Software Application Profile

Software Application Profile: exposomeShiny— a toolbox for exposome data analysis

Xavier Escriba-Montagut,[1] **Xavier Basagaña** [ID] **,**[1,2] **Martine Vrijheid**[1,2,3,4] **and Juan R Gonzalez** [ID] [1,2,3,4]*

[1]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain and [4]Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Bellaterra (Barcelona), Spain

*Corresponding author. Barcelona Biomedical Research Park (PRBB), Doctor Aiguader, 88. 08003 Barcelona, Spain.
E-mail: juanr.gonzalez@isglobal.org

Editorial decision 17 September 2021; Accepted 17 September 2021

Abstract

Motivation: Studying the role of the exposome in human health and its impact on different omic layers requires advanced statistical methods. Many of these methods are implemented in different R and Bioconductor packages, but their use may require strong expertise in R, in writing pipelines and in using new R classes which may not be familiar to non-advanced users. ExposomeShiny provides a bridge between researchers and most of the state-of-the-art exposome analysis methodologies, without the need of advanced programming skills.

Implementation: ExposomeShiny is a standalone web application implemented in R. It is available as source files and can be installed in any server or computer avoiding problems with data confidentiality. It is executed in RStudio which opens a browser window with the web application.

General features: The presented implementation allows the conduct of: (i) data pre-processing: normalization and missing imputation (including limit of detection); (ii) descriptive analysis; (iii) exposome principal component analysis (PCA) and hierarchical clustering; (iv) exposome-wide association studies (ExWAS) and variable selection ExWAS; (v) omic data integration by single association and multi-omic analyses; and (vi) post-exposome data analyses to gain biological insight for the exposures, genes or using the Comparative Toxicogenomics Database (CTD) and pathway analysis.

Availability: The exposomeShiny source code is freely available on Github at [https://github.com/isglobal-brge/exposomeShiny], Git tag v1.4. The software is also available as a Docker image [https://hub.docker.com/r/brgelab/exposome-shiny], tag v1.4. A user guide with information about the analysis methodologies as well as information on how to use exposomeShiny is freely hosted at [https://isglobal-brge.github.io/exposome_bookdown/].

Introduction

The exposome concept was first described by Wild in 2005.[1] This concept refers to the totality of environmental and lifestyle exposures of an individual over the lifespan, i.e. not just a cross-sectional evaluation of exposures but also their change over time. Wild later proposed. in 2012.[2] a three-domain division of the exposome: internal, general external and specific external. The general external exposures are the wider exposures that an individual can be exposed to such as mental stress, climate, living environment etc. The specific external exposures cover a large range of exposures that the individual could be exposed to, such as lifestyle factors, medical interventions, environmental pollutants etc. The internal domain refers to how the biological/toxicological processes respond to the external exposures (both general and specific).

The exposome can include traditional measures of exposure (e.g. traditional biomonitoring, environmental monitoring) but also includes untargeted discovery of unknown chemicals of biological importance.[3] In this application we focus on targeted approach.

From the initial proposal of the concept to nowadays, the amount of research focused on the exposome has grown significantly, as shown by the numerous funded projects that focus on the exposome analysis, including HELIX,[4] LifeCycle,[5] HERCULES,[6] EXPOsOMICS,[7] CHEAR[8] and ATHLETE[9] among many others.

Assessing the association between the exposome and health still raises many exposure assessment issues and statistical challenges, as described in Beggs *et al.*[10] and Zota *et al.*[11] However, there are several statistical methods currently available that can be used to address different scientific questions in the context of exposome and health relationships.[12] Some of these methods are already implemented in an R framework called rexposome, which includes three Bioconductor packages: rexposome, omicRexposome[13] and CTDquerier.[14] These packages are based on a new class of R objects called ExposomeSet, which serves as the input for the main functions designed to describe the exposome, to perform exposome/environment wide association analyses (ExWAS), to integrate the exposome with different omic layers and to provide some biological insights using Comparative Toxicogenomic Database (CTD).[15] CLAIRE is another R package to perform ExWAS which is mainly designed to clean environmental data from public repositories such as the Health and Nutrition Examination Survey (NHANES).[16] Other general R packages can be used to integrate the exposome with other omic data using multivariate methods.[17]

Most of the epidemiological researchers use SPSS, Stata, SAS or R to perform their data analyses. Even R users may not be familiar with Bioconductor classes.[18] In order to overcome this difficulty and also to help researchers who are not familiar with R, we developed exposomeShiny, a user-friendly web-based tool that bundles different methodologies required when performing exposome data analyses.[12]

We have implemented several procedures to perform both data pre-processing and statistical analyses. Still, data checks for potential errors or outliers must be conducted beforehand. The following list summarizes the methods currently implemented. A more detailed description and references are provided in an online tutorial [https://isglobal-brge.github.io/exposome_bookdown/methods.html].

Data pre-processing:

i. Missing imputation. Visualization of missing data. Imputation of missing data due to limit of detection (LOD) using either LOD/sqrt(2)[19] or by imputing left-censored data using random draws from a truncated distribution.[20] Missing values due to LOD should be coded using value other than NA (i.e. 'lod'). Multiple imputation of regular missing data is performed using the method of chained equations.[21]

ii. Normalization. Shapiro-Wilks test to assess normality. Different transformations to achieve normality are allowed (log, sqrt and x^1/3).

Data analysis:

i. Descriptive analysis. Description of exposures (median, mean, standard error etc.) and graphical description using box plots with the option of grouping with the categorical exposures.

ii. Exposures clustering. Clusters of individuals which share similar exposure patterns using hierarchical clustering

iii. Principal component analysis (PCA). Used to reduce the dimensionality of the exposome. We allow the user to run normal PCA, that is including all continuous exposures and FAMD method that allow incorporation of both categorical and continuous exposures.[22]

iv. Exposure correlation. Assesses correlation among all exposures. It uses Pearson for numerical-to-numerical

correlation, Cramer's V for categorical-to-categorical correlation and linear models for categorical-to-numerical correlation. Circos plots can be created.

v. Single association analysis between exposures and health outcomes (ExWAS). Generalized linear regression models of each exposure against the outcome, possibly adjusting for confounders, using separate models for each exposure and with correction for multiple testing. Although the term EWAS was previously used to refer to exposome-wide association studies,[23] we advocate for the use of ExWAS[24] to avoid confusion with epigenome-wide association studies.

vi. Variable selection ExWAS. Variable selection (e.g. exposures) procedure that finds a final model for the outcome which includes multiple exposures in the same model. In particular, the package implements generalized linear models via penalized maximum likelihood using the lasso or elastic net penalty.[25]

vii. Omic association analysis. Association analyses between omic features (gene expression, CpGs) and exposures (as predictors, with optional phenotypes as covariates) are assessed using linear models with the option of using surrogate variable analysis (SVA) to account for unobserved variability.[26]

viii. Multi-omic data analysis. Integration of multiple omic layers can be performed using multiple coinertia analysis (MCIA), generalized canonical correlation (GCCA) and partial least squares (PLS) methods. These methods support the sub-setting of the exposures domain by families, but not by the results of previous methods (such as variable selection ExWAS).

ix. Post-exposome data analysis. We allow the user to acquire biological insights for the exposures, genes or diseases of interest obtained from the different analyses using the CTD database, which contains manually curated information about chemical-gene/protein interactions and chemical-disease and gene-disease relationships, integrated with functional and pathway data.[14] A recent update and description of this database can be found elsewere.[27] We also allow the performance of enrichment analysis after omic data analysis using GO and KEGG databases.

Implementation

The application can be downloaded or installed into any computer from [https://github.com/isglobal-brge/exposomeShiny]. The required R, R packages and RStudio versions are listed on the bookdown. For simplicity, and to avoid problems with R versions or installing issues, we also provide the software as a Docker image. We notice that since the application runs on the user's computer, no data need to be sent to external web servers, avoiding problems with data confidentiality. We also present a detailed online user's guide that includes several real examples and flowcharts describing the main analyses at [https://isglobal-brge.github.io/exposome_bookdown/analysis-flowcharts.html].

The toolbox has been designed as a bundle of boxes that each perform one specific analysis task. All boxes have a standardized input and output scheme to ensure structural integrity of the software. This modular design aims at developing a future-proof software that can be easily upgraded and improved in the future, as improving an existing functionality is only a matter of changing one box while respecting the current inputs and outputs, and adding a new functionality is done by just creating a new box and connecting it properly to the already existent structure of boxes. This is crucial since, although we have implemented several methods currently used for exposome data analysis, this is an active line of research and other methods will need to be added to our tool.

ExposomeShiny has been developed using the R programming language using the Shiny package. Shiny is a package developed by RStudio that is aimed at developing user interfaces with R code which facilitate implementing analysis pipelines.[28] On first use, users need to run a script that will install all the required packages; this information is detailed on the available online user's guide linked on the abstract. In order to launch exposomeShiny, the integrated development environment (IDE) created by RStudio has a built-in button that reads 'Run App'. By clicking on it, the researcher is ready to use the application. The R code of exposomeShiny can be downloaded from GitHub at [https://github.com/isglobal-brge/exposomeShiny].

A scheme of the box structure of exposomeShiny and the connections between the different boxes is shown in Figure 1. There are three main functional blocks: the exposome analysis block performs descriptive analysis, PCA, ExWAS and mExWAS;The omic block runs single omic association analysis and cross-omic integration; and the biological insights block carries out enrichment analyses using CTD, GO and KEGG databases.

We first need to upload exposome data into exposomeShiny. There are two options. First, the user can have three different files corresponding to the phenotypes (i.e. outcomes and covariates), the exposures and the annotation of the exposures (i.e. information about families of exposures, units etc.). The tool also allows the user to upload data in a single file having the phenotypes and exposures. In this second case, the family structure of the exposure will be manually entered if needed (see Section 3
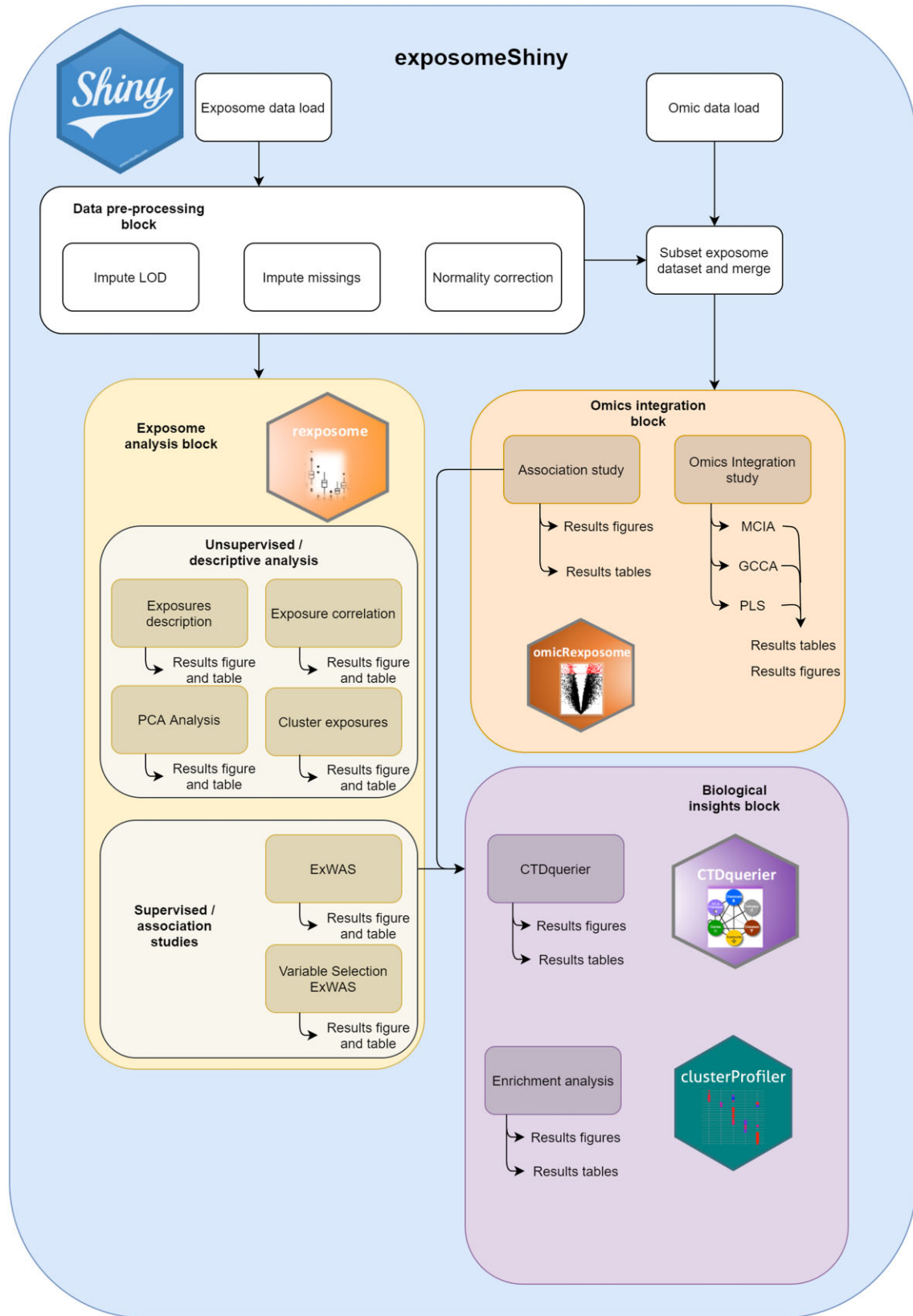
Figure 1 Block structure of exposomeShiny. Arrows connecting blocks can be interpreted as data pipes, otherwise arrows represent outputs shown on the graphical user interface. Each functional block that is powered by a certain R package contains the hexagon logotype of the R package, otherwise the blocks are powered by the R core functions

of the bookdown). The current version only deals with cross-sectional data, which implies that we only consider one time point for exposures and omic data. The exposures can be pre-processed using the Shiny application. However, if different methodologies are required, they can be applied to the dataset before entering the application, skipping the pre-processing steps. Regarding the omics data, these should be normalized before being used in the application.

The input datasets are checked for consistency at the data entry stage. In the case of uploading data in three different datasets, the user needs to indicate the variables that are used to link them. The user can also indicate how the missing values due to LOD are encoded (e.g. using the character 'lod') (see Figure 2). Regular missing values must be encoded using NA. All non-numerical variables are encoded as factors. The user can also analyse numerical variables as factors in those cases where data are discrete values. In that case, the user must indicate what is the maximum number of different values that a variable must have to be considered as factor (n selected by the user); otherwise they are encoded as continuous variables.

Use

This section illustrates a simple use case of exposomeShiny. The dataset used for this example (exposures.csv, description.csv and phenotypes.csv) can be downloaded from: [https://github.com/isglobal-brge/exposomeShiny/tree/master/data]. Data contain information on 109 different individuals, 88 different exposures grouped into sdven exposure families and eight phenotypes; they contain missings at random but not LOD missings. The illustration consists of an exposome-wide association study (ExWAS), i.e. an analysis in which all exposures are regressed separately with the outcome while controlling for multiple testing. The analysis will require a previous step in which missing exposure data will be imputed, and the subsequent analysis will be conducted according to multiple imputation rules to incorporate the imputation uncertainty in the results.

First, one needs to load the exposome datasets (remember that there is the option of having data on a single file, check section 3.2 of the bookdown). All three files are selected from the exposomeShiny tool and loaded into the toolbox by the click of a button (Figure 2). Afterwards, the input files can be visualized, and configuration fields

Figure 2 Data entry page of exposomeShiny

appear to check data integrity, variable encoding and LOD missing detection (right column in Figure 2). If there were LOD missings on the introduced data, automatically the different imputation options would appear. More information on that can be found in section 5.1.2 of the bookdown.

Once the dataset is loaded, the researcher can visually inspect the amount of missing data in the dataset using the 'Missing data' tab (Figure 3). A plot will show the percentage of missing values for every exposure. The button 'Impute missing values using mice' will create 20 imputed datasets using the method of chained equations. Once the imputed datasets are created, all subsequent analyses (*Exposome analysis block and Omic block* in Figure 1) are performed by using the imputed datasets. If a different imputation methodology is needed, the researcher has to do it outside of the application, upload the new data and skip this step.

After data imputation, ExWAS is performed as illustrated in Figure 4. In this use case, the interest is focused on assessing the association between the exposures and the phenotype 'blood pressure' while adjusting for sex of the individuals. The researcher needs to select the ExWAS tab and input the desired study variables. In this case, we choose 'blood_pre' in the '*Choose outcome variable*'

drop-down list, we specify 'sex' in the '*Choose the covariable(s)*' box and we set the field '*Choose the output family*' to 'Gaussian'. The ExWAS is automatically performed and the plot with the results is shown on the same ExWAS page of exposomeShiny (Figure 4). Stratified analyses can be performed by selecting a desired variable in the '*Stratified analysis by*' box.

With a few clicks and no lines of code, the researcher is able to perform an ExWAS study on multiply imputed datasets and can download the resulting figure to use in a paper if required. After that, the researcher could be interested in obtaining biological insights on some exposures that the ExWAS study highlighted as relevant. To this end, the Biological insights block (Figure 1) of exposomeShiny can be used. Information on how to run these analyses as well others including omic data are described in the chapter 'Analysis flowcharts' in our bookdown [https://isglobal-brge.github.io/exposome_bookdown/analysis-flowcharts.html].

Discussion

ExposomeShiny is a toolbox that aims to close the gap between epidemiologists interested in performing studies
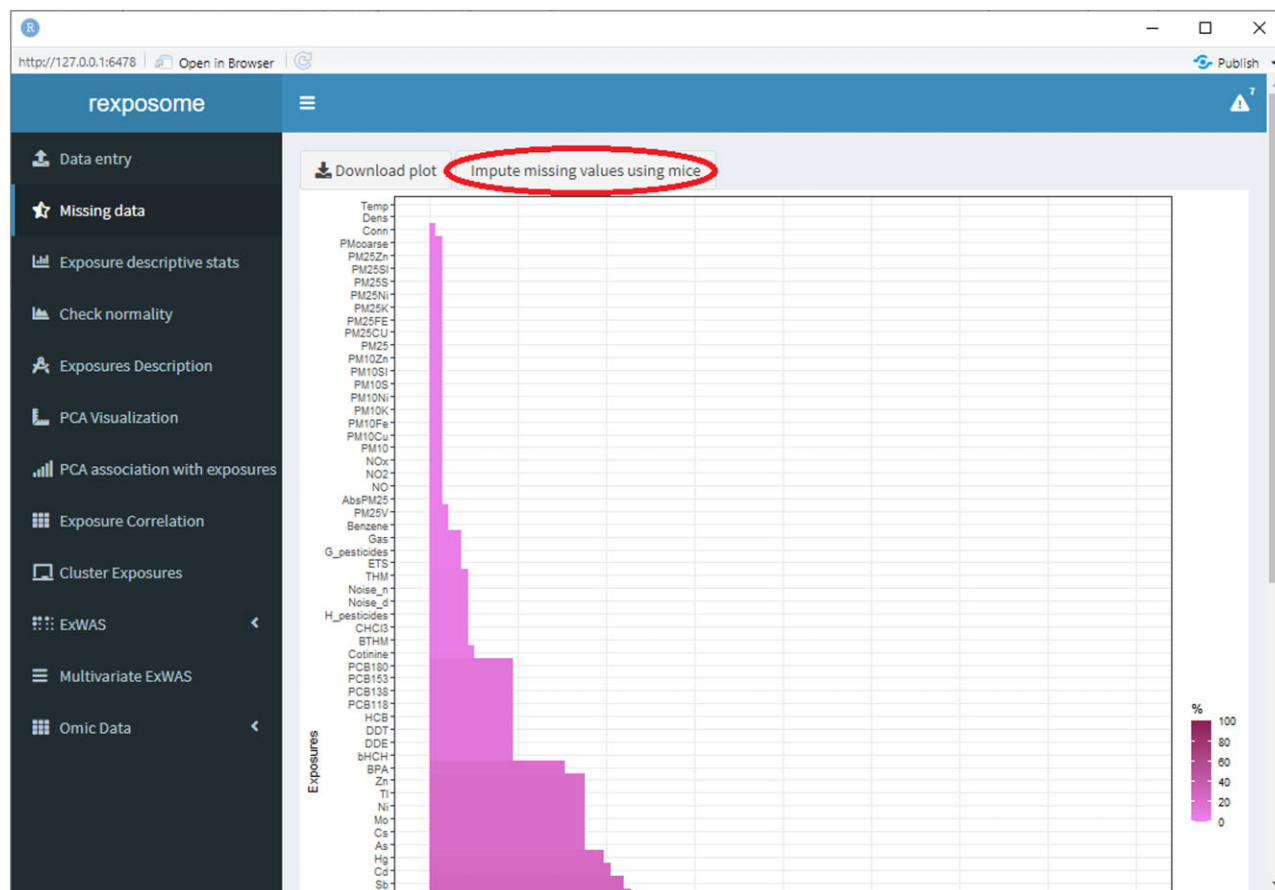

Figure 3 Missing data page of exposomeShiny. Image edited to highlight the 'Impute missing values using mice' button, which triggers the multiple imputation by chained equations (MICE) imputation of the exposome dataset loaded into the application
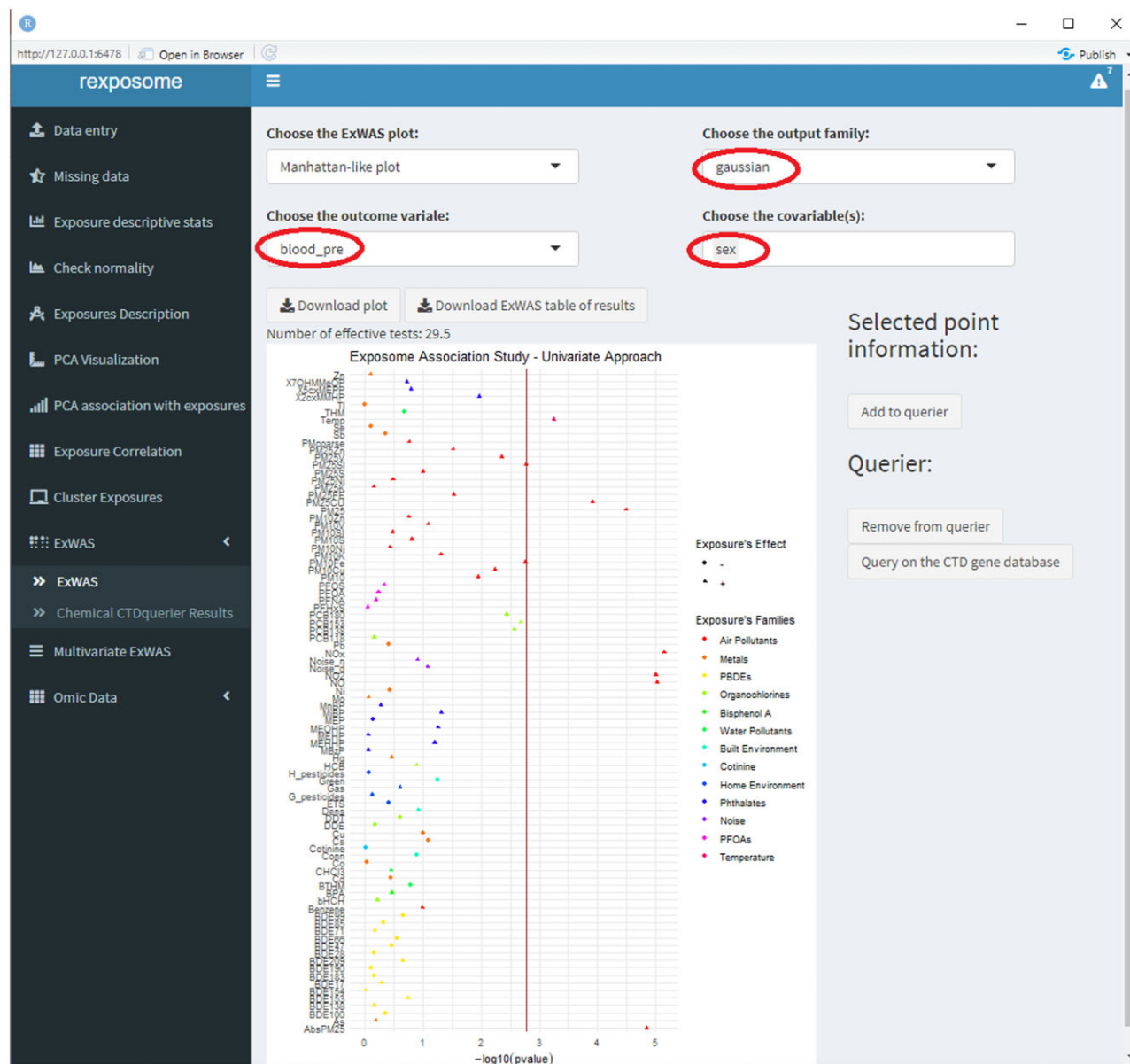
Figure 4 Exposome-wide association studies (ExWAS) page of exposomeShiny. Image edited to highlight the three input fields to configure the ExWAS analysis. The red line corresponds to the threshold for significance after correction for multiple testing by dividing the nominal level (i.e. 0.05) by the exact number of tests obtained using the method proposed by Li J *et al.*[29]

using exposome datasets and the required R knowledge needed to do so, such as being familiar with Bioconductor and all the different packages needed to perform different analyses on exposome datasets. Furthermore, it also helps epidemiologists familiar with those topics to accelerate the pipeline of their studies by providing a toolbox that reduces a whole exposome analysis to just a few clicks, saving time for the researchers. However, the presented application does not free the researcher from the burden of designing a proper analysis, deciding which of the implemented methods are the most appropriate for each case and performing regression diagnostics or sensitivity analyses. Hence, the researcher needs statistical knowledge to

use our solution. The toolbox is especially suited to perform exposome data analyses, allowing the researchers to quickly detect any interesting patterns in the data. On some occasions the statistical method may require tuned analyses. If so, processed data can be downloaded to be later studied in depth outside the toolbox.

ExposomeShiny includes many of the required statistical methods often used when analysing the exposome.[12] However, other more sophisticated methods can be required such as those to detect interactions in exposome-health studies[30] or repeated measurements. This new methodology, as well as others that may appear in the future, can readily be implemented in exposomeShiny if needed, given the modularity nature of our

implementation. Currently, there are methodologies to obtain exposome data based on mass spectometry, such as gas chromatography[31] (GC-MS) and liquid chromatography[32] (LC-MS). Data from these new technologies could also be included in exposomeShiny when methods and protocols of analyses are well defined. There are some Bioconductor packages aimed at performing analysis with this type of data, such as metaRbolomics[33] and MWASTools[34] that could be used to this end. In particular, exposomeShiny could integrate the functionalities of such packages to offer support for those analyses and extend their capabilities. Moreover, new databases with exposome data are being created currently as exposome study is attracting more attention. ExposomeShiny could be adapted in the future to work with the information that those new databases hold. Some examples of new databases are: 'Exposome Explorer' [http://exposome-explorer.iarc.fr/] and 'Blood Exposome' [http://bloodexposome.org]. The connection to new databases will be done using *ad hoc* solutions that will arrange data into the required input format of exposomeShiny. New functionalities and database connections can be added easily to the existing software and will be added upon request.

In conclusion, the paradigm of the exposome research is complex and wide. Providing researchers with an upgradeable toolbox that contains all the required tools and methodologies, ensures not only the long-term success of many studies but also its quality, as researchers can be freed of learning the technicalities of how to use new tools and can just focus on the interpretation of results. For that the presented toolbox, exposomeShiny, is intended to be maintained and upgraded accordingly in the years to come, since it is one of the tasks defined in the recently funded ATHLETE project [https://athleteproject.eu/].

Author contributions

Creation of the Shiny app: X.E. Data analysis: X.E., X.B. and J.R.G. Study concept and design: M.V. Manuscript writing: X.E. and J.R.G. Project supervision: J.R.G. Interpretation of the data and critical revision and editing of the manuscript: all authors. All authors read and approved the final manuscript.

Conflict of interest

We have no conflicts of interest.

References

1. Wild CP. Complementing the genome with an "'Exposome'": the outstanding challenge of environmental exposure measurement in molecular epidemiology.Cancer Epidemiol Biomarkers Prev 2005;**14**:1847–50.
2. Wild CP. The exposome: from concept to utility. *Int J Epidemiol* 2012;**41**:24–32.
3. Dennis KK, Marder E, Balshaw DM et al. Biomonitoring in the era of the exposome. *Environ Health Perspect* 2017;**125**:502–10.
4. Vrijheid M, Slama R, Robinson O et al. The Human Early-Life Exposome (HELIX): project rationale and design. *Environ Health Perspect* 2014;**122**:535–44.
5. Jaddoe VWV, Felix JF, Andersen AMN et al.; LifeCycle Project Group. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. *Eur J Epidemiol* 2020;**35**:709–24.
6. Niedzwiecki MM, Miller GW. HERCULES: an academic center to support exposome research. In: *Unraveling the Exposome: A Practical View*. Basel: Springer International Publishing, 2018, pp.339–48.
7. Vineis P, Chadeau-Hyam M, Gmuender H et al.; EXPOsOMICS Consortium. The exposome in practice: design of the EXPOsOMICS project. *Int J Hyg Environ Health* 2017;**220**:142–51.
8. Balshaw DM, Collman GW, Gray KA, Thompson CL. The Children's Health Exposure Analysis Resource: enabling research into the environmental influences on children's health outcomes. *Curr Opin Pediatr* 2017;**29**:385–89.
9. ATHLETE Project | H2020 | CORDIS | European Commission. Advancing Tools for Human Early Lifecourse Exposome Research and Translation. 2020.https://cordis.europa.eu/project/id/874583/es (26 March 2020, date last accessed).
10. Beggs PJ. Climate change, aeroallergens, and the aeroexposome. *Environ Res Lett* 2021;**16**:35006.
11. Zota AR, Vannoy BN. Integrating intersectionality into the exposome paradigm: a novel approach to racial inequities in uterine fibroids. Am J Public Health 2021;**111**:104–109.
12. Santos S, Maitre L, Warembourg C et al. Applying the exposome concept in birth cohort research: a review of statistical approaches. *Eur J Epidemiol* 2020;**35**:193–204.
13. Hernandez C, Wellenius G, Tamayo I et al. Comprehensive study of the exposome and omic data using rexposome Bioconductor Packages. *Bioinformatics* 2019;**35**:5344–45.
14. Hernandez-Ferrer C, Gonzalez JR. CTDquerier: a bioconductor R package for Comparative Toxicogenomics Database TM data extraction, visualization and enrichment of environmental and toxicological studies. Bioinformatic 2018;**34**:3235–37.
15. Mattingly CJ, Colby GT, Forrest JN, Boyer JL. The comparative toxicogenomics database (CTD). *Environ Health Perspect* 2003;**111**:793–95.

16. Lucas AM, Palmiero NE, McGuigan J *et al.* CLARITE facilitates the quality control and analysis process for EWAS of metabolic-related traits. *Front Genet* 2019;**10**:1240.

17. Chadeau-Hyam M, Campanella G, Jombart T *et al.* Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen* 2013; **54**:542–57.

18. Huber W, Carey VJ, Gentleman R *et al.* Orchestrating high-throughput genomic analysis with Bioconductor HHS Public Access. *Nat Methods* 2015;**12**:115–21.

19. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol* 2003;**157**:355–63.

20. Pleil JD. Imputing defensible values for left-censored "below level of quantitation" (LoQ) biomarker measurements. *J Breath Res* 2016;**10**:045001.

21. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;**45**:1–67.

22. Pagès J. *Revue de Statistique Appliquée Analyse Factorielle de Données Mixtes Article Numérisé Dans Le Cadre Du Programme Numérisation de Documents Anciens Mathématiques.* Vol. **52**. 2004. http://www.sfds.asso.fr/publi cat/rsa.htm (15 February 2021, date last accessed).

23. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* 2010;**5**:e10746.

24. Juarez PD, Matthews-Juarez P. Applying an Exposome-Wide (ExWAS) approach to cancer research. *Front Oncol* 2018;**8**:313.

25. Agier L, Portengen L, Chadeau-Hyam M *et al.* A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. *Environ Health Perspect* 2016; **124**:1848–56.

26. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;**28**:882–83.

27. Davis AP, Grondin CJ, Johnson RJ *et al.* Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res* 2021;**49**:D1138–43.

28. Di Filippo L, Righelli D, Gagliardi M, Matarazzo MR, Angelini C. HiCeekR: a novel shiny app for Hi-C data analysis. *Front Genet* 2019;**10**:1079.

29. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* 2005; **95**:221–27. doi:10.1038/sj.hdy.6800717

30. Barrera-Gómez J, Agier L, Portengen L *et al.* A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ Health* 2017;**16**:74.

31. Weggler BA, Gruber B, Focant JF. Comprehensive two-dimensional gas-chromatography to study the human exposome: current trends and perspectives. *Curr Opin Environ Sci Health* 2020;**15**:16–25.

32. Preindl K, Braun D, Aichinger G *et al.* A generic liquid chromatography-tandem mass spectrometry exposome method for the determination of xenoestrogens in biological matrices. *Anal Chem* 2019;**91**:11334–42.

33. Stanstrup J, Broeckling C, Helmus R *et al.* The metaRbolomics toolbox in bioconductor and beyond. *Metabolites* 2019;**9**:200.

34. Rodriguez-Martinez A, Posma JM, Ayala R *et al.* MWASTools: An R/bioconductor package for metabolome-wide association studies. *Bioinformatics* 2018;**34**:890–92.