

# Chapter 11

## Statistical Models to Explore the Exposome: From OMICs Profiling to ‘Mechanome’ Characterization



Marc Chadeau-Hyam and Roel Vermeulen

**Abstract** Over the past decade, high-resolution molecular profiles using OMICS technologies have accumulated and have given rise to an unprecedented source of information to explore the effective biological effects of external stressors and to detect drivers of subsequent disease risk. Although the volume, dimensionality, and complexity of OMICs data are constantly increasing, several methods enabling their analysis are now available. The exploration of these data relies on statistical approaches including univariate models coupled with multiple testing correction, dimensionality reduction techniques, and variable selection approaches. While these methods are established, their application in an exposome context is raising specific methodological challenges. In addition, the isolated exploration of an OMIC profile offers the possibility to capture stressor-induced biological/biochemical alterations, potentially impacting individual risk profiles, but this may only yield a fractional picture of the complex molecular events involved, therefore limiting our understanding of the effective mechanisms mediating the effect of the exposome. Despite efficient developments over systems biological approaches, such integrations remain at best data-specific, usually disease-specific, and more systematically restricted to the exploration of (few) predefined hypotheses. The challenging task of exploring the ‘mechanome’ as defined by the ensemble of stressor-induced molecular mechanisms occurring throughout the life course and determining the individual’s risk of developing adverse conditions can be decomposed in three interdependent streams

---

M. Chadeau-Hyam (✉)

MRC/PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics,  
School of Public Health, Imperial College London, London, UK

Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands  
e-mail: [m.chadeau@imperial.ac.uk](mailto:m.chadeau@imperial.ac.uk)

R. Vermeulen

MRC/PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics,  
School of Public Health, Imperial College London, London, UK

Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands  
Department of Molecular Epidemiology, Julius Center, University Medical Center Utrecht,  
Utrecht, The Netherlands

focusing on (1) OMICs profiling, (2) OMICs data integration, and (3) the exploration of molecular mechanisms involved in the exposure effect mediation towards (chronic) disease development.






**Keywords** Statistical models · Omics · Mechanome · Bioinformatics

## Exposome and Statistical Challenges

The original definition of the exposome (Wild 2005) is broad and comprises all external stressors, as defined by any nongenetic factor affecting, possibly at different life stages, the individual risk of chronic diseases. Following-up on the exposome concept, Rappaport and Smith (2010) complemented this definition by assuming that any effective component of the exposome should be detectable in the internal environment, hence defining the internal exposome. The internal exposome comprises biological responses to the external environment and interactions with exogenous compounds as well as exogenous compounds themselves entering the internal environment.

With the rapid advances in molecular biology, cost-effective technologies have emerged and enable the generation of high-resolution and high-quality individual molecular profiles in large-scale studies. Resulting OMICs profiles can be defined as high-throughput biochemical measurements of the abundance and/or structural features of molecules involved in the main biological processes such as metabolism and its regulation. As depicted in Fig. 11.1, OMICs data are supported by a large range of molecules including DNA, RNA, proteins, as well as (potentially inorganic) small molecules. By nature, OMICs data are heterogeneous in terms of their

1. *Dimension*: From tens to millions of features being measured
2. *Nature*: OMICs data are either binary, categorical, counts, or continuous.

	Supporting Structure	Platforms (log <sub>10</sub> order of magnitude)	Features
 <b>Genome</b>	DNA	Microarrays (6) Sequencing (9)	Categorical data Distance-driven correlation Extremely stable over time
 <b>Epigenome</b>	DNA methylation Histone modifications Non-coding RNA	Microarrays (5) Bisulfite sequencing (1)	Continuous data Affected by time and exposures (with reduced pleiotropy)
 <b>Transcriptome</b>	mRNA	Microarrays (5) RNA sequencing (9)	Continuous data Affected by time and exposures Strong measurement noise
 <b>Proteome</b>	Proteins	Microarrays (5) Mass spectrometry (5)	Continuous data Affected by time and exposures
 <b>Metabolome</b>	Small molecules	Mass spectrometry (5) NMR spectroscopy (4)	Continuous data Structured correlation Strongly affected by exposures

**Fig. 11.1** Overview of the main types of OMICs data available (from Chadeau-Hyam et al. 2013)

3. *Complexity*: Both the intensity and the complexity of the correlation patterns between features assayed within single OMICs can be simply distance-driven, or involve distant and multivariate correlations.
4. *Stability/volatility*: OMICs data do not respond with the same intensity and dynamics to external stresses.

These sources of heterogeneity define specific statistical challenges to enable the full exploration of OMICs data. However, it also confers OMICs data a large level of biological complementarity, which, coupled with their high dimensionality, provides an agnostic view of the cellular activity and its regulation at different molecular levels, and therefore has the potential to identify effective and functional alterations induced by external stressors.

The characterization of the internal exposome then first relies, through OMICs profiling, on the identification of biological/physiological signals that are associated to the development of adverse health outcomes and, second, on the identification of potential (sets of) exposures driving these alterations. Hence, to fully characterize the internal exposome it needs to be complemented by the external exposome, which includes the full set of external exposures and experiences potentially triggering the biology and health-relevant factors.

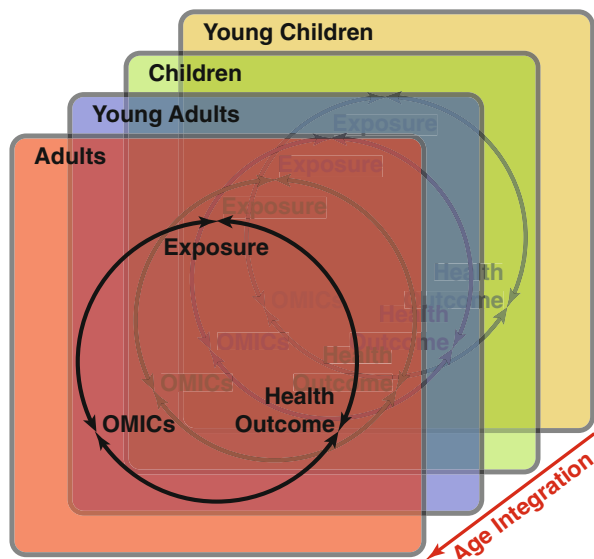
The analysis of OMICs profiles in relation to exposome factors imposes specific methodological constraints first relating to the fact that (sets of) exposures are expected to have subtle and complex effects, hence challenging the statistical power of the screening models. To improve statistical power, complex study designs featuring multiple observations per participants such as experimental crossover designs are warranted (Vineis et al. 2016). In addition, exposure mixtures are supposed to play a more important role than each of the exposures taken separately (Carlin et al. 2013; Dominici et al. 2010; Rider et al. 2013), which calls for these models to be able to handle multivariate predictors, to account for the complex biological responses to sets of exposures, multivariate outcomes, and possibly include interactions between exposures.

As a result, a full exposome dataset is complex and, for a given health outcome and a given age range, it comprises three compartments: one health outcome, one set of exposures, and one set of OMICs profiles (Fig. 11.2).

The full exploration of an exposome dataset involves three main sets of analyses. First, Exposures–health outcome analyses aiming at identifying (sets of) exposures that are involved in the development of an adverse condition. Similarly, OMICs–health outcome profiling is seeking to identify molecular alterations that are related to the development of the condition of interest. Resulting associations comprise putative early disease manifestations, effective physiological changes affecting the risk of the outcome, or exposure-related marks that are (directly or within a pathway) affecting the health outcome of interest. Finally, OMICs–exposure profiling aims at characterizing the internal exposome as defined by the physiological response to sets of external stressors.

To fully exploit the entire dataset, several levels of integration are warranted. First, when multiple OMICs profiles are available in the same individuals, ‘cross-

**Fig. 11.2** Schematic representation of an exposome dataset for a given health outcome



OMICs' analyses investigate how disease or exposure-related signals found at one molecular level correlate to those found at another level. As such, these multi-OMIC analyses have the potential to disentangle redundant and complementary molecular signals, and have the potential to inform the molecular cascades triggered by exposure(s) and/or involved in the development of an adverse condition. A second level of integration involves the investigation of the common information between markers of exposures and markers of the health outcome. This idea has been formalized as the “meet-in-the-middle” concept (Assi et al. 2015; Vineis and Perera 2007; Chadeau-Hyam et al. 2011) and may help identifying intermediate biomarkers that are on the molecular pathway linking exposure to health outcome.

The complexity of the exposome characterization is further challenged by its dynamic nature. First, exposure levels naturally vary in time and at different time resolution scales: for example, chronically, within a day or following environmental changes, or across historical periods. Furthermore, the effect of exposures may respond to different dynamics (e.g., acute or chronic effect), and there are possible age-related effect modifications and susceptibility functions that may regulate the effect of exposures during the life course, defining critical life stages where exposures have a greater effect.

To account for the dynamics of the exposome, study designs must be adapted. This includes the implementation of experimental designs in which the participants are exposed to different levels of exposure, to enable the exploration of acute effects of exposures (Font-Ribera et al. 2010; McCreanor et al. 2007). Personal exposure monitoring campaigns capture effect of exposures at the scale of days and weeks, and long-term effect of exposure relies on exposure modeling applied to cohort studies.

Age-related susceptibility functions and effect modification could be formally modeled, but in the absence of either a single study covering all ages or studies with multiple measurements in the same individuals at different life stages, it could be restricted to investigating OMICs-Exposure-Health outcomes relationship in each age class separately (see Fig. 11.2) and subsequently seeking for signals that are common or specific to certain age classes.

Altogether, the analysis of a full exposome dataset can be decomposed in three main analytical streams which will be detailed in the remainder of this chapter: OMICs and Exposure profiling techniques, methods to improve results interpretation notably through the integration of prioritized OMICs signals, and methodological developments to ensure higher dimensional OMICs integration. In a concluding section of this chapter, we will define the notion of the “mechanome” and will present possible analytical framework enabling its exploration.

## Main Approaches for OMIC-Exposure Profiling

Over the past decade, technological developments in molecular biology have given rise to a large amount of complex OMICs datasets enabling in-depth investigation of both physiological responses to external exposures (e.g., biomarkers of exposure and early effect), and of internal signatures of health outcomes (e.g., biomarkers of disease risk and onset). Concurrently to the emergence of this high-resolution data, strong methodological efforts have been carried out to enable their exploration. Resulting methods are now established and have been reviewed (Balding 2006; Chadeau-Hyam et al. 2013; Agier et al. 2016).

As mentioned above, one key feature to be accounted for while analyzing OMICs and exposome data is the correlation structure existing across the variables. For OMICs data, both the strength and the complexity of correlation structure are heterogeneous across different types of OMICs data. For instance, correlation between genetic markers is mainly distance-driven resulting in nearby genetic variants being strongly correlated. For other OMICs data such as NMR-based metabolomics data, the correlation patterns respond to more complex patterns including (1) a local component resulting in two nearby features relating to the same compound, (2) a nonlocal component resulting from the fact that a single compound could be reflected at different spectral regions, and (3) a functional component reflecting that an observed biological phenomenon is unlikely to be driven by a single compound. Similarly, external exposome data comprise large ranges of data sets including for instance biomarkers, exposure measurements, and behavioral factors. As a result, the correlation patterns within and across families of exposures are complex (Fig. 11.3).

Despite their specificity, most OMICs data, and in a lesser extent, external exposome data share a high-dimensional nature corresponding to the so-called “small  $n$ , large  $p$ ” situation, in which the number of measurements  $p$  is large and can even exceed the number of observations  $n$  (Fig. 11.4).

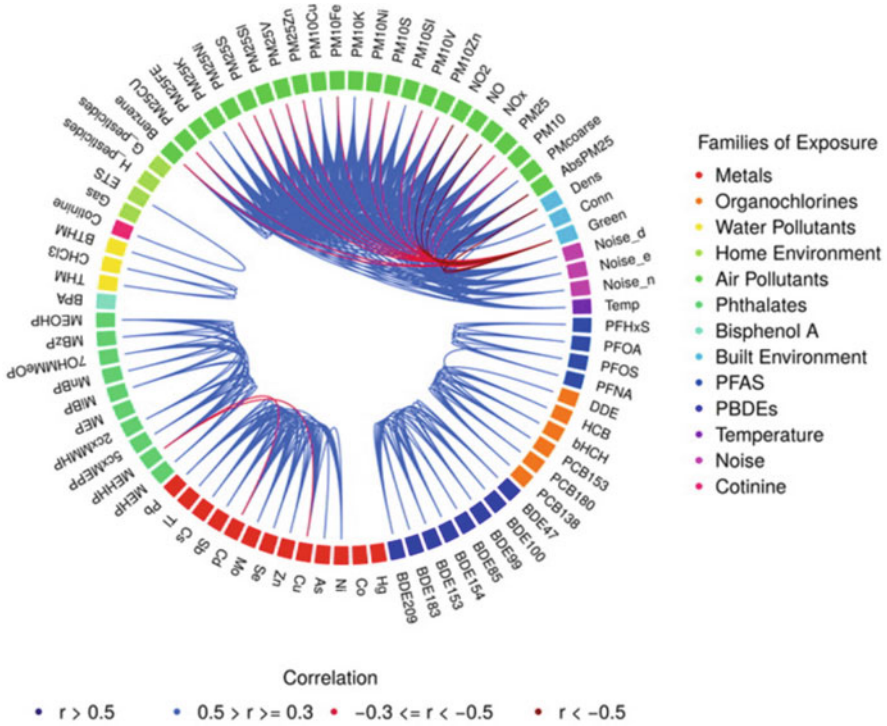


Fig. 11.3 Circos plot representation of the correlations within external exposome data (from Robinson et al. 2015)

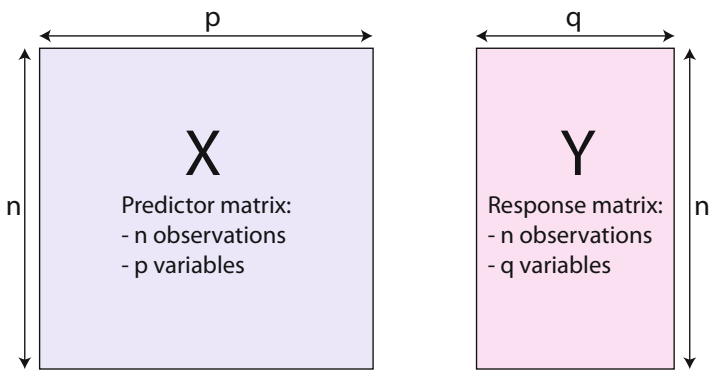


Fig. 11.4 Illustration of the “small  $n$ , large  $p$ ” situation (from Chadeau-Hyam et al. 2013)

In this setup, statistical inferences based on classical approaches are at best biased, and generally numerically intractable. Three main approaches have been proposed to accommodate this situation:

1. *Univariate approaches* assessing separately the association between each variable in the predictor matrix (exposures or OMICs data) and the outcome of interest. These are coupled with multiple testing correction strategies.
2. *Dimensionality reduction techniques* building upon the correlation within the data to construct a summary of lower dimension at the cost of a minimal loss of information.
3. *Variable selection techniques* assuming that not all predictors are relevant to the outcome of interest and seek for a sparse subset of predictors mostly related to the outcome.

### *Univariate Models and Multiple Testing Correction*

A first approach to accommodate the “large  $p$ , small  $n$ ” situation consists in considering each variable in the predictor matrix separately. The association between each predictor and outcomes of interest is assessed and tested using the same statistical model. For a set of  $p$  predictors regressed against a single outcome, a total of  $p$  tests are then being performed, each with the same per-test significance level  $\alpha'$  measuring the risk of rejecting the null hypothesis of no association while it is true (type I error). Conducting  $p$  tests results in an inflated number of expected false positive findings, which, assuming independence of the tests, equals  $\alpha' \times p$ . In order to control the inflated number of associations that are falsely declared as statistically significant across all performed tests, techniques to control the Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR) have been proposed.

The outcome of  $p$  tests can be summarized by the following table, where  $V$  is the number of false positive findings: the number of times the null hypothesis has been rejected while it was true,  $S$  is the number of true positive findings, and  $R$  is the number of positive calls across the  $p$  tests performed.

	$H_0$ True	$H_0$ False	Total
$H_0$ rejected	$V$	$S$	$R$
$H_0$ not rejected	$U$	$T$	$W$

The FWER is defined as

$$\text{FWER} = p(V \geq 1).$$

It represents the probability of drawing at least one false positive conclusion across the  $p$  tests. Methods to control FWER can involve the calculation of the per-test significance level  $\alpha'$  to be applied to each of the  $p$  tests to ensure that the actual  $\text{FWER} \leq \alpha$ , which is equivalent to ensure that  $1 - \text{FWER}$ , the probability not to draw any false positive, is greater than  $1 - \alpha$ . In order to control the FWER, one intuitive approach is to adjust (i.e., reduce) the per-test significance level  $\alpha'$  to be applied to each test, as proposed by the Bonferroni or Šidák corrections where

$\alpha' = \frac{\alpha}{p}$  and  $\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}}$ , respectively. Both correction strategies are known to be stringently protecting against the type I error. However, due to the correlations existing across the predictors (OMICs data or exposome features), across the actual  $p$  tests performed, the same information is (at least partially) tested several times. Less conservative corrections account for this redundancy and some rely on calculation of the effective number of tests (ENT) as defined by the virtual number of independent tests performed across the actual  $p$  tests performed. One direct way to approach the ENT estimation is to perform eigen analysis of the variance–covariance matrix and subsequently apply either a Bonferroni or a Šidák correction using the estimated ENT (Patterson et al. 2006). While this approach is computationally efficient, it is numerically limited as the estimated ENT will be upper-bounded by the number of observations (Schafer and Strimmer 2005). As an alternative method, which is scalable to smaller sample sizes, estimation via resampling procedures such as permutations have been proposed (Castagne et al. 2017; Chadeau-Hyam et al. 2010; Hoggart et al. 2008; Westfall and Young 1993). In practice, for a given set of predictors and a given statistical model, the responses(s) are randomly shuffled. The resulting permuted data set mimics the null hypothesis of no association. Regressing the  $p$  predictors against the permuted (set of) outcome, the minimal  $p$ -value (noted  $q$ ) represents the maximal per-test significance to be applied not to draw any false positive conclusion. Repeating that permutation procedure, one can estimate the distribution of  $q$ , from which the per-test significance to ensure a control of the FWER at a desired level can be derived. This type of approach has been successfully applied in genetics (Zou et al. 2004; Dudbridge and Gusnanto 2008; Hoggart et al. 2008) and metabolomics (Chadeau-Hyam et al. 2010) but can become computationally intensive.

The False Discovery Rate is the proportion of false positive findings among the significant association:

$$\text{FDR} = E\left(\frac{V}{R}\right),$$

and its control defines the expected proportion of positive findings that are allowed to be false. Methods to control the FDR are iterative procedures looking at the ordered set of  $p$   $p$ -values and comparing them to a cutoff value which depends on the rank of the considered  $p$ -value. For instance, Benjamini and Hochberg (1995) procedure compares  $p$ -values sorted in descending order to the increasingly stringent cutoff value. It returns a list of associations declared statistically significant ensuring that the FDR is upper-bounded by the desired value.

It can be shown that controlling the FWER also provides control of the FDR, and, intuitively, the FDR control is less stringent than that of the FWER. If one runs 100 experiments controlling the FWER at a 0.05 level, on average, less than 5 experiments will result in one or more false positive findings, while the FDR control over these experiments will allow that all experiments can include on average 5 false positives.



The use of univariate approaches using (generalized) linear models to analyze high dimensional data has been extremely successful mainly because (1) they are computationally efficient; (2) they are extremely flexible and can accommodate a large range of parametric and nonparametric relationships; (3) they can accommodate all types of predictors and outcomes; (4) they are readily available in most statistical packages.

However, by design, these approaches only assess the marginal effect of each predictor on the outcome of interest, and do not consider a potential joined effect of the predictors.

### ***Multivariate Models: Dimensionality Reduction and Variable Selection***

Owing to the complexity of the effect of exposures and the possible pleiotropy of their downstream consequences, there is a need to model jointly the OMICs or exposome data in relation to exposures and/or health outcomes.

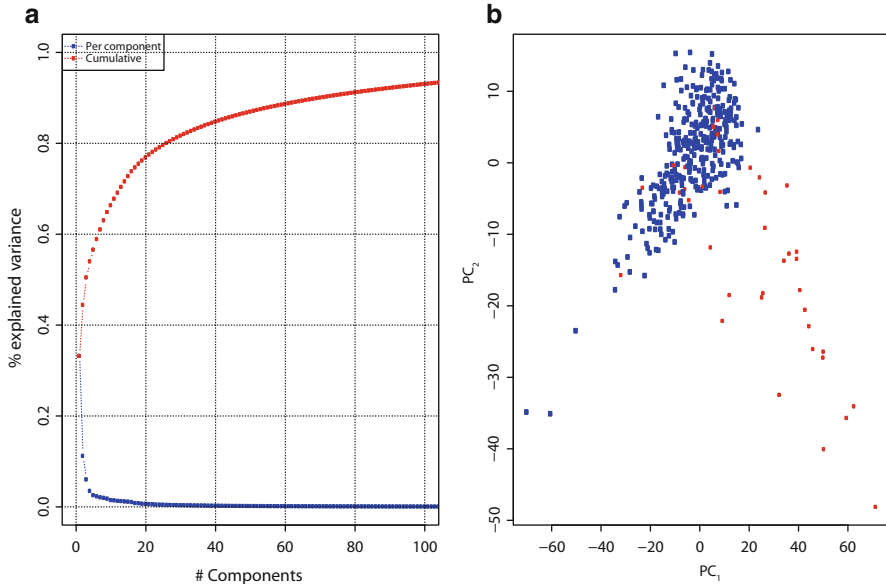
#### **Dimensionality Reduction Techniques**

Dimensionality reduction techniques use the correlation among the predictors to define summary of the original variables into a few(er) synthetic variables (the principal components, PCs) that capture the latent structure of the data. This is achieved by searching for linear combinations of variables that optimize some measure of diversity among observations. The  $i$ th principle component  $PC_i$  is then defined as

$$PC_i = \alpha_{i1}X_1 + \alpha_{i2}X_2 + \dots + \alpha_{ip}X_p,$$

where  $X_1, \dots, X_p$  denote the  $p$  original variables in the predictor matrix and  $\alpha_{i1}, \dots, \alpha_{ip}$ , the vector of linear coefficients or weights defining the contribution of each of the original variables to the  $i$ th component.

Hence, the use of dimensionality reduction techniques can be reduced to the search of the  $p$  weights vectors, which relies on eigen analysis. While considering  $p$  components, a simple rotation of the original dataset is performed (i.e., no loss of information). The principle of dimensionality reduction techniques is to identify the fewest components that minimally distort the original dataset. Across the dimensionality reduction techniques, different measures of the information are considered: the variance for Principal Component Analysis (PCA) (Hotelling 1933a, b; Pearson 1901), or the  $\chi^2$  distance for Correspondence Analysis (CA) (Greenacre 1984). From the eigen decomposition, it is possible to rank the components with respect to the proportion of information they explain. For instance, for PCA, scree plots (see



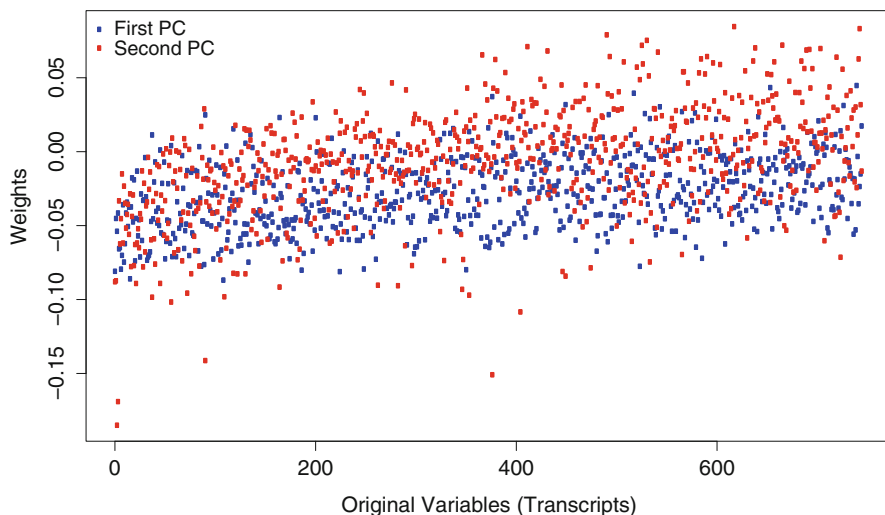
**Fig. 11.5** Scree plot of a PCA for 745 transcripts found associated with risk of chronic lymphocytic leukemia (a). In the score plot (b), individuals are colored according to their prospective disease status: cases in red and controls in blue (from Chadeau-Hyam et al. 2014b)

Fig. 11.5) represents the information restitution, as measured by the proportion of the variance of the original dataset explained by each component.

From Fig. 11.5a, it is then possible to identify the number of PCs that are required to explain more than a certain proportion of the original variance: 25 and 72 components are necessary to explain 80% and 90% of the variance, respectively. The individual contribution of each PC shows that from the 4th component onward, the proportion of the variance explained is marginal. Projection of the data on the two first components (score plot in Fig. 11.5b) suggests that the 45% explained variance by these two sole components yield good discrimination between cases and controls.

Those latent variables can subsequently be used in a (possibly multivariable) regression model to assess how the main drivers of the variation in the original set of variables are related to an outcome of interest. While these may result in numerically tractable inferences, the characterization of potential associations remains conditional on the interpretability of the components. Loadings plots (Fig. 11.6) precisely represent the contribution of each of the original variables to the PCs and can help in understanding the latent structure captured by each component.

In the example presented in Fig. 11.6, while the first two component discriminated cases from controls, none of the first two components were driven by a specific set of transcripts, which hampers the biological understanding of the underlying associations. One way to improve interpretability of the components is to ensure sparsity in the loadings coefficients which can be done through penalization (see below).



**Fig. 11.6** Loadings plots for the first two components of the PCA of the 745 CLL-associated transcripts (from Chadeau-Hyam et al. 2014b)

Principal Components Analysis (PCA) (Hotelling 1933a, b; Pearson 1901) has proved useful in genetics (Jombart et al. 2009) and has become a standard in genome-wide association studies (GWAS), where it is used to correct for population stratification (Price et al. 2006; Reich et al. 2008). PCA can accommodate both continuous and discrete data, and is not affected by the potential correlation between predictors or by a larger number of variables than observations.

However, while PCA has proven efficient in summarizing large datasets in (far) fewer dimensions, representing the latent structures in the data driving most of the variability in the original dataset, nothing guarantees that this variation is relevant to an outcome of interest. As a supervised alternative to PCA, Partial Least-Square (PLS) approaches have been proposed (Wold et al. 1984). PLS components are defined such that they maximize the covariance between the predictors and response variables. As such, PLS components not only capture as much variance of the original variable as possible, but also focus on the variance that is relevant to the outcome of interest. PLS-based methods are extremely popular in chemometrics and have been successfully applied in metabolomics (Holmes et al. 2008; Fonville et al. 2010; Yap et al. 2010). Application to other OMICs data were successful in epigenomics (Belshaw et al. 2010), transcriptomics (Musumarra et al. 2011; Fasoli et al. 2012), and proteomics (Wang et al. 2011).

Irrespective of the type of analysis, the use of latent variables in a regression context requires the specification of the number of latent variables to be considered, which usually relies on cross-validation procedures aiming at the identification of the number of components that optimizes both interpretability and prediction error.

## Penalization

As mentioned above, one intuitive way to improve results interpretability in multivariate models is to ensure that the number of variables found relevant to the outcome is as sparse as possible. Inducing sparsity hence relies on favoring a minimal set of nonredundant variables jointly associated to the outcome and penalizes irrelevant and/or irrelevant ones.

Penalization techniques have been introduced in regression models to induce sparsity in the vector of regression coefficients. The principle of penalized regression is to estimate all regression coefficients under the constraint that a function of these regression coefficients, the penalty, is bounded by a fixed value.

Of the different penalized regression approaches, Ridge Regression uses as penalty the  $L^2$  norm (Hoerl and Kennard 1970):

$$L^2 = \sum_{i=1}^p \beta_i^2,$$

where  $\beta_i^2$  is the  $i$ th regression coefficient linking the outcome and the  $i$ th variable in the predictor matrix. It can be demonstrated that the least influential predictors will see their effect size estimate shrunk toward 0, while the estimates for the most important variables will remain unchanged.

LASSO models use the  $L^1$  norm as defined by the sum of the absolute value of the  $p$  regression coefficients (Tibshirani 1996):

$$L^1 = \sum_{i=1}^p |\beta_i|$$

Intuitively, upper-bounding the  $L^1$  penalty is more demanding than constraining the  $L^2$  norm (in which regression coefficients are squared). It can be demonstrated that LASSO penalization, for geometric reasons, is able to perform variable selection and shrinks irrelevant regression coefficients exactly to 0. Conversely, while ridge regression provides stable coefficient estimates for ill-posed problems (i.e., when  $n < p$ ), it does not guarantee sparsity. Moreover, one major limitation of LASSO methods is that the number of non-penalized variables is upper-bounded by the number of observations: LASSO model cannot select more than  $n$  variables (i.e., provide more than  $n$  variables with nonzero regression coefficients). To provide a more general approach, combining advantages of both Ridge and LASSO regression, the Elastic Net has been developed (Zou and Hastie 2005) and uses a penalty which is a weighted sum of the  $L^1$  and  $L^2$  norms according to a calibration parameter  $\lambda$ .

In practice, penalized regression requires a preliminary calibration of the penalty parameter(s), which directly affects the number of selected variables, the estimates of the regression coefficients, and thus the statistical performance of the models.

Calibration procedures usually involve the minimization of the mean square error of prediction by cross-validation.

The calibrated model will return a list of shrunk regression coefficients, and for LASSO and Elastic Net, the variables with non-null shrunk coefficients are those found to be jointly associated to the outcome. Generalized version of both LASSO and Elastic Net accommodates linear, logistic (binary responses) and multinomial (categorical responses), Poisson (count data as response), and Cox (survival models) regression models (Friedman et al. 2010; Simon et al. 2011).

The penalization paradigm applies outside the scope of linear regression models and has been used to provide shrunk loadings coefficients for dimensionality reduction techniques, hence defining sparse versions of both PCA and PLS. These models ensure a sparse definition of the latent variables, thus improving their interpretability. Sparse PCA (sPCA) and sparse PLS (sPLS) have been used to analyze OMICs data (Zou et al. 2006; Shen and Huang 2008; Witten et al. 2009; Boulesteix and Strimmer 2007; Le Cao et al. 2008, 2009; Chun and Keles 2009, 2010).

### Bayesian Variable Selection Approaches

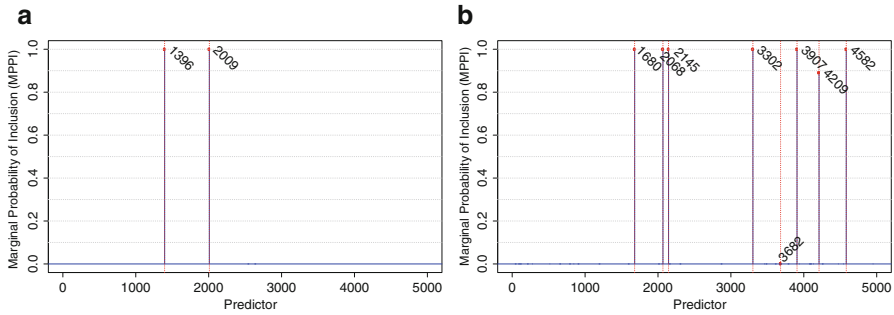
Full Bayesian variable selection (BVS) procedures have been proposed to cope with the “large  $p$ , small  $n$ ” framework (West 2003), and these rely on the estimation of the posterior distribution of the following latent binary vector:

$$\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p) \in \{0, 1\}^p,$$

where  $\gamma_i$  is a binary variable indicating if the  $i$ th variable is in the model.

In that setup, a model is defined by the subset of variables it includes, and hence by a given vector gamma (e.g., the null model corresponds to a vector  $\gamma$  comprising  $p$  0's).

The objective of BVS is to identify the best sets of variables that jointly associate to the outcome of interest, and the main challenge underlying BVS inference resides in the dimensionality of the space in which to search for the best models, which grows exponentially with  $p$  ( $2^p$ ). While there is an extensive theoretical literature on the parameterization and numerical estimation of BVS models, few implementations offer a ready-to-use interface for high-dimensional profiling. These rely on linear or generalized linear models. Shotgun Stochastic Search (SSS) was one of the first packages to be made available that enabled Bayesian large-scale genome screening through BVS (Hans et al. 2007). The search in the model space is iteratively performed such that, if at a given iteration  $q$  features are selected, all the models (1) of size  $(q-1)$ ; (2) of size  $q$  (i.e., replacing any of the variables in the current model by any of the  $p-k$  remaining ones); (3) of size  $(q + 1)$  will be evaluated and compared. In order to favor the evaluation of meaningful models, piMASS (Guan and Stephens 2011) optimizes the search strategy by specifying a proposal distribution (i.e., defining, for a given



**Fig. 11.7** Marginal Posterior Probability of Inclusion (MPPI) of the variables in one of the best models, from the R2GUESS model. Results are presented for 2 simulations using 5,000 SNP of which 2 (a) and 8 (b) are assumed to have an effect on the simulated outcome (from Liquet et al. 2016a)

predictor, the probability to be proposed for inclusion in a model) that accounts for the correlation between the predictor and outcome.

Both SSS and piMASS are able to accommodate binary, categorical, and continuous outcomes, and SSS additionally implements survival models.

GUESS and its R implementation R2GUESS is an alternative approach which accommodates multivariate outcomes through the use of a multivariate linear model (Liquet et al. 2016a; Bottolo et al. 2011). Its search algorithm is based on multiple-chain genetic algorithms, and its latest version makes use of graphics processing unit (GPU) linear algebra libraries, thus enabling its application to genome-wide scale analyses (hundreds of thousands of predictors simultaneously) (Bottolo et al. 2013).

In practice, BVS models return a list of models that have higher posterior probability. Integrating the posterior probabilities of the best models, it is also possible to infer the marginal probability of inclusion (MPPI) of a given variable in the best models. These MPPIs (see Fig. 11.7) can be viewed as a measure of the strength of association of the (set of) predictors with the outcome of interest.

In Fig. 11.7a, it is apparent that both SNPs used for the simulation were detected by the model and were always included in the best models. For the simulation using eight SNPs (Fig. 11.7b), only seven SNPs were recovered (with MPPI > 0.95), and one SNP was missed, due, notably, to its correlation with the other included SNPs.

## OMIC–Exposure Profiling in Practice: Use and Extensions

### *Quantitative Assessment of the Profiling Techniques*

In order to quantitatively assess statistical performances of the main profiling approaches, simulation studies were conducted (Agier et al. 2016; Liquet et al.

2016a). Of these, one exposome-focused study used real correlation matrices among 237 prioritized exposome features from the INMA study (Guxens et al. 2012). The wide range of simulation scenarios comprised (1) several levels of explained variance, (2) different number of “causal” exposures  $k = 1, 2, 3, 5, 10,$  or  $25,$  and (3) different effect sizes.

Across all scenarios investigated, each of the simulated dataset was analysed using a preselected set of 6 methods borrowed from the three main families defined in the previous section:

1. Univariate models:

- (a) “GWAS”-type approaches, EWAS with multiple testing correction via FDR control
- (b) EWAS followed by a multivariate regression step to limit confounding: EWAS-MLR

2. Dimensionality Reduction Techniques:

- (a) Sparse Partial Least-Square s-PLS

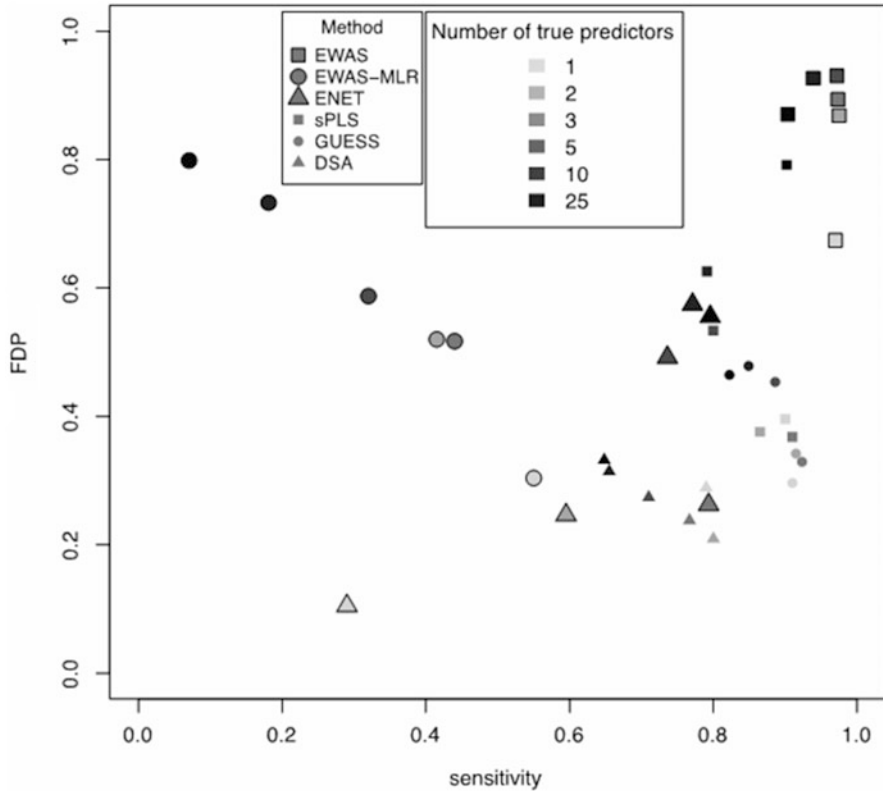
3. Variable Selection Approaches:

- (a) Elastic Net E-NET
- (b) Bayesian Variable Selection R2GUESS
- (c) Deletion/Substitution/Addition algorithm DSA, a penalized stepwise model selection approach (Haight et al. 2010).

In that simulation context, the statistical performances of the different models can be quantified and compared based on both their ability to identify the “true” predictors, used to simulate the outcome (sensitivity), and their ability not to include irrelevant exposures (specificity). In Fig. 11.8, the latter is measured using the False Discovery Proportion (FDP).

Simulation showed that across the large number of scenarios investigated, multivariate methods systematically outperformed univariate approaches to explore external exposome–outcome relationships in the presence of complex correlations (Fig. 11.8). Although these approaches did not achieve low false discovery performances, they yielded a better balance between sensitivity and FDP. Based on refined performance metrics (i.e., accounting for the correlation among predictors), DSA and R2GUESS were identified as providing somewhat better performances. From this work, it was also concluded that in real case analyses, methodological choices should also be guided by computational complexity and feasibility, as well as flexibility considerations such as the ability to accommodate confounders.

These results are in-line with additional simulation studies carried out for OMICs data (i.e., larger number of variables) in relation to simpler outcomes (Liquet et al. 2016a). These simulations showed the superiority of R2GUESS over multivariate alternatives but showed better performances overall of methods under investigation, in the absence of complex correlation across predictors (Fig. 11.9).

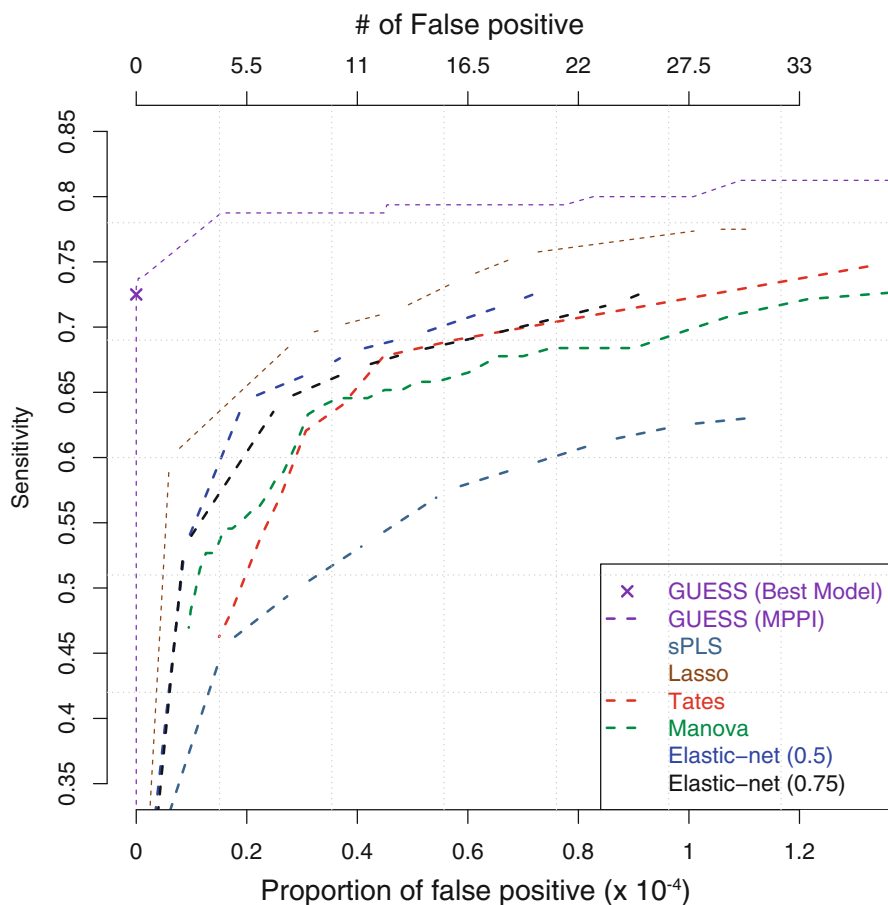


**Fig. 11.8** Sensitivity and False Discovery Proportion (FDP) yielded by six different approaches to perform exposome–outcome association study. Results are based on simulated data using realistic exposome data and considering a number of “true” predictors ranging from 1 to 25 (from Agier et al. 2016)

### *Modeling and Correcting for Nuisance Variation*

As mentioned above, both the acquisition of OMICs profiles and exposure measurements are prone to measurement error, resulting in part of the observed variability being related to technical factors during data generation rather than to the variables of interest. As a consequence, such technical variability has the potential to dilute the effects of interest and is referred to as nuisance variation. Several preventive approaches can be implemented to limit such dilution effects. First, during data acquisition through careful randomization of the samples, and notably ensuring that analytical entities (e.g., case–control pairs) are processed within the same analytical batch, and that entities are randomly distributed across batches. Second, through the preprocessing of the data, using specific normalization techniques, generally relying on quality control samples.





**Fig. 11.9** ROC curves comparing the statistical performances of established multivariate OMICs profiling approaches. Results were obtained using 20 independent simulated datasets including 273,675 SNPs in 3122 individuals and setting 8 “true” predictors (from Liquet et al. 2016a)

Despite these preventive measures, technically induced noise can persist in the data, and may therefore be accounted for during the statistical analyses of the data.

Linear mixed models can be implemented to account for nuisance variation (McHale et al. 2011; Chadeau-Hyam et al. 2014b). These include random effects in the statistical model to account for a structure in the variance of observations. This structure and its strength are assumed to depend on nuisance parameters (e.g., technical covariates) and enter the linear model in the form of additional terms that depend on the nuisance parameters. These additional terms can be estimated by likelihood or restricted likelihood maximization (Lindstrom and Bates 1990). One possible formulation of mixed models for a given variable ( $Y$ ) (e.g., one OMICs measurement) is, for one individual  $i$ :

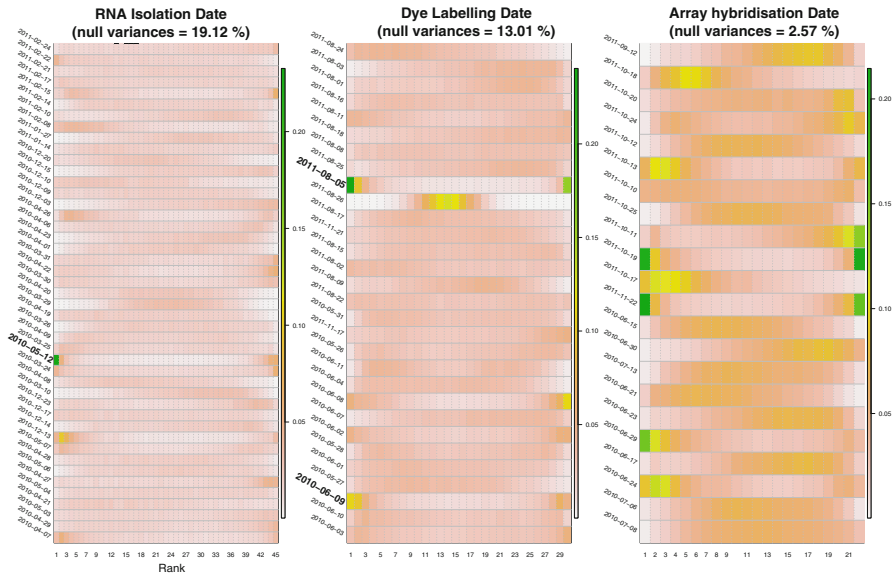
$$Y^i = (\alpha + u_{A^i}) + (\beta_1 + \beta_{A^i})X^i + \beta_2 FE^i + \varepsilon^i \quad (1),$$

where  $\alpha$  is the intercept of the model,  $\varepsilon^i$  is the residual error, and  $X^i$  is the outcome of interest (e.g., case-control status). The resulting effect size estimate  $\beta_1$  can be interpreted as the change in the response variable ( $Y^i$ ) per unit change in the variable of interest  $X^i$ .  $FE^i$  is a vector of fixed effects (typically confounders) observed for individual  $i$  and corresponding regression coefficients are compiled in the vector  $\beta_2$ . Nuisance variation is modeled through a random intercept  $u_{A^i}$  and a random slope  $\beta_{A^i}$ , where the grouping factor  $A^i$  compiles the technical factors describing how data from sample  $i$  were generated. In that setup, the random intercept captures a systematic shift in the measurement of  $Y^i$ , which is related to experimental conditions, while the random slopes would account for a systematic experimentally induced attenuation (or amplification) of the relationship linking the measurement and the variable of interest. Nuisance variation is generally modeled through a random intercept (i.e., neglecting the random slope), and the random intercept  $u_{A^i}$  represents the shift associated to  $A^i$ , the value of the random effect variable(s)  $A$  observed for individual  $i$ . For example, in a study of microarray-based gene expression data (Chadeau-Hyam et al. 2014b), the dates of the three main steps of sample processing were used as random effect variables: RNA isolation, hybridization, and dye labeling.

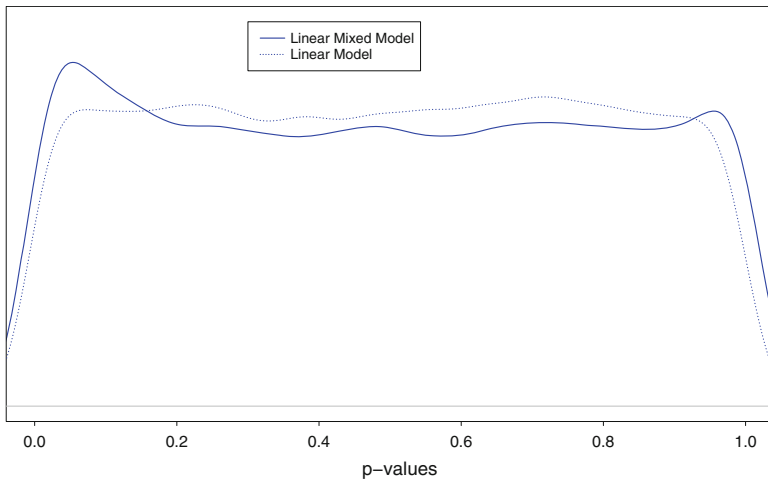
Random intercept estimates over all assayed transcripts can be summarized by their variance, and more specifically the number of times each of the isolation, hybridization, and labeling steps was estimated to generate marginal noise (i.e., null variance). In this example, the proportion of null variance estimates was 19, 13, and 2% for isolation, labeling, and hybridization, respectively, suggesting that hybridization generated more noise than the two other processing steps. Random effect estimates can be further investigated by analyzing, for each of the three random effect covariates, the ranking of each date with respect to the estimated random intercept, as depicted in Fig. 11.10.

Estimates suggest that some dates were yielding higher variances (i.e., more noise), for instance, it seems that samples whose RNA was isolated on the May 12, 2010, were associated to higher noise. The impact of the nuisance variation on subsequent inferences can be assessed by comparing the distributions of  $p$ -values from the linear mixed models to those from a linear model (i.e., setting the random intercept to 0). In Fig. 11.11, linear models exhibit a typically null  $p$ -value distribution while the inclusion of a random intercept provides a stronger support for the alternative hypothesis, with a sharper peak at smaller  $p$ -values.

The inclusion of random effects to model and correct for nuisance variation is not restricted to univariate models and can be used for penalized regression, and BVS. While theoretically, these models could natively be used to account for nuisance variation, they may become computationally demanding and may yield convergence and calibration issues. Moreover, there is no integrative solution to explicitly model technical confounding while using dimensionality reduction techniques. In both cases, one possibility is to adopt a two-step strategy first fitting a linear mixed model as defined in (1). Once fitted, the model will return estimates of the random effects, and subtracting from the observed value ( $Y^i$ ) the random effect estimates  $u_{A^i}$



**Fig. 11.10** Summary of the random intercept estimates from a linear mixed model investigating the relationship between ( $N = 29,662$ ) transcripts and future onset of lymphoma. Random effects variables are the dates of the three main processing steps of each biosample underwent. Estimates are summarized by the proportion (across the 29,662 estimates) of null variance estimated, and modality of each grouping factor (i.e., dates) is characterized by the rank of their estimated variance (from Chadeau-Hyam et al. 2014b)



**Fig. 11.11**  $P$ -value distribution from a linear mixed model (plain line) and a linear model (dotted line) assessing the relationship between gene expression level and future onset of lymphoma. Results are based on the 29,662  $p$ -values from both models (from Chadeau-Hyam et al. 2014b)

provides measurement in which the effect of the potential technical confounder has been removed. Resulting “denoised” data can subsequently enter any statistical model and will provide results corrected for nuisance variation (Castagne et al. 2016; Chadeau-Hyam et al. 2014b; Guida et al. 2015).

### *Accommodating Complex Study Designs*

In order to improve statistical power to detect the possibly complex (and hence multivariate) biological responses to external stressors, more complex study designs are warranted. These include intervention studies where participants are placed in several controlled environments, presenting exposure contrasts. Developments over these designs can also account for differential response time and may feature in each environment, multiple bio-sampling at different time points.

Irrespective of the detailed design, all such study features repeated measures (of exposure and/or OMICs data) for each participant. One natural way of modeling repeated measurement data is to adopt, in a univariate context, a linear mixed model approach, setting the participant ID as a random effect variable. This model will decompose the within and across individual variation to identify OMICs changes related to changes in exposure. In that context,  $k$  observations per participants are considered. The linear mixed model setting the individual ID as random effect assumes a simple variance–covariance structure across the  $k$  observations which only depends on the individual (see below for  $k = 6$ ).

$$\begin{pmatrix} \sigma^2 & \delta & \delta & \delta & \delta & \delta \\ \delta & \sigma^2 & \delta & \delta & \delta & \delta \\ \delta & \delta & \sigma^2 & \delta & \delta & \delta \\ \delta & \delta & \delta & \sigma^2 & \delta & \delta \\ \delta & \delta & \delta & \delta & \sigma^2 & \delta \\ \delta & \delta & \delta & \delta & \delta & \sigma^2 \end{pmatrix}$$

The advantage of the linear mixed model approach is that it can be scaled to any other type of models using linear mixed versions of univariate models or variable selection approaches, or adopting a multilevel extension for dimensionality reduction techniques (Liquet et al. 2012).

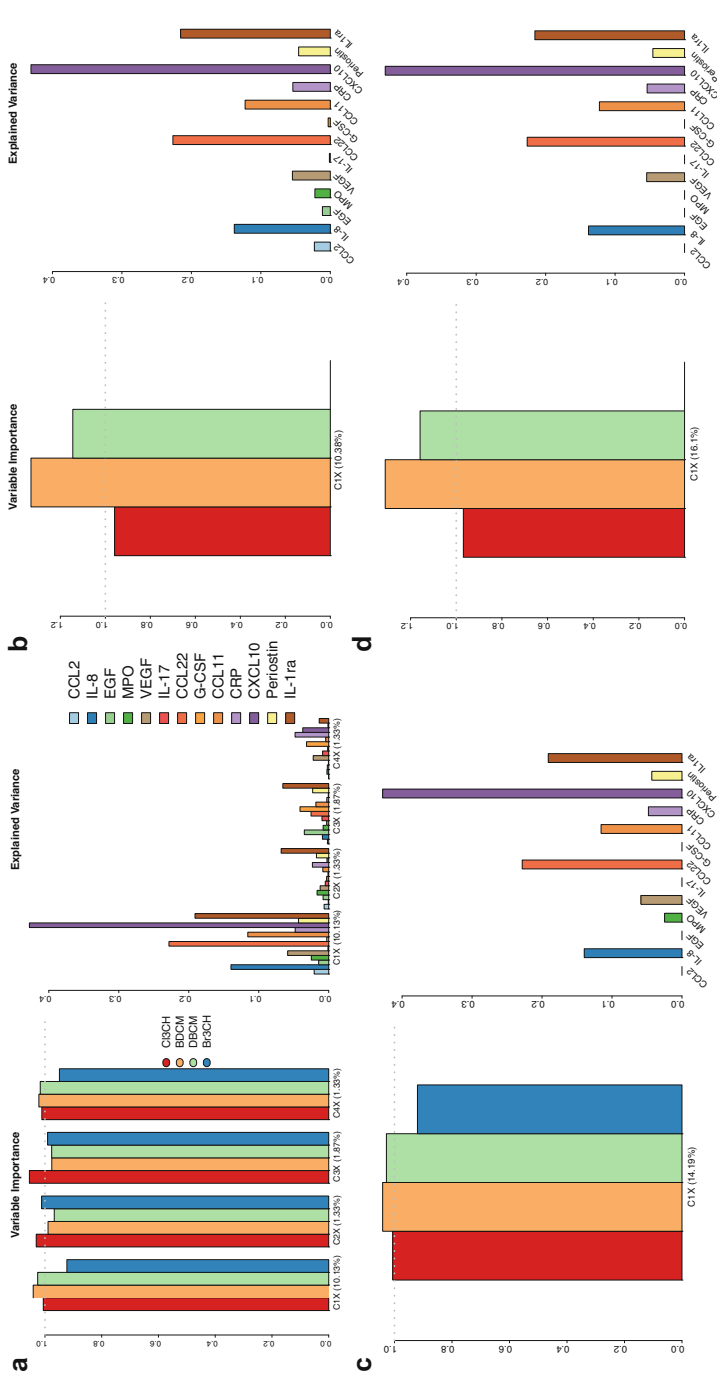
Additional flexibility could be gained by adopting a multivariate normal (MVN) model where the variance–covariance matrix depends not only on individuals, but also on each of the  $k$  experimental conditions:

$$\begin{pmatrix} \sigma_{11}^2 & \delta_{12} & \delta_{13} & \delta_{14} & \delta_{15} & \delta_{16} \\ \delta_{21} & \sigma_{22}^2 & \delta_{23} & \delta_{24} & \delta_{25} & \delta_{26} \\ \delta_{31} & \delta_{32} & \sigma_{33}^2 & \delta_{34} & \delta_{35} & \delta_{36} \\ \delta_{41} & \delta_{42} & \delta_{43} & \sigma_{44}^2 & \delta_{45} & \delta_{46} \\ \delta_{51} & \delta_{52} & \delta_{53} & \delta_{54} & \sigma_{55}^2 & \delta_{56} \\ \delta_{61} & \delta_{62} & \delta_{63} & \delta_{64} & \delta_{65} & \sigma_{66}^2 \end{pmatrix}$$

Both approaches have been successfully used to identify OMICs biomarkers related to acute changes in environmental exposures. Despite limited sample sizes, the MVN approach is showed to be efficient in identifying metabolomic, transcriptomic, and inflammatory changes due to acute experimentally induced changes in exposure to water disinfection by-products (van Veldhoven et al. 2017; Vlaanderen et al. 2017; Espin-Perez et al. 2018).

Exposome studies investigate complex effects of exposures that may act jointly to generate a complex biological response. Such exploration calls for the use of methods handling multivariate exposures and multivariate responses. PLS approaches are able to handle that situation, and multilevel extensions of these approaches further accommodate multiple observations per participant. In essence, multilevel PLS first decomposes the observed variability into within- and between-individual variability (Liquet et al. 2012). The former captures differences between individuals, including confounding factors, and the within-individual variability is measuring the changes in both exposures and responses between the different measurements, hence measuring the effect of the experiment. The latter is subsequently included into a standard PLS model to identify linear combinations of exposures that are best able to explain effects on responses. A recent proof of principle analysis focused on the inflammatory response (measured through blood levels of 13 inflammatory-related proteins) to acute exposure to disinfection by-products (DBP) while swimming in a chlorinated pool (Jain et al. 2018).

PLS identifies the latent variable in  $X$  capturing most of the variability in exposures that is relevant to the inflammatory profile. The PLS components of exposures can be ordered as per their relevance to the full set of proteins, and symmetrically, the components of proteins are defined and ordered according to their covariance with exposures. The relevance of each component can be quantified by the proportion of variation it explains. The proportion of variance in  $X$  (or  $Y$ ) explained by a given component of  $X$  (or  $Y$ ) measures how accurately that single component summarizes the entire information contained in the original  $X$  (or  $Y$ ) matrix. The percentage of variance of  $Y$  explained by the components of  $X$  measures the relevance of the information summary provided by the PLS components to the outcome matrix. This does not only depend on the quality of the component summary but also on the correlation between  $X$  and  $Y$ . In addition, Variable Importance in Projection (VIP) scores quantify the contribution of each original predictor (here exposure) to the overall explanatory performances of a given PLS component. Empirically, a VIP score  $< 1$  indicates low-to-moderate contribution of a variable (Fig. 11.12a). Sparsity can be induced through penalization in the definition of the PLS components (Chun and Keles 2010). When penalization is applied to PLS



**Fig. 11.12** Results from a multilevel PLS analysis of the exposure to 4 DBP while swimming in a chlorinated pool in relation to the blood level of 13 inflammatory-related proteins. Variable Importance in Projection (VIP) plots and proportion of variance explained by protein. Results are presented for PLS model (a), for sparse PLS performing variable selection on exposures (b), on proteins (c), and both on exposures and proteins (d) (from Jain et al. Submitted)

components of  $X$  (Fig. 11.12b), the resulting sparse PLS (sPLS) model shrinks the loadings coefficients towards zero for the least informative variables (exposures) and hence helps identifying the exposures mostly affecting inflammatory profiles. Symmetrically, variable selection can be performed on the responses (proteins, Fig. 11.12c), in order to identify the proteins whose expression is mostly affected by exposures. In a final step, variable selection can be performed on both exposures and proteins (Fig. 11.12d).

These results suggest that BrCH3 contributed less than other exposures to the inflammatory response to the swimming experiment and that among the 13 assayed proteins, 8 were more affected by the experiment. The application of multilevel-PLS models was successful, despite strong correlation and co-occurrence of the exposures, in identifying the most relevant exposures, and the proteins mostly affected by exposures.

## From Improved Interpretability to Mechanome Characterization

Full exploitation of the rich sets of results yielded by OMICs and exposome profiling techniques relies on their biological interpretation (and potentially, validation). While for some molecular entities, interpretation can be eased by knowing the functional role of the molecule (e.g., proteins, or in a lesser extent, transcripts), interpretability can become challenging when the function of the putative biomarker is unknown. Interpretability becomes even more challenging when OMICs profiles are related to complex exposures as they encompass a multivariate combination of environmental and biochemical factors.

To address this issue, a natural approach is to adopt the “meet-in-the-middle” paradigm (Vineis and Perera 2007) and explore as exhaustively as possible which factors may affect the level of exposure-related biomarkers. This two-step strategy can help in identifying molecular alterations that are associated to external stressors and health outcomes. While a natural way to identify “meet-in-the-middle” associations is through univariate models, several examples and recent developments used multivariate approaches (Assi et al. 2015; Chadeau-Hyam et al. 2011).

Biological interpretability of molecular alterations/features identified in high-throughput profiling is highly dependent on the functional characterization of the assayed molecules. Typically, interpretation of the results of a targeted proteomic assay is easier as the biological functionality is documented. While gene expression profiles govern RNA translation, the overall gene expression regulation may be multivariate (i.e., involve several transcripts) and pleiotropic (i.e., through a complex cascade involving other genes and transcripts). Hence, direct interpretation of the results from a transcriptome-wide association study (i.e., based only on the genomic location of the identified transcripts) should be complemented by an investigation of

biological pathways potentially affected by differential expression. Ontology-based tools interrogating existing databases are rich sources of information to infer biological pathways corresponding to the candidate biomarkers identified. Specifically, gene enrichment analyses assess if, and to what extent, the list of candidate transcripts is enriched for specific pathways, basically checking if the distribution of identified transcripts is significantly different from what would be expected if candidate transcripts were chosen at random (The Gene Ontology Consortium 2017; Ashburner et al. 2000; Huang et al. 2008).

For metabolomic data, feature annotation and signal interpretation can also rely on database interrogation and pathway identification. Recent developments include an efficient and reliable tool (Li et al. 2013), which has proven efficient in identifying molecules and corresponding pathways from full-resolution mass-spectrometry profiles.

For other OMIC profiles, such as DNA methylation data, results interpretation can be more challenging, as there is, to date, no established database linking the CpG site-specific levels to their downstream consequences or to sets of general biological pathways. In the absence of such information, biological interpretation of CpG sites found differentially methylated can be done by linking them to other OMIC data measured in the same individuals, and whose functional role is better characterized.

### ***OMICs Integration: An Intuitive Approach***

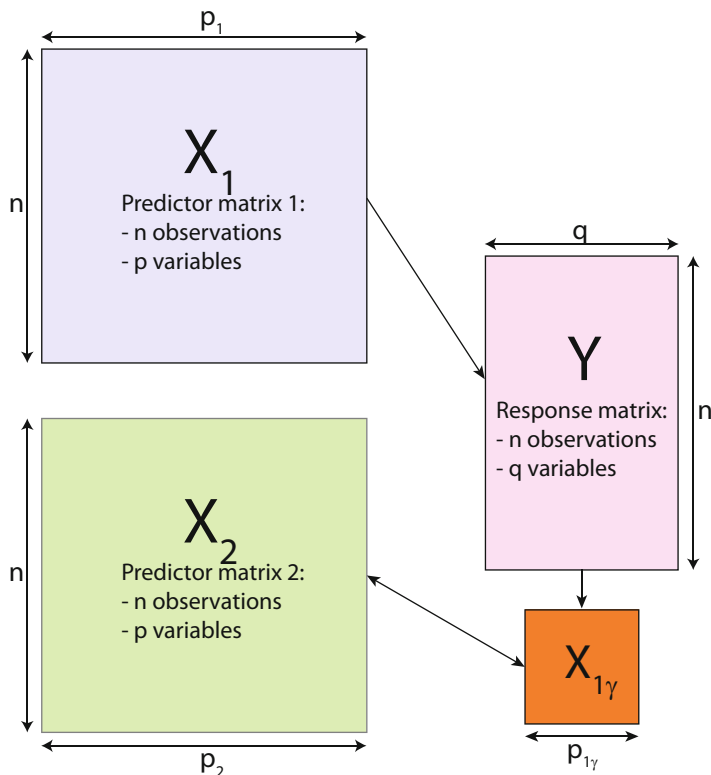
One intuitive approach to OMICs integration relies on the following two-step strategy corresponding to a targeted OMIC integration (Fig. 11.13):

1. Regressing the first matrix ( $X_1$ , e.g., methylation data) against the outcome ( $Y$ ) in order to identify  $X_{1Y}$ , a subset of variable of size  $p_{1Y}$  associated to  $Y$ . That list of  $p_{1Y}$  candidate biomarkers can be derived from  $p_1$  univariate models or a single multivariate model.
2. Regressing the  $p_{1Y}$  outcome-associated OMICs (e.g., smoking-related CpG sites) against the full-resolution second OMIC profile ( $X_2$ , e.g., transcripts). Adopting a univariate approach that would correspond to  $p_{1Y} \times p_2$  tests.

This strategy has been applied in several studies, and in particular in the study of smoking-induced DNA methylation alterations (Guida et al. 2015), where blood-derived DNA methylation profiles from Illumina Infinium HumanMethylation450 BeadChip from ( $N = 745$ ) participants of EPIC and NOWAC cohorts were used to identify 751 differentially methylated CpG sites in relation to smoking history. In a subset of that study population ( $N = 271$ ), genome-wide gene expression profiles obtained from the Illumina HumanWG-6 array (assaying  $N = 8952$  genes) were also available.

Using a univariate linear approach, each of the  $751 \times 8952 = 6.72 \times 10^6$  pairwise associations were tested and 5,636 CpG-transcript pairs (corresponding to





**Fig. 11.13** OMICs integration, a two-step strategy

265 unique CpG sites, and 426 genes) were found significantly associated. Most CpG-transcript pairs were inversely associated suggesting that hypermethylation is associated to a gene down-regulation, and in case a CpG site was associated to several transcripts, the sign on these associations is generally consistent across transcripts.

The list of transcripts found associated to the smoking-related differentially methylated sites can be used as input for gene enrichment analyses, which, in that example, identified relevant biological pathways involved in the effect mediation (through methylation changes) of the exposure to tobacco smoke.

Another striking result from this analysis is that of the 5,636 statistically significant CpG transcript pairs, only 5 involved a CpG and a transcript located on the same gene. This suggests that regulatory cascades affected by exposure-induced methylation changes are complex and involve distant (trans) associations.

This supports the fact that there is no justification for reducing OMICs integration to local interactions or correlations, and there is a clear need to extend screening

approaches to explore long-distance relationships. In that context, dimensionality reduction techniques, and Bayesian variable selection approaches handling multivariate  $X$  and  $Y$ , may be considered, and already exist in a sparse version. However, while interpretability is classically sought for by inducing sparsity, it may not be sufficient to ensure a detailed understanding of the complex patterns exhibited by OMICs data integration, and the inclusion of prior knowledge about functionally relevant structures within and across OMICs profiles may be necessary.

### Further Approaches to OMICs Integration

PLS algorithms have proven efficient in the task of selecting correlated sets of signals across two blocks of high-throughput data (Le Cao et al. 2008; Parkhomenko et al. 2009). In order to exploit prior knowledge on the structure existing in the data, potential grouping of the covariates within each block of data can be envisaged (Zhou et al. 2010). Recently, based on a novel penalty function controlling the number of groups to be selected and the sparsity within each group, a group and sparse group PLS (gPLS and sgPLS, respectively) method has been proposed that improves both sparsity and interpretability (Liquet et al. 2016b). In practice, for sparse PLS models, the components are defined as

$$C^X = \underbrace{\alpha_1 X_1}_{\neq 0} + \underbrace{\alpha_2 X_2}_{=0} + \dots + \underbrace{\alpha_k X_k}_{\neq 0} + \underbrace{\alpha_{k+1} X_{k+1}}_{\neq 0} + \dots + \underbrace{\alpha_k X_k}_{=0}$$

where loadings for the unimportant variables are shrunk to 0 (e.g.,  $X_2$  and  $X_k$  in the example). For group PLS, the model is given *a priori* a group structure and will select the entire group (groups  $i$  and  $j$  in the example below):

$$C^X = \overbrace{\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \dots + \alpha_k X_k + \alpha_{k+1} X_{k+1}}^{\text{group } i} + \dots + \overbrace{\alpha_{p-2} X_{p-2} + \alpha_{p-1} X_{p-1} + \alpha_p X_p}^{\text{group } j}$$

$\underbrace{\alpha_1 X_1}_{=0} \quad \underbrace{\alpha_2 X_2}_{=0} \quad \underbrace{\alpha_3 X_3}_{=0} \quad \underbrace{\alpha_k X_k}_{\neq 0} \quad \underbrace{\alpha_{k+1} X_{k+1}}_{\neq 0}$   
 $\underbrace{\alpha_{p-2} X_{p-2}}_{\neq 0} \quad \underbrace{\alpha_{p-1} X_{p-1}}_{\neq 0} \quad \underbrace{\alpha_p X_p}_{\neq 0}$

For sparse group PLS models (sgPLS), the component will be defined by selecting or not a given group, and within the selected groups it will only select the most relevant variables ( $X_k$  in group  $i$  and  $X_p$  in group  $j$  in the example below):

$$\begin{aligned}
 C^X = & \underbrace{\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3}_{\substack{\text{group 1} \\ =0}} + \dots + \underbrace{\alpha_k X_k + \alpha_{k+1} X_{k+1}}_{\substack{\text{group } i \\ \neq 0}} \\
 & + \dots + \underbrace{\alpha_{p-2} X_{p-2} + \alpha_{p-1} X_{p-1} + \alpha_p X_p}_{\substack{\text{group } j \\ =0}} .
 \end{aligned}$$

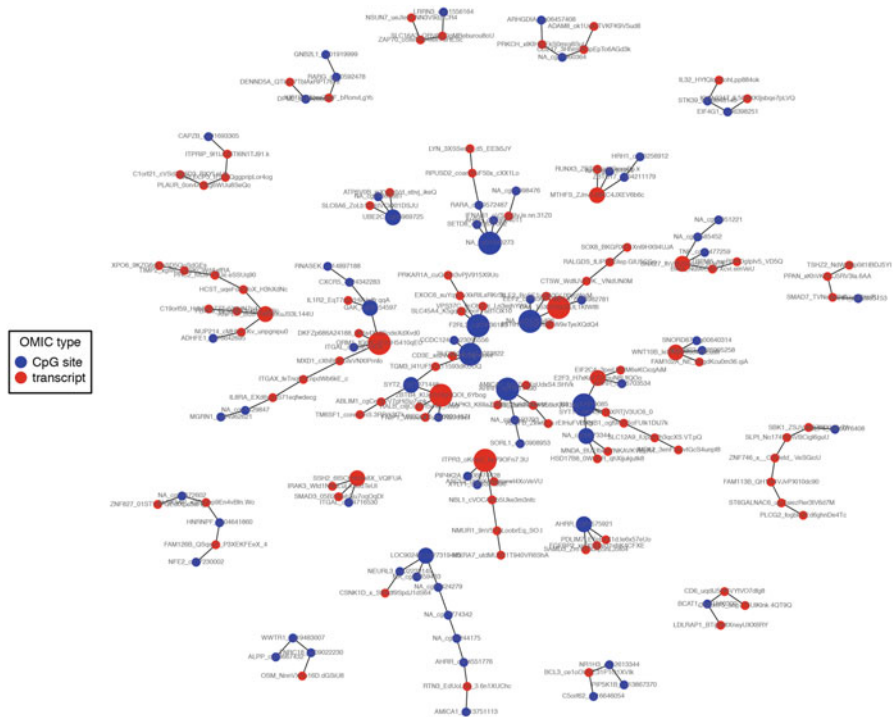
These features are key to improve the interpretability of results from OMICs profiling and can facilitate the integration of data arising from different platforms. In order to realistically and functionally define these groups, developments can be considered to use external information (e.g., empirical information from external studies), or to automatically define discriminant groupings of the features within each of the OMICs data. That could involve a preliminary topological investigation of the network within each type of OMICs data. The combined use of variable grouping and penalization can also be applied to penalized regression (e.g., through the group LASSO) (Simon et al. 2013).

OMICs profiling and their integrative alternatives produce a prioritized list of (multi-)OMIC markers that jointly reflect the molecular effects of exposures. Insight into their mode of action could be gained by exploring their inter-connections, the regulatory cascades they are involved in.

The inference of network topologies can identify nodes, as defined by (combination of) exposure-related biomarkers, which will be interconnected if they are related (typically with high pairwise correlations). Supervised alternatives, as defined by differential networks (Salamanca Beatriz et al. 2014; Valcarcel et al. 2011, 2014), will account for differences in subpopulations by linking two nodes if their relation is not the same in the two populations (e.g., cases and controls).

As a proof-of-principle example, we applied these models to the smoking-related markers (265 differentially methylated sites and 425 associated transcripts) of smoking exposures presented above (Chadeau-Hyam, personal communication). The application of the differential networks requires to carefully choose the metrics used to measure pairwise correlation (e.g., Spearman correlation, partial correlation, shrinkage correlation), as well as the way to select influential edges (e.g., significance assessment via permutations, or stability analyses). Once these choices are set, differential networks provide a visualization of the features that are differentially related in two subpopulations of interest. After strong shrinkage of the network topology, the analysis of smoking-related CpG sites and associated transcripts showed sets of independent modules combining CpG sites and transcripts (Fig. 11.14), hence facilitating their functional interpretation.

This type of methodology can directly be generalized to experimental studies. For instance, differential network methodology can be applied to the metabolites and transcripts found associated to DBP exposure while swimming in a chlorinated pool (van Veldhoven et al. 2017). The preliminary screening for these two sets of OMICs



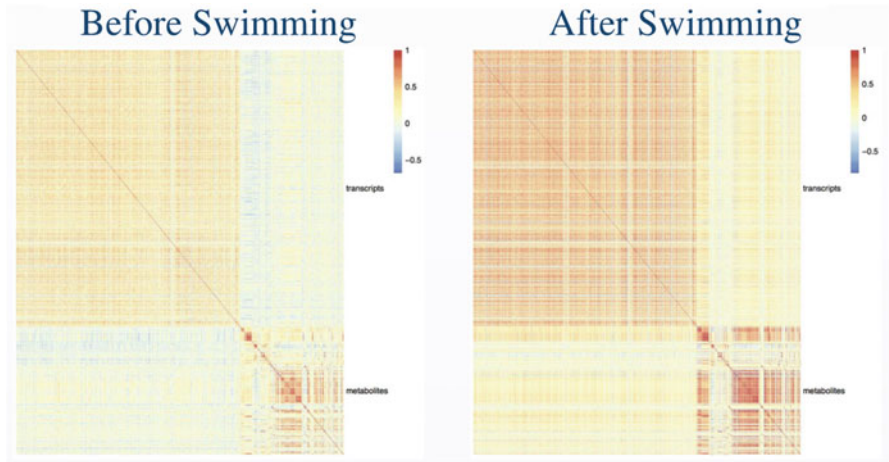
**Fig. 11.14** Differential network including 265 smoking-related differentially methylated CpG sites and 426 correlated transcripts

profiles typically gives rise to two correlation heatmaps before and after swimming. In both heatmaps, correlation levels within each class of biomarkers (metabolic features on the one hand, and transcripts on the other hand) are higher than across classes and seem to be strengthened after the swimming experiment (Fig. 11.15).

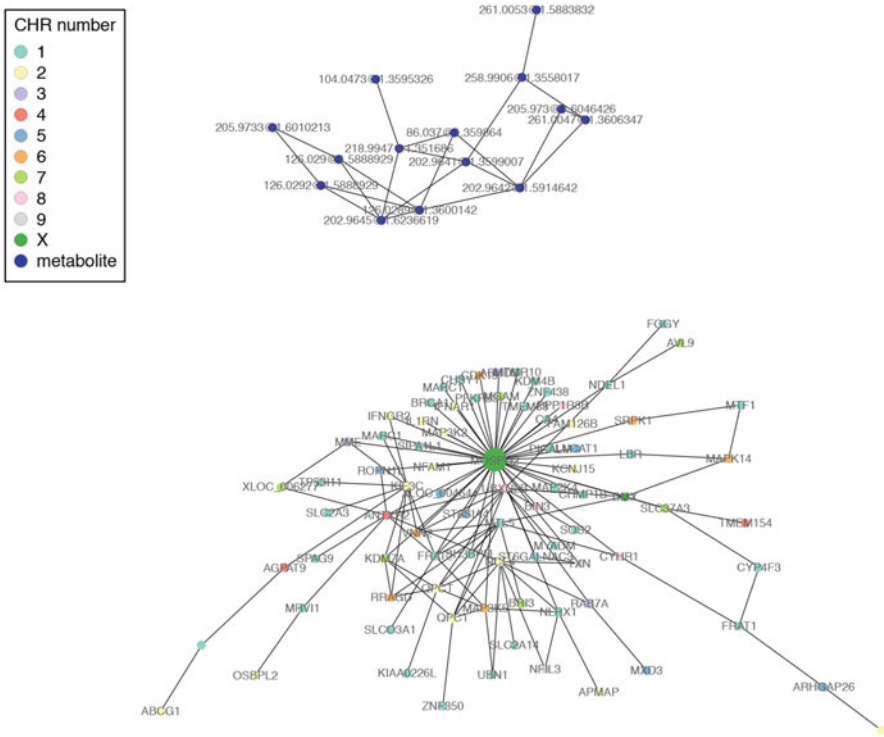
The application of differential network comparing the correlation before and after the swim clearly shows no evidence of correlated/interacting responses at both molecular levels to the experimentally induced changes in exposure (Fig. 11.16).

### *Perspectives: Toward Mechanome Characterization*

The set of methods described in this chapter represents a non-exhaustive list of approaches that have successfully been used in exposome studies. One primary methodological challenge raised by exposome characterization relates to high dimensionality of the data and the complexity of the effects of interest, which are usually multivariate and pleotropic. Models handling the dimensionality of such data are now established and have been successfully applied. Models to integrate the

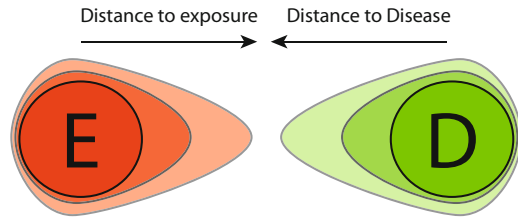


**Fig. 11.15** Correlation heatmaps of the ( $N = 293$ ) metabolic features and ( $N = 721$ ) transcripts found associated to DBP exposure during an experimental swim in a chlorinated pool



**Fig. 11.16** Differential network of the metabolic features and transcripts found associated to DBP exposure during an experimental swim in a chlorinated pool

**Fig. 11.17** Schematic representation of the sequential approach to explore the mechanome. The approach can start from exposures (E) and/or disease endpoint (D)



diversity of exposome data are developing and have been applied in a targeted setting (i.e., applied to preselected sets of exposures and biomarkers). While, from a computational standpoint, these approaches could be scaled up to full-resolution data, interpretability of the results will remain rate limiting in the absence of additional information describing the functional role of molecular markers of exposures.

The mechanome can be defined as the ensemble of exposure-triggered regulatory cascades affecting the individual's risk of developing an adverse condition. While its full exploration is not feasible, the combined use of profiling techniques and network inference as presented in this chapter have the potential to provide a prioritized list of biomarkers involved in the mechanome of a given condition.

Assuming that the strength of correlation/association among the OMICs signals is reflecting their functional proximity in the molecular pathway, sequential approaches can be used to identify a prioritized, sparse and nonredundant subset of OMICs signals potentially involved in the molecular mechanisms of interest (Fig. 11.17).

Starting with either the exposure or the health outcome, using the same pool of profiling techniques, one can identify a core set of OMICs markers of exposure and outcome, respectively. By using conditional modeling approaches, "first order" sets of biomarkers can be defined as those associated to at least one core biomarker but not directly to either exposure or disease endpoint. Repeating this sequential procedure will generate a list of ordered sets of biomarkers with respect to their "distance" to exposure or disease, which can potentially inform the structure about causal structures and relationships.

These structures can be explored by means of network topological investigations first within classes of markers to inform about the multivariate physiological (1) response to exposure, and (2) changes leading to disease onset, at different stages of the molecular pathway and at different molecular levels, through the identification hubs playing a pivotal role in functional translation of the effect of exposure or leading to increased risk of the outcome.

Topological investigations can be extended across classes of biomarkers by considering each order of biomarkers as a distinct subnetwork and seeking for the most likely (or the shortest) path across classes of markers. The identified path (s) linking central nodes across classes from core exposure, to core disease biomarkers, can provide a visualization of the molecular pathways involved in the exposure-induced development of the outcome. In that setting, OMICs integration

will be achieved by either (1) running the full approach on each set of OMICs data separately and combining each OMIC specific network using, for instance, a multilayer network (Kivelä et al. 2013), or (2) pooling OMICs data (or sets of prioritized thereof) in a single network.

The resulting list of prioritized and ordered sets of biomarkers can subsequently be fed into probabilistic graphical models, where the inclusion of directed edges within the graph can help in addressing causality. In this setting, the mechanistic exploration of the data can be viewed as a search for the most relevant causal graph (s). To ensure computational feasibility, numerical algorithms to efficiently explore such a vast model space (e.g., stochastic search algorithms, and importance sampling) can be implemented.

Longitudinal data provide a gold-standard setting to investigate mechanisms as they enable the explicit modeling of the processes leading to the observation and allow formal causal assessment. Longitudinal models include multistate models, which are defined by a set of ordered states reflecting the evolution of the health status, and can be fully characterized by the set of transition probabilities between each compartment. Model estimation aims at quantifying the transitions ensuring the best reconstruction of the pathological trajectories in each subject, hence adding to the classification problem (discriminating healthy and diseased subjects) a dynamic component (estimating the time of onset). While these models were initially developed to accommodate data from longitudinal studies, they can fruitfully be generalized to cross-sectional data and history of exposure to external stressors (e.g., smoking history), as exemplified by a recent proof-of-principle publication on smoking-induced lung cancer (Chadeau-Hyam et al. 2014a). Including biomarkers in the definition of the transition probabilities may help identify the step(s) of the pathological pathway they may exert their effect on, and may therefore help in understanding their functional role.

This methodological framework will also be able to accommodate OMICs trajectories in case of repeated measurement of OMICs profiles. As such, this approach will define a quantitative complement to trajectory classification procedures (e.g., manifold and dynamic time warping algorithms) to identify OMICs evolution patterns that are characteristic of the exposure to external stressors and/or of future disease risk, by leveraging off the information brought about by the auto-correlation structure embedded in OMICs trajectories.

Overall the characterization of the mechanome relies on the elucidation of causal structures existing among prioritized sets of exposure-related biomarkers. Because formal causal assessment can only be achieved in a longitudinal setup, a deeper understanding of the mechanisms involved in the exposome, as formalized in the mechanome concept, will undoubtedly rely on the generation and exploitation of exposome data measured in the same individuals at different life stages. The inclusion of this temporal component in exposome data will define a new set of statistical challenges that should represent one of the key methodological priorities of exposome research in the coming years.

## References

- Agier L, Portengen L, Chadeau-Hyam M, Basagana X, Giorgis-Allemand L, Siroux V, Robinson O, Vlaanderen J, Gonzalez JR, Nieuwenhuijsen MJ, Vineis P, Vrijheid M, Slama R, Vermeulen R (2016) A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ Health Perspect* 124(12):1848–1856. <https://doi.org/10.1289/EHP172>
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25:25. <https://doi.org/10.1038/75556>
- Assi N, Pages A, Vineis P, Chadeau-Hyam M, Stepien M, Duarte-Salles T, Byrnes G, Boumaza H, Knüppel S, Kühn T, Palli D, Bamia C, Boshuizen H, Bonet C, Overvad K, Johansson M, Travis R, Gunter M, Lund E, Dossus L, Elena-Herrmann B, Riboli E, Jenab M, Viallon V, Ferrari P (2015) A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study. *Mutagenesis* 30(6):743–753
- Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7(10):781–791. <https://doi.org/10.1038/nrg1916>
- Belshaw NJ, Pal N, Tapp HS, Dainty JR, Lewis MPN, Williams MR, Lund EK, Johnson IT (2010) Patterns of DNA methylation in individual colonic crypts reveal aging and cancer-related field defects in the morphologically normal mucosa. *Carcinogenesis* 31(6):1158–1163. <https://doi.org/10.1093/carcin/bgq077>
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* 57:289–300
- Botto L, Chadeau-Hyam M, Hastie DI, Langley SR, Petretto E, Tired L, Tregouet D, Richardson S (2011) ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics* 27(4):587–588. <https://doi.org/10.1093/bioinformatics/btq684>
- Botto L, Chadeau-Hyam M, Hastie DI, Zeller T, Liqueur B, Newcombe P, Yengo L, Wild PS, Schillert A, Ziegler A, Nielsen SF, Butterworth AS, Ho WK, Castagne R, Munzel T, Tregouet D, Falchi M, Cambien F, Nordestgaard BG, Fumeron F, Tybjaerg-Hansen A, Froguel P, Danesh J, Petretto E, Blankenberg S, Tired L, Richardson S (2013) GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet* 9(8):e1003657. <https://doi.org/10.1371/journal.pgen.1003657>
- Boulesteix AL, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* 8(1):32–44. <https://doi.org/10.1093/bib/bb1016>
- Carlin DJ, Rider CV, Woychik R, Birnbaum LS (2013) Unraveling the health effects of environmental mixtures: an NIEHS priority. *Environ Health Perspect* 121(1):A6–A8
- Castagne R, Kelly-Irving M, Campanella G, Guida F, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Kleinjans J, de Kok T, Kyrtopoulos SA, Lang T, Stringhini S, Vermeulen R, Vineis P, Delpierre C, Chadeau-Hyam M (2016) Biological marks of early-life socioeconomic experience is detected in the adult inflammatory transcriptome. *Sci Rep* 6:38705. <https://doi.org/10.1038/srep38705>
- Castagne R, Boulange CL, Karaman I, Campanella G, Santos Ferreira DL, Kaluarachchi MR, Lehne B, Moayyeri A, Lewis MR, Spagou K, Dona AC, Evangelos V, Tracy R, Greenland P, Lindon JC, Herrington D, Ebbels TMD, Elliott P, Tzoulaki I, Chadeau-Hyam M (2017) Improving visualization and interpretation of metabolome-wide association studies: an application in a population-based cohort using untargeted 1h nmr metabolic profiling. *J Proteome Res* 16(10):3623–3633. <https://doi.org/10.1021/acs.jproteome.7b00344>
- Chadeau-Hyam M, Ebbels TMD, Brown IJ, Chan Q, Stemler J, Huang CC, Daviglus ML, Ueshima H, Zhao L, Holmes E, Nicholson JK, Elliott P, De Iorio M (2010) Metabolic profiling



- and the metabolome-wide association study: significance level for biomarker identification. *J Proteome Res* 9(9):4620–4627. <https://doi.org/10.1021/pr1003449>
- Chadeau-Hyam M, Athersuch TJ, Keun HC, De Iorio M, Ebbels TMD, Jenab M, Sacerdote C, Bruce SJ, Holmes E, Vineis P (2011) Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers* 16 (1):83–88. <https://doi.org/10.3109/1354750x.2010.533285>
- Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, Liquet B, Vermeulen RC (2013) Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen* 54(7):542–557. <https://doi.org/10.1002/em.21797>
- Chadeau-Hyam M, Tubert-Bitter P, Guihenneuc-Jouyau C, Campanella G, Richardson S, Vermeulen R, De Iorio M, Galea S, Vineis P (2014a) Dynamics of the risk of smoking-induced lung cancer: a compartmental hidden Markov model for longitudinal analysis. *Epidemiology* 25(1):28–34. <https://doi.org/10.1097/EDE.0000000000000032>
- Chadeau-Hyam M, Vermeulen RC, Hebel DG, Castagne R, Campanella G, Portengen L, Kelly RS, Bergdahl IA, Melin B, Hallmans G, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, de Kok TM, Smith MT, Kleijnans JC, Vineis P, Kyrtopoulos SA, EnviroGenoMarkers project consortium (2014b) Prediagnostic transcriptomic markers of chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Ann Oncol* 25(5):1065–1072. <https://doi.org/10.1093/annonc/mdu056>
- Chun H, Keles S (2009) Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics* 182(1):79–90. <https://doi.org/10.1534/genetics.109.100362>
- Chun H, Keles S (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Series B Stat Methodol* 72:3–25
- Dominici F, Peng RD, Barr CD, Bell ML (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology* 21(2):187–194. <https://doi.org/10.1097/EDE.0b013e3181cc86e8>
- Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 32(3):227–234. <https://doi.org/10.1002/gepi.20297>
- Espin-Perez A, Font-Ribera L, van Veldhoven K, Krauskopf J, Portengen L, Chadeau-Hyam M, Vermeulen R, Grimalt JO, Villanueva CM, Vineis P, Kogevinas M, Kleijnans JC, de Kok TM (2018) Blood transcriptional and microRNA responses to short-term exposure to disinfection by-products in a swimming pool. *Environ Int* 110:42–50. <https://doi.org/10.1016/j.envint.2017.10.003>
- Fasoli M, Dal Santo S, Zenoni S, Torielli GB, Farina L, Zamboni A, Porceddu A, Venturini L, Bicego M, Murino V, Ferrarini A, Delledonne M, Pezzotti M (2012) The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell* 24(9):3489–3505. <https://doi.org/10.1105/tpc.112.100230>
- Font-Ribera L, Kogevinas M, Zock JP, Gomez FP, Barreiro E, Nieuwenhuijsen MJ, Fernandez P, Lourencetti C, Perez-Olabarria M, Bustamante M, Marcos R, Grimalt JO, Villanueva CM (2010) Short-term changes in respiratory biomarkers after swimming in a chlorinated pool. *Environ Health Perspect* 118(11):1538–1544. <https://doi.org/10.1289/ehp.1001961>
- Fonville JM, Richards SE, Barton RH, Boulange CL, Ebbels TMD, Nicholson JK, Holmes E, Dumas ME (2010) The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *J Chemom* 24(11–12):636–649. <https://doi.org/10.1002/cem.1359>
- Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33(1):1–22
- Greenacre M (1984) Theory and applications of correspondence analysis. Academic Press, London
- Guan YT, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat* 5(3):1780–1815. <https://doi.org/10.1214/11-aos455>

- Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, Georgiadis P, Vermeulen RCH, Lund E, Vineis P, Chadeau-Hyam M (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet* 24(8):2349–2359. <https://doi.org/10.1093/hmg/ddu751>
- Guxens M, Ballester F, Espada M, Fernandez MF, Grimalt JO, Ibarluzea J, Olea N, Rebagliato M, Tardon A, Torrent M, Vioque J, Vrijheid M, Sunyer J, Project I (2012) Cohort profile: the INMA--Infancia y Medio Ambiente--(environment and childhood) project. *Int J Epidemiol* 41(4):930–940. <https://doi.org/10.1093/ije/dyr054>
- Haight TJ, Wang Y, van der Laan MJ, Tager IB (2010) A cross-validation deletion-substitution-addition model selection algorithm: application to marginal structural models. *Comput Stat Data Anal* 54(12):3080–3094. <https://doi.org/10.1016/j.csda.2010.02.002>
- Hans C, Dobra A, West M (2007) Shotgun stochastic search for “large p” regression. *J Am Stat Assoc* 102(478):507–516. <https://doi.org/10.1198/016214507000000121>
- Hoerl AE, Kennard RW (1970) Ridge regression—biased estimation for nonorthogonal problems. *Technometrics* 12(1):661–676. <https://doi.org/10.2307/1267351>
- Hoggart CJ, Clark TG, De Lorio M, Whittaker JC, Balding DJ (2008) Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol* 32(2):179–185
- Holmes E, Loo RL, Stamler J, Bictash M, Yap IK, Chan Q, Ebbels T, De Iorio M, Brown II, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao L, Nicholson JK, Elliott P (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 453(7193):396–400
- Hotelling H (1933a) Analysis of complex statistical variables into principal components. *J Educ Psychol* 24(6):417–441
- Hotelling H (1933b) Analysis of complex statistical variables into principal components. *J Educ Psychol* 24(7):498–520
- Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44. <https://doi.org/10.1038/nprot.2008.211>
- Jain P, Vineis P, Liquet B, Vlaanderen J, Bodinier B, van Veldhoven K, Kogevinas M, Athersuch TJ, Font-Ribera L, Villanueva CM, Vermeulen R, Chadeau-Hyam M (2018) A multivariate approach to investigate the combined biological effects of multiple exposures. *J Epidemiol Community Health* 72(7):564–571. <https://doi.org/10.1136/jech-2017-210061>
- Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. *Heredity* 102(4):330–341. <https://doi.org/10.1038/hdy.2008.130>
- Kivelä M, Arenas A, Barthelemy M, Gleeson J, Moreno Y, Porter M (2013) Multilayer networks. *J Complex Netw* 2(3):203–271
- Le Cao KA, Rossouw D, Robert-Granie C, Besse P (2008) A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol* 7(1):35
- Le Cao KA, Martin PGP, Robert-Granie C, Besse P (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10:34. <https://doi.org/10.1186/1471-2105-10-34>
- Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, Jones DP, Pulendran B (2013) Predicting network activity from high throughput metabolomics. *PLoS Comput Biol* 9(7):e1003123. <https://doi.org/10.1371/journal.pcbi.1003123>
- Lindstrom MJ, Bates DM (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics* 46(3):673–687. <https://doi.org/10.2307/2532087>
- Liquet B, Le Cao K-A, Hocini H, Thiebaut R (2012) A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics* 13(1):325
- Liquet B, Bottolo L, Campanella G, Richardson S, Chadeau-Hyam M (2016a) R2GUESS: a graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses. *J Stat Softw* 69(2). <https://doi.org/10.18637/jss.v069.i02>

- Liquet B, Lafaye de Micheaux P, Hejblum B, Thiebaut R (2016b) Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics* 32(1):35–42
- McCreanor J, Cullinan P, Nieuwenhuijsen MJ, Stewart-Evans J, Malliarou E, Jarup L, Harrington R, Svartengren M, Han IK, Ohman-Strickland P, Chung KF, Zhang J (2007) Respiratory effects of exposure to diesel traffic in persons with asthma. *N Engl J Med* 357(23):2348–2358. <https://doi.org/10.1056/NEJMoa071535>
- McHale CM, Zhang LP, Lan Q, Vermeulen R, Li GL, Hubbard AE, Porter KE, Thomas R, Portier CJ, Shen M, Rappaport SM, Yin SN, Smith MT, Rothman N (2011) Global gene expression profiling of a population exposed to a range of benzene levels. *Environ Health Perspect* 119(5):628–634. <https://doi.org/10.1289/ehp.1002546>
- Musumarra G, Condorelli DF, Fortuna CG (2011) OPLS-DA as a suitable method for selecting a set of gene transcripts discriminating RAS- and PTPN11-mutated cells in acute lymphoblastic leukaemia. *Comb Chem High Throughput Screen* 14(1):36–46
- Parkhomenko E, Tritchler D, Beyene J (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol* 8:1. <https://doi.org/10.2202/1544-6115.1406>
- Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190. <https://doi.org/10.1371/journal.pgen.0020190>
- Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos Mag* 2(6):559–572
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909. <https://doi.org/10.1038/ng1847>
- Rappaport SM, Smith MT (2010) Environment and disease risks. *Science* 330(6003):460–461. <https://doi.org/10.1126/science.1192603>
- Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. *Nat Genet* 40(5):491–492. <https://doi.org/10.1038/ng0508-491>
- Rider CV, Carlin DJ, Devito MJ, Thompson CL, Walker NJ (2013) Mixtures research at NIEHS: an evolving program. *Toxicology* 313(2–3):94–102. <https://doi.org/10.1016/j.tox.2012.10.017>
- Robinson O, Basagana X, Agier L, de Castro M, Hernandez-Ferrer C, Gonzalez JR, Grimalt JO, Nieuwenhuijsen M, Sunyer J, Slama R, Vrijheid M (2015) The pregnancy exposome: multiple environmental exposures in the INMA-Sabadell birth cohort. *Environ Sci Technol* 49(17):10632–10641. <https://doi.org/10.1021/acs.est.5b01782>
- Salamanca Beatriz V, Ebbels Timothy MD, Iorio Maria D (2014) Variance and covariance heterogeneity analysis for detection of metabolites associated with cadmium exposure. *Stat Appl Genet Mol Biol* 13:191–201. <https://doi.org/10.1515/sagmb-2013-0041>
- Schafer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol* 4:32. <https://doi.org/10.2202/1544-6115.1175>
- Shen HP, Huang JHZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J Multivar Anal* 99(6):1015–1034. <https://doi.org/10.1016/j.jmva.2007.06.007>
- Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for cox’s proportional hazards model via coordinate descent. *J Stat Softw* 39(5):1–13
- Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. *J Comput Graph Stat* 22(2):231–245. <https://doi.org/10.1080/10618600.2012.681250>
- The Gene Ontology Consortium (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res* 45(D1):D331–D338. <https://doi.org/10.1093/nar/gkw1108>
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58(1):267–288. <https://doi.org/10.2307/2346178>
- Valcarcel B, Wurtz P, al Basatena NKS, Tukiainen T, Kangas AJ, Soininen P, Jarvelin MR, Ala-Korpela M, Ebbels TM, de Iorio M (2011) A differential network approach to exploring differences between biological states: an application to prediabetes. *PLoS One* 6(9):e24702. <https://doi.org/10.1371/journal.pone.0024702>

- Valcarcel B, Ebbels TMD, Kangas AJ, Soininen P, Elliot P, Ala-Korpela M, Jarvelin MR, de Iorio M (2014) Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity. *J R Soc Interface* 11 (94):20130908. <https://doi.org/10.1098/rsif.2013.0908>
- van Veldhoven K, Keski-Rahkonen P, Barupal DK, Villanueva CM, Font-Ribera L, Scalbert A, Bodinier B, Grimalt JO, Zwiener C, Vlaanderen J, Portengen L, Vermeulen R, Vineis P, Chadeau-Hyam M, Kogevinas M (2017) Effects of exposure to water disinfection by-products in a swimming pool: a metabolome-wide association study. *Environ Int* 111:60–70. <https://doi.org/10.1016/j.envint.2017.11.017>
- Vineis P, Perera F (2007) Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol Biomark Prev* 16(10):1954–1965
- Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, Kogevinas M, Kyrtopoulos S, Nieuwenhuijsen M, Phillips DH, Probst-Hensch N, Scalbert A, Vermeulen R, Wild CP (2016) The exposome in practice: design of the EXPOsOMICS project. *Int J Hyg Environ Health* 220(2 Pt A):142–151. <https://doi.org/10.1016/j.ijheh.2016.08.001>
- Vlaanderen J, van Veldhoven K, Font-Ribera L, Villanueva CM, Chadeau-Hyam M, Portengen L, Grimalt JO, Zwiener C, Heederik D, Zhang X, Vineis P, Kogevinas M, Vermeulen R (2017) Acute changes in serum immune markers due to swimming in a chlorinated pool. *Environ Int* 105:1–11. <https://doi.org/10.1016/j.envint.2017.04.009>
- Wang H, Gottfries J, Barrenäs F, Benson M (2011) Identification of novel biomarkers in seasonal allergic rhinitis by combining proteomic, multivariate and pathway analysis. *PLoS One* 6(8): e23563. <https://doi.org/10.1371/journal.pone.0023563>
- West M (2003) Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian statistics 7*. Clarendon Press, Oxford
- Westfall P, Young S (1993) Resampling-based multiple testing: examples and methods for *p*-value adjustment (Wiley Series in Probability and Statistics). Wiley-Interscience
- Wild CP (2005) Complementing the genome with an ‘exposome’: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev* 14(8):1847–1850. <https://doi.org/10.1158/1055-9965.EPI-05-0456>
- Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3):515–534. <https://doi.org/10.1093/biostatistics/kxp008>
- Wold S, Ruhe A, Wold H, Dunn WJ (1984) The collinearity problem in linear-regression - the partial least-squares (PLS) approach to generalized inverses. *SIAM J Sci Stat Comput* 5 (3):735–743. <https://doi.org/10.1137/0905052>
- Yap IKS, Brown IJ, Chan Q, Wijeyesekera A, Garcia-Perez I, Bictash M, Loo RL, Chadeau-Hyam M, Ebbels T, Iorio MD, Maibaum E, Zhao L, Kesteloot H, Daviglus ML, Stamler J, Nicholson JK, Elliott P, Holmes E (2010) Metabolome-wide association study identifies multiple biomarkers that discriminate north and south chinese populations at differing risks of cardiovascular disease: INTERMAP study. *J Proteome Res* 9(12):6647–6654. <https://doi.org/10.1021/pr100798r>
- Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19):2375–2382. <https://doi.org/10.1093/bioinformatics/btq448>
- Zou F, Fine JP, Hu J, Lin DY (2004) An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* 168(4):2307–2316. <https://doi.org/10.1534/genetics.104.031427>
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 67(2):301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
- Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J Comput Graph Stat* 15 (2):265–286. <https://doi.org/10.1198/106186006x113430>