

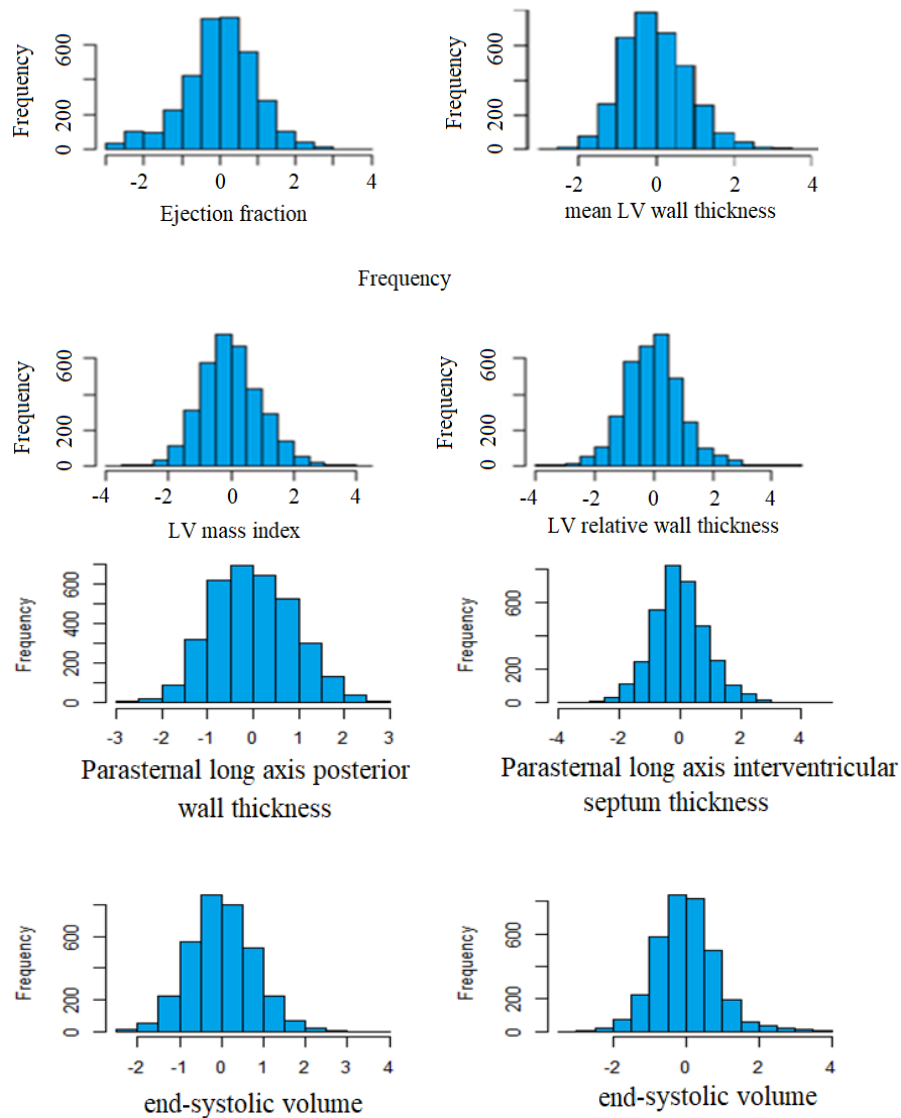
A Multi-Trait Approach Identified Genetic Variants Including a Rare Mutation in RGS3 with Impact on Abnormalities of Cardiac Structure/Function

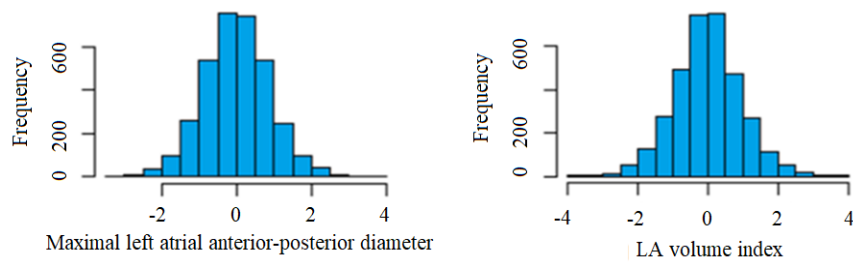
Akram Yazdani*, Azam Yazdani, Raúl Méndez Giráldez, David Aguilar, Luca Sartore

Supplementary Figures

Supplementary Figure 1

The histogram of the traits in Table1 after winsorization, standardization, and normal transformation

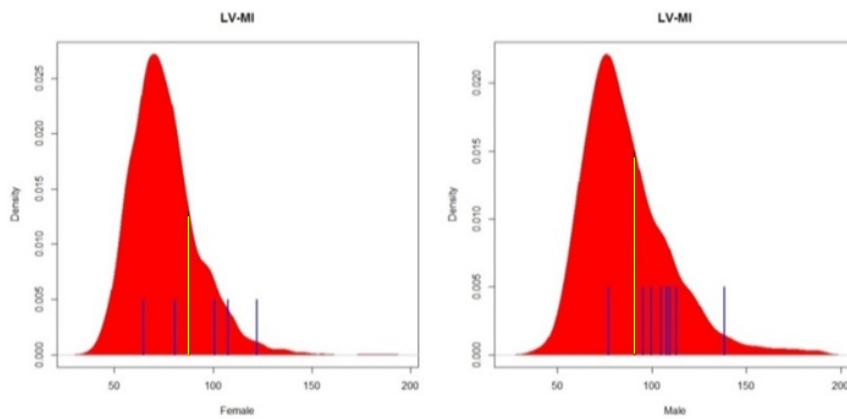




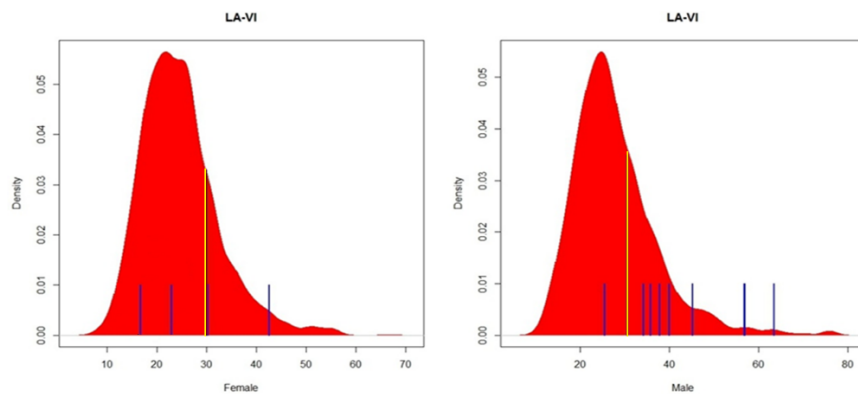
Supplementary Figure 2

Empirical Distributions of the traits in Table 3 for individuals with Reference allele: The yellow vertical line shows the third quartiles of the distribution. Individuals with alternative allele are represented with blue vertical line.

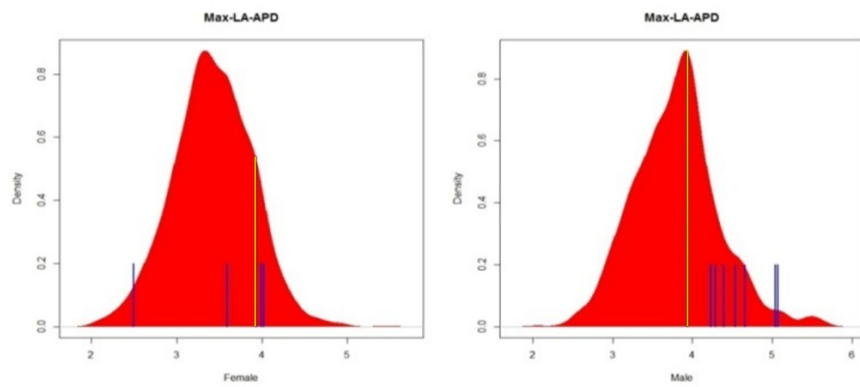
NC_000009.11:g.116346115C>A in *RGS3*



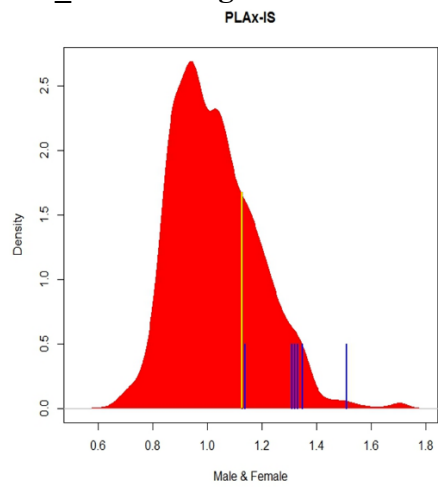
NC_000009.11:g.116346115C>A in *RGS3*



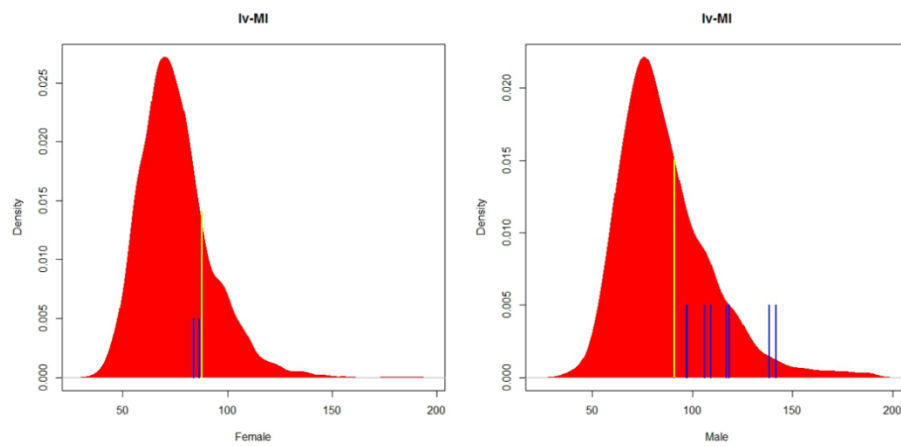
NC_000009.11:g.116346115C>A in *RGS3*



NC_000017.10:g.7802658C>T in *CHD3*



NC_000017.10:g.73897977C>T in *MRPL38*



Supplementary Tables

The VEP analysis results for identified genetic variants are summarized in the following tables where

- **HGVS name:** description of sequence variation in genomic established by the Human Genome Variation Society
- **Consequence:** consequence type of the allele on transcript
- **HGVSp:** HGVS protein sequence name
-

Supplementary Table 1

HGVS name	Allele	Consequence	Gene symbol	Exon	Intron	HGVSp	Amino acids	Codons
NC_000009.11: g.116346115C>A	A	missense_variant	RGS3	2/7	-	NP_001263189.1: p.Thr129Asn	T/N	aCc/aAc
NC_000009.11: g.116346115C>A	A	intron_variant	RGS3	-	10/14	-	-	-
NC_000009.11: g.116346115C>A	A	missense_variant	RGS3	3/8	-	NP_001269851.1: p.Thr129Asn	T/N	aCc/aAc
NC_000009.11:g. 116346115C>A	A	missense_variant	RGS3	18/23	-	NP_001269852.1: p.Thr698Asn	T/N	aCc/aAc
NC_000009.11: g.116346115C>A	A	missense_variant	RGS3	2/7	-	NP_001309144.1: p.Thr147Asn	T/N	aCc/aAc
NC_000009.11: g.116346115C>A	A	missense_variant	RGS3	11/16	-	NP_570613.2: p.Thr527Asn	T/N	aCc/aAc
NC_000009.11: g.116346115C>A	A	intron_variant	RGS3	-	1/4	-	-	-
NC_000009.11: g.116346115C>A	A	missense_variant	RGS3	21/26	-	NP_652759.3: p.Thr808Asn	T/N	aCc/aAc
NC_000009.11: g.116346115C>A	A	non_coding_ transcript_ exon_variant	RGS3	18/23	-	-	-	-
NC_000009.11: g.116346115C>A	A	non_coding_ transcript_ exon_variant	RGS3	3/8	-	-	-	-

Supplementary Table 2

HGVS name	Allele	Consequence	Gene symbol	Biotype	Exon	Intron	HGVSp	Amino acids	Codons
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/39	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/39	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/38	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/39	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/38	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/39	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/33	-	-	-
NC_000017.10: g.7802658C>T	T	splice_region_ variant,intron_ variant	CHD3	protein_ coding	-	14/21	-	-	-

Supplementary Table 3

HGVS name	Allele	Consequence	Gene symbol	Biotype	Exon	Intron	HGVSp	Amino acids	Codons
NC_000017.10: g.73897977C>T	T	upstream_gene_ variant	TRIM65	protein_ coding	-	-	-	-	-
NC_000017.10: g.73897977C>T	T	missense_variant	MRPL38	protein_ coding	4/9	-	-	R/Q	cGg/cAg
NC_000017.10: g.73897977C>T	T	upstream_gene_ variant	TRIM65	protein_ coding	-	-	-	-	-
NC_000017.10: g.73897977C>T	T	non_coding_ transcript_ exon_variant	MRPL38	misc_RNA	4/8	-	-	-	-

Supplementary statistical methods

Let assume that we measured q phenotypic traits for n individuals with p recorded genotypic variants. The multi-trait polygenic mixed model for this data can be represented as

$$\mathbf{Y}_{nq \times 1} = \boldsymbol{\mu}_{nq \times 1} + (\mathbf{X}_{n \times p} \otimes \mathbf{I}_q) \boldsymbol{\beta}_{pq \times 1} + \mathbf{U}_{nq \times 1} + \boldsymbol{\epsilon}_{qn \times 1} \quad (1)$$

where

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma), \Sigma_{nq \times nq} = \text{diag}[\Sigma_{ii}]_{q \times q}, \Sigma_{ii} = \text{diag}[\sigma_i]_{n \times n}.$$

Each entry of $\mathbf{Y} = \{y_{ij}\}_{i=1, \dots, q, j=1, \dots, n}$ represents i th trait recorded for j th individual, β_{ik} is the effect of

k th genomic variants on i th trait in coefficient vector $\boldsymbol{\beta} = \{\beta_{ik}\}_{i=1, \dots, q, k=1, \dots, p}$, vector $\mathbf{U} = \{u_{ij}\}_{i=1, \dots, q, j=1, \dots, n}$

includes random effects corresponding to vector $\mathbf{Y}$ as

$$\mathbf{U} \sim N(\mathbf{0}, \Psi)$$

where

$$\Psi = \begin{bmatrix} \Psi_{11} & \cdots & \Psi_{1q} \\ \vdots & \ddots & \vdots \\ \Psi_{q1} & \cdots & \Psi_{qq} \end{bmatrix},$$

and $\Psi_{il} = \{\psi_{jg}\}_{j=1,\dots,n, g=1,\dots,n}, \quad i, l \in \{1, \dots, q\}.$

Without loss of generality, we assume y_{ij} s are standardized and write the likelihood function as

$$\ell(\boldsymbol{\beta}, \mathbf{U}, \Sigma, \Psi | \mathbf{Y}) \propto \exp \left\{ -\frac{1}{2} \mathbf{e}^T \Sigma^{-1} \mathbf{e} \right\}$$

where $\mathbf{e} = \mathbf{Y} - (\mathbf{X} \otimes \mathbf{I}_q) \boldsymbol{\beta} - \mathbf{U}.$

Bayesian network over traits: The large sample size (n) and large number of traits (q) introduce a large number of parameters into the model through Ψ and make the numerical algorithm of model (1) unstable and slow to converge. Therefore, to reduce the number of parameters in multi-trait polygenic mixed model and provide larger precision for parameter estimations, we propose to estimate the sparse structure of precision matrix Ψ^{-1} in priori using phenotypic data (individuals without genotype record) which is accessible in majority of medical studies. Here in particular, we apply a Bayesian network that relies on probabilistic graphical models to estimate sparse relationship among traits.

Bayesian networks represent underlying relationship among traits based on efficient and effective representation of their joint probability distribution (1)(2)(3) where nodes represent traits and links represent partial correlations. A missing edge between two traits reveals that the two corresponding traits do not have a significant relationship after excluding effect of the other traits in the analysis. To learn about sparsity of precision matrix Ψ^{-1} , we incorporate independencies inferred from Bayesian networks to reduce the number of parameters in the model and efficiently compute posterior probabilities. In the matrix Ψ^{-1} , we replace zero with the parameters corresponding to missing arrows in the identified Bayesian network over the traits. To avoid overfitting, we identify the Bayesian network using the data on the set of individuals without genotype record.

Prior Specification: There have been many studies on setting a prior distribution on precision matrix Ψ^{-1} . The most common approach is to set Wishart distribution as prior on Ψ^{-1} while it is conjugate prior for multivariate normal model. Wishart distribution is fully parameterized with a single degree of freedom parameter and scale matrix parameter. The degree of freedom parameter that needs to be larger than dimension of relatedness matrix (Ψ) represents the strength of information surrounded around scale matrix parameter. Therefore, in large scale problems, this choice of prior yields to highly concentrated distribution about the scale matrix(4)(5)(6).

We here assume that individuals are independent which can be readily satisfied in population studies and clinical studies that are not family based. The assumption of independent individuals leads to a sparse relatedness matrix Ψ as

$$\Psi = \begin{bmatrix} \Psi_{11} & \cdots & \Psi_{1q} \\ \vdots & \ddots & \vdots \\ \Psi_{q1} & \cdots & \Psi_{qq} \end{bmatrix}, \quad \Psi_{il} = \text{diag}[\psi_{jg}], \quad i, l \in \{1, \dots, n\}, \quad j, g \in \{1, \dots, q\}.$$

To set prior on this matrix, we first rearrange vector $\mathbf{U}$ in order to be partition with individual's index j . If we denote the rearranged random effect vector with $\mathbf{U}^*$ such that $\mathbf{U}^* \sim N(\mathbf{0}, \Psi^*)$, Ψ^* is a block diagonal as

$$\Psi^* = \begin{bmatrix} \Psi_{11}^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_{nn}^* \end{bmatrix}$$

where each Ψ_{jj}^* is a $q \times q$ dense matrix. Ψ_{jj}^* takes into account correlation among different traits for j th individual. While Ψ^{*-1} is a block diagonal matrix and q is a small number, the use of Wishart prior for Ψ_{jj}^{*-1} is appropriate. Furthermore, we incorporate the identified relationship among the traits using Bayesian network and estimate the sparsity of Ψ_{jj}^{*-1} and set G-Wishart (GW) distribution (7) as prior on Ψ_{jj}^{*-1} . G-Whishart distribution, which is a restricted domain of

Wishart given constraints of network structure, provides sparse block diagonal matrix and reduces computational burden time. The density of G-Wishart is

$$p(\Psi_{jj}^{*-1} | \text{BN}_T) = \left(I_{GT}(\nu, \Lambda) \right)^{-1} |\Psi_{jj}^{*-1}|^{\frac{\nu-2}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\Lambda \Psi_{jj}^{*-1}) \right\}$$

where BN_T stands for the Bayesian Networks over the traits, and

$$I_{GT}(\nu, \Lambda) = \int |\Psi_{jj}^{*-1}|^{(\nu-2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\Lambda \Psi_{jj}^{*-1}) \right\} d\Psi_{jj}^{*-1}$$

is the normalizing constant, which is finite for $\nu > 2$. This results in a reduction of the number of parameters in multi-trait polygenic mixed model, and consequently expedites the convergence of the algorithm.

We set conjugate prior for other hyperparameters of the model as

$$\boldsymbol{\beta} \sim N(0, \Omega)$$

where

$$\Omega_{pq \times pq} = \text{diag} [\Omega_{kk}]_{p \times p}, \quad \Omega_{kk} = \text{diag} [\sigma_j]_{p \times p}$$

and

$$\sigma_i \sim IG \left(\frac{a_1}{2}, \frac{b_1}{2} \right)$$

The aforementioned prior specification leads to Gibbs sampling scheme for the model with the graphical representation shown in the following Figure.

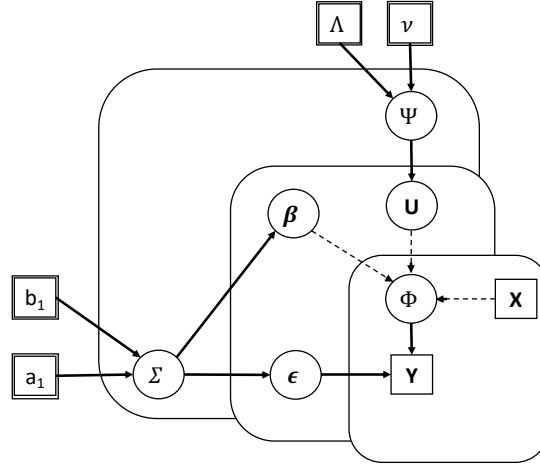


Figure: Graphical representation of the model; $\mathbf{Y} = \Phi + \epsilon$ and $\Phi = (\mathbf{X} \otimes \mathbf{I})\boldsymbol{\beta} + \mathbf{U}$

In the graphical representation of the model, circle nodes represent unobserved stochastic variables, square nodes with a single border represent observed stochastic variables (i.e., data), squares with double border denote fixed quantities in prior distributions. The arrows represent the dependencies between variables.

Gibbs Sampling Scheme of IBMT:

$$\boldsymbol{\beta}_i \mid \mathbf{y}_i, \mathbf{u}_i, \sigma_i \sim \mathcal{N} [(\mathbf{X}^T \mathbf{X} + \mathbf{I})^{-1} \mathbf{X}^{-1} (\mathbf{y}_i - \mathbf{u}_i), \sigma_i (\mathbf{X}^T \mathbf{X} + \mathbf{I})^{-1}]$$

$$\mathbf{u}_j^* \mid \mathbf{z}_j, \Psi_{jj}^*, \Sigma_{jj} \sim \mathcal{N} \left[(\Psi_{jj}^* + \Sigma_{jj}^{-1})^{-1} \Sigma_{jj}^{-1} \mathbf{z}_j, (\Psi_{jj}^* + \Sigma_{jj}^{-1})^{-1} \right]$$

$$\sigma_i \mid \mathbf{y}_i, \boldsymbol{\beta}_i, \mathbf{u}_i \sim \text{ING} \left[\frac{n+p+a_1}{2}, \frac{1}{2} (\mathbf{e}_i^T \mathbf{e}_i + \boldsymbol{\beta}_i^T \boldsymbol{\beta}_i + b_1) \right]$$

$$\Psi_{jj}^{*-1} \mid \mathbf{u}_j \sim \text{GW} [\Lambda + \mathbf{u}_j^T \mathbf{u}_j, \nu + q]$$

where $\mathbf{e}_i = (\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{U}_i)$ and vector $\mathbf{Z}$ is obtained after reordering entries of vector $(\mathbf{Y} - \mathbf{X} \otimes \mathbf{I}_q) \boldsymbol{\beta}$ in order to be partitioned by each individual similar to $\mathbf{U}^*$.

Supplementary references

1. Pearl J. Probabilistic Reasoning in Intelligent Systems. Probabilistic Reason Intell Syst. 1988;552.
2. Yazdani A, Yazdani A, Samiei A, Boerwinkle E. Erratum to: A causal network analysis in an observational study identifies metabolomics pathways influencing plasma triglyceride levels. J Biomed Inform. 2016;63:337–43.
3. Yazdani A, Yazdani A, Boerwinkle E. A Causal Network Analysis of the Fatty Acid Metabolome in African-Americans Reveals a Critical Role for Palmitoleate and Margarate. Omi A J Integr Biol. 2016;20(8):480–4.
4. Xiang R, Khare K, Ghosh M. High dimensional posterior convergence rates for decomposable graphical models. Electron J Stat. 2015;9(2):2828–54.
5. Banerjee S. Posterior convergence rates for high-dimensional precision matrix estimation using G-Wishart priors. Stat. 2017;6(1):207–17.
6. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. 2006;1(3):515–34.
7. Roverato A. Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models. Scand J Stat. 2002;29(1993):391–411.