DAP-PIR: Efficient Genetic Association Analysis through Pseudo Importance Resampling

Bo Wang[1] and Xiaoquan Wen[1*]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

*Corresponding author(s). E-mail(s): xwen@umich.edu;
Contributing authors: wnbo@umich.edu;

Abstract

In genome-wide association studies (GWAS), identifying causal variants from a vast number of genetic variants is challenging. This paper presents DAP-PIR, a method combining Deterministic Approximation of Posteriors (DAP) with Pseudo-Importance Resampling (PIR), to enhance computational efficiency and improve the accuracy of fine-mapping in genetic studies. By selecting a limited number of high-probability models, DAP-PIR effectively narrows down the candidate models, maintaining accuracy in posterior estimation while reducing computational costs. Simulations and real data analyses show the superiority of DAP-PIR over existing methods, including SuSiE, particularly in managing high-dimensional genetic data.

Keywords: Genetic fine mapping, variable selection, importance sampling, variational inference

1 Introduction

Genome-wide association studies (GWAS) have successfully uncovered thousands of genetic variants associated with complex traits and diseases [1]. However, due to the intricate linkage disequilibrium (LD) structure in the human genome, many variants identified in GWAS are merely correlated with true causal variants rather than causative themselves [2]. Fine mapping analysis is therefore essential to refine these associations and identify specific genetic variants likely to be causal [3, 4].

To address this challenge, Bayesian variable selection regression (BVSR) methods are widely employed to account for the uncertainty in identifying candidate causal variants. Traditional methods like CAVIAR rely on exhaustive searches [5], while DAP-G employs a boosting approach to prioritize variants, which, although effective, can sometimes be slow in practice [6, 7]. Alternatively, the SuSiE method offers an efficient, variational approach to prioritize SNPs based on posterior inclusion probabilities (PIP) [8]. However, SuSiE can occasionally experience accuracy issues with PIP, particularly in complex LD structures.

Here, we propose DAP-PIR, which combines the strengths of both DAP and SuSiE to achieve efficient and accurate deterministic approximation of posteriors. By leveraging SuSiE's variational approximation results, DAP-PIR efficiently prioritizes models for deterministic posterior estimation. This approach ensures computational efficiency while providing a robust solution to fine mapping challenges in GWAS. Through both simulations and real data applications, we demonstrate DAP-PIR's advantages over existing methods in terms of both speed and accuracy, making it an effective tool for fine mapping causal variants in GWAS.

2 Results

2.1 Overview of the DAP-PIR algorithm

2.2 Comparison of fine mapping methods

We conduct several simulation studies to show the accuracy and efficiency of the DAP-PIR algorithm. Comparing the (1) power; (2) coverage; (3) PIP calibration; (4) computational efficiency among fine mapping methods.

2.3 Real data results

We analyze some real data using various fine mapping methods.

3 Discussion

4 Methods

4.1 DAP-PIR

4.1.1 Bayesian variable selection model

We consider the context of association analysis for a single quantitative trait. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ denote an n-dimensional vector representing the trait of interest, $\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_p)$ represent an $n \times p$ genotype matrix, where each column $\mathbf{g}_j$ is an n-dimensional vector of genotypes for the j-th variant. We assume that $\mathbf{y}$ and the columns of $\mathbf{G}$ are centered to have a mean of zero to avoid the need for an intercept term. The model is then defined as:

$$\mathbf{y} = \sum_{j=1}^{p} \beta_j \mathbf{g}_j + \mathbf{e},\ \mathbf{e} \sim N_n\left(0, \tau^{-1} I_n\right), \tag{1}$$

where β_j denotes the genetic effect of the j-th variant, and τ represents the precision. The goal of GWAS fine mapping is to prioritize causal variants rather than estimate their effect sizes, therefore we introduce an inclusion indicator $\gamma_j = I(\beta_j \neq 0)$ for each variant, where $\gamma_j = 1$ indicates that the j-th variant is causal. The primary aim is infer the posterior distribution of the causal status vector: $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$. While evaluating each possible model configuration $\boldsymbol{\gamma}$ is straightforward as it only requires integrating out β_j's and τ, summing over all 2^p model configurations becomes computationally intractable as p increases. This makes traditional Markov Chain Monte Carlo (MCMC) methods inefficient for fine mapping tasks involving even a moderate number of variants. Therefore, we utilize the efficient pseudo importance resampling method to explore the model space and approximate the posterior distribution of $\boldsymbol{\gamma}$.

4.1.2 Pseudo importance resampling

For a discrete random variable $\mathbf{X}$ with target distribution $p(\mathbf{x})$, the objective is to approxiamte the expectation $\mathbb{E}_p[f(\mathbf{X})]$ for a function $f(\mathbf{x}) = \mathbf{x}$. Due to the high-dimensional sample space, exact computation is impractical. We employ Importance Sampling to estimate this expectation by sampling from a proposal distribution $q(\mathbf{x})$ and assigning importance weights $w\left(\mathbf{x}^i\right) = p\left(\mathbf{x}^i\right)/q\left(\mathbf{x}^i\right)$ for samples $\mathbf{x}^i \sim q(\mathbf{x})$. The IS estimate is:

$$\tilde{\boldsymbol{\mu}}_{\text{IS}} = \frac{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right) f\left(\mathbf{x}^i\right)}{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right)}. \tag{2}$$

To obtain a reliable estimate, $q(\mathbf{x})$ should closely approximate $p(\mathbf{x})$, particularly in regions where $f(\mathbf{x})\, p(\mathbf{x})$ has significant values.

We use Variational Approximation (VA) to construct an effective proposal distribution by minimizing the Kullback-Leibler (KL) divergence between $p(\mathbf{x})$ to $q(\mathbf{x})$. VA provides a tractable approxiamtion, balancing efficiency and accuracy. The Variational Importance Sampling (VIS) method integrates VA with IS, leveraging VA's computational efficiency while reducing bias through IS.

Sampling Importance Resampling (SIR) selects IS-generated samples proportional to their importance weights to approximate $p(\mathbf{x})$ [9, 10]. Given n samples $\left\{\left(\mathbf{x}^i, w\left(\mathbf{x}^i\right)\right) : i = 1, \cdots, n\right\}$, SIR resamples m values from them according to their importance weights:

$$\Pr\left[\mathbf{x}^* = \mathbf{x}^k\right] = \frac{w\left(\mathbf{x}^k\right)}{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right)}. \tag{3}$$

The resampled set $\left\{\mathbf{x}^{*j} : j = 1, \cdots, m\right\}$ has equal weights and represents the target distribution p. The SIR algorithm is outlines in Algorithm 2. The rationale behind SIR is based on the fact that as n/m tends to infinity, the m samples are drawn with

Algorithm 1 Variational Importance Sampling (VIS)

```
1. Target distribution: derive the analytical expression of $p'(\mathbf{x})$ up to a normalizing constant.
2. Proposal distribution: obtain the variational approximation $q(\mathbf{x})$ to $p(\mathbf{x})$ within the family $\mathcal{Q}$.
3. Sampling: for $i \in \{1, \ldots, n\}$,
   (a) draw $\mathbf{x}^i$ from the proposal distribution $q(\mathbf{x})$.
   (b) calculate the function value $f\left(\mathbf{x}^i\right)$.
   (c) calculate the importance weight $w\left(\mathbf{x}^i\right) = p'\left(\mathbf{x}^i\right)/q\left(\mathbf{x}^i\right)$.
4. Estimation: estimating $\mathbb{E}_p[f\left(\mathbf{X}\right)]$ by $\tilde{\boldsymbol{\mu}}_{\mathrm{IS}} = \frac{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right) f\left(\mathbf{x}^i\right)}{\sum_{i=1}^{n} w\left(\mathbf{x}^i\right)}$.
```

probabilities given by:

$$p^*\left(\mathbf{x}\right) \propto q\left(\mathbf{x}\right) w\left(\mathbf{x}\right) \propto q\left(\mathbf{x}\right) \frac{p\left(\mathbf{x}\right)}{q\left(\mathbf{x}\right)} = p\left(\mathbf{x}\right),$$

which shows that the SIR algorithm generates independent and identically distributed (i.i.d) samples from the target distribution $p\left(\mathbf{x}\right)$, as desired. However, drawing such a large number of samples can be computationally expensive. To balance computational efficiency and performance, Rubin suggested using $n/m = 20$ as a practical ratio for resampling, which provides adequate performance without requiring excessive duplicates in the resampling [9].

Algorithm 2 Sampling Importance Resampling (SIR)

```
1. Target distribution: the distribution $p\left(\mathbf{x}\right)$ with the expression of $p'\left(\mathbf{x}\right)$ up to a normalizing constant.
2. Proposal distribution: the variational approximation $q\left(\mathbf{x}\right)$ to $p\left(\mathbf{x}\right)$ within the family $\mathcal{Q}$.
3. Sampling: for $i \in \{1, \ldots, n\}$,
   (a) draw $\mathbf{x}^i$ from the proposal distribution $q\left(\mathbf{x}\right)$.
   (b) calculate the importance weight $w\left(\mathbf{x}^i\right) = p'\left(\mathbf{x}^i\right)/q\left(\mathbf{x}^i\right)$.
4. Resampling: for $j \in \{1, \ldots, m\}$,
   (a) resample $\mathbf{x}^{*j}$ from the samples $\left\{\mathbf{x}^i : i = 1, \cdots, n\right\}$ with probabilities proportional to the importance weights.
   (b) include $\mathbf{x}^{*j}$ in the resampled set.
5. Result: the resampled set $\left\{\mathbf{x}^{*j} : j = 1, \cdots, m\right\}$ which is approximately distributed according to the target distribution $p\left(\mathbf{x}\right)$.
```

To enhance efficiency, we introduce Pseudo Importance Resampling (PIR), which eliminates explicit resampling by shrinking the proposal distribution. PIR focuses on regions where $p\left(\mathbf{x}\right)$ has meaningful contributions, reducing computational cost in high-dimensional settings.

In practice, if a value of $\mathbf{x}$ exhibits a near-zero probability density under the target distribution $p(\mathbf{x})$, its contribution to the final result becomes negligible. In such cases, the proposal distribution can be effectively reduced to zero for that region, excluding it from further consideration. This reduction in the number of regions to explore allows us to focus computational resources on areas where $p(\mathbf{x})$ significantly impacts the result. Prioritizing these high-priority regions improves efficiency in high-dimensional settings, making the process more computational feasible. The KL divergence penalizes cases where p is large while $q(\mathbf{x})$ is small, implying that when $q(\mathbf{x})$ is small, it is highly probable that $p(\mathbf{x})$ is also small. As a result, we can safely reduce the sample space by identifying and excluding values of $\mathbf{x}$ that are negligible under the proposal distribution. Even in cases where $p(\mathbf{x})$ is small but $q(\mathbf{x})$ is large (and thus not excluded), the contribution to the overall result remains minimal due to the small value of $p(\mathbf{x})$.

For a specific point $\mathbf{x}^{(k)}$ in the target distribution, the proposal distribution shrinkage is defined as:

$$q'\left(\mathbf{x}^{(k)}\right) = \begin{cases} 0, & \text{if } q\left(\mathbf{x}^{(k)}\right) < \epsilon, \\ q\left(\mathbf{x}^{(k)}\right), & \text{if } q\left(\mathbf{x}^{(k)}\right) \geq \epsilon, \end{cases}$$

where ϵ is a threshold value. The normalized proposal distribution is given by:

$$q^*\left(\mathbf{x}^{(k)}\right) = \frac{q'\left(\mathbf{x}^{(k)}\right)}{\sum_{i=1}^{N} q'\left(\mathbf{x}^{(i)}\right)} = \frac{q'\left(\mathbf{x}^{(k)}\right)}{\sum_{j=1}^{M} q\left(\mathbf{x}^{(j)}\right)},$$

where $\mathbf{x}^{(j)}$ corresponds to points with non-zero proposal density, N is the total number of possible values, and M is the number of points with non-zero proposal density. During the SIR process, the probability of selecting a value $\mathbf{x}^{(k)}$ in the reduced sample space is calculated as:

$$\begin{aligned}
\Pr\left(\mathbf{x}^* = \mathbf{x}^{(k)}\right) &= \frac{w\left(\mathbf{x}^{(k)}\right) \#\left(\mathbf{x}^{(k)}\right)}{\sum_{i=1}^{N} w\left(\mathbf{x}^{(i)}\right) \#\left(\mathbf{x}^{(i)}\right)} \\
&= \frac{w\left(\mathbf{x}^{(k)}\right) \#\left(\mathbf{x}^{(k)}\right)}{\sum_{j=1}^{M} w\left(\mathbf{x}^{(j)}\right) \#\left(\mathbf{x}^{(j)}\right)} \\
&\xrightarrow{n\to\infty} \frac{\frac{p'\left(\mathbf{x}^{(k)}\right)}{q^*\left(\mathbf{x}^{(k)}\right)} q^*\left(\mathbf{x}^{(k)}\right) \times n}{\sum_{j=1}^{M} \frac{p'\left(\mathbf{x}^{(j)}\right)}{q^*\left(\mathbf{x}^{(j)}\right)} q^*\left(\mathbf{x}^{(j)}\right) \times n} \\
&= \frac{p'\left(\mathbf{x}^{(k)}\right)}{\sum_{j=1}^{M} p'\left(\mathbf{x}^{(j)}\right)},
\end{aligned}$$

where $\#\left(\mathbf{x}^{(k)}\right)$ is the number of samples with value equal to $\mathbf{x}^{(k)}$. Thus, instead of drawing an large number of n samples and resampling an adequate number of m from them, we can directly approximate the probability of each value in the target distribution by focusing only on the M points with non-negligible probability. This process, based on SIR but avoiding the separate steps of sampling and resampling, is referred to as Pseudo Importance Resampling (PIR) outlined in Algorithm 3. PIR effectively reduces the sample space by shrinking the proposal distribution and concentrating on the regions that matter most, providing a good approximation to the target distribution.

Algorithm 3 Pseudo Importance Resampling (PIR)

```
1. Target distribution: the distribution $p(\mathbf{x})$ with the expression of $p'(\mathbf{x})$ up to a normalizing constant.
2. Proposal distribution: the variational approximation $q(\mathbf{x})$ to $p(\mathbf{x})$ within the family $\mathcal{Q}$.
3. Proposal distribution shrinkage: shrink $q(\mathbf{x})$ to obtain $q^*(\mathbf{x})$, reducing the sample space to $M$ non-negligible partitions.
4. Pseudo resampling: for $j \in \{1, \ldots, M\}$,
   (a) calculate the probability score $p'\left(\mathbf{x}^{(j)}\right)$.
   (b) calculate the approximation $p'\left(\mathbf{x}^{(j)}\right) / \sum_{j=1}^{M} p'\left(\mathbf{x}^{(j)}\right)$.
```

4.1.3 Deterministic approximation of poseteriors through PIR

The posterior probability of one model configuration $\boldsymbol{\gamma}^{(i)}$ can be derived using Bayes' theorem:

$$p\left(\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(i)} \mid \mathbf{G}, \mathbf{y}\right) = \frac{\pi\left(\boldsymbol{\gamma}^{(i)}\right) \mathrm{BF}\left(\boldsymbol{\gamma}^{(i)}\right)}{\sum_{\boldsymbol{\gamma}' \in \Gamma} \pi\left(\boldsymbol{\gamma}'\right) \mathrm{BF}\left(\boldsymbol{\gamma}'\right)},$$

where Γ denotes all 2^p possible model configurations, $\pi\left(\boldsymbol{\gamma}^{(i)}\right)$ is the prior probability of model $\boldsymbol{\gamma}^{(i)}$, and $\mathrm{BF}\left(\boldsymbol{\gamma}^{(i)}\right)$ is the Bayes factor, which measures the marginal likelihood of $\boldsymbol{\gamma}$ evaluated at $\boldsymbol{\gamma}^{(i)}$. The posterior inclusion probability (PIP) for each variant is then calculated by marginalizing the posterior probabilities across all model configurations:

$$\mathrm{PIP}_j := \Pr\left(\gamma_j = 1 | \mathbf{G}, \mathbf{y}\right) = \sum_{\boldsymbol{\gamma}' : \gamma_j = 1} p(\boldsymbol{\gamma} = \boldsymbol{\gamma}' | \mathbf{G}, \mathbf{y}).$$

Under the D_2 prior, the prior distribution for τ is given by:

$$\tau \sim \Gamma\left(\kappa/2, \lambda/2\right),$$

where the limiting form of the prior is obtained as $\kappa, \lambda \to 0$. And we assign a spike-and-slab prior for the effect size β_j:

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\delta_0 + \gamma_j N\left(0, \phi^2/\tau\right),$$

where δ_0 refers to Dirac's delta function, and ϕ^2 is the prior variance of the effect size of the causal variant. In the computation of the Bayes factor, we use $\phi^2 = 0.6^2$. For a specific value of ϕ^2 and a model configuration $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)$, the limiting of the Bayes factor is calculated by:

$$\lim_{\substack{\kappa \to 0 \\ \lambda \to 0}} \mathrm{BF}(\boldsymbol{\gamma}) = \frac{1}{|\phi^{-2} I + \mathbf{X}^T \mathbf{X}|^{1/2}} \frac{1}{\phi} \left[\frac{\mathbf{y}^T \left(I - \mathbf{X} \left(\phi^{-2} I + \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \right) \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \right]^{-n/2}, \quad (4)$$

where $\mathbf{X}$ represents genotypes indicated by non-zero entry in $\boldsymbol{\gamma}$.

4.2 Simulations

4.2.1 Simulation with normal distribution

To evaluate the performance of DAP-PIR, we conduct a simulation study with a normal distribution for the genotype data. Specifically, we consider a scenario with $n = 500$ individuals and $p = 10$ SNPs to relive the computational burden of exact computation of the posterior. Each SNP g_{ij} is generated independently from a standard normal distribution. We randomly select one to five variants and specify them as causal variants with the setting:

$$g_{ij} \sim \mathrm{N}(0, 1), \quad (5)$$

$$y_i = \sum_{j=1}^{10} \beta_j g_{ij} + \epsilon_i, \quad (6)$$

$$\epsilon_i \sim \mathrm{N}(0, \tau^{-1}), \quad (7)$$

$$\beta_j \sim (1 - \gamma_j)\delta_0 + \gamma_j \, \mathrm{N}\left(0, \sigma_0^2/\tau\right), \quad (8)$$

where β_j represents the effect size of the j-th SNP, ϵ_i is the residual term with varaince τ^{-1}, and σ_0^2/τ is the prior effect size variance. The effect sizes β_j are determined by a binary indicator γ_j: if $\gamma_j = 0$, β_j is set to zero and indicates a non-causal SNP; if $\gamma_j = 1$, β_j is drawn from a normal distribution with mean zero and variance σ_0^2/τ, representing a causal SNP. The prior probability of each SNP being causal is set to 1/10 by default, and we randomly assign S causal variants. The indicator vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{10})$ is a 10-vector of binary variables such that only S entries are 1 and the other entries are 0. In the simulation, we set $\tau = 1$, $\sigma_0^2 = 0.6^2$, and choose $S \in \{1, 2, 3, 4, 5\}$. For each setting, we simulate 1000 times. Then we have simulated $1,000 \times 5 = 5,000$ datasets for further investigation.

4.2.2 Simulation on GTEx V8 data

Using the GTEx V8 data [11] with $n = 670$ individual, we select 1,000 genes and $p = 1,000$ SNPs near the region of each selected gene. We simulate the phenotypes with various combinations of the number of effects, S, and proportion of variance explained (PVE) by genotypes, ϕ. The simulation settings are as follows:

$$\begin{aligned}
y_i &= \sum_{j=1}^{1000} \beta_j g_{ij} + \epsilon_i, && (9) \\
\epsilon_i &\sim \mathrm{N}(0, \sigma^2), && (10) \\
\sigma^2 &= \frac{1-\phi}{\phi} \mathrm{Var}(\mathbf{G}\boldsymbol{\beta}), && (11) \\
\beta_j &\sim (1-\gamma_j)\delta_0 + \gamma_j \, \mathrm{N}\left(0, 0.6^2\right). && (12)
\end{aligned}$$

Specifically, for each gene, we randomly assign S variants to be causal, while their effect sizes are independently drawn from $\mathrm{N}(0, 0.6^2)$ and set the other effect sizes to zero. The variance of the error term σ^2 is given by $\frac{1-\phi}{\phi}\mathrm{Var}(\mathbf{G}\boldsymbol{\beta})$ and the simulated phenotypes follow $\mathrm{N}(\mathbf{G}\boldsymbol{\beta}, \sigma^2)$. We follow the SuSiE setting and generate data with pairwise combinations of $S \in \{1, 2, 3, 4, 5\}$ and $\phi \in \{0.05, 0.1, 0.2, 0.4\}$. Then we have simulated $1,000 \times 5 \times 4 = 20,000$ datasets for further investigation.

4.3 Real data application

5 Figures and tables

Algorithm 4 Calculate $y = x^n$

```
Require: $n \geq 0 \vee x \neq 0$
Ensure: $y = x^n$
$y \Leftarrow 1$
if $n < 0$ then
    $X \Leftarrow 1/x$
    $N \Leftarrow -n$
else
    $X \Leftarrow x$
    $N \Leftarrow n$
end if
while $N \neq 0$ do
    if $N$ is even then
        $X \Leftarrow X \times X$
        $N \Leftarrow N/2$
    else[$N$ is odd]
        $y \Leftarrow y \times X$
        $N \Leftarrow N - 1$
    end if
end while
```

References

[1] Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J.: 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics **101**(1), 5–22 (2017)

[2] Ardlie, K.G., Kruglyak, L., Seielstad, M.: Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics **3**(4), 299–309 (2002)

[3] Spain, S.L., Barrett, J.C.: Strategies for fine-mapping complex traits. Human molecular genetics **24**(R1), 111–119 (2015)

[4] Schaid, D.J., Chen, W., Larson, N.B.: From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**(8), 491–504 (2018)

[5] Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E.: Identifying causal variants at loci with multiple signals of association. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 610–611 (2014)

[6] Wen, X., Lee, Y., Luca, F., Pique-Regi, R.: Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. The American Journal of Human Genetics **98**(6), 1114–1129 (2016)

[7] Lee, Y., Luca, F., Pique-Regi, R., Wen, X.: Bayesian multi-snp genetic association analysis: control of fdr and use of summary statistics. BioRxiv, 316471 (2018)

[8] Wang, G., Sarkar, A., Carbonetto, P., Stephens, M.: A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society Series B: Statistical Methodology **82**(5), 1273–1300 (2020)

[9] Rubin, D.B.: The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. Journal of the American Statistical Association **82**(398), 543–546 (1987)

[10] Rubin, D.B.: Using the sir algorithm to simulate posterior distributions. In: Bayesian Statistics 3. Proceedings of the Third Valencia International Meeting, 1-5 June 1987, pp. 395–402 (1988). Clarendon Press

[11] Consortium, G., Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., *et al.*: The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science **348**(6235), 648–660 (2015)