly, from HER2 status (X) to methylation (M), and secondly, from methylation (M) to risk score (Y) (Fig 2D). This observation suggests a complex interplay between the exposure, mediators, and outcome, where the presence of the mediator counterbalance the negative effect of the exposure on the outcome. This phenomenon highlights the complex relationships within biological pathways involved in tumor growth and patient survival. This result highlights the importance of considering mediators to understand the mechanisms underlying seemingly simple observed associations. Remarkably, most of the top 10 identified mediators were associated with genes known to be involved in breast cancer biology, thus supporting the biological relevance of our approach (see Table 1).

ID	Chromosome	Start	Gene Symbol	Pubmed hits
cg11911951	chr16	1380215	AL031721.1;UNKL	0;0
cg26530341	chr8	23225840	RP11-1149O23.3;TNFRSF10A	0;54
cg01804429	chr4	68350359	YTHDC1	7
cg03684977	chr17	39737550	GRB7	123
cg11203041	chr12	16347252	MGST1	6
cg00138126	chr20	57711641	PMEPA1;RP5-1059L7.1	17;0
cg13760253	chr8	66020465	DNAJC5B	0
cg03070194	chr1	109668062	GSTM2;GSTM4	9;7
cg15852891	chr5	77639121	OTP	8
cg00400263	chr20	59939146	FAM217B;PPP1R3D	0;0

Table 1 – Top10 mediators. ID corresponds to the CpG probe name. Chromosome and Start are coordinates provided by the Illumina services. Gene Symbol corresponds to genes known to be associated with the CpG probes (Illumina annotation file). Pubmed hits corresponds to the number of output from the search "(Breast cancer) AND (‘Gene Symbol‘)".

Second use case : Gender and multiple sclerosis subtypes. In this second case study, we conducted mediation analysis to explore the role of gene expression in the pathway linking patient gender to the occurrence of Multiple Sclerosis (MS) specific subtypes: Clinically Isolated Syndrome (CIS) and Relapsing-Remitting (RR). We used the publicly available dataset GSE137143 (Kim et al., 2021). This dataset comprises peripheral blood mononuclear cells (PBMCs) from healthy individuals and patients with MS. Upon observing a significant decrease in CIS-MS occurrence among women (see Fig 3A), we sought to investigate this phenomenon further. Although the prevalence of MS is known to be higher in women (Harbo et al., 2013), men tends indeed to present worst disease progression (Voskuhl et al., 2020). We found that being female was associated with a 1.58-fold lower risk of developing Relapsing-Remitting Multiple Sclerosis (Wald-test, p=0.0036, sd = 0.54). In the original dataset, gene expression was assessed in CD4+, CD8+ T cells, and monocytes. To avoid potential confounding effect from paired data (i.e. different measures on the same individual), we focused solely on CD4+ T cell transcriptome to examine the effect of gene expression in the path between gender and MS subtypes. Our analytical model incorporated X as a binary exposure variable representing gender (0 = male, 1 = female), M encompassing 18,010 coding gene expressions, and Y denoting MS subtypes (0 = CIS, 1 = RR).

We conducted a mediation analysis on gene expression to assess the indirect effect of gender on MS subtypes. Due to the small cohort size, we did not identify any significant mediators when applying an FDR control strategy. Alternatively, we opted to perform Gene Set Enrichment Analysis (GSEA) (Korotkevich et al., 2021) to detect biological pathways enriched in our mediation analysis (see Fig 3B). We ranked the mediators using the max-squared test P-values and screened the mSigDB Hallmark collection of gene sets (Liberzon et al., 2015). In the Fig 3B, we present the top 10 upregulated pathways identified. It was expected to find hallmarks associated with blood in PBMCs, such as coagulation or heme metabolism. Given the inflammatory nature of MS, it was also reassuring to find immune-related gene sets, such as complement activity or allograft rejection, which correlates with cytolytic activity. Interestingly, we also identified several pathways previously associated with MS disease that thus appear promising for elucidating

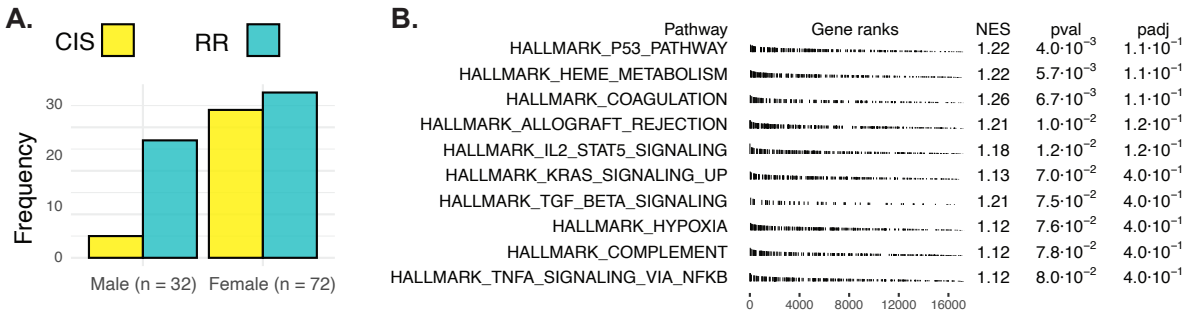

Figure 3 – Effect of gender on MS subtypes. A. Barplot ot subtype occurence according to gender. Total number of individual: 104. **B**. The pathways represent the top 10 selected gene sets. Gene ranks were determined using the -log10 of the max-squared *P*-values. NES (Normalized Enrichment Scores) were computed using the GSEA function. Additionally, *P*-values (pval) and adjusted *P*-values (padj) via Benjamini-Hochberg correction were associated with each pathway NES.

the causal relationship between gender, gene expression, and MS subtypes: (i) several studies previously established a link between hypoxia and MS disease progression (Halder and Milner, 2020); (ii) some studies have demonstrated the pivotal role of the interleukin 2 receptor pathway in autoimmune response, particularly in MS progression (Peerlings et al., 2021), and (iii) p53 has been associated with immune regulation (Muñoz-Fontela et al., 2016). All of these findings present intriguing avenues for further investigation into the gender impact on MS subtypes.

To further investigate the role of top mediators, despite the low level of significance in the max-squared test, we focused on the top two mediators: *NT5E* (HDMAX2 max-squared *P*-value = 5.10^{-06}) and *MVB12B* (HDMAX2 max-squared *P*-value = 2.10^{-06}). Interestingly, these mediators exhibit opposite effects on the pathway between gender and MS subtypes (Fig 4). NT5E is an ecto-5'-nucleotidase known to play a role in immune response deregulation in MS. It has been reported that ectonucleotidases are associated with RR MS in relapsing patients (Álvarez-Sánchez et al., 2019). Notably, *NT5E* expression appears to be elevated in women, thereby increasing the risk of developing RR MS in this population (ACME = +0.06 with a CI of [0.02,0.14]; PM = 19% with a CI of [4,97]). On the other hand, MVB12B is a member of the ESCRT-1 complex. This protein was recently identified in a genome-wide protein quantitative trait locus study aimed at identifying drivers of immune-related diseases (The SCALLOP consortium et al., 2023). In our study, *MVB12B* was associated with a reduced risk of RR MS in women (ACME = -0.09 with a CI of [-0.18,-0.02]; PM = 29% with a CI of [5,100]). Note that in this case study, the small sample size poses challenges in interpreting the estimated confidence intervals, particularly when calculating mediated proportions. Nonetheless, the mediation study offers valuable indications into the mediation mechanisms at play. These findings underscore the ability of high-dimensional analysis to provide insightful understanding into the biological mechanisms underlying observed statistical associations, even with a small cohort size.

Discussion

In this article, we introduced the `hdmax2` R package, dedicated to high-dimensional mediation analysis. The HDMAX2 method includes unobserved latent factors through a latent factor mixed model approach. It supports the use of exposures of various types and the

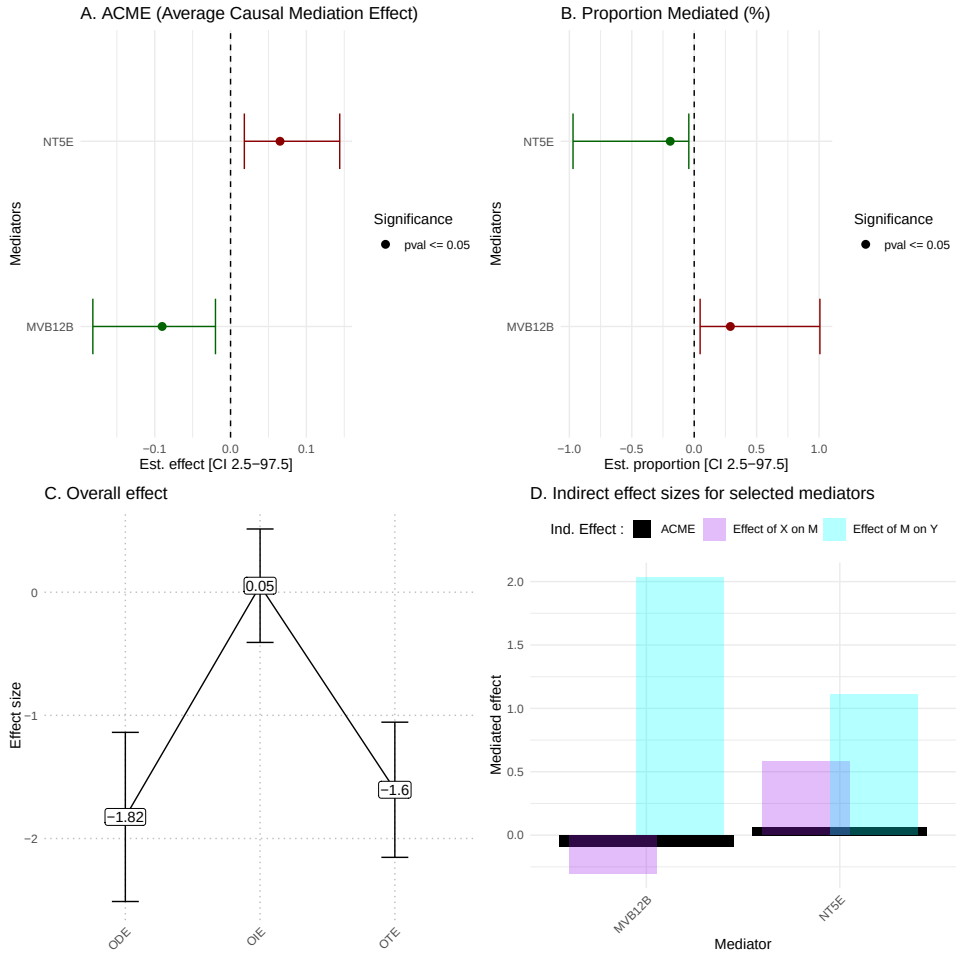

Figure 4 – **Summary of the multiple sclerosis use case. A** Estimates of indirect effect (ACME) and **A** proportions of mediated effect (PM) for the top2 mediators. The effect estimate is represented by a dot and its 95% CI by the bar. Symbols correspond to the significance cut off of 5% (square for P-value ≥ 0.05, circle P-value < 0.05). Colors correspond to the sign of the effect (green for estimated effect ≤ 0 , red for estimated effect > 0). **C** Effect sizes of Overall Direct Effect (ODE), Overall Indirect Effect (OIE) and Overall Total Effect (OTE). Error bars correspond to standart deviation (ODE and OTE) or confidence interval (OIE).**D** Indirect effect sizes for the selected mediators. Black corresponds to the ACME, violet to the effect of exposure X on mediator M in the model $X \sim M$, and blue corresponds to the effect of mediator M on outcome Y in the model $Y \sim M + X$.

consideration of both continuous and binary outcomes. We provide a detailed usage vignette and showcase the relevance of our approach through transcriptome and methylome dataset analyses use-cases. The package is available on GitHub for easy access and contribution.

While the HDMAX2 method has been optimized for high-dimension analysis, it can naturally be applied outside of this framework, as long as the number of mediators considered is superior to K (the number of confounders estimated by the latent factor mixed models). The use cases we propose have been selected for demonstration purposes, taking advantage of publicly available data. The choice of mediators to consider should be made by the user based on their specific needs and scientific questions. We provide helper functions in a separate vignette, allowing for conducting mediator selection under FDR control, or for the aggregation of P-values to study regions of interest. Aggregated Methylated Regions (AMR) are a typical

example of regions of interest, when studying DNA methylation mediated effect. AMR are made of CpG with significant *P*-value, in a given genomic region, they can be viewed as the parallel of Differentially Methylated Regions in classical EWAS. Although mediation analyses can establish a statistical association between an exposure, a mediator and an outcome, they do not guarantee a causal role in the biological processes observed. The user must then be careful not to over-interpret his results, and to build his models taking into account the sequentiality of the elements observed and the interactions between the different variables included in the model.

We believe that the HDMAX2 method and its implementation as a package could find wide applications in the fields of epidemiology, as well as in computational biology approaches aimed at gaining deeper insights into the biological background of diseases such as cancer, for which omics datasets are readily available in the public domain. It is worth noting that recent studies have proposed methods for conducting mediation analysis in the context of survival outcomes (Clark-Boucher et al., 2023; Luo et al., 2020; Zhang et al., 2021) . This is a significant demand in healthcare, and we are considering including this capability in the `hdmax2` package once we have conducted a comprehensive analysis of the methodological approach to implement.

Acknowledgements

We thank Florent Chuffart for his critical reading of the manuscript. We are grateful to the discussion group *Meth*^{2}, Lucile Broséus, Ariane Guilbert and Claire-Cécile Barrot for their helpful discussions.

Fundings

This work was funded by the project THEMA, from "Appel à projets IRGA 2021-2022" from the University Grenoble Alpes and by the French Agency for National Research (ETAPE // ANR-18-CE36-0005 and CauseHet // ANR-22-CE45-0030). It has also been carried out with financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm (ACACIA project AAP-MIC-2021).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, script, code, and supplementary information availability

Data are available online on public repositories. Script and codes are available online: https://github.com/bcm-uga/hdmax2.

References

Álvarez-Sánchez N, Cruz-Chamorro I, Díaz-Sánchez M, Lardone PJ, Guerrero JM, Carrillo-Vico A (2019). *Peripheral CD39-expressing T regulatory cells are increased and associated with relapsing-remitting multiple sclerosis in relapsing patients*. *Scientific Reports* **9**. Publisher: Nature Publishing Group, 2302. https://doi.org/10.1038/s41598-019-38897-w. URL: https://www.nature.com/articles/s41598-019-38897-w (visited on 03/07/2024).

Baron RM, Kenny DA (1986). *The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations*. *Journal of Personality and Social Psychology* **51**, 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173.

Blum MGB, Valeri L, François O, Cadiou S, Siroux V, Lepeule J, Slama R (2020). *Challenges Raised by Mediation Analysis in a High-Dimension Setting*. *Environmental Health Perspectives* **128**, 55001. https://doi.org/10.1289/EHP6240.

Caye K, Jumentier B, Lepeule J, François O (2019). *LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies*. *Molecular Biology and Evolution* **36**, 852–860. https://doi.org/10.1093/molbev/msz008. URL: https://doi.org/10.1093/molbev/msz008 (visited on 11/29/2021).

Clark-Boucher D, Zhou X, Du J, Liu Y, Needham BL, Smith JA, Mukherjee B (2023). *Methods for mediation analysis with high-dimensional DNA methylation data: Possible choices and comparisons*. *PLoS genetics* **19**, e1011022. https://doi.org/10.1371/journal.pgen.1011022.

Dai JY, Stanford JL, LeBlanc M (2022). *A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses*. *Journal of the American Statistical Association* **117**. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1765785, 198–213. https://doi.org/10.1080/01621459.2020.1765785. URL: https://doi.org/10.1080/01621459.2020.1765785 (visited on 02/12/2024).

Djordjilović V, Hemerik J, Thoresen M (2022). *On optimal two-stage testing of multiple mediators*. *Biometrical Journal* **64**. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.202100190, 1090–1108. https://doi.org/10.1002/bimj.202100190. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202100190 (visited on 02/12/2024).

Dolgalev I (2022). *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*. R package version 7.5.1.9001. URL: https://igordot.github.io/msigdbr/.

Halder SK, Milner R (2020). *Hypoxia in multiple sclerosis; is it the chicken or the egg? Brain* **144**, 402–410. https://doi.org/10.1093/brain/awaa427. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8453297/ (visited on 03/07/2024).

Harbo HF, Gold R, Tintoré M (2013). *Sex and gender issues in multiple sclerosis*. *Therapeutic Advances in Neurological Disorders* **6**, 237–248. https://doi.org/10.1177/1756285613488434. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3707353/ (visited on 03/07/2024).

Imai K, Keele L, Tingley D (2010). *A general approach to causal mediation analysis*. *Psychological Methods* **15**, 309–334. https://doi.org/10.1037/a0020761. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0020761 (visited on 11/29/2021).

Jumentier B, Barrot CC, Estavoyer M, Tost J, Heude B, Fran çO, Lepeule J (2023). *High-Dimensional Mediation Analysis: A New Method Applied to Maternal Smoking, Placental DNA*

Methylation, and Birth Outcomes. *Environmental Health Perspectives* **131** (). Publisher: Environmental Health Perspectives, 047011. https://doi.org/10.1289/EHP11559. URL: https://ehp.niehs.nih.gov/doi/10.1289/EHP11559 (visited on 09/19/2023).

Kim K, Pröbstel AK, Baumann R, Dyckow J, Landefeld J, Kogl E, Madireddy L, Loudermilk R, Eggers EL, Singh S, Caillier SJ, Hauser SL, Cree BAC, UCSF MS-EPIC Team, Schirmer L, Wilson MR, Baranzini SE (2021). *Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis*. *Brain* **144**, 450–461. https://doi.org/10.1093/brain/awaa421. URL: https://doi.org/10.1093/brain/awaa421 (visited on 03/01/2024).

Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A (2021). *Fast gene set enrichment analysis*. Pages: 060012 Section: New Results. https://doi.org/10.1101/060012. URL: https://www.biorxiv.org/content/10.1101/060012v3 (visited on 03/07/2024).

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P (2015). *The Molecular Signatures Database (MSigDB) hallmark gene set collection*. *Cell systems* **1**, 417–425. https://doi.org/10.1016/j.cels.2015.12.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707969/ (visited on 03/07/2024).

Love MI, Huber W, Anders S (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. *Genome Biology* **15**, 550. https://doi.org/10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (visited on 06/10/2022).

Luo C, Fa B, Yan Y, Wang Y, Zhou Y, Zhang Y, Yu Z (2020). *High-dimensional mediation analysis in survival models*. *PLOS Computational Biology* **16**. Ed. by Benjamin Althouse, e1007768. https://doi.org/10.1371/journal.pcbi.1007768. URL: https://dx.plos.org/10.1371/journal.pcbi.1007768 (visited on 09/19/2023).

Muñoz-Fontela C, Mandinova A, Aaronson SA, Lee SW (2016). *Emerging roles of p53 and other tumour-suppressor genes in immune regulation*. *Nature Reviews Immunology* **16**. Publisher: Nature Publishing Group, 741–750. https://doi.org/10.1038/nri.2016.99. URL: https://www.nature.com/articles/nri.2016.99 (visited on 03/07/2024).

Orrantia-Borunda E, Anchondo-Nuñez P, Acuña-Aguilar LE, Gómez-Valles FO, Ramírez-Valdespino CA (2022). *Subtypes of Breast Cancer*. In: *Breast Cancer*. Ed. by Harvey N. Mayrovitz. Brisbane (AU): Exon Publications. URL: http://www.ncbi.nlm.nih.gov/books/NBK583808/ (visited on 03/19/2024).

Peerlings D, Mimpen M, Damoiseaux J (2021). *The IL-2 - IL-2 receptor pathway: Key to understanding multiple sclerosis*. *Journal of Translational Autoimmunity* **4**, 100123. https://doi.org/10.1016/j.jtauto.2021.100123.

Richiardi L, Bellocco R, Zugna D (2013). *Mediation analysis in epidemiology: methods, interpretation and bias*. *International Journal of Epidemiology* **42**, 1511–1519. https://doi.org/10.1093/ije/dyt127. URL: https://doi.org/10.1093/ije/dyt127 (visited on 09/30/2021).

Sampson JN, Boca SM, Moore SC, Heller R (2018). *FWER and FDR control when testing multiple mediators*. *Bioinformatics* **34**, 2418–2424. https://doi.org/10.1093/bioinformatics/bty064. URL: https://doi.org/10.1093/bioinformatics/bty064 (visited on 02/12/2024).

Sobel ME (1982). *Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models*. *Sociological Methodology* **13**. Publisher: [American Sociological Association, Wiley, Sage Publications, Inc.], 290–312. https://doi.org/10.2307/270723. URL: https://www.jstor.org/stable/270723 (visited on 11/29/2021).

The SCALLOP consortium, Zhao JH, Stacey D, Eriksson N, Macdonald-Dunlop E, Hedman ÅK, Kalnapenkis A, Enroth S, Cozzetto D, Digby-Bell J, Marten J, Folkersen L, Herder C, Jonsson L, Bergen SE, Geiger C, Needham EJ, Surendran P, Estonian Biobank Research Team, Paul DS, et al. (2023). *Mapping pQTLs of circulating inflammatory proteins identifies drivers of immune-related disease risk and novel therapeutic targets*. preprint. Genetic and Genomic Medicine. https://doi.org/10.1101/2023.03.24.23287680. URL: http://medrxiv.org/lookup/doi/10.1101/2023.03.24.23287680 (visited on 03/07/2024).

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). ***mediation*** *: R Package for Causal Mediation Analysis*. *Journal of Statistical Software* **59**. https://doi.org/10.18637/jss.v059.i05. URL: http://www.jstatsoft.org/v59/i05/ (visited on 02/09/2024).

Voskuhl RR, Patel K, Paul F, Gold SM, Scheel M, Kuchling J, Cooper G, Asseyer S, Chien C, Brandt AU, Meyer CE, MacKenzie-Graham A (2020). *Sex differences in brain atrophy in multiple sclerosis*. *Biology of Sex Differences* **11**, 49. https://doi.org/10.1186/s13293-020-00326-3. URL: https://doi.org/10.1186/s13293-020-00326-3 (visited on 03/07/2024).

Yin Q, Ma H, Dong Y, Zhang S, Wang J, Liang J, Mao L, Zeng L, Xiong X, Chen X, Wang J, Zheng X (2024). *The integration of multidisciplinary approaches revealed PTGES3 as a novel drug target for breast cancer treatment*. *Journal of Translational Medicine* **22**. Number: 1 Publisher: BioMed Central, 1–16. https://doi.org/10.1186/s12967-024-04899-0. URL: https://translational.medicine.biomedcentral.com/articles/10.1186/s12967-024-04899-0 (visited on 02/06/2024).

Zeng P, Shao Z, Zhou X (2021). *Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges*. *Computational and Structural Biotechnology Journal* **19**, 3209–3224. https://doi.org/10.1016/j.csbj.2021.05.042. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8187160/ (visited on 12/02/2021).

Zhang H, Zheng Y, Hou L, Zheng C, Liu L (2021). *Mediation analysis for survival data with high-dimensional mediators*. *Bioinformatics (Oxford, England)* **37**, 3815–3821. https://doi.org/10.1093/bioinformatics/btab564.

Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, Zhang W, Schwartz J, Just A, Colicino E, Vokonas P, Zhao L, Lv J, Baccarelli A, Hou L, Liu L (2016). *Estimating and testing high-dimensional mediation effects in epigenetic studies*. *Bioinformatics* **32**, 3150–3154. https://doi.org/10.1093/bioinformatics/btw351. URL: https://doi.org/10.1093/bioinformatics/btw351 (visited on 02/12/2024).