

Supplementary Material

1 Supplementary Figures and Tables

1.1 Supplementary Tables

Supplementary Table 1. Echocardiographic imaging variables of post-myocardial infarction patients in CDCS study

Supplementary Table 2. Partial correlations associated with the edges connecting at least one of the echocardiographic imaging variables and clinical biomarkers in the CDCS cohort.

Supplementary Table 3. Hypothesis testing (two-sample t-test) of data features between MACE and event-free patients and between the HF subgroup and event-free patients in the CDCS cohort.

Supplementary Table 4. MACE signature derived by the supervised analysis module of iOmicsPASS+ using the CDCS data, with edge-specific *d-scores* of the two groups (MACE and event-free).

Supplementary Table 5. Biological function enrichment analysis of the proteins in the MACE signature, accounting for tissue-specific mRNA expression.

Supplementary Table 6. HF signature derived by the supervised analysis module of iOmicsPASS+ using the CDCS data, with edge-specific *d-scores* of the two groups (HF and event-free).

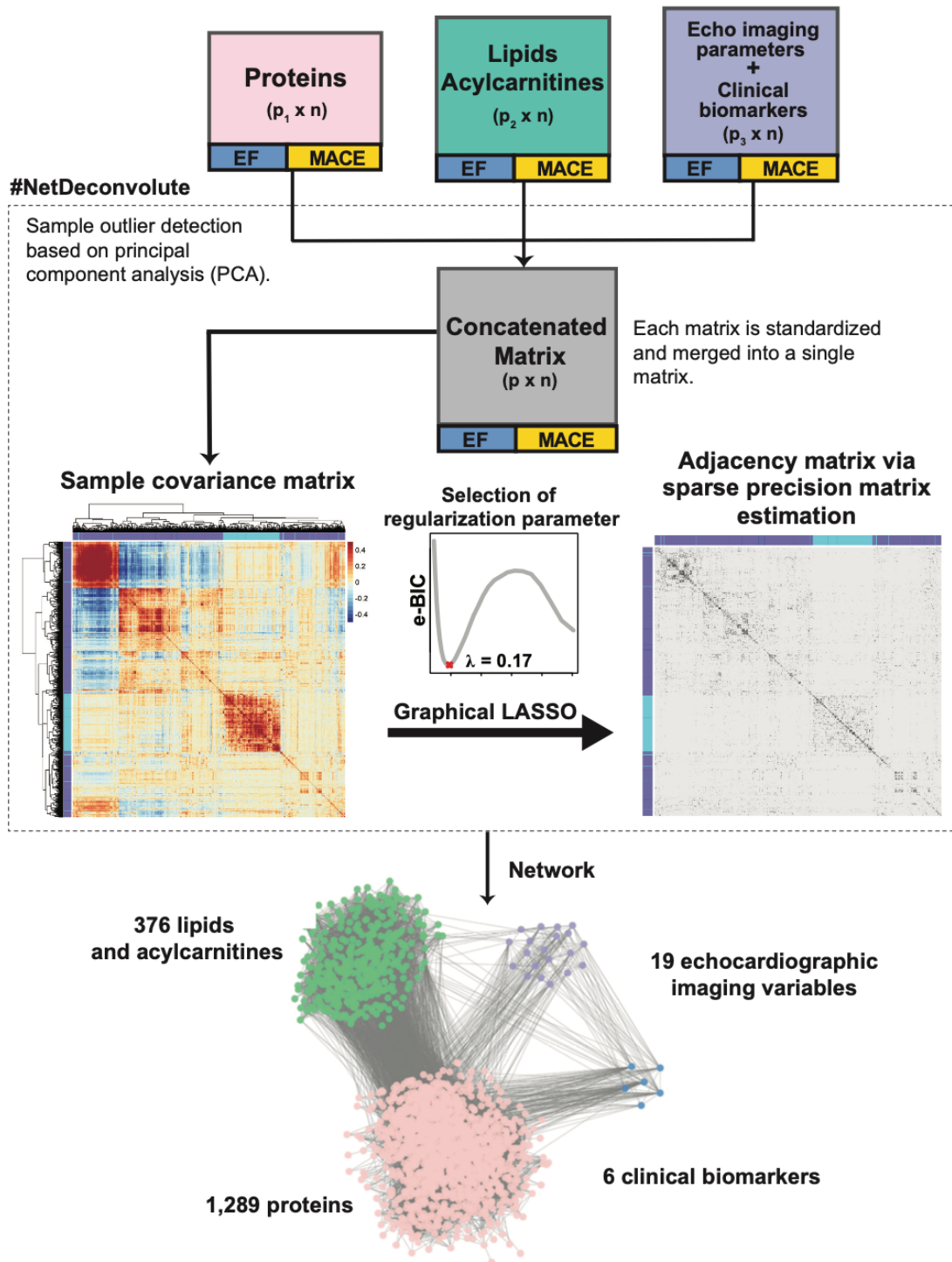
Supplementary Table 7. Biological function enrichment analysis of the proteins in the HF signature, accounting for tissue-specific mRNA expression.

Supplementary Table 8. Summary table of tissue-specific over-expression of individual genes encoding the plasma proteins at their respective, potential sites of origin.

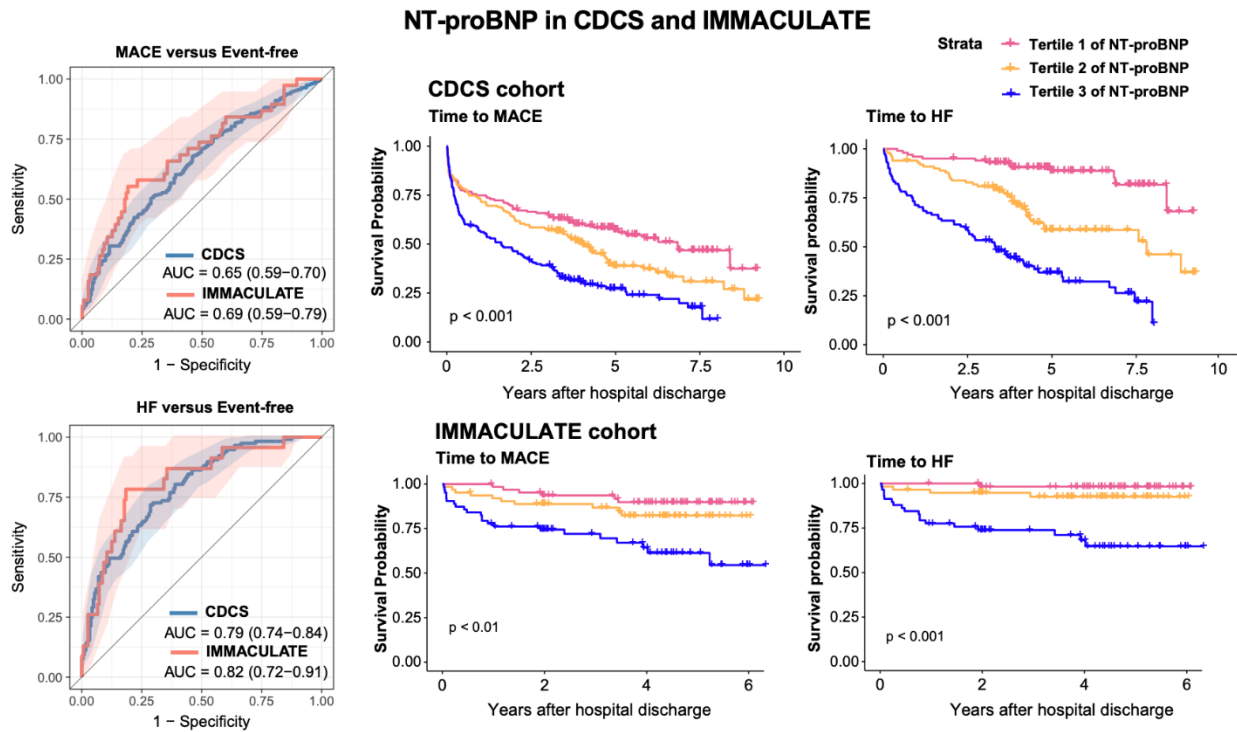
Supplementary Table 9. Echocardiographic imaging variables of post-myocardial infarction patients in IMMACULATE study

Supplementary Table 10. Statistical summary of the risk of MACE and HF adjusted for medication use after hospital discharge in CDCS and IMMACULATE cohort using Cox proportional hazards model.

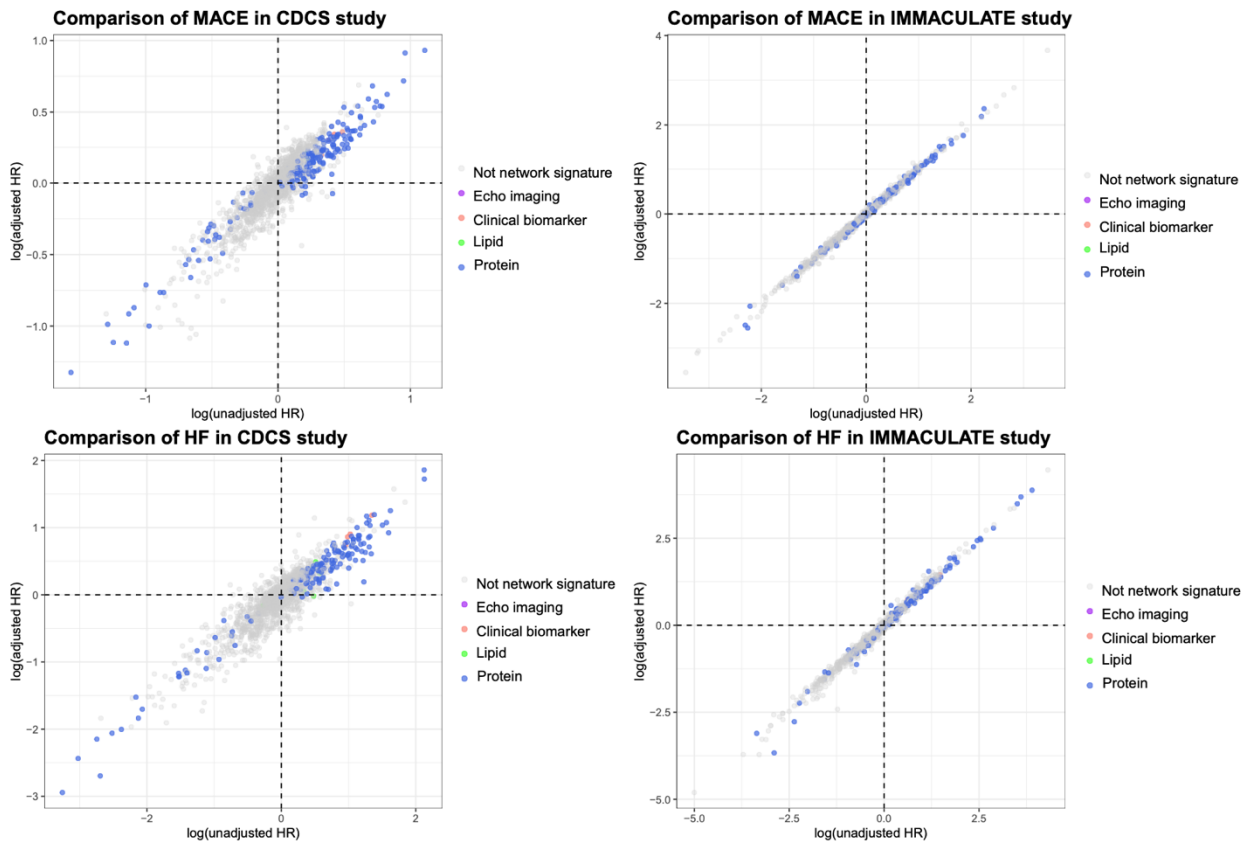
1.2 Supplementary Figures



Supplementary Figure 1. Data features are standardized and outliers are removed prior to integration of multi-modal data sets. In CDCS, 1,289 proteins from SOMALOGIC and 376 lipids from targeted MS were integrated with 19 cardiac imaging variables and 6 clinical biomarkers through this network inference.



Supplementary Figure 2. The prognostic value of NT-proBNP for MACE and HF in the two cohorts, measured by the ROC analysis as well as the separation of risk profiles via tertile-based stratification.



Supplementary Figure 3. The comparison of the unadjusted and adjusted log hazard ratios to evaluate the effect of medication use during hospital stay to baseline for MACE (top panel) and incident HF (bottom panel) in CDCS and IMMACULATE study. In CDCS, the log hazard ratios are adjusted for the use of beta-blockers, clopidogrel, calcium channel antagonist, long-acting nitrates, ACE inhibitors, aspirin, diuretics, statin and warfarin. For IMMACULATE, the log hazard ratios are adjusted for the use of beta-blockers, ACE inhibitors and angiotensin receptor blockers (ARB). The markers from the identified network signatures are colored in blue (protein), green (lipid), salmon (clinical biomarker) and purple (echo imaging). Markers that are not a part of the signature are shown in grey.

2 Supplementary Information

2.1 The Coronary Disease Cohort Study (CDCS)

The CDCS cohort consists of 2,140 patients admitted into two tertiary hospitals (Christchurch Hospital and Auckland City Hospital) in New Zealand (NZ) for an ACS event, recruited from 2002 and 2009. They were invited to return to the hospital one month after discharge for baseline measurements. Patients were excluded from the study if they had a severe comorbidity that limited their life expectancy to less than 3 years. Clinical measurements, blood sample collection, and echocardiogram were all obtained at baseline (30 days from discharge) and 4 months, 6 months and 12 months after discharge. These individuals were followed up until the end of the study on 31st January 2012, death or lost-to-follow-up. Subsequent clinical events and mortality information were obtained from the NZ Nation Health Information System. The study was approved by the New Zealand Multi-region Ethics Committee (CTY/02/02/018) and all participants had provided a written informed consent prior to their participation. More information on the study can be found in Prickett *et al* (1).

Out of these 2140 patients, a nested sub-cohort of 750 patients matched by age and gender was selected, and both aptamer-based SOMALOGIC protein assay and multiple reaction monitoring (MRM)-based lipidomics were used to measure proteins and lipids, respectively. Among these, 250 patients were readmitted for HF after their hospital discharge, 250 patients were readmitted for myocardial infarction (MI) episode and 250 patients were selected as controls without any recurrence of MI or incidence of heart failure hospitalization. The controls were further defined as free of adverse cardiac remodelling and any cardiovascular-related death over the follow-up period. They were selected by matching each HF patient to a MI patient and a control by age and gender using the MatchIT package in R (2).

Following a quality control step for the molecular data, 741 patients were retained for network inference. For the supervised analysis using the iOmicsPASS method, 286 patients were removed due to the following exclusion criteria: 69 patients with cardiac remodelling, 81 patients with a medical history of heart failure, 97 patients with history of stroke, 78 patients with a prior MI within the previous five years, and 19 patients with missing or unknown status to minimize possible error in the outcomes. A total of 464 patients was retained for predictive subnetwork identification for MACE and HF. Of those, 185 patients remained event-free, and 279 patients had a secondary MACE, including 117 patients hospitalised for HF during follow-up.

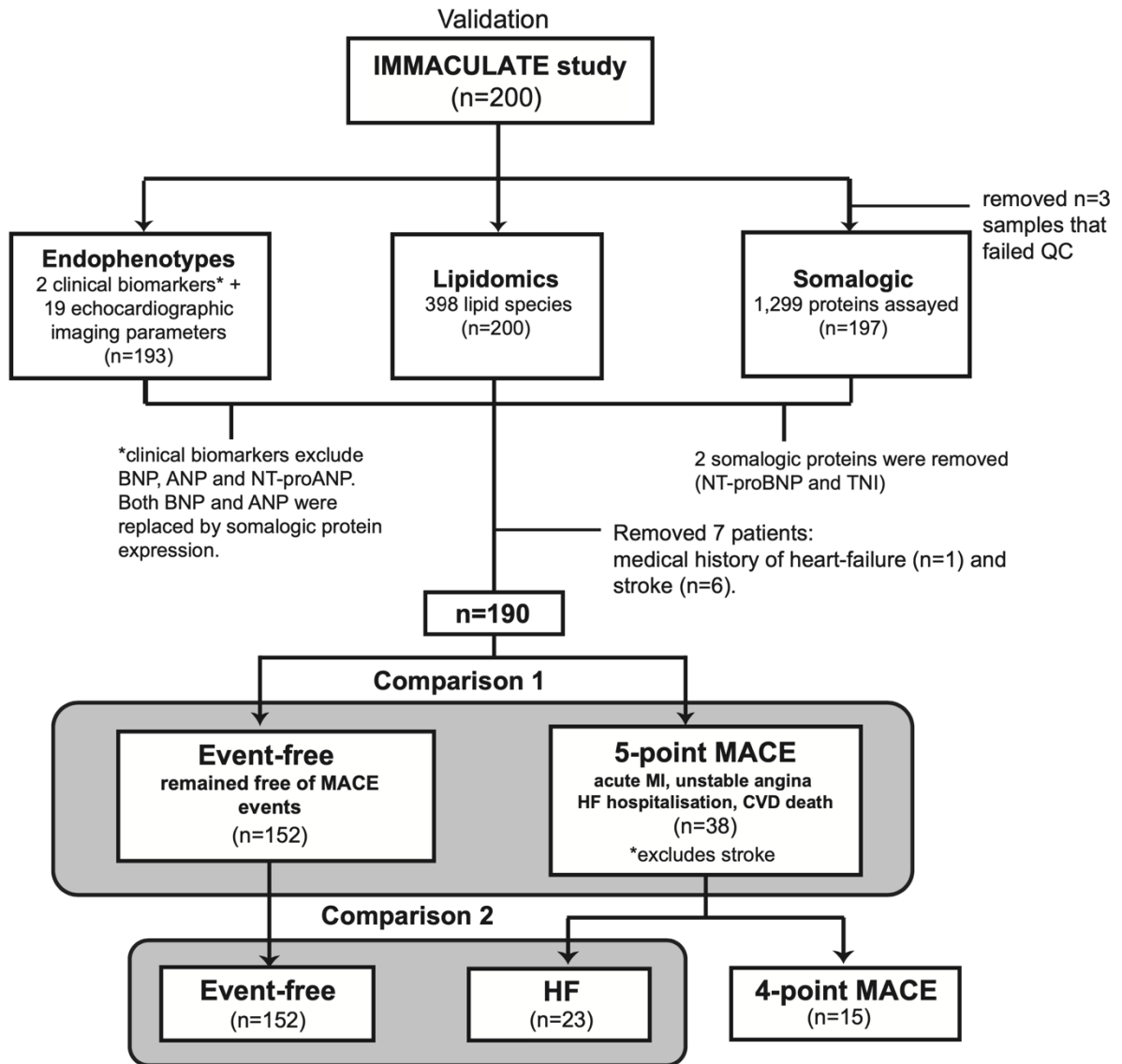
2.2. The Improving Outcomes in Myocardial Infarction through Reversal of Cardiac Remodelling (IMMACULATE) registry

The IMMACULATE registry consists of 859 patients who were admitted for MI into three local hospitals in Singapore (National University Hospital, Tan Tock Seng Hospital and Singapore General Hospital) from 2011 to 2014. The patients were followed up for a median of 3.9 years (interquartile range IQR: 2.0-4.8 years).

Inclusion criteria were patients with a ST-segment elevation MI (STEMI) or non ST-segment elevation MI (NSTEMI) with ischemic chest pain or angina symptoms, concentrations of troponin exceeding 99th percentile URL and >50% occlusion of one or more coronary arteries based on angiography. Patients were excluded if they were anaemic (haemoglobin <8 g/dL in men and <7 g/dL in women), had severe renal impairment (estimated glomerular filtration rate, eGFR<15 ml/min/m²), were in cardiogenic shock not able to be weaned off inotropes or intra-aortic balloon pump, history of malignancy or other conditions limiting life expectancy diagnosed within the last 12 months.

Hospital records were used to capture readmissions due to a recurrent ACS event or other morbidity. Mortality data and primary cause of death was obtained from the National Registry of Death Office (NRDO). The study was approved by the institutional review board and the ethics committee at Singapore's National Healthcare Group Domain Specific Review Board (DSRB 2013/00248 and 2013/00635). All participants provided written informed consent.

We selected 200 IMMACULATE patients with clinical information and blood samples obtained at three time points: (i) within 24-72 hours of admission, (ii) 1 month after hospital discharge and (iii) 6 months after hospital discharge. At all three time points, plasma lipids and proteins were profiled and echocardiography undertaken. In the current report, the IMMACULATE study serves as validation and only measurements one month from hospital discharge post-MI, a temporal match with CDCS, have been used. Among these, 190 subjects had all relevant biochemistry data for cardiac markers, proteomics, and lipidomics data. We further removed one patient with self-reported history of HF and six patients with history of stroke. However, we did not remove any patients with self-reported prior MI as the duration of prior MI to primary ACS event was unknown, as well as due to the constraint in the sample size. As a result, we recorded 152 patients who remained event-free and 38 patients with secondary 5-point MACE (15 4-point MACE and 23 HF). Here, we use IMMACULATE cohort as a validation for the evaluation of classification performance.



2.3. SOMALOGIC Protein Assay

Plasma samples (50 μ L) collected 30 days after the index event were analysed using a Slow Off-rate Modified Aptamer (SOMAmer)-based capture array, “SOMAscan” (somaLogic, Inc, Boulder, CO, USA) (3). The abundance of a protein is reported using relative fluorescent units (RFU). A total of 1,305 human proteins were quantified (47% secreted proteins, 28% extracellular domains and 25% intracellular proteins). Details of the experimental methods and data analysis protocol are described in Chan *et al* (4). We note that the protein names in all diagrams and tables are given by the gene symbols of the encoding genes (all italicised), with the exception of NT-proBNP and cardiac troponin I, which are measured by both immunoassay and SOMALOGIC.

2.4. Lipidomics Assay

Samples were randomized into 10 analytical batches and batch quality control (BQC) samples were prepared by pooling equal plasma aliquots from all the samples. To assess analytical performances, one BQC sample was included for every ten study samples. Plasma (10 μ L) was mixed with 190 μ L 1-butanol/methanol (1:1, v/v), [BuMe], containing 4.5 μ L of SPLASH™ Lipidomix® Mass Spec Standard I (#330707) and 4.5 μ L Cer/Sph Mixture I (#LM6002) from Avanti Polar Lipids, Inc. The mixture was vortexed for 30 seconds, then sonicated for 30 min (ice was added into the sonication bath) and then centrifuged at 14,000 x g for 10 min at 10°C. The supernatant was transferred into autosampler vials for mass spectrometry. Extracted blanks were prepared using the same extraction protocol but without plasma samples. Technical quality control (TQC) samples were generated by pooling lipid extracts from the study samples and were used to monitor the analytical system performance.

Analysis of plasma extracts was performed on an Agilent 6495A QQQ mass spectrometer coupled to an Agilent 1290 Infinity-II UHPLC system and the MRM data acquired in positive ion mode were processed using the MRMkit tool (5). Peaks were integrated and the concentration data were further normalized by a time-trend and batch effect correction algorithm built in MRMkit. Only lipid species with the overall coefficient-of-variation smaller than 25% were included for the downstream analysis. We did not perform multi-point calibration to derive molar concentrations for this analysis since the current peak area values normalized by the internal standard provide the same quantitative information as the final absolute concentration values (linear transforms).

2.5. Quality control for proteomics and lipidomics data

For both CDCS and IMMACULATE studies, only the measurements collected one month after hospital discharge (baseline) were used. For the CDCS study, a total of 376 unique lipids were used for the subsequent analysis. For the IMMACULATE study, a total of 418 lipid species were retained for validation analysis for relaxed criteria (CoV 50%). Internal standards were used to normalize the raw peak areas in the corresponding lipid class and concentrations were further normalized to the protein concentration in the original sample. Due to the methodology used in our study, we do not report absolute concentration values. Endogenous species were quantified using one standard per lipid class, hence our method can only deliver relative quantification results, which does not hamper testing of concentration differences between subjects within any given lipid species. For the proteome analysis in both studies, a standard SOMALOGIC 1.3k array was used. For the purpose of the analysis, we removed 5 proteins that were from human viruses (i.e. human papillomavirus) and samples that failed QC, resulting in 747 samples in CDCS and 197 in IMMACULATE study.

2.6. Biological function enrichment analysis incorporating tissue specific gene expression

The proteins included in the final predictive signatures were annotated in terms of their biological processes in Gene Ontology (GO) using gProfiler (6). Enrichment analysis was repeated with an additional filtering of query and background gene lists with regard to the minimal expression level (transcript per million reads, or TPM > 5) in different tissues, heart (atrial appendage, left ventricle), arteries (aorta, coronary), kidneys (cortex, medulla), liver, lungs and skeletal muscle, in GTEx (7).

2.7. Software packages for data analysis and visualisation

All analyses were carried out in R studio (R version 4.2.0) (8), using external R packages huge (9), gplots (10), mice (11), pROC (12), clusterProfiler (13), as well as iOmicsPASS+ presented in this paper. Visualization of networks were done using Cytoscape (version 3.8.2) (14).

2.8. iOmicsPASS+

iOmicsPASS has been expanded and reimplemented as iOmicsPASS+ to allow for greater flexibility and generalizability to other types of omics data. First, the tool now allows for only one network and one data as input. If only one data (X) is used as input, then the network input is assumed to be one which links the features within the data. If more than one data is used as input (e.g. X, Y and Z), then a second network should be provided to link the features between the data in X to the data in Y. Data Z is used to normalize against matching features in data Y as in the original iOmicsPASS.

Second, the network input now allows for direction of association for the pair of interacting or co-varying data features. The user can specify an additional column in the network file to include whether the association between the pair of features is positive (i.e. 1) or negative (i.e. -1). For the latter, the interaction scores are calculated as ratios (differences of two Z-scores), instead of the usual product (sum of two Z-scores). Within iOmicsPASS+, a sample correlation matrix is first calculated to determine if the direction of the association is consistent with the indicated direction by the user in the network. An edge with interaction score is created only if the sign of the correlation between two data features is consistent with the specified direction of association. Otherwise, if inconsistent, the pair of data features is not used as predictive features for the subsequent analysis. This enables the application to other omics data such as microRNAs and DNA methylation where the inhibition of gene expression occurs as a result of interaction with the target genes.

Last, several new modules are added to the R package to facilitate users to build, install, create input parameter file and run the software. One of the most important module is a network inference module which estimates the best co-varying network between the different types of data, using an existing R-package “huge” (9), which carries out graphical LASSO (GLASSO) estimation. Within the module, we propose two different approaches for estimating the network: a supervised approach and a hybrid approach. In the supervised approach, the network is fully estimated from the precision matrix while the latter fuses a user-provided network

input based on prior knowledge with the estimated GLASSO-derived partial correlation network to produce a combined network.

Another essential module in the R package is the prediction module that uses the iOmicsPASS results to assign new samples to the phenotypic groups based on the set of predictive signatures. The prediction module uses a discriminant model to assign each sample to the phenotype group with the highest discriminant score. In this module, the discriminant score is modified such that the prior probability incorporates other clinical information such as age, gender and BMI, that may influence the classification.

The following sections will describe the network inference method in greater detail.

2.8.1. Notation

We first define a few notations for the subsequent sections. Consider a p -dimensional data, $X = (X_1, X_2, \dots, X_p)^T$, with mean vector μ and covariance matrix $\Sigma_{p \times p} = (\sigma_{ij})_{1 \leq i, j \leq p}$.

Let

$$X_i = (X_{1i}, X_{2i}, \dots, X_{ni}), \quad i = 1, \dots, p$$

denote cross-sample observations of data feature i , where n is the sample size of the data. The covariance matrix Σ is estimated by the sample covariance matrix S , where the $(i, j)^{th}$ entry represents the sample covariance of X_i and X_j calculated as:

$$s_{ij} = \frac{1}{n-1} \sum_{r=1}^n (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) \quad (1)$$

Then, the precision matrix, $\Omega = (\omega_{ij})_{1 \leq i, j \leq p}$, is the inverse of the covariance matrix Σ and the $(i, j)^{th}$ entry describes the conditional dependency of X_i and X_j on all other $(p-2)$ variables. If $\omega_{ij} = 0$, then X_i and X_j are said to be conditionally independent given other variables present. If $\omega_{ij} \neq 0$, then X_i and X_j are not conditionally independent and an edge can be formed between the two in a network setting.

2.8.2. Network inference module

In this module, there are two proposed ways to create a pseudo network, linking one –omics to another. First, a supervised approach is completely driven by the estimates from the data. Second, a hybrid approach, which uses a mix of what is known as a prior and what is derived from the data by the supervised approach to produce a

combined network adjacency matrix. The subsections below provide a detailed description of the method in the two approaches.

Supervised approach

In the supervised approach, the input data is first standardized to unit standard deviation so that the sample covariance matrix and the sample correlation matrix are equivalent. Before the sample covariance matrix is computed, PCA is used to identify any outliers, defined to be more than 4 SDs away from the median of the first 2 PCs derived from concatenating all the data into a single matrix. After filtering outliers, the sample covariance matrix S is computed and corrected to the nearest PSD matrix if any of its eigenvalues are negative.

To convert a matrix A to the nearest PSD matrix, the approach by Nicholas J. Higham (15) is used:

$$A_{PSD} = \frac{B+H}{2} \quad (2)$$

where $B = \frac{(A+A^T)}{2}$, $H = V\Sigma V^*$ and $B = W\Sigma V^*$.

Here, H is the symmetric polar factor of matrix B where the polar decomposition of $B = UH$ ($U^T U = I, H = H^T \geq 0$). U represents the unitary matrix and H represents the hermitian matrix with PSD property and $B = W\Sigma V^*$ using singular value decomposition (SVD) where $V^* = V^T$ is the conjugate transpose of matrix V .

To ensure that the corrected sample covariance matrix is not too far from the actual sample covariance matrix, the difference of the Frobenius norm, $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, of the two matrices is reported in the software. Upon ensuring that the covariance matrix satisfies the PSD property, graphical LASSO in “huge” R package is implemented. The function requires a specification of lambda values to tune the L₁-penalty term in the model and computes the corresponding penalized log-likelihoods at all lambda values. Model selection criteria, including AIC, BIC, e-BIC and cross-validation (CV), are used to help users to select an optimal regularization parameter λ . Users may choose to specify their own vector of lambda values, otherwise, the software automatically generates a grid of 30 lambda values that is exponentially decreasing from 1 to 0.01:

$$\lambda_{grid} = \{\lambda_1, \lambda_2, \dots, \lambda_{30}\} = \exp\{\log(1), \log(0.853), \dots, \log(0.01)\} \quad (3)$$

As a visual aid, a calibration plot is produced as a portable document format (PDF) with all four model selection criteria plotted against the vector of lambda values and users can recalibrate the lambda values and specify the model selection criterion to use to select the optimal regularization parameter.

After selecting an optimal regularization parameter λ , the final GLASSO model is fitted and the estimated sparse inverse covariance matrix, $\widehat{\Omega}$, is produced. The non-zero entries in this matrix are then converted into an edge-level network file to be used for the predictive analysis module. In addition, a partial correlation matrix is produced as output to provide a measure of confounding-free correlation for every pair of data features present in the network. For entries where the estimated inverse covariance matrix are shrunken to zero, the partial correlations remain as zero. The regularized partial correlation estimate between data features i and j is calculated as follows (16, 17):

$$\tilde{\rho}_{i,j} = \frac{-\widehat{\omega}_{i,j}}{\sqrt{\widehat{\omega}_{i,i} \widehat{\omega}_{j,j}}} \quad (4)$$

where $\widehat{\omega}_{i,j}$ represents the $(i, j)^{th}$ entry in the estimated precision matrix, $\widehat{\Omega}$. Then, heat maps illustrating the derivation of the sample covariance matrix to the estimated nearest PD covariance matrix, the estimated precision matrix and the corresponding adjacency matrix, are also produced in PDF files.

Hybrid approach

The hybrid approach is a semi-supervised approach where the prior knowledge of the network is combined with the estimated network by the supervised approach above into a single adjacency matrix. First, a network file is required as input as the prior (matrix \mathbf{P}), then the supervised approach is carried out to obtain an estimated precision matrix (matrix $\widehat{\Omega}$). Let us first define the following matrices:

Prior network: Matrix $\mathbf{P}_{p \times p} = (p_{ij}) \in \{-1, 0, 1\}$

Estimated Precision network: Matrix $\widehat{\Omega}_{p \times p} = (\widehat{\omega}_{i,j}) \in [-1, 1]$

Sample covariance matrix: Matrix $\mathbf{S}_{p \times p} = (s_{ij}) \in [-1, 1]$.

Then we can create the following matrices:

$$\text{Matrix } \mathbf{A}: (a_{ij}) = \begin{cases} \text{sign}(\tilde{\rho}_{i,j}), & \text{if } i \neq j \\ 0, & \text{if } i = j, \end{cases}$$

$$\text{Matrix } \mathbf{U}: (u_{ij}) = \text{sign}(a_{ij} + p_{ij})$$

where the entries in \mathbf{A} denote the sign of the partial correlation and the entries in \mathbf{U} represent the agreement in the signs of the non-zero entries in matrix \mathbf{A} and \mathbf{P} . If the direction of association between feature i and j is

different in the prior from what is observed in the data (i.e. $a_{ij} = 1, p_{ij} = -1$ or $a_{ij} = -1, p_{ij} = 1$), the entry becomes zero.

We extract the corresponding elementwise sample covariances in \mathbf{S} from the non-zero entries in matrix \mathbf{U} by defining the Hadamard product of the two matrices to result in a non-negative matrix:

$$\mathbf{B} = \mathbf{U} \circ \mathbf{S} \quad (5)$$

A score matrix \mathbf{S}^* , constrained between 0 to 1, can be computed as follows:

$$\text{Matrix } \mathbf{S}^* = (s_{ij}^*) \begin{cases} -\frac{|b'_{ij}|}{b_{\max}} & \text{if } a_{ij} = 0, p_{ij} \neq 0 \\ \frac{|b'_{ij}|}{b_{\max}} & \text{otherwise,} \end{cases} \quad (6)$$

where $\mathbf{B}' = (b'_{ij}) = \mathbf{B}\mathbf{W}\mathbf{B}^T$ and $b_{\max} = \mathbf{max}_{i,j} b'_{ij}$. Here, the matrix $\mathbf{W} = \frac{1}{2}(|\mathbf{A} + \mathbf{P}|)$ is a weight matrix taking values 0, 0.5 and 1.

Lastly, a fused matrix \mathbf{F} is computed by:

$$\mathbf{F} = \mathbf{U} \circ (\mathbf{P} + \mathbf{S}^*) \quad (7)$$

adding to the prior if the edge is supported by what is estimated in $\hat{\Omega}$ (i.e. $\hat{\omega}_{i,j} \neq 0$), or penalizing the prior if the edge is present in the prior but not supported by the estimated network (i.e. $\hat{\omega}_{i,j} = 0$).

The final network is generated by forming an edge between feature i and j if the absolute value in $(i, j)^{th}$ entry of matrix \mathbf{F} is at least 0.5 and the direction of association of the edge is determined by the sign of the entry. Heatmaps illustrating the derivation of the sample covariance matrix to the estimated precision matrix, the input prior matrix, the product fused matrix and the corresponding adjacency matrix, are produced as a graphical output at the end.

2.8.3. Model selection criterion

To optimize the regularization parameter, λ , for the penalised log-likelihood in GLASSO that controls for the sparsity of the precision matrix, several model selection criteria are used, including Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (18, 19). However, when dealing with molecular data, we often encounter the curse of the high dimensionality where the size of the feature space, p , is much greater than the sample size, n . Hence, other criteria such as extended Bayesian information criterion (e-BIC) and K -fold CV may be more relevant.

Akaike Information Criterion, AIC

AIC is a model selection criterion which penalises for overfitting of the model. The formula for AIC is:

$$AIC_{\lambda} = 2k - 2\ln(\mathcal{L}) \quad (8)$$

where k is the number of non-zero elements in the estimated precision matrix ($\hat{\Omega}$) and \mathcal{L} is the likelihood function without the penalty term. The best regularised parameter would correspond to the model which yield the lowest AIC value.

Bayesian Information Criterion, BIC

BIC is similar to AIC but penalizes more against over-parameterization. Due to the heavier penalty, linear regression models chosen by BIC tends to be the same or simpler models than the one chosen by AIC (20). The formula for BIC is:

$$BIC_{\lambda} = k \ln(n) - 2\ln(\mathcal{L}) \quad (9)$$

where n is the sample size. One underlying assumption of BIC is that the observations should be independent and identically distributed (21). Otherwise, the effective sample size, n , should be adjusted accordingly. Then, the regularized parameter should be picked with the model which has the lowest BIC value.

Extended Bayesian Information Criterion, e-BIC

Another recently proposed model selection criterion is the extended BIC, which came about after Foyel *et al.* and Chen *et al.* showed that AIC and BIC lose consistency when the feature space exceeds the sample size (i.e. $p \gg n$). They propose a modification to the BIC formula as follows (19, 22):

$$e\text{-}BIC_{\lambda} = k \ln(n) - 2\ln(\mathcal{L}) + 4ky \log(p) \quad (10)$$

where $0 \leq \gamma \leq 1$, with higher values yielding simpler and more parsimonious models. When $\gamma = 0$, the formula reverts to BIC and $\gamma = 0.5$ is suggested to be used for more consistent models (19). Similar to both AIC and BIC, the best regularized parameter corresponds to the model with the smallest e-BIC value.

K-fold Cross-Validation, CV

An alternative approach to estimate a regularized parameter is through learning directly from the data. In 2009, Fan *et al.* (23) proposed a K -fold cross validation score given by:

$$CV(\lambda) = \sum_{k=1}^K n_k \log |\widehat{\Omega}_{-k}(\lambda)| - \sum_{i \in T_k} (x^{(i)})^T \widehat{\Omega}_{-k}(\lambda) x^{(i)} \quad (11)$$

where $\log |\Omega| = \log (\det(\Omega))$, n_k denotes the sample size and the set T_k represents the test dataset in the k^{th} fold. Here, $\widehat{\Omega}_{-k}(\lambda)$ is the precision matrix estimated by using the training dataset, $\cup_{j \neq k} T_j$. Typically, a five-fold cross-validation is recommended unless the sample size is too small and the regularised parameter is chosen at the value with the largest CV score.

2.8.4. Software architecture

iOmicsPASS+ is developed as an R package and the software is publicly available on GitHub (<https://github.com/cssblab/iOmicsPASSplus>). The package includes several modules to guide users to install, create an input parameter file containing the directory to which all the necessary files are and execute the program. The main software iOmicsPASS is still however coded in C++, and the functions enable users to execute the program from the R console, making it more user-friendly for individuals who are less comfortable with the command-line option. Below is the list of the functions included in the R package:

----- INSTALL.iOmicsPASS() -----

This function calls system to build and compile the C++ program from the specified directory. In Windows, it creates a Windows executable (.exe) and in Mac OS, it creates a compiled binary in the */bin* directory.

----- NetDeconvolute() -----

This function carries out the network inference module by first standardizing the input data (i.e. inputDat) and combining multiple datasets into a single matrix. PCA is then carried out to screen for outlying samples that are more than 4 SDs away from the median of the first 2 PCs. After this, a supervised or hybrid approach is performed to estimate the adjacency matrix from the precision matrix or fused matrix, respectively, and to form the resulting network.

----- createPrior() -----

This function fits a logistic model for a binary outcome or a multiple logistic model for a phenotypic group with more than 2 outcomes, adjusting for the user-specified clinical variables. User has to specify which variables are categorical and thus need to be converted to factor variables, otherwise they are treated as continuous variables. It then calculates the probability of every sample's membership to each phenotypic outcome based on the fitted model. The probabilities calculated can be fed into run.iOmicsPASS() to be used as prior in the discriminant model for classification.

The data can also be split into a training and test dataset where the logistic model is fitted on the training dataset and the classification probabilities are calculated on the test dataset, by using the option “predict=TRUE”. The probabilities calculated can be fed into Predict.iOmicsPASS() function as prior probabilities used in the discriminant model to assign test samples to the phenotypic outcome groups in the training model.

----- createInputParam() -----

This function creates an input parameter file needed to run iOmicsPASS with user’s specifications and the directory containing the required input files.

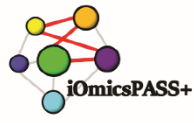
----- run.iOmicsPASS() -----

This function calls system and execute the compiled binary, performing predictive analysis to identify network signatures that can separate the phenotypic groups of interest. The software may take some time if the dimensionality of the network is very large. It is advised to first run CV once to create a misclassification error plot by setting Cross.Validate=TRUE, then turn off the CV option and re-specify a suitable threshold that strikes a good balance between change in mean CV error and the number of selected edges. Its function also outputs the key model parameters required for prediction module below.

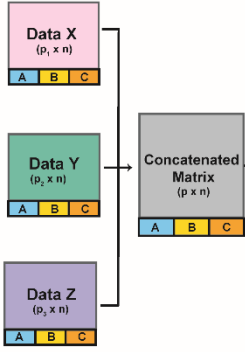
----- Predict.iOmicsPASS() -----

This function requires the model parameters associated with the network signatures obtained from a training dataset. Using the same network features in the test data to compute a co-expression matrix, class probabilities are computed to assign each new sample to a phenotypic group.

Overview of iOmicsPASS+



Each matrix is standardized and put together into a single matrix after screening for outliers



Other R functions

#INSTALL.iOmicsPASS()

Compiles the C++ program into /bin folder for running the subnetwork discovery module.

#createPrior()

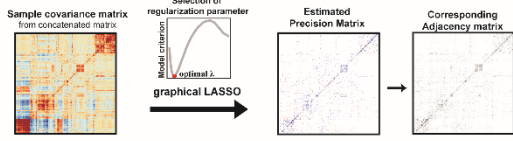
Creates prior class probabilities for the discriminant model by fitting a logistic/multinomial regression model with given clinical information or variables.

#createInputParam()

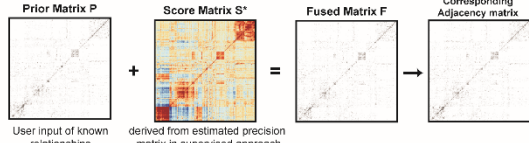
Helps user to create an input file, specifying the parameters needed to run the subnetwork discovery module.

Network Prediction module #NetDeconvolute()

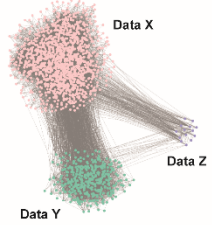
(A) Supervised Approach



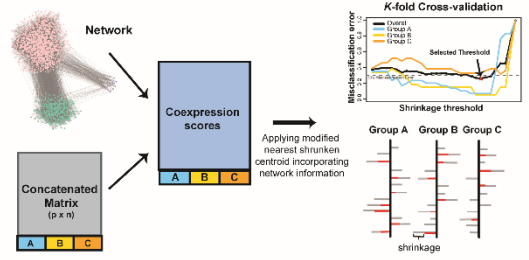
(B) Hybrid Approach



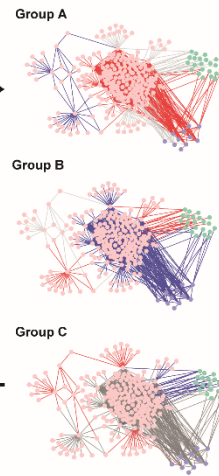
Estimated Correlation Network



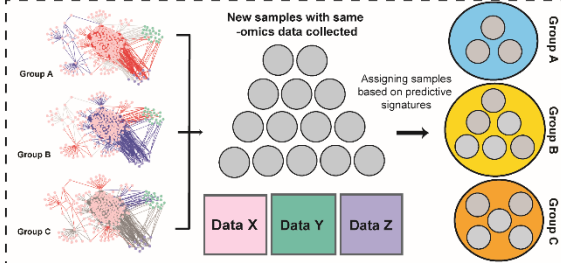
Subnetwork Discovery module #run.iOmicsPASS()



Identification of Predictive subnetwork signatures



Classification module #Predict.iOmicsPASS()



References

1. Prickett TC, Doughty RN, Troughton RW, Frampton CM, Whalley GA, Ellis CJ, et al. C-Type Natriuretic Peptides in Coronary Disease. *Clin Chem*. 2017;63(1):316-24.
2. Daniel E. Ho KI, Gary King, Elizabeth A. Stuart. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*. 2011;42(8):1-28.
3. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One*. 2010;5(12):e15004.
4. Chan MY, Efthymios M, Tan SH, Pickering JW, Troughton R, Pemberton C, et al. Prioritizing Candidates of Post-Myocardial Infarction Heart Failure Using Plasma Proteomics and Single-Cell Transcriptomics. *Circulation*. 2020;142(15):1408-21.
5. Teo G, Chew WS, Burla BJ, Herr D, Tai ES, Wenk MR, et al. MRMkit: Automated Data Processing for Large-Scale Targeted Metabolomics Analysis. *Anal Chem*. 2020;92(20):13677-82.
6. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*. 2007;35(Web Server issue):W193-200.
7. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-5.
8. Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
9. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. The huge Package for High-dimensional Undirected Graph Estimation in R. *J Mach Learn Res*. 2012;13:1059-62.
10. Warnes G, Bolker B, Bonebakker L, Gentleman R, Liaw W, Lumley T, et al., editors. gplots: Various R Programming Tools for Plotting Data 2015.
11. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(3):1 - 67.
12. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
13. Yu G WL, Han Y, He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*. 2012;16(6):264-87.
14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504.
15. Higham NJ. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*. 1988;103:103-18.
16. Mohsen P. Covariance estimation: the GLM and regularization perspectives. *Statist Sci*. 2011;26(3):369-87.
17. Lee KW, Lip GY, Blann AD. Plasma angiopoietin-1, angiopoietin-2, angiopoietin receptor tie-2, and vascular endothelial growth factor levels in acute coronary syndromes. *Circulation*. 2004;110(16):2355-60.
18. Yuan M, Lin Y. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*. 2007;94(1):19-35.
19. Foygel R. D, M. Extended Bayesian Information Criteria for Gaussian Graphical Models. Culotta JDL, CKI, WaJS-TaRSZaA, editor: Curran Associates, Inc.; 2010.
20. Shao J. An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*. 1997;7(2):221-42.
21. Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics*. 1978;6(2):461-4.
22. Chen J. CZ. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*. 2008;95(3):759-71.
23. Fan J, Lv J, Qi L. Sparse High Dimensional Models in Economics. *Annu Rev Econom*. 2011;3:291-317.