Supplemental data

Table of Content

Supplementary note

A simple example of bias due to covariate adjustment in genetic association study

Consider four variables, Y a given outcome, C a potential covariate, U an unmeasured variable, and G a genetic variant. For simplicity all four variables are assumed to be normally distributed with mean 0 and variance 1. Consider a model where the Y and C are correlated because they are both influenced by U, and G is associated with C only with effect β_C (as illustrated in **Figure 1d**). Under such model Y and C can be written as:

$$Y = \gamma_1 \times U + \sqrt{1-\gamma_1^2} \times \varepsilon_Y$$

$$C = \gamma_2 \times U + \beta_C \times G + \sqrt{1-\gamma_2^2-\beta_C^2} \times \varepsilon_C$$

Consider the case where γ_1 and γ_2 are positives for simplicity, the correlation between Y and C equals $\gamma_1\gamma_2$. It follows that Y adjusted for C equals:

$$Y_{adjusted.C} = Y - \gamma_1\gamma_2 \times C$$

$$= \gamma_1 \times U + \sqrt{1-\gamma_1^2} \times \varepsilon_Y - \gamma_1\gamma_2\left(\gamma_2 \times U + \beta_C \times G + \sqrt{1-\gamma_2^2-\beta_C^2} \times \varepsilon_C\right)$$

$$= \gamma_1\gamma_2 \times (\gamma_1 - \gamma_2) \times U - \gamma_1\gamma_2 \times \beta_C \times G - \gamma_1\gamma_2 \times \sqrt{1-\gamma_2^2-\beta_C^2} \times \varepsilon_C + \sqrt{1-\gamma_1^2} \times \varepsilon_Y$$

Thus, after the adjustment, the outcome $Y_{adjusted.C}$ depends on G with an effect equal to $\gamma_1\gamma_2 \times \beta_C$.

Consider now a slightly different model, where the SNP is associated with both Y and C with effect β_Y and β_C, respectively (as illustrated in **Figure 1c**), so that:

$$Y = \gamma_1 \times U + \beta_Y \times G + \sqrt{1-\gamma_2^2-\beta_Y^2} \times \varepsilon_Y$$

$$C = \gamma_2 \times U + \beta_C \times G + \sqrt{1-\gamma_2^2-\beta_C^2} \times \varepsilon_C$$

One can similarly show that Y adjusted for C equals:

$$Y_{adjusted.C} = (\beta_Y - \gamma_1\gamma_2 \times \beta_C) \times G + \varepsilon_{Y_{adj}}$$

where $\varepsilon_{Y_{adj}}$is a random variable not associated with G. In such case the effect of G on Y would be either overestimated if β_C and β_Y have opposite direction, and underestimated if β_C and β_Y have effect in the same direction.

The above results illustrated the bias induced in a two-step adjustment strategy (i.e. adjusting the outcome for covariates in a first step, and then testing for association between the residual of the outcome and the predictor). However a similar bias exists if adjustment and test for association with the predictor are performed in a single framework. Indeed, we showed in a recent study[1] that the expected value of the joint least square estimates of $\hat{\gamma}$ and $\hat{\beta}$, the effect of C and G on Y, respectively equal:

$$\mathbb{E}[\hat{\gamma}] = \frac{1}{1-(\beta_C)^2}(\gamma - \beta_Y\beta_C)$$

$$\mathbb{E}\left[\hat{\beta}_{adj}\right] = \frac{1}{1-(\beta_C)^2}(\beta_Y - \beta_C\gamma)$$

<u>Conditional effect estimate of the covariates</u>

For simplicity, we consider all variables in question are standardized to have mean 0 and variance 1. In standard ordinary least squared, $\hat{\delta}$, the estimated effect of X on C from the linear model $C \sim \delta X + \varepsilon$ has an expected value of $\mathbb{E}\left[\hat{\delta}\right] = \delta$ and a variance $\sigma^2_{\hat{\delta}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, were σ^2 is the variance of ε. When the sample size N is large and δ is small, $\sigma^2_{\hat{\delta}} \approx 1/N$.

Now consider the more general case assuming C, Y, and X follow a multivariate normal distribution, i.e.

$$\begin{bmatrix} C \\ Y \\ X \end{bmatrix} \sim MVN\left[\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \gamma & \delta \\ \gamma & 1 & \beta \\ \delta & \beta & 1 \end{bmatrix}\right] = MVN\left[\vec{0}, \Sigma\right]$$

The maximum likelihood for Σ is asymptotically similar to the method-of-moments estimator, so that for a sample size n, we have the following approximation:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{N}\left(\begin{bmatrix} c_i - \bar{c} \\ y_i - \bar{y} \\ x_i - \bar{x} \end{bmatrix}[c_i - \bar{c} \quad y_i - \bar{y} \quad x_i - \bar{x}]\right) = \begin{bmatrix} \hat{\sigma}^2_C & \hat{\gamma} & \hat{\delta} \\ \hat{\gamma} & \hat{\sigma}^2_Y & \hat{\beta} \\ \hat{\delta} & \hat{\beta} & \hat{\sigma}^2_X \end{bmatrix} = \frac{1}{n}S$$

Where S follows a 3 degrees of freedom Wishart distribution, $S \sim W_3(\Sigma, n)$, which has an expected value $\mathbb{E}[S] = n\Sigma$.

The matrix S can be partition as:

$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$ where $S_{11} = n\hat{\sigma}^2_C$; $S_{22} = n\begin{pmatrix} \hat{\sigma}^2_Y & \hat{\beta} \\ \hat{\beta} & \hat{\sigma}^2_X \end{pmatrix}$; and $S_{21} = n\begin{pmatrix} \hat{\gamma} \\ \hat{\delta} \end{pmatrix}$

It follows that the distribution of S_{21} conditional on S_{22} equals $S_{21} \mid S_{22} \sim N(\mu, \Lambda)$ with $\mu = S_{22}\Sigma_{22}^{-1}\Sigma_{21}$ and $\Lambda = S_{22} \otimes \Sigma_{11.2}$, and:

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - [\gamma \quad \delta]\frac{1}{1-\beta^2}\begin{bmatrix} 1 & -\beta \\ -\beta & 1 \end{bmatrix}\begin{bmatrix} \gamma \\ \delta \end{bmatrix}$$

$$= 1 - \frac{1}{1-\beta^2}[\gamma - \beta\delta \quad -\beta\gamma + \delta]\begin{bmatrix} \gamma \\ \delta \end{bmatrix}$$

$$= 1 - \frac{\gamma(\gamma - \beta\delta) + \delta(-\beta\gamma + \delta)}{1-\beta^2}$$

$$= 1 - \frac{\gamma^2 + \delta^2 - 2\delta\gamma\beta}{1-\beta^2}$$

so that:

$$\Lambda = S_{22} \bigotimes \Sigma_{11.2} = n\Sigma_{11.2}\begin{pmatrix}\hat{\sigma}_Y^2 & \hat{\beta} \\ \hat{\beta} & \hat{\sigma}_X^2\end{pmatrix} = \left(1 - \frac{\gamma^2 + \delta^2 - 2\delta\gamma\beta}{1-\beta^2}\right) n\begin{pmatrix}\hat{\sigma}_Y^2 & \hat{\beta} \\ \hat{\beta} & \hat{\sigma}_X^2\end{pmatrix}$$

and,

$$\begin{aligned}\mu &= S_{22}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \left(n\begin{bmatrix}\hat{\sigma}_Y^2 & \hat{\beta} \\ \hat{\beta} & \hat{\sigma}_X^2\end{bmatrix}\right)\left(\frac{1}{1-\beta^2}\begin{bmatrix}1 & -\beta \\ -\beta & 1\end{bmatrix}\right)\begin{bmatrix}\gamma \\ \delta\end{bmatrix} \\ &= \frac{n}{1-\beta^2}\begin{bmatrix}\hat{\sigma}_Y^2 - \beta\hat{\beta} & \hat{\beta} - \beta\sigma_Y^2 \\ \hat{\beta} - \beta\hat{\sigma}_X^2 & \hat{\sigma}_X^2 - \beta\hat{\beta}\end{bmatrix}\begin{bmatrix}\gamma \\ \delta\end{bmatrix} \\ &= \frac{n}{1-\beta^2}\begin{bmatrix}\gamma\left(\hat{\sigma}_Y^2 - \beta\hat{\beta}\right) + \delta\left(\hat{\beta} - \beta\hat{\sigma}_Y^2\right) \\ \gamma\left(\hat{\beta} - \beta\hat{\sigma}_X^2\right) + \delta\left(\hat{\sigma}_X^2 - \beta\hat{\beta}\right)\end{bmatrix}\end{aligned}$$

Hence, in the special case where $\delta = \beta = 0$, we have :

$$\mu = n\begin{bmatrix}\gamma\hat{\sigma}_Y^2 \\ \gamma\hat{\beta}\end{bmatrix}$$

When sample size is large enough and $|\gamma| \gg 0$, we assume that $\hat{\sigma}_Y^2 \approx \hat{\sigma}_X^2 \approx 1$ and $\gamma \approx \hat{\gamma}$. It follows that the mean can be approximated by:

$$\mu \approx \begin{bmatrix}n\hat{\gamma} \\ n\hat{\gamma}\hat{\beta}\end{bmatrix}$$

From above, we have $\mathbb{E}\left[\hat{\Sigma}_{21} \mid S_{22}\right] = \frac{1}{n}\mathbb{E}[S_{21} \mid S_{22}]$, so that for the same special case, we have:

$$\begin{aligned}\mathbb{E}\begin{bmatrix}\hat{\gamma} \\ \hat{\delta}\end{bmatrix} \mid S_{22}\Big]_{\delta=\beta=0} &= \frac{1}{n}\mathbb{E}\begin{bmatrix}n\hat{\gamma} \\ n\hat{\delta}\end{bmatrix} \mid \hat{\sigma}_Y^2, \hat{\sigma}_X^2, \hat{\beta}\Big]_{\delta=\beta=0} \\ &\approx \frac{1}{n}\mathbb{E}\begin{bmatrix}n\hat{\gamma} \\ n\hat{\delta}\end{bmatrix} \mid \hat{\beta}\Big]_{\delta=\beta=0} \\ &\approx \begin{bmatrix}\hat{\gamma} \\ \hat{\gamma}\hat{\beta}\end{bmatrix}\end{aligned}$$

In other words:

$$\mathbb{E}\left(\hat{\delta}|\hat{\beta}\right)_{\delta=\beta=0} \approx \hat{\gamma}\hat{\beta}$$

And the variance of $n\hat{\delta}$ conditional on S_{22} under the same assumption ($\hat{\sigma}_Y^2 \approx 1$, $\hat{\sigma}_X^2 \approx 1$ and $\gamma \approx \hat{\gamma}$), and for the special case where $\delta = \beta = 0$, it follows that:

$$var\left(n\hat{\delta}|S_{22}\right)_{\delta=\beta=0} = var\left(n\hat{\delta}|\hat{\sigma}_Y^2, \hat{\sigma}_X^2, \hat{\beta}\right)_{\delta=\beta=0}$$

$$= \left[\left(1 - \frac{\gamma^2 + \delta^2 - 2\delta\gamma\beta}{1-\beta^2}\right) n\hat{\sigma}_X^2\right]_{\delta=\beta=0}$$
$$\approx (1-\hat{\gamma}^2)n$$

so that:

$$var(\hat{\delta}|\hat{\beta})_{\delta=\beta=0} \approx= \frac{1}{n^2}(1-\hat{\gamma}^2)n \approx \frac{(1-\hat{\gamma}^2)}{n}$$

<u>Type I error inflation when filtering covariates based on their p-values for association with the predictor</u>

Consider the three standardized variables: an outcome Y, a predictor X and a covariate C. The p-value based filtering consists in including C as a covariate in the model $Y \sim \beta X + \varepsilon_Y$ only if its p-value for association with X from the marginal linear model $C \sim \delta X + \varepsilon_C$ is greater than a threshold α. Filtering out a covariate based on its *p*-value is equivalent to filtering based on $\chi^2_{\hat{\delta}}$, the 1 degree of freedom chi-squared statistic for the test of $\hat{\delta}$, i.e. $pvalue \leq \alpha \Leftrightarrow \chi^2_{\hat{\delta}} \geq t$ and conversely $p > \alpha \Leftrightarrow \chi^2_{\hat{\delta}} < t$. Where $t = \phi^{-1}(\alpha)$ and ϕ^{-1} is the inverse of a 1 degree of freedom chi-squared cumulative distribution function.

In parallel, and using the same derivation as in the previous section and assuming $\hat{\sigma}_Y^2 \approx 1$ and $\gamma \approx \hat{\gamma}$, one can show that $\mathbb{E}(\hat{\beta}|\hat{\delta})_{\delta=\beta=0} \approx \hat{\gamma}\hat{\delta}$ and $var(\hat{\beta}|\hat{\delta})_{\delta=\beta=0} \approx (1-\hat{\gamma}^2)/n$, where n is the sample size.

When sample size is large so that $var(\hat{\delta}) \approx var(\hat{\beta}) \approx 1/n$, the expected value of $\chi^2_{\hat{\beta}}$, the 1 degree of freedom chi-squared statistics for the test of $\hat{\beta}$, can be expressed as a function of $\chi^2_{\hat{\delta}}$:

$$\mathbb{E}\left(\chi^2_{\hat{\beta}}|\hat{\delta}\right)_{\delta=\beta=0} \approx n\mathbb{E}\left(\frac{\hat{\beta}^2}{\hat{\sigma}_Y^2}|\hat{\delta}\right)_{\delta=\beta=0}$$
$$\approx n\left[\mathbb{E}(\hat{\beta}|\hat{\delta})^2_{\delta=\beta=0} + var(\hat{\beta}|\hat{\delta})_{\delta=\beta=0}\right]$$
$$\approx n\left[(\hat{\gamma}\hat{\delta})^2 + \frac{(1-\hat{\gamma}^2)}{n}\right]$$
$$\approx 1 - \hat{\gamma}^2 + \hat{\gamma}^2\chi^2_{\hat{\delta}}$$

Adding boundaries on $\hat{\delta}$, the expected value of $\chi^2_{\hat{\beta}}$ has the following form, which depends on the first moment of a truncated chi-squared distribution:

$$\mathbb{E}\left(\chi^2_{\hat{\beta}}|\chi^2_{\hat{\delta}} > t\right)_{\delta=\beta=0} = 1 - \hat{\gamma}^2 + \hat{\gamma}^2 \int_t^{+\infty} \chi^2_{\hat{\delta}}\, f(\chi^2_{\hat{\delta}}) d\chi^2_{\hat{\delta}} = 1 - \hat{\gamma}^2 + \hat{\gamma}^2 \mathbb{E}(\chi^2_{\hat{\delta}})_{\chi^2_{\hat{\delta}} \in [t,+\infty]}$$

where f is the probability distribution function of a 1 degree of freedom central chi-squared distribution. Although solutions might exists, solving this integral is non-trivial[2] and of limited interest for the purpose of our study. However, the monotonicity of f, and the positivity of f and $\chi^2_{\hat{\delta}}$ implies when $t \in (0, +\infty)$:

$$\int_{t}^{+\infty} \chi^2_{\hat{\delta}} f(\chi^2_{\hat{\delta}}) d\chi^2_{\hat{\delta}} > \int_{0}^{+\infty} \chi^2_{\hat{\delta}} f(\chi^2_{\hat{\delta}}) d\chi^2_{\hat{\delta}} \Leftrightarrow \mathbb{E}(\chi^2_{\hat{\delta}})_{\chi^2_{\hat{\delta}} \in [t,+\infty]} > \mathbb{E}(\chi^2_{\hat{\delta}})$$

$$\Leftrightarrow \mathbb{E}(\chi^2_{\hat{\delta}})_{\chi^2_{\hat{\delta}} \in [t,+\infty]} > 1$$

$$\Leftrightarrow 1 - \hat{\gamma}^2 + \hat{\gamma}^2 \mathbb{E}(\chi^2_{\hat{\delta}})_{\chi^2_{\hat{\delta}} \in [t,+\infty]} > 1 - \hat{\gamma}^2 + \hat{\gamma}^2$$

$$\Leftrightarrow \mathbb{E}\left(\chi^2_{\hat{\beta}} | \chi^2_{\hat{\delta}} > t\right)_{\delta=\beta=0} > 1$$

$$\Leftrightarrow \mathbb{E}\left(\chi^2_{\hat{\beta}} | pvalue_{\hat{\delta}} < \alpha\right)_{\delta=\beta=0} > 1$$

This shows that subset of $\chi^2_{\hat{\beta}}$ selected based on $pvalue_{\hat{\delta}} < \alpha$ is inflated.

In parallel, we previously demonstrated in Aschard et al[1] that the estimated effect of X on Y from the adjusted model, $Y \sim \beta_{adj} X + \gamma C + \varepsilon_Y$, can be approximated by $\hat{\beta}_{adj} \approx \hat{\beta} - \hat{\gamma}\hat{\delta}$. It follows that under the same assumption described above ($|\gamma| \gg 0$ so that $\gamma = \hat{\gamma}$), we have:

$$\mathbb{E}(\hat{\beta}_{adj} | \hat{\delta})_{\delta=\beta=0} \approx \mathbb{E}(\hat{\beta} - \hat{\gamma}\hat{\delta} | \hat{\delta})_{\delta=\beta=0}$$

$$\approx \mathbb{E}(\hat{\beta} | \hat{\delta})_{\delta=\beta=0} - \hat{\gamma}\hat{\delta}$$

$$\approx 0$$

And the variance of $\hat{\beta}_{adj}$ conditional on $\hat{\delta}$ equals:

$$var(\hat{\beta}_{adj} | \hat{\delta})_{\delta=\beta=0} \approx var(\hat{\beta} - \hat{\gamma}\hat{\delta} | \hat{\delta})_{\delta=\beta=0}$$

$$\approx var(\hat{\beta} | \hat{\delta})_{\delta=\beta=0} + \hat{\gamma}^2 var(\hat{\delta} | \hat{\delta})$$

$$\approx \frac{(1 - \hat{\gamma}^2)}{n}$$

The adjusted model, $\chi^2_{\hat{\beta}_{adj}}$, the 1 degree of freedom chi-squared statistics for the test of $\hat{\beta}_{adj}$, has expected value:

$$\mathbb{E}\left(\chi^2_{\hat{\beta}_{adj}} | \hat{\delta}\right)_{\delta=\beta=0} \approx \mathbb{E}\left(\frac{n\hat{\beta}^2_{adj}}{\hat{\sigma}^2_Y - \hat{\gamma}^2} | \hat{\delta}\right)_{\delta=\beta=0}$$

$$\approx \frac{n}{1 - \hat{\gamma}^2}\left[\mathbb{E}(\hat{\beta}_{adj} | \hat{\delta})^2_{\delta=\beta=0} + var(\hat{\beta}_{adj} | \hat{\delta})_{\delta=\beta=0}\right]$$

$$\approx \frac{n}{1 - \hat{\gamma}^2}\left[\frac{1 - \hat{\gamma}^2}{n}\right]$$

$$\approx 1$$

So that $\mathbb{E}\left(\chi^2_{\hat{\beta}_{adj}}|\hat{\delta}\right)_{\delta=\beta=0}$ is independent of $\hat{\delta}$:

Hence, the *p*-value based filtering in a complete null model ($\delta = \beta = 0$) consists in merging chi-squared statistics from the adjusted test ($\chi^2_{\hat{\beta}_{adj}}$) which are distributed under the null independently of $\hat{\delta}$ with a mean of 1, and chi-squared statistics from the unadjusted test ($\chi^2_{\hat{\beta}}$) which have an expected value that increases with decreasing α threshold when $|\gamma| \neq 0$, being systematically larger than 1 when $\alpha > 0$.

We verified this result empirically in **Supplementary Fig. 2**.

The CMS algorithm

We develop an algorithm to select relevant covariates when testing for association between a predictor X and an outcome Y. For a set of candidate covariates $\boldsymbol{C} = (C_1, C_2, \dots C_m)$, the filtering is applied on $\hat{\delta}_l$ and p_l, the estimated marginal effect of the predictor X on C_l and its associated *p*-value, respectively. It uses four major features: i) r^2_C the total amount of variance of Y explained by the $\boldsymbol{C}$; ii) $(\widehat{\gamma}^2_{lu}, \hat{\gamma}^2_{lm})$the estimated effect of each $C_{l\in 1\dots m}$ on Y and their joint effect respectively; iii) $\hat{\beta}$, the estimated effect of X on Y from the marginal model $Y \sim \alpha + \beta X$; and iv) p_{MUL}, the *p*-value for the multivariate test of all $C_{l=1\dots m}$ and X, which is estimated using a standard multivariate approach (i.e. MANOVA).

Filtering is applied in two steps using the aforementioned features and additional parameters described thereafter. Step 1 is an iterative procedure that filters out candidate covariate until p_{MUL} reach a given threshold. Step 2 is also iterative and uses covariates pre-selected at step 1. It consists in deriving two confidence intervals $\Delta_{\mathrm{l.cond}}$ and $\Delta_{\mathrm{l.un}}$, for the expected distribution of $\hat{\delta}_l$ conditional on $\hat{\beta}$ under a complete null model ($\delta_l = 0$ and $\beta = 0$), and the unconditional distribution of $\hat{\delta}_l$, respectively. The unconditional distribution of $\hat{\delta}_l$ can be approximated as $\mathcal{N}(0, \sqrt{1/n})$, while the conditional distribution equals $\mathcal{N}(\hat{\gamma}\hat{\beta}, \sqrt{(1-\hat{\gamma}^2)/n})$, where $\hat{\gamma}$ is the estimated correlation between Y and C. The final inclusion area for each $\hat{\delta}_l$ is defined as the union of $\Delta_{l.cond}$ and $\Delta_{l.un}$, after applying stringency weights w_u and w_c,respecctively.

The proposed multi-step algorithm is defined as follows:

For each predictor X and Y

1. Univariate association
    - 1.1. Standardized all variables (Y, X, $\boldsymbol{C}$) to have mean 0 and variance 1
    - 1.2. Initialize $L = 1 \dots m$, the list of selected covariates, with all available covariates
    - 1.3. Derive for each $l \in L$, $\hat{\gamma}_{lu}$and $\hat{\gamma}_{lm}$ the marginal effect estimates from the univariate regression $Y \sim C_{l=1\dots m}$, and multivariate model $Y \sim \boldsymbol{C}$, respectively.
2. Filter 1: multivariate
    - 2.1. Perform a marginal association test between X and each $C_{l=1\dots m}$
        - 2.1.1. Derive all $\hat{\delta}_l$ and p_l from $C \sim \delta_0 + \delta_l \times X$
    - 2.2. Set p_{MUL}=0

2.3. While $p_{MUL} < t_{MUL}$

2.3.1. Derive p_{MUL} from $\boldsymbol{C}_L \sim X$ using a multivariate test, where $\boldsymbol{C}_L$ is the data matrix $\boldsymbol{C}$ including only $l \in L$ covariates.

2.3.2. Update L by removing the C_l that match $p_{l \in L} = min(p_{l \in L})$ from the set of candidate covariates

3. if $\mathrm{L} \neq 0$, filter2: univariate

3.1. while $\mathrm{L} \neq 0$ and $\mathrm{L}_{t+1} \neq \mathrm{L}_t$

3.1.1. Update for each $l \in L$ $\hat{\gamma}_{lm}$ the effect estimates from the multivariate model $Y \sim \boldsymbol{C}_L$.

3.1.2. Derive r_C^2 the variance of Y explained by $\boldsymbol{C}_L$ from the model in 3.1.1

3.1.3. Derive for each $l \in L$ the stringency of the inclusion area $w_{ST} = 0.1 \times p_{MUL} \times (1 - r_C^2) \times (1 - \hat{\gamma}_{lu}^2)/\hat{\gamma}_{lm}^2$

3.1.4. Derive the specific conditional and unconditional weights using the threshold functions $w_c = \min(w_{ST}, f_c(\chi_\beta^2))$ and $w_u = \min(w_{ST}, f_u(\chi_\beta^2))$, where $\chi_\beta^2 = N \times \hat{\beta}^2/\sigma_X^2$:

a. $$f_c = \begin{cases} \chi_\beta^2/8 & if \chi_\beta^2 < 16 \\ 2 - \chi_\beta^2/8 & if \chi_\beta^2 > 16 \text{ and } \chi_\beta^2 < 32 \\ 0 & \text{Otherwise} \end{cases}$$

b. $$f_u = \begin{cases} \chi_\beta^2/8 & if \chi_\beta^2 < 16 \\ 2 & \text{Otherwise} \end{cases}$$

3.1.5. Derive the mean $\mu_{l.un} = 0$ and standard deviation $\sigma_{l.un} = \sqrt{\frac{1}{N}}$ of the unconditional distribution of $\hat{\delta}_l$ and the associated inclusion area: $\Delta_{l.un} = [\mu_{l.un} - \sigma_{l.un} \times w_u,\ \mu_{l.un} + \sigma_{l.un} \times w_u]$.

3.1.6. Derive the mean $\mu_{l.cond} = \hat{\gamma}_l \times \hat{\beta}$ and standard deviation $\sigma_{l.cond} = \sqrt{\frac{(1-\hat{\gamma}_l^2)}{N}}$ of the conditional null distribution of $\hat{\delta}_l$, and the associated inclusion area: $\Delta_{l.cond} = [\mu_{l.cond} - \sigma_{l.cond} \times w_c,\ \mu_{l.cond} + \sigma_{l.cond} \times w_c]$

3.1.7. Update L by removing all l which $\hat{\delta}_l$ is not included in $\Delta_{l.cond} \cup \Delta_{l.un}$

4. Perform the test of association between X and Y, while adjusting for the selected covariates

4.1. Estimate $\hat{\beta}_{CMS}$ and derive the associated *p*-value from the multivariate model including all $l \in L$ covariate from $Y \sim \beta_0 + \beta_{CMS} \times \mathrm{X} + \boldsymbol{\beta_L} \times \boldsymbol{C_L}$

Extensive simulation models

We simulated series of 10,000 replicates under null models where a predictor of interest –here a single nucleotide polymorphism (SNP)– is not associated with the primary outcome but is associated with a fraction π of the covariates and under the alternative where the predictor is associated with the primary outcome only. For the null model we considered π=[0%, 15%, 35%], while we focused on the case π=[0%] for the alternative. We

generated effects of the predictor on the primary outcome (β) and on the covariate (δ) so that the expected noncentrality chi-square parameter (*ncp*, which equals the median χ^2 or the average χ^2 minus 1) goes from very low to moderately high ($ncp = [3, 5, 10, 20]$). More precisely, for the SNP effect on the outcome we used either of the four ncp so that $= \sqrt{ncp/N}$. For the SNP effect on the covariates we draw uniformly δ from $\left[0, 2 \times \sqrt{ncp/N}\right]$ so that expected effect across replicates equals $\sqrt{ncp/N}$. We assigned negative effects to δ with a probability of 10%. We considered sample sizes N of 300, 2,000 and 6,000, value of r_C^2, the variance of Y explained by $\boldsymbol{C}$, in [25%, 50%, 75%] and total number of phenotypes in [10, 40, 80]. For the null models we derived the genomic inflation factor λ, while for the alternative model we estimated power at an α threshold of $5x10^{-7}$, to correct for 100,000 tests.

Supplementary Figures 19-45 present the QQplots for these 27 scenarios, each scenario including 16 series of 10,000 replicates and four tests: standard marginal univariate test (LR); the optimally adjusted test (OPT) that includes as covariates only the outcomes not associated with the predictor ; the *CMS* approach ; and a univariate test that include as covariate all outcome with a *p*-value for association with the predictor above 0.1 (FT). A more succinct summary including type I error rate is presented in **Supplementary Tables 2-4**

Comparison with alternatives methods

We compared the performances of the *CMS* algorithm against a range of alternatives approaches. We first assessed strategies that consist in capturing shared "hidden factors" and use these factors as covariates. In particular we considered adjustment for principal component and adjustment for PEER factors, both being derived from the entire dataset (i.e. outcome and covariates). We then considered a penalized regression approach as it is commonly used for selection of variables and the assessment of their relative importance. Among many possibilities, we arbitrarily choose to use LASSO[3] in combination with the Exact post-selection inference to compute association *P*-values for each selected predictor.[4] Finally, Bayesian methods have the possibility of assessing the true relation between genotype and phenotype in the presence of correlated covariates. In particular we considered the recent method *mvBIMBAM*[5] which, given a single predictor and multiple outcome, allows to derive posterior probabilities of a genetic variant being either directly, indirectly, or not associated with each outcome.

We first generated series of 500 replicates of 10,000 individuals under a null model in order to compare the robustness of the hidden factor adjustment and the LASSO approach. More specifically, for each replicate, we generated 50 correlated variables from a multivariate normal distribution, with correlation matrix defined so that pairwise correlation varies in [-0.4; 0.4] and a single genotype randomly drawn from a binomial with a coded allele frequency in [0.05; 0.95]. We randomly picked one of the variables as the primary outcome, and considered the remaining as secondary outcomes – and refer further to this variable as covariates. We then added an effect of the genetic variant on a random subset of the covariates, with effect drawn uniformly from [-0.07; 0.07], so that variance of the covariates explained by the genetic variant was always smaller than 0.25%. For PEER and PCs, we tested the association between the genetic variant and the primary outcome while adjusting for an increasing number of either PCs or PEER factors. For LASSO, we provided the model containing the genotype plus all the covariates as independent variables, and let the algorithm estimate the relative importance of each predictor.

First, as shown **Supplementary Figure 7**, the LASSO approach shows severe type I error rate inflation. This is partly expected, as the goal of such methods is to build model with improved prediction accuracy, whether or not the predictors are related or not. Second, concordant with other recent work from our group[6], we observed increasing type I error rates when increasing the number of PCs or PEER factors in the model (**Supplementary Fig. 7a,b,c**). The *CMS* algorithm showed correct type I error rates on the same data. Hence, while these approaches have the main advantage of being computationally efficient (i.e. the hidden factor are

derived once for the complete dataset), we found that they induce false positives. While it is possible to learn factors while adjusting for individual genetic variants[7], these approaches do not scale computationally and still introduced biases in our simulations (**Supplementary Fig. 7c**).

For *mvBIMBAM*, we used the same simulation framework, except that we decreased the total number of variable to 10, as *mvBIMBAM* was intractable for larger number of variables. Also, as *mvBIMBAM* is a Bayesian method, it does not output *P*-value. To allow for fair comparison we had to use a metric accounting for type I error rate relative to power. We choose the ROC curves and AUC and derived sensibility and specificity using p-values for *CMS*, and posterior probability that the phenotype is not affected by the genetic factor for mvBIMBAM. Significance in AUC difference was assessed using the approach proposed by A DeLong et al.[8] The *CMS* approach was more than 100 fold faster than *mvBIMBAM* and the two methods showed similar accuracy when compared using ROC curves (**Supplementary Fig. 6**). Although we noted a slightly better AUC for *mvBIMBAM* (AUC_{bimbam}= 0.834, AUC_{CMS}=0.810, *P*=0.02), the improvement was due to a higher sensitivity at very low specificity (i.e. for a false discovery rate>0.5), which is of limited interest in the context of association studies.

Unsuccessful approaches explored

In the process of defining our algorithm we considered a few alternative strategies to those used in the proposed final version. Below is a brief overview of the major ones we assessed but ended up being unsuccessful:

1. Instead of trying to decipher *type I* and *type II covariates* (i.e. covariate truly associated with the predictor, and those independent of a predictor X, respectively), we considered removing the predictor's effect from the covariate. While several approaches might be considered, we only explored the most naïve one, which consists in using the residual of the covariate after adjusting it for the predictor X. In brief, in step 1) we fit the model $C_l \sim \delta_l X$ and derive $C_l^* = C_l - \hat{\delta}_l X$, then in step 2) we performed the test of $\hat{\beta}$ from the model $Y \sim \beta G + \gamma^* C_l^*$. While the approach is intuitive in theory, in real data, it faces the problem of finite sample sizes. Indeed, $\hat{\delta}_l$ is only an approximation of δ_l, and adjusting Y for C_l^* actually results in introducing a major residual effects of X on Y.

2. The first feature of our algorithm is a multivariate test for association. This test was added to the algorithm to address bias due to linear combination of covariates –i.e. in cases where multiple covariate have small association with the predictor, so that univariate test will have very low power. Before using this approach, we first explored a stepwise selection procedure where 1) all covariates were ranked based on their association the predictor, 2) the covariate were then added to the model $Y \sim \beta X + \gamma_1 C_1 + \gamma_2 C_2 + \cdots$ one by one, and at each loop, we tested the overall association $G \sim \gamma_1 C_1 + \gamma_2 C_2 + \cdots$ The selection process stopped when the later test was significant. As compared to the implemented multivariate approach, this alternative was both time-consuming and inefficient in selecting *type II covariates*.

3. To limit the issue of multicollinearity, we considered first deriving principal components (PCs) of candidate covariates. The resulting PCs being orthogonal, adding or removing a PCs in the model does not affect the estimated effect of other PCs. While appealing this approach showed limited performances. The reason is that PCs are linear combinations of the raw phenotypes, and any genetic effect on a single outcome will tend to be disseminated on all PCs. It follows that most PCs end up being *type I covariates* (i.e. are truly associated with the predictor).

Leveraging the architecture of genomic data

While the proposed *CMS* approach can be applied on any type of data, genomic data have several advantages, one of them being that the underlying structure is relatively well understood[9-12]. Consider the example of **Supplementary Figure 17** where a phenotype *P1* is associated with two genotypes *G1* and *G2*. The association with *G1* and *G2* go through two (almost) independent causal pathway *path1* and *path2* which involve two different sets of RNAs, proteins, metabolites and environmental exposures. Any other phenotypes that depend on *path1* but not *path2* can be used to improve the detection of the association between *P1* and *G2*, and conversely phenotypes depending on *path2* can be leverage to detect *G1* (**Supplementary Figure 17B**). Also, one can see from this figure that variable downstream on the causal pathway cannot be used to identify association between upstream variables, i.e. the "end-of-chain" phenotype *P0* to *P4* cannot a priori be used to improve the detection of the association between the genotype and the metabolites (**Supplementary Figure 17C**). Accounting for this information, we can identify which set of outcomes would be the most relevant for *CMS*. In particular, it allows excluding variables that would be *a priori* a source of bias.

Supplementary Table 1. Type I error of the conditional and unconditional filtering for a single covariate

Threshold	Approach	%sel	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$
T=0.05	UnCond	95.1%	0.0017	0.016	0.068	0.13	0.24	0.53
	Cond	94.5%	0.0004	0.004	0.028	0.07	0.17	0.48
	Mix	98.6%	0.0013	0.011	0.049	0.10	0.21	0.51
T=0.10	UnCond	90.3%	0.0017	0.017	0.075	0.14	0.25	0.55
	Cond	89.3%	0.0003	0.003	0.022	0.06	0.15	0.46
	Mix	96.8%	0.0013	0.010	0.048	0.10	0.21	0.51
T=0.20	UnCond	80.4%	0.0016	0.017	0.079	0.15	0.28	0.57
	Cond	79.2%	0.0003	0.002	0.019	0.05	0.13	0.43
	Mix	92.1%	0.0012	0.010	0.048	0.10	0.21	0.51
T=0.30	UnCond	70.6%	0.0014	0.016	0.078	0.15	0.28	0.59
	Cond	69.2%	0.0003	0.003	0.020	0.05	0.12	0.41
	Mix	86.2%	0.0010	0.010	0.048	0.10	0.21	0.50

Abbreviation: %sel is the percentage of time the covariate is included in the model.

Type I error rate is derived based on a series of 10,000 simulated datasets. For each dataset, a primary outcome and a secondary outcome are generated with a correlation of 0.8, and an independently generated predictor is tested for association with the primary outcome while adding the secondary outcome as a covariate based on its conditional or unconditional distribution.

Supplementary Table 2. Type I error rate for 300 individuals

Nsamp.	Nphe.	r^2	π	α=0.05				α=0.001				α=0.0001			
				LR	OPT	CMS	FT	LR	OPT	CMS	FT	LR	OPT	CMS	FT
300	10	0.25	0.00	0.051	0.051	0.047	0.064	$1.1x10^{-3}$	$1.1x10^{-3}$	$9.2x10^{-4}$	$1.6x10^{-3}$	$7.5x10^{-5}$	$1.5x10^{-4}$	$1.2x10^{-4}$	$2.2x10^{-4}$
300	10	0.25	0.15	0.052	0.051	0.048	0.065	$1.1x10^{-3}$	$1.0x10^{-3}$	$1.2x10^{-3}$	$2.2x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$1.5x10^{-4}$	$2.7x10^{-4}$
300	10	0.25	0.35	0.052	0.051	0.048	0.066	$1.0x10^{-3}$	$1.1x10^{-3}$	$1.1x10^{-3}$	$2.1x10^{-3}$	$1.7x10^{-4}$	$1.5x10^{-4}$	$2.5x10^{-4}$	$4.0x10^{-4}$
300	10	0.50	0.00	0.052	0.050	0.046	0.080	$1.0x10^{-3}$	$1.1x10^{-3}$	$1.1x10^{-3}$	$2.5x10^{-3}$	$1.0x10^{-4}$	$1.0x10^{-4}$	$1.5x10^{-4}$	$3.7x10^{-4}$
300	10	0.50	0.15	0.050	0.052	0.046	0.086	$1.2x10^{-3}$	$1.1x10^{-3}$	$1.8x10^{-3}$	$4.8x10^{-3}$	$1.2x10^{-4}$	$1.5x10^{-4}$	$2.7x10^{-4}$	$1.4x10^{-3}$
300	10	0.50	0.35	0.051	0.052	0.048	0.086	$1.3x10^{-3}$	$1.2x10^{-3}$	$1.9x10^{-3}$	$5.4x10^{-3}$	$1.5x10^{-4}$	$1.0x10^{-4}$	$6.5x10^{-4}$	$1.7x10^{-3}$
300	10	0.75	0.00	0.052	0.052	0.046	0.106	$1.0x10^{-3}$	$9.7x10^{-4}$	$1.4x10^{-3}$	$3.7x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$2.2x10^{-4}$	$4.5x10^{-4}$
300	10	0.75	0.15	0.052	0.051	0.049	0.116	$1.2x10^{-3}$	$1.1x10^{-3}$	$1.8x10^{-3}$	$1.5x10^{-2}$	$5.0x10^{-4}$	$1.7x10^{-4}$	$3.7x10^{-4}$	$7.0x10^{-3}$
300	10	0.75	0.35	0.049	0.050	0.046	0.105	$1.0x10^{-3}$	$9.2x10^{-4}$	$1.6x10^{-3}$	$1.2x10^{-2}$	$5.0x10^{-5}$	$7.5x10^{-5}$	$4.7x10^{-4}$	$5.0x10^{-3}$
300	40	0.25	0.00	0.052	0.051	0.042	0.066	$1.5x10^{-3}$	$1.3x10^{-3}$	$8.0x10^{-4}$	$2.0x10^{-3}$	$2.2x10^{-4}$	$1.5x10^{-4}$	$1.2x10^{-4}$	$4.5x10^{-4}$
300	40	0.25	0.15	0.051	0.050	0.043	0.066	$1.2x10^{-3}$	$1.1x10^{-3}$	$8.0x10^{-4}$	$2.2x10^{-3}$	$1.7x10^{-4}$	$2.0x10^{-4}$	$1.0x10^{-4}$	$4.2x10^{-4}$
300	40	0.25	0.35	0.051	0.050	0.044	0.068	$7.5x10^{-4}$	$1.2x10^{-3}$	$8.7x10^{-4}$	$2.1x10^{-3}$	$2.5x10^{-4}$	$1.2x10^{-4}$	$5.0x10^{-5}$	$1.7x10^{-4}$
300	40	0.50	0.00	0.051	0.050	0.040	0.085	$1.3x10^{-3}$	$9.7x10^{-4}$	$1.0x10^{-4}$	$3.1x10^{-3}$	$1.7x10^{-4}$	$1.2x10^{-4}$	$2.5x10^{-5}$	$5.0x10^{-4}$
300	40	0.50	0.15	0.050	0.052	0.041	0.087	$1.3x10^{-3}$	$1.4x10^{-3}$	$1.0x10^{-3}$	$4.0x10^{-3}$	$2.2x10^{-4}$	$1.0x10^{-4}$	$2.0x10^{-4}$	$8.7x10^{-4}$
300	40	0.50	0.35	0.050	0.052	0.044	0.084	$1.0x10^{-3}$	$1.1x10^{-3}$	$9.7x10^{-4}$	$4.1x10^{-3}$	$7.5x10^{-4}$	$1.2x10^{-4}$	$3.0x10^{-4}$	$8.5x10^{-4}$
300	40	0.75	0.00	0.051	0.052	0.041	0.105	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$4.7x10^{-3}$	$5.0x10^{-4}$	$7.5x10^{-5}$	$1.2x10^{-4}$	$7.2x10^{-4}$
300	40	0.75	0.15	0.051	0.050	0.040	0.100	$1.1x10^{-3}$	$8.5x10^{-4}$	$1.1x10^{-3}$	$6.7x10^{-3}$	$2.7x10^{-4}$	$1.2x10^{-4}$	$1.7x10^{-4}$	$2.1x10^{-3}$
300	40	0.75	0.35	0.051	0.052	0.043	0.091	$9.7x10^{-4}$	$8.5x10^{-4}$	$9.2x10^{-4}$	$5.0x10^{-3}$	$7.5x10^{-5}$	$7.5x10^{-5}$	$1.7x10^{-4}$	$1.0x10^{-3}$
300	80	0.25	0	0.051	0.051	0.036	0.069	$1.2x10^{-3}$	$1.3x10^{-3}$	$6.7x10^{-4}$	$1.8x10^{-3}$	$7.5x10^{-5}$	$1.5x10^{-4}$	$2.5x10^{-5}$	$2.0x10^{-4}$
300	80	0.25	0.15	0.051	0.051	0.037	0.068	$1.3x10^{-3}$	$1.2x10^{-3}$	$6.5x10^{-4}$	$2.3x10^{-3}$	$2.0x10^{-4}$	$1.5x10^{-4}$	$7.5x10^{-5}$	$4.2x10^{-4}$
300	80	0.25	0.35	0.051	0.053	0.039	0.070	$1.0x10^{-3}$	$1.2x10^{-3}$	$6.0x10^{-4}$	$2.3x10^{-3}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$5.0x10^{-5}$	$2.7x10^{-4}$
300	80	0.50	0.00	0.050	0.053	0.033	0.086	$8.2x10^{-4}$	$7.5x10^{-4}$	$4.2x10^{-4}$	$3.0x10^{-3}$	$1.0x10^{-4}$	$1.5x10^{-4}$	$1.0x10^{-4}$	$4.5x10^{-4}$
300	80	0.50	0.15	0.051	0.051	0.036	0.087	$1.2x10^{-3}$	$1.0x10^{-3}$	$8.0x10^{-4}$	$3.7x10^{-3}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$1.0x10^{-4}$	$6.5x10^{-4}$
300	80	0.50	0.35	0.048	0.051	0.037	0.081	$8.5x10^{-4}$	$1.1x10^{-3}$	$5.0x10^{-4}$	$3.0x10^{-3}$	$5.0x10^{-5}$	$1.7x10^{-4}$	$5.0x10^{-5}$	$5.2x10^{-4}$
300	80	0.75	0.00	0.051	0.052	0.035	0.099	$9.0x10^{-4}$	$1.2x10^{-3}$	$8.0x10^{-4}$	$4.8x10^{-3}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$1.0x10^{-4}$	$8.0x10^{-4}$
300	80	0.75	0.15	0.050	0.053	0.036	0.094	$1.4x10^{-3}$	$1.2x10^{-3}$	$7.2x10^{-4}$	$5.6x10^{-3}$	$2.7x10^{-4}$	$1.7x10^{-4}$	$1.2x10^{-4}$	$1.2x10^{-3}$
300	80	0.75	0.35	0.051	0.050	0.039	0.088	$1.1x10^{-3}$	$1.2x10^{-3}$	$8.5x10^{-4}$	$3.9x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$6.7x10^{-4}$

Abbreviations: r^2 is the total outcome variance that can be explained by other variables ; π is the proportion of variables associated with the predictor ; LR= standard (unadjusted) marginal univariate test ; OPT= the optimally adjusted test that includes as covariates only the variables not associated with the predictor ; and FT=a univariate test that include as covariate all variable with a p-value for association with the predictor above 0.1.

Supplementary Table 3. Type I error rate for 2000 individuals

Nsamp.	Nphe.	r^2	π	α=0.05				α=0.001				α=0.0001			
				LR	OPT	CMS	FT	LR	OPT	CMS	FT	LR	OPT	CMS	FT
2000	10	0.25	0.00	0.051	0.051	0.049	0.065	$1.0x10^{-3}$	$9.2x10^{-4}$	$1.1x10^{-3}$	$1.6x10^{-3}$	$5.0x10^{-5}$	$5.0x10^{-5}$	$7.5x10^{-5}$	$1.2x10^{-4}$
2000	10	0.25	0.15	0.052	0.050	0.049	0.067	$1.0x10^{-3}$	$9.5x10^{-4}$	$9.5x10^{-4}$	$1.8x10^{-3}$	$1.0x10^{-4}$	$1.0x10^{-4}$	$1.2x10^{-4}$	$2.2x10^{-4}$
2000	10	0.25	0.35	0.050	0.049	0.049	0.067	$9.0x10^{-4}$	$6.5x10^{-4}$	$1.1x10^{-3}$	$2.0x10^{-3}$	$1.5x10^{-4}$	$7.5x10^{-5}$	$1.5x10^{-4}$	$3.2x10^{-4}$
2000	10	0.50	0.00	0.050	0.053	0.047	0.082	$9.0x10^{-4}$	$1.1x10^{-3}$	$1.2x10^{-3}$	$2.7x10^{-3}$	$1.7x10^{-4}$	$1.2x10^{-4}$	$2.0x10^{-4}$	$3.7x10^{-4}$
2000	10	0.50	0.15	0.051	0.053	0.049	0.088	$1.1x10^{-3}$	$9.7x10^{-4}$	$1.6x10^{-3}$	$4.7x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$3.0x10^{-4}$	$1.0x10^{-4}$
2000	10	0.50	0.35	0.049	0.052	0.048	0.086	$1.3x10^{-3}$	$1.4x10^{-3}$	$1.7x10^{-3}$	$5.1x10^{-3}$	$1.5x10^{-4}$	$1.2x10^{-4}$	$3.7x10^{-4}$	$1.0x10^{-3}$
2000	10	0.75	0.00	0.051	0.052	0.047	0.107	$1.1x10^{-3}$	$1.0x10^{-3}$	$1.3x10^{-3}$	$3.7x10^{-3}$	$1.5x10^{-4}$	$5.0x10^{-5}$	$3.2x10^{-4}$	$5.5x10^{-4}$
2000	10	0.75	0.15	0.049	0.051	0.046	0.112	$1.1x10^{-3}$	$9.7x10^{-4}$	$1.9x10^{-3}$	$1.3x10^{-2}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$5.7x10^{-4}$	$5.9x10^{-3}$
2000	10	0.75	0.35	0.049	0.048	0.047	0.105	$9.0x10^{-4}$	$1.0x10^{-3}$	$1.9x10^{-3}$	$1.2x10^{-2}$	$7.5x10^{-5}$	$1.2x10^{-4}$	$5.7x10^{-4}$	$5.0x10^{-3}$
2000	40	0.25	0.00	0.048	0.050	0.046	0.063	$8.7x10^{-4}$	$1.1x10^{-3}$	$8.7x10^{-4}$	$1.9x10^{-3}$	$7.5x10^{-5}$	$1.5x10^{-4}$	$1.5x10^{-4}$	$2.0x10^{-4}$
2000	40	0.25	0.15	0.049	0.051	0.047	0.066	$1.0x10^{-3}$	$1.0x10^{-3}$	$9.0x10^{-4}$	$2.3x10^{-3}$	$7.5x10^{-5}$	$2.5x10^{-5}$	$2.5x10^{-5}$	$1.2x10^{-4}$
2000	40	0.25	0.35	0.052	0.051	0.050	0.069	$1.0x10^{-3}$	$1.2x10^{-3}$	$1.1x10^{-3}$	$2.1x10^{-3}$	$1.2x10^{-4}$	$1.2x10^{-4}$	$2.0x10^{-4}$	$3.2x10^{-4}$
2000	40	0.50	0.00	0.049	0.049	0.044	0.084	$8.2x10^{-4}$	$1.2x10^{-3}$	$9.5x10^{-4}$	$3.0x10^{-3}$	$5.0x10^{-5}$	$5.0x10^{-5}$	$1.5x10^{-4}$	$5.2x10^{-4}$
2000	40	0.50	0.15	0.049	0.051	0.045	0.084	$9.5x10^{-4}$	$1.3x10^{-3}$	$1.2x10^{-3}$	$3.9x10^{-3}$	$1.7x10^{-4}$	$1.5x10^{-4}$	$3.0x10^{-4}$	$1.0x10^{-3}$
2000	40	0.50	0.35	0.050	0.049	0.046	0.082	$6.0x10^{-4}$	$8.5x10^{-4}$	$8.0x10^{-4}$	$3.1x10^{-3}$	$2.5x10^{-5}$	0	$1.7x10^{-4}$	$5.0x10^{-4}$
2000	40	0.75	0.00	0.049	0.049	0.045	0.105	$9.2x10^{-4}$	$9.0x10^{-4}$	$9.2x10^{-4}$	$4.3x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$1.0x10^{-4}$	$7.7x10^{-4}$
2000	40	0.75	0.15	0.050	0.050	0.046	0.097	$1.4x10^{-3}$	$1.1x10^{-3}$	$1.4x10^{-3}$	$7.1x10^{-3}$	0	$1.0x10^{-4}$	$2.0x10^{-4}$	$1.6x10^{-3}$
2000	40	0.75	0.35	0.051	0.051	0.048	0.090	$1.3x10^{-3}$	$1.1x10^{-3}$	$1.3x10^{-3}$	$5.2x10^{-3}$	$1.0x10^{-4}$	$1.7x10^{-4}$	$3.7x10^{-4}$	$1.1x10^{-3}$
2000	80	0.25	0.00	0.052	0.051	0.048	0.067	$8.0x10^{-4}$	$1.0x10^{-3}$	$7.5x10^{-4}$	$1.5x10^{-3}$	$7.5x10^{-5}$	$5.0x10^{-5}$	$7.5x10^{-5}$	$1.5x10^{-4}$
2000	80	0.25	0.15	0.050	0.049	0.047	0.067	$9.0x10^{-4}$	$1.2x10^{-3}$	$8.7x10^{-4}$	$1.9x10^{-3}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$5.0x10^{-5}$	$2.5x10^{-4}$
2000	80	0.25	0.35	0.050	0.052	0.048	0.068	$8.7x10^{-4}$	$1.1x10^{-3}$	$8.2x10^{-4}$	$1.9x10^{-3}$	$5.0x10^{-5}$	$1.5x10^{-4}$	$1.5x10^{-4}$	$3.0x10^{-4}$
2000	80	0.50	0.00	0.049	0.050	0.043	0.083	$7.5x10^{-4}$	$8.7x10^{-4}$	$8.0x10^{-4}$	$3.0x10^{-3}$	$1.0x10^{-4}$	$2.5x10^{-5}$	$1.7x10^{-4}$	$3.2x10^{-4}$
2000	80	0.50	0.15	0.051	0.050	0.047	0.085	$8.5x10^{-4}$	$8.x10^{-4}$	$7.5x10^{-4}$	$3.5x10^{-3}$	$5.0x10^{-5}$	$1.2x10^{-4}$	$1.7x10^{-4}$	$4.5x10^{-4}$
2000	80	0.50	0.35	0.049	0.050	0.046	0.081	$1.1x10^{-3}$	$1.2x10^{-3}$	$1.4x10^{-3}$	$3.7x10^{-3}$	$7.5x10^{-5}$	$7.5x10^{-5}$	$3.5x10^{-4}$	$7.7x10^{-4}$
2000	80	0.75	0.00	0.048	0.050	0.043	0.099	$1.1x10^{-3}$	$8.7x10^{-3}$	$9.7x10^{-4}$	$4.6x10^{-3}$	$7.5x10^{-5}$	$1.7x10^{-4}$	$1.0x10^{-4}$	$7.7x10^{-4}$
2000	80	0.75	0.15	0.048	0.051	0.044	0.091	$1.1x10^{-3}$	$9.5x10^{-4}$	$1.1x10^{-3}$	$4.7x10^{-3}$	$2.0x10^{-4}$	$2.0x10^{-4}$	$2.0x10^{-4}$	$7.7x10^{-4}$
2000	80	0.75	0.35	0.050	0.051	0.047	0.087	$1.3x10^{-3}$	$1.1x10^{-3}$	$1.5x10^{-3}$	$4.3x10^{-3}$	$1.2x10^{-4}$	$5.0x10^{-5}$	$1.7x10^{-4}$	$7.0x10^{-4}$

Abbreviations: r^2 is the total outcome variance that can be explained by other variables ; π is the proportion of variables associated with the predictor ; LR= standard (unadjusted) marginal univariate test ; OPT= the optimally adjusted test that includes as covariates only the variables not associated with the predictor ; and FT=a univariate test that include as covariate all variable with a p-value for association with the predictor above 0.1.

Supplementary Table 4. Type I error rate for 6000 individuals

Nsamp.	Nphe.	r^2	π	α=0.05				α=0.001				α=0.0001			
				LR	OPT	CMS	FT	LR	OPT	CMS	FT	LR	OPT	CMS	FT
6000	10	0.25	0.00	0.052	0.051	0.049	0.065	$9.2x10^{-4}$	$9.5x10^{-4}$	$9.0x10^{-4}$	$1.6x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$1.x10^{-4}$	$2.2x10^{-4}$
6000	10	0.25	0.15	0.051	0.048	0.049	0.064	$7.7x10^{-4}$	$9.7x10^{-4}$	$7.7x10^{-4}$	$1.9x10^{-3}$	$2.5x10^{-5}$	$1.2x10^{-4}$	$1.5x10^{-4}$	$2.2x10^{-4}$
6000	10	0.25	0.35	0.050	0.050	0.049	0.067	$9.2x10^{-4}$	$1.0x10^{-3}$	$9.2x10^{-4}$	$1.9x10^{-3}$	$7.5x10^{-5}$	$7.5x10^{-5}$	$2.0x10^{-4}$	$3.0x10^{-4}$
6000	10	0.50	0.00	0.050	0.052	0.047	0.082	$8.0x10^{-4}$	$9.7x10^{-4}$	$1.3x10^{-3}$	$2.8x10^{-3}$	$1.0x10^{-4}$	$1.2x10^{-4}$	$2.0x10^{-4}$	$4.5x10^{-4}$
6000	10	0.50	0.15	0.052	0.051	0.049	0.085	$1.2x10^{-3}$	$1.3x10^{-3}$	$1.9x10^{-3}$	$5.3x10^{-3}$	$1.5x10^{-4}$	$1.2x10^{-4}$	$5.7x10^{-4}$	$1.3x10^{-3}$
6000	10	0.50	0.35	0.050	0.049	0.048	0.084	$9.7x10^{-4}$	$9.5x10^{-4}$	$1.7x10^{-3}$	$4.8x10^{-3}$	$7.5x10^{-5}$	$1.0x10^{-4}$	$3.0x10^{-4}$	$9.7x10^{-4}$
6000	10	0.75	0.00	0.050	0.049	0.048	0.103	$1.0x10^{-3}$	$8.7x10^{-4}$	$1.5x10^{-3}$	$3.5x10^{-3}$	$1.2x10^{-4}$	$1.0x10^{-4}$	$2.7x10^{-4}$	$6.5x10^{-4}$
6000	10	0.75	0.15	0.050	0.051	0.047	0.112	$9.5x10^{-4}$	$1.1x10^{-3}$	$1.7x10^{-3}$	$1.4x10^{-2}$	$2.5x10^{-5}$	$1.5x10^{-4}$	$4.5x10^{-4}$	$6.0x10^{-3}$
6000	10	0.75	0.35	0.048	0.050	0.048	0.105	$9.7x10^{-4}$	$8.0x10^{-4}$	$1.9x10^{-3}$	$1.2x10^{-2}$	$1.0x10^{-4}$	$2.5x10^{-5}$	$5.2x10^{-4}$	$5.2x10^{-3}$
6000	40	0.25	0.00	0.051	0.051	0.050	0.066	$1.0x10^{-3}$	$1.1x10^{-3}$	$7.2x10^{-4}$	$1.8x10^{-3}$	$5.0x10^{-5}$	$1.0x10^{-4}$	$1.0x10^{-4}$	$1.2x10^{-4}$
6000	40	0.25	0.15	0.051	0.050	0.049	0.066	$7.0x10^{-4}$	$8.0x10^{-4}$	$7.7x10^{-}$	$1.6x10^{-3}$	$2.5x10^{-5}$	$5.0x10^{-5}$	$5.0x10^{-5}$	$1.5x10^{-4}$
6000	40	0.25	0.35	0.052	0.052	0.051	0.067	$1.2x10^{-3}$	$1.1x10^{-3}$	$1.3x10^{-3}$	$2.3x10^{-3}$	$2.2x10^{-4}$	$1.0x10^{-4}$	$2.5x10^{-4}$	$3.7x10^{-4}$
6000	40	0.50	0.00	0.049	0.050	0.046	0.086	$1.1x10^{-3}$	$8.0x10^{-4}$	$9.2x10^{-4}$	$2.7x10^{-3}$	$1.7x10^{-4}$	0	$1.2x10^{-4}$	$4.2x10^{-4}$
6000	40	0.50	0.15	0.050	0.049	0.047	0.083	$9.5x10^{-4}$	$8.0x10^{-4}$	$9.7x10^{-4}$	$4.0x10^{-3}$	$2.5x10^{-5}$	$5.0x10^{-5}$	$1.5x10^{-4}$	$6.0x10^{-4}$
6000	40	0.50	0.35	0.052	0.051	0.049	0.084	$1.3x10^{-3}$	$1.2x10^{-3}$	$1.4x10^{-3}$	$3.9x10^{-3}$	$1.5x10^{-4}$	$7.5x10^{-5}$	$2.7x10^{-4}$	$5.7x10^{-4}$
6000	40	0.75	0.00	0.052	0.050	0.048	0.108	$1.2x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$4.8x10^{-3}$	$1.0x10^{-4}$	$5.0x10^{-5}$	$7.5x10^{-5}$	$5.7x10^{-4}$
6000	40	0.75	0.15	0.050	0.051	0.046	0.098	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.1x10^{-3}$	$6.0x10^{-3}$	$1.2x10^{-4}$	$1.2x10^{-4}$	$2.0x10^{-4}$	$1.3x10^{-3}$
6000	40	0.75	0.35	0.052	0.051	0.049	0.088	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.3x10^{-3}$	$4.8x10^{-3}$	$2.0x10^{-4}$	$1.5x10^{-4}$	$2.7x10^{-4}$	$1.1x10^{-3}$
6000	80	0.25	0.00	0.049	0.047	0.047	0.065	$8.5x10^{-4}$	$9.5x10^{-4}$	$9.0x10^{-4}$	$1.7x10^{-3}$	$5.0x10^{-5}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$1.7x10^{-3}$
6000	80	0.25	0.15	0.051	0.051	0.048	0.067	$9.5x10^{-4}$	$1.1x10^{-3}$	$1.1x10^{-3}$	$2.1x10^{-3}$	$7.5x10^{-5}$	$1.2x10^{-4}$	$1.2x10^{-4}$	$1.7x10^{-4}$
6000	80	0.25	0.35	0.051	0.052	0.049	0.069	$1.1x10^{-3}$	$1.2x10^{-3}$	$1.3x10^{-3}$	$2.4x10^{-3}$	$1.2x10^{-4}$	$1.2x10^{-4}$	$1.7x10^{-4}$	$3.2x10^{-4}$
6000	80	0.50	0.00	0.050	0.051	0.047	0.087	$1.2x10^{-3}$	$8.5x10^{-4}$	$1.2x10^{-3}$	$3.4x10^{-3}$	$1.0x10^{-4}$	$5.0x10^{-5}$	$1.7x10^{-4}$	$7.0x10^{-4}$
6000	80	0.50	0.15	0.050	0.049	0.047	0.085	$9.2x10^{-4}$	$9.7x10^{-4}$	$1.2x10^{-3}$	$3.3x10^{-3}$	$5.0x10^{-5}$	$1.0x10^{-4}$	$7.5x10^{-5}$	$5.5x10^{-4}$
6000	80	0.50	0.35	0.051	0.050	0.049	0.080	$9.7x10^{-4}$	$9.0x10^{-4}$	$1.1x10^{-3}$	$3.4x10^{-3}$	$7.5x10^{-5}$	$1.0x10^{-4}$	$2.5x10^{-4}$	$6.5x10^{-4}$
6000	80	0.75	0.00	0.051	0.051	0.047	0.098	$8.2x10^{-4}$	$9.5x10^{-4}$	$9.5x10^{-4}$	$4.7x10^{-3}$	$1.5x10^{-4}$	$5.0x10^{-5}$	$7.5x10^{-5}$	$7.0x10^{-4}$
6000	80	0.75	0.15	0.050	0.047	0.048	0.089	$8.5x10^{-4}$	$1.2x10^{-3}$	$9.5x10^{-4}$	$4.4x10^{-3}$	$2.5x10^{-5}$	$2.5x10^{-5}$	$1.5x10^{-4}$	$1.0x10^{-3}$
6000	80	0.75	0.35	0.048	0.052	0.046	0.091	$8.7x10^{-4}$	$1.1x10^{-3}$	$1.1x10^{-3}$	$3.8x10^{-3}$	$1.3x10^{-4}$	$1.0x10^{-4}$	$2.7x10^{-4}$	$8.0x10^{-4}$

Abbreviations: r^2 is the total outcome variance that can be explained by other variables ; π is the proportion of variables associated with the predictor ; LR= standard (unadjusted) marginal univariate test ; OPT= the optimally adjusted test that includes as covariates only the variables not associated with the predictor ; and FT=a univariate test that include as covariate all variable with a p-value for association with the predictor above 0.1.

Supplementary Table 5. Replication of the metabolite results

		Discovery			KORA+TwinsUK		FHS		Finnish	
SNP	**Gene**	**Outcome**	P_{LR}	P_{CMS}	**Best SNP**	**pval**	**Best SNP**	**pval**	**Best SNP**	**pval**
rs477992	PHGDH	serine	$6.2x10^{-5}$	$1.4x10^{-7}$	rs1163251	$7.0x10^{-27}$	rs677277	$6.5x10^{-5}$	-	-
rs2216405	CPS1	glycine	$4.1x10^{-26}$	$2.3x10^{-33}$	rs715	$1.6x10^{-147}$	rs7422339	$2.4x10^{-58}$	-	-
		serine	$3.7x10^{-5}$	$6.4x10^{-10}$		$2.7x10^{-21}$	rs4513234	$1.5x10^{-5}$	-	-
		creatine	$7.6x10^{-8}$	$4.8x10^{-9}$		$9.6x10^{-25}$	rs7422339	$2.5x10^{-11}$	-	-
		acetylglycine	$2.2x10^{-8}$	$3.1x10^{-9}$		$6.7x10^{-58}$	-	-	-	-
rs780094	GCKR	alanine	$6.1x10^{-5}$	$4.0x10^{-8}$	rs1260326	$5.6x10^{-14}$	rs1260326	$7.6x10^{-12}$	rs1728918	$5.7x10^{-11}$
rs1352844	GC	lactose	$6.1x10^{-7}$	$6.3x10^{-6}$	-	-	-	-	-	-
rs7094971	SLC16A9	carnitine	$2.9x10^{-10}$	$1.1x10^{-15}$	rs12356193	$3.7x10^{-63}$	rs1171617	$5.9x10^{-26}$	-	-
		acetylcarnitine	$1.4x10^{-6}$	$9.4x10^{-13}$	rs1171614	$3.4x10^{-23}$	-	-	-	-
rs2657879	GLS2	glutamine	$3.1x10^{-5}$	$4.2x10^{-10}$	rs2694917	$1.6x10^{-29}$	-	-	rs2638315	$1.1x10^{-32}$
rs6499165	SLC7A6	lysine	$2.6x10^{-5}$	$7.5x10^{-10}$	rs8056893	$2.5x10^{-39}$	-	-	-	-

Abbreviations: P_{LR} is the p-value for the standard unadjusted univariate test of each single phenotype with each single SNP; P_{CMS} is the p-value from the CMS approach. N equals 8,330, 7,824, and 2,076 for Finnish[13], KORA+TwinsUK[10,14], and FHS[15],

Supplementary Table 6. Summary of results when adjusting for PCs of metabolites

#PC	SNP	Gene	Outcome	P_{LR}	P_{CMS}	SS_{incr}	
5PCS	rs1352844	GC	lactose	**$7.3x10^{-7}$**	**$7.3x10^{-7}$**	1.00	-
	rs174547	FADS1	tryptophan	**$9.6x10^{-5}$**	**$1.2x10^{-8}$**	2.33	FHS[15]
			creatinine	**$7.2x10^{-7}$**	**$2.6x10^{-8}$**	5.56	-
	rs2216405	CSP1	glycine	**$9.6x10^{-26}$**	**$4.3x10^{-23}$**	1.75	KORA+TwinsUK[10] / FHS[15]
			acetylglycine	**$1.0x10^{-7}$**	**$3.8x10^{-5}$**	1.75	KORA+TwinsUK[10]
	rs2657879	SPRYD4	alanine	$1.2x10^{-4}$	**$2.4x10^{-8}$**	2.56	KORA+TwinsUK[10] / FHS[15] / Finnish[13]
			glutamine	**$2.3x10^{-11}$**	**$2.6x10^{-13}$**	2.38	Finnish[13]
	rs2764886	PHGDH	serine	$5.1x10^{-3}$	**$2.3x10^{-8}$**	2.13	KORA+TwinsUK[10]
	rs6499165	SLC7A6	lysine	**$3.3x10^{-8}$**	**$3.3x10^{-8}$**	1.00	KORA+TwinsUK[10]
	rs7094971	SLC16A9	carnitine	**$1.8x10^{-14}$**	**$9.5x10^{-14}$**	1.96	KORA+TwinsUK[10] / FHS[15]
			acetylcarnitine	**$2.4x10^{-7}$**	**$1.3x10^{-9}$**	1.72	KORA+TwinsUK[10]
	rs780094	GCKR	alanine	**$1.7x10^{-8}$**	**$6.2x10^{-6}$**	1.79	KORA+TwinsUK[10] / FHS[15] / Finnish[13]
	rs8103135	TPRX2P	carnitine	$1.6x10^{-5}$	**$1.8x10^{-7}$**	5.88	-
10 PCS	rs1352844	GC	lactose	**$6.0x10^{-7}$**	**$6.0x10^{-7}$**	1.00	-
	rs17288067	ACO1	urate	$2.8x10^{-3}$	**$8.3x10^{-8}$**	5.00	-
	rs174547	FADS1	tryptophan	$1.3x10^{-5}$	**$8.9x10^{-7}$**	12.50	FHS[15]
			creatinine	**$1.6x10^{-8}$**	**$4.7x10^{-7}$**	4.17	-
			taurochenodeoxycholic acid	$1.8x10^{-2}$	**$1.4x10^{-7}$**	10.00	-
	rs2216405	CSP1	glycine	**$2.8x10^{-28}$**	**$4.4x10^{-27}$**	1.69	KORA+TwinsUK[10] / FHS[15]
			creatine	**$5.6x10^{-8}$**	**$4.4x10^{-9}$**	2.33	KORA+TwinsUK[10] / FHS[15]
	rs2657879	SPRYD4	glutamine	**$7.0x10^{-12}$**	**$5.8x10^{-12}$**	2.38	Finnish[13]
	rs3760776	FUT6	butyrobetaine	**$6.7x10^{-7}$**	**$4.2x10^{-8}$**	4.17	-
	rs477992	PHGDH	serine	$9.2x10^{-7}$	**$9.0x10^{-8}$**	2.86	KORA+TwinsUK[10]
	rs6499165	SLC7A6	lysine	**$2.8x10^{-9}$**	**$6.1x10^{-7}$**	2.63	KORA+TwinsUK[10]
	rs7094971	SLC16A9	dimethylarginine	$9.7x10^{-4}$	**$1.2x10^{-9}$**	2.27	-
			carnitine	**$1.4x10^{-10}$**	**$2.4x10^{-13}$**	2.70	KORA+TwinsUK[10] / FHS[15]
			acetylcarnitine	**$5.9x10^{-8}$**	**$2.9x10^{-6}$**	2.94	KORA+TwinsUK[10]
	rs780094	GCKR	alanine	**$7.6x10^{-8}$**	**$1.6x10^{-3}$**	2.00	KORA+TwinsUK[10] / FHS[15] / Finnish[13]
			threonine	$2.2x10^{-6}$	**$7.8x10^{-8}$**	1.43	Suhre[14]

	rs7855483	ACO1	4-Dimethylaminopyridine	$5.4x10^{-3}$	**$4.4x10^{-9}$**	3.85	-
20 PCS	rs10493380	LEPR	Trimethylamine N-oxide	$4.4x10^{-3}$	**$1.1x10^{-7}$**	4.00	FHS[15]
			alpha-hydroxybutyrate dehydrogenase	$4.7x10^{-4}$	**$5.6x10^{-9}$**	3.03	-
	rs17277546	TRIM4	allantoin	$8.5x10^{-3}$	**$5.3x10^{-8}$**	8.33	-
	rs174547	FADS1	tryptophan	$7.0x10^{-6}$	**$9.0x10^{-7}$**	2.70	FHS[15]
			creatinine	**$8.7x10^{-8}$**	**$5.1x10^{-8}$**	4.00	-
	rs2216405	CSP1	glycine	**$1.3x10^{-20}$**	**$2.2x10^{-20}$**	1.75	KORA+TwinsUK[10] / FHS[15]
			creatine	$4.5x10^{-5}$	**$5.3x10^{-11}$**	2.13	KORA+TwinsUK[10] / FHS[15]
	rs2657879	SPRYD4	glutamine	**$9.0x10^{-10}$**	**$2.4x10^{-13}$**	2.63	Finnish[13]
	rs272889	AC034220.3	2-aminoadipate	$5.7x10^{-4}$	**$1.5x10^{-7}$**	7.69	-
	rs2764886	PHGDH	serine	$2.9x10^{-5}$	**$7.8x10^{-11}$**	2.27	KORA+TwinsUK[10] / FHS[15]
	rs3760776	FUT6	butyrobetaine	$1.5x10^{-5}$	**$8.5x10^{-8}$**	2.50	-
			sarcosine	$2.0x10^{-3}$	**$5.0x10^{-7}$**	3.85	-
	rs477992	PHGDH	serine	$2.2x10^{-6}$	**$7.4x10^{-10}$**	4.17	KORA+TwinsUK[10]
			aconitate	$7.9x10^{-4}$	**$5.2x10^{-7}$**	2.17	-
	rs6499165	SLC7A6	lysine	**$1.8x10^{-9}$**	**$1.1x10^{-7}$**	3.33	KORA+TwinsUK[10]
	rs7094971	SLC16A9	carnitine	**$1.0x10^{-7}$**	**$1.7x10^{-10}$**	2.63	KORA+TwinsUK[10] / FHS[15]
	rs780094	GCKR	thyroxine	$2.6x10^{-4}$	**$3.7x10^{-7}$**	2.44	-

Abbreviation: P_{LR} is the p-value for the standard unadjusted univariate test of each single phenotype with each single SNP; P_{CMS} is the p-value from the CMS algorithm; SS_{incr} is the equivalent sample size increase achieved after adjusting for covariates selected by the CMS algorithm.

There was 79 metabolites tested for association with 668 SNPs, so a total of 52104 tests. P-value threshold accounting for multiple testing is 9.5x10-7. Significant p-values are indicated in bold.

Supplementary Table 7. Summary of GEUVADIS results when adjusting for an increasing number of principal component of expression

	No PC	5 PCs	30 PCs
LR	515 (6.2%)	771 (6.33%)	2186 (17.6%)
CMS	996 (12%)	1599 (13.1%)	2420 (19.5%)
Overlap	515 (6.2%)	767 (6.3%)	2162 (17.4%)
Combined	996 (12%)	1603 (13.2%)	2444 (19.7%)

Abbreviation: LR= standard (unadjusted) marginal linear regression.

Supplementary Table 8. Replication rate of established Cis-eQTL between existing LCL studies after excluding European GEUVADIS data

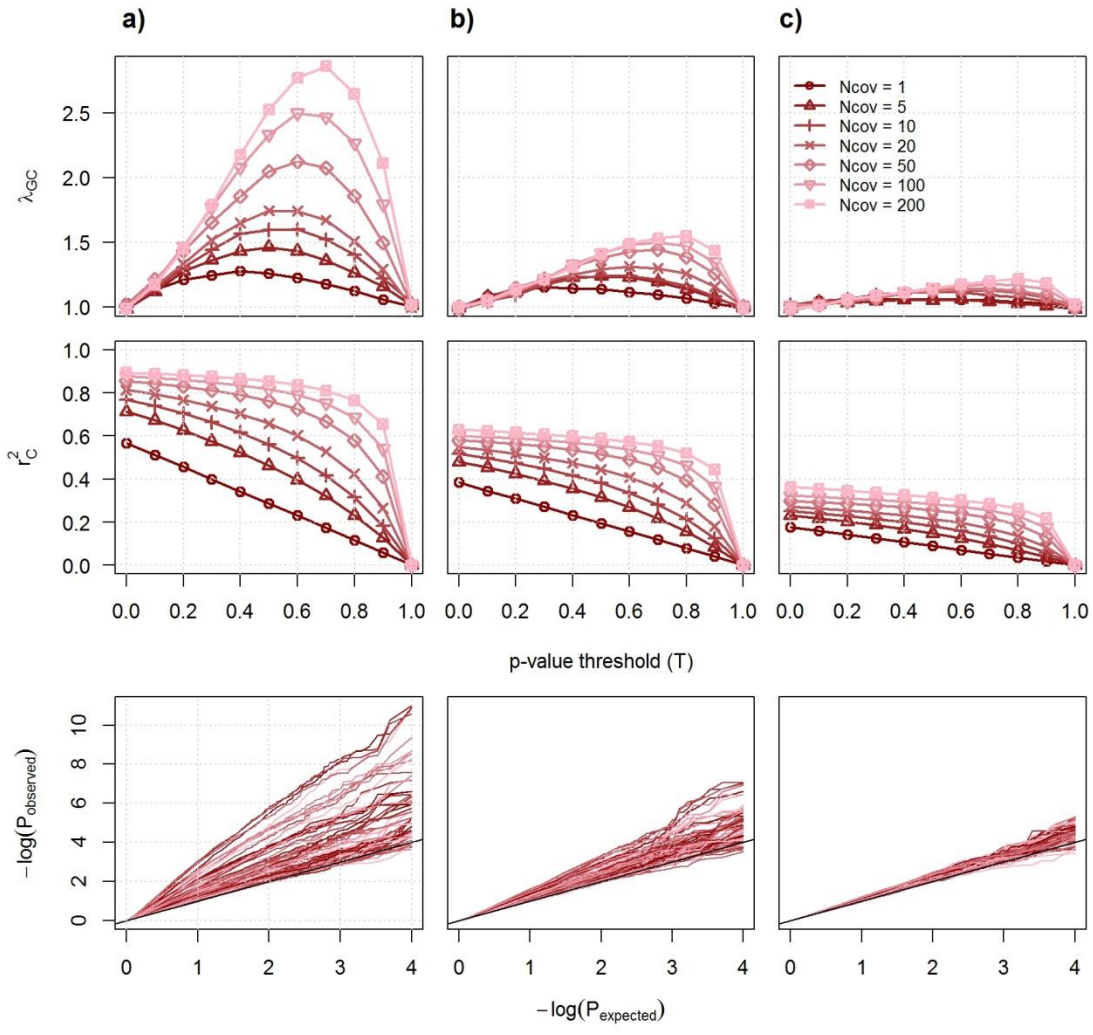
Study	Population ancestry	Sample size	All SNP-gene pairs		Top SNP per gene	
			#hits	***replicated (%)***	***#hits***	***replicated (%)***
3CL	*CEU*	75	560	283 *(51%)*	439	215 *(49%)*
AS	*CEU*	206	26,427	9,928 *(38%)*	1,793	695 *(39%)*
EGEUV_YRI	*YRI*	89	19,621	333 *(2%)*	498	15 *(3%)*
HA_CEU	*CEU*	30	6,934	4,526 *(65%)*	353	196 *(56%)*
HA_CHB	*CHB*	45	6,829	5,095 *(75%)*	355	215 *(61%)*
HA_JPT	*JPT*	45	8,266	5,910 *(71%)*	406	234 *(58%)*
HA_YRI	*YRI*	30	5,495	2,907 *(53%)*	451	196 *(43%)*
HA2_CEU	*CEU*	30	4,453	1,276 *(29%)*	106	39 *(37%)*
HA2_YRI	*YRI*	30	5,027	398 *(8%)*	165	29 *(18%)*
HRC	*CEU*	60	9,370	911 *(10%)*	956	18 *(2%)*
HRY	*YRI*	69	883	150 *(17%)*	867	149 *(17%)*
MRC	*CEU*	950	181,065	37,893 *(21%)*	1,279	233 *(18%)*
MuTHER_LCL	*CEU*	160	215,407	30,101 *(14%)*	4,012	406 *(10%)*

** count of replication count identified cis-eQTL from each single study that replicate in at least one of the other 12 studies listed in the first column of the table.*

Study references: HA[16], AS[17], 3C[18], HRC[19], HRY[20], MuTHER[21], MRC[22], E-GEUV[23].

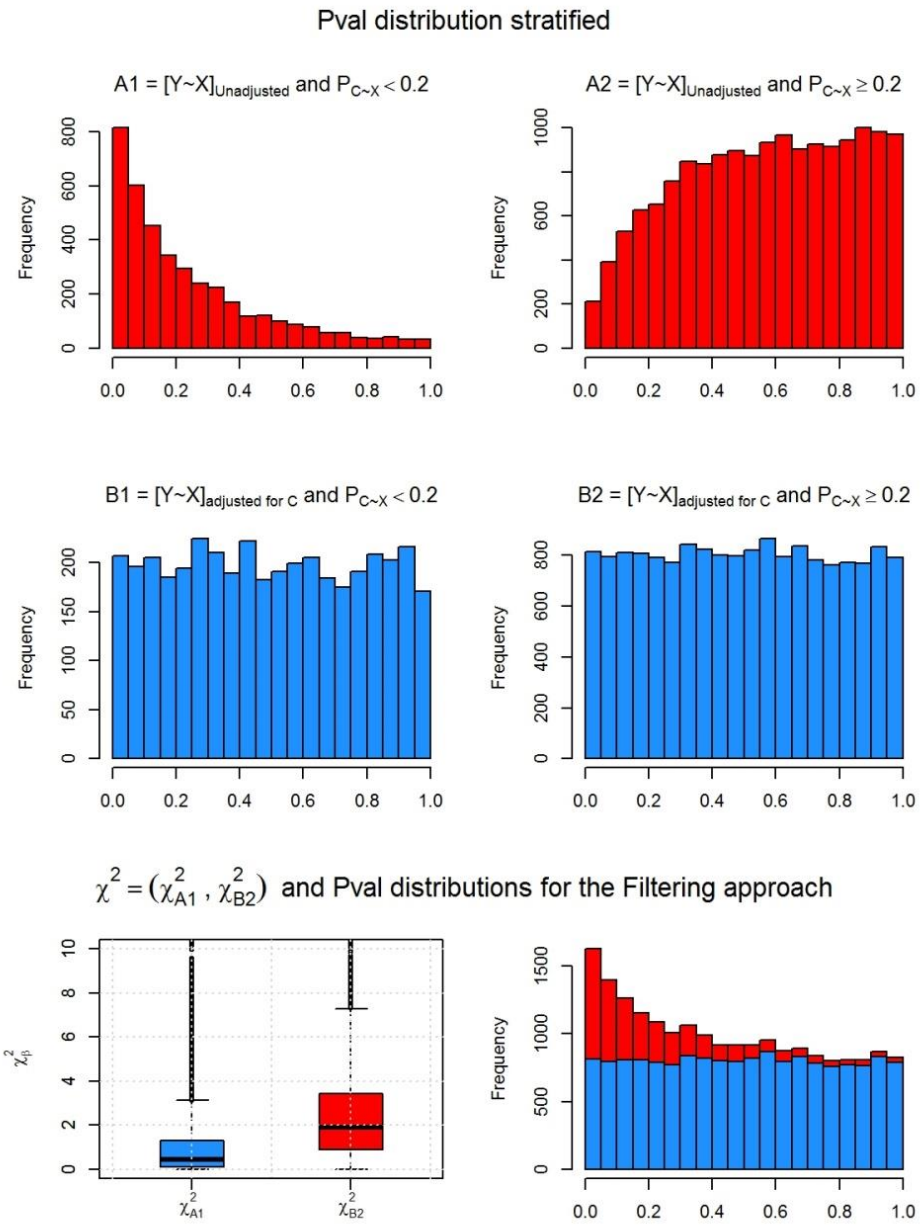
Population ancestry: YRI=Yoruba in Ibadan Nigeria, CEU=Utah Residents with Northern and Western European Ancestry, CHB=Han Chinese in Beijing, JPT=Japanese in Tokyo.

Supplementary Figure 1. Genomic inflation factor for p-values-based filtering.
We generated series of 10,000 replicates each including 2 to 200 correlated variables and a predictor under the complete null –i.e. not associated with any of the correlated variables. We randomly defined one of those variables as the primary outcome and tested it for association with the predictor while including other variables as covariates if their association test with the predictor had a *p*-value above a threshold T. Upper panel shows the genomic inflation factor λ_{GC} of the *p*-values from this test while increasing T from 0 to 1. We considered either strong (a), moderate (b) or low (c) correlation between variables, as measured in the middle panel by r_C^2, the variance of the primary outcome explained by covariate included in the model. The QQplots of each of these experiments (lower panels) show an overall inflation of the test without marked outliers at the tail of the distribution.



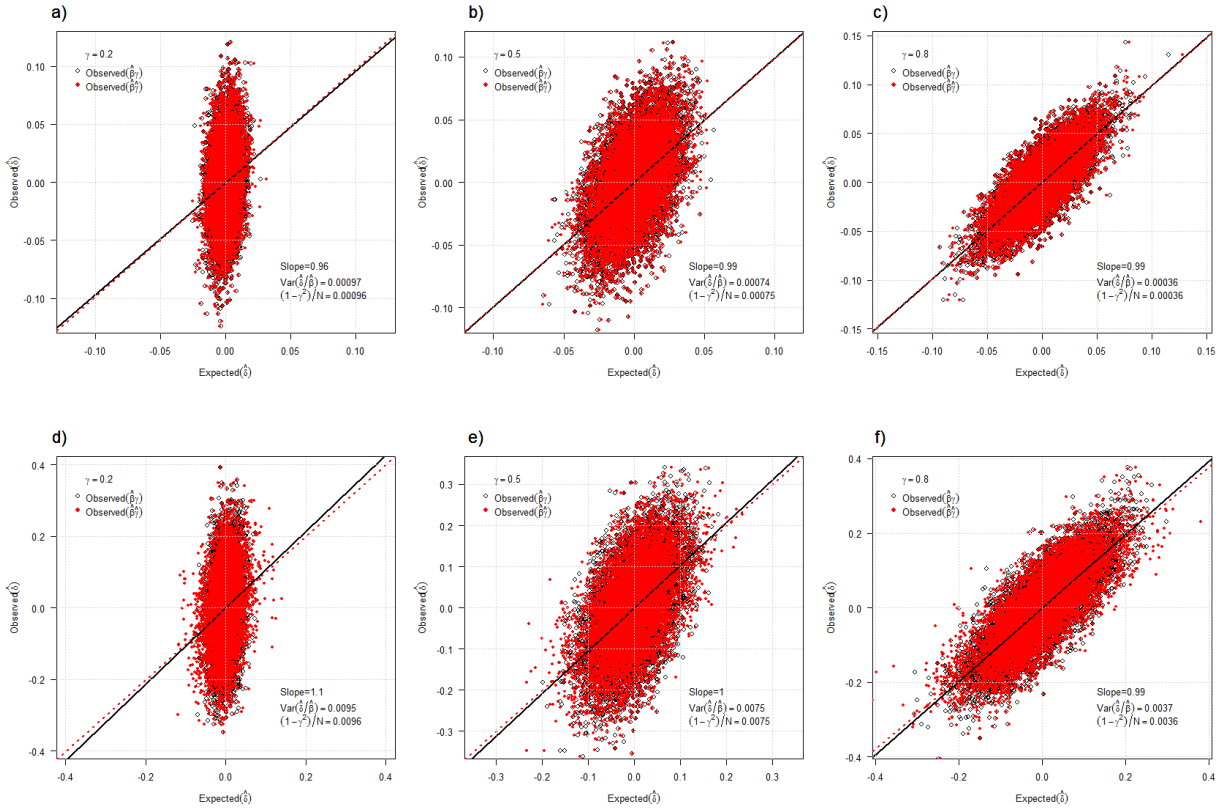
Supplementary Figure 2. Illustration of the type I error rate inflation for p-values based filtering.

We simulated series of 10,000 replicates, each including three variables, Y, C, and X, corresponding to an outcome, a candidate covariate, and a predictor, respectively. The three variables were generated using a multivariate normal distribution with mean 0, variance 1, and covariance $cov(Y,C) = \gamma = 0.8$, while $cov(Y,X) = cov(C,X) = 0$. Top histogram plots (red) show the p-value distribution for Y-X association obtain from standard marginal regression for the subset of replicates where the p-value for association for C-X is lower than 0.2 (left) or larger or equals than 0.2 (right). Middle plot (blue) show the same distribution but when Y-X regression are adjusted for C. Bottom plot shows the resulting chi-square (left) and p-value distribution for the resulting p-value-based filtering approach, that merges top left and middle right tests.



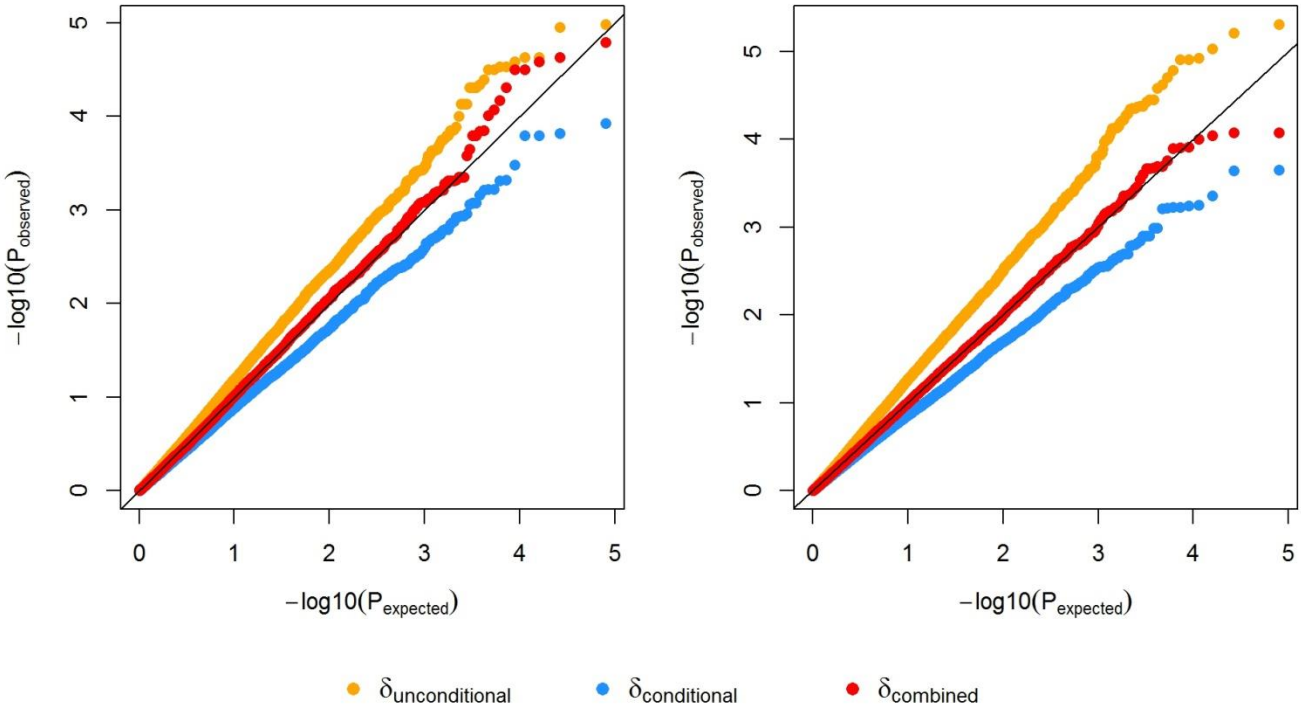
Supplementary Figure 3. Conditional mean and variance of the predictor-covariate regression coefficient.

We simulated series of 10,000 replicates, each including three variables, Y, C, and X, corresponding to an outcome, a candidate covariate, and a predictor, respectively. The three variables were generated using a multivariate normal distribution with mean 0, variance 1, and covariance $cov(Y,C) = \gamma$, $cov(Y,X) = \beta$, and $cov(X,C) = \delta$. We explored the special case where $\delta = \beta = 0$, while γ is relatively large and equals 0.2 (a, d), 0.5 (b, e), or 0.8 (c, f). For each series we derived and plotted the observed $\hat{\delta}$ against its expected value defined either as $\hat{\beta}\gamma$ (black circles) or $\hat{\beta}\hat{\gamma}$ (red dots) and estimated the regression coefficient between the 2 terms (i.e. the slope parameter, red dashed line). We also estimated the variance of $\hat{\delta}$ conditional on $\hat{\beta}$, $var(\hat{\delta}|\hat{\beta})$, defined as $\hat{\delta} - \mathbb{E}(\hat{\delta}|\hat{\beta}) = var(\hat{\delta} - \hat{\beta}\hat{\gamma})$, which we compared against its expected value, defined as $(1-\gamma^2)/N$, where N is the sample size. Upper (a, b, c) and lower (d, e, f) panels correspond to the case where $N = 100$ and $N = 1000$, respectively.



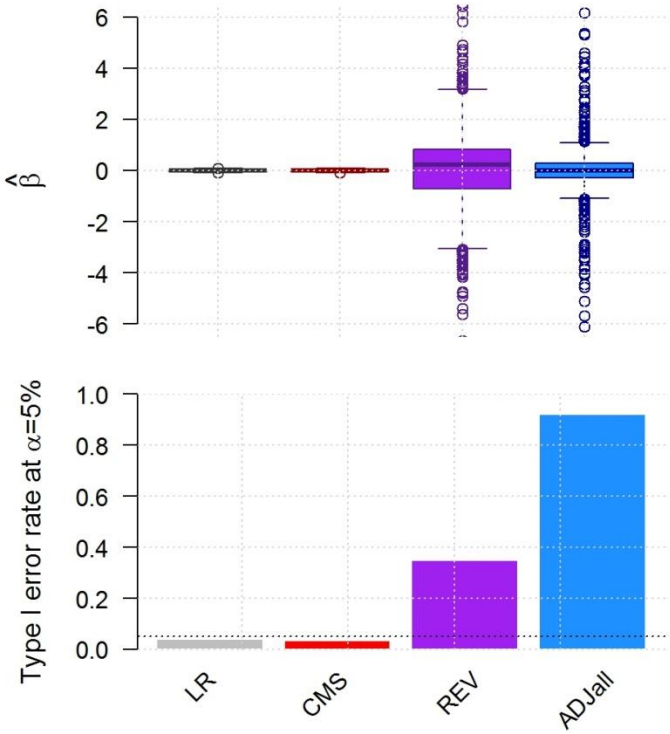
Supplementary Figure 4. QQplots for the predictor-covariate regression coefficient under the null.

We generated series of 10,000 replicates including 2,000 individuals. For each individual we generated independently a single predictor X, and 20 (left panel) or 50 variables (right panel) from a multivariate normal distribution with covariance matrix defined so that the pairwise correlation varies in [-0.4, 0.4]. We randomly defined one of those variables as the primary outcome and tested it for association with the predictor while including other variables as covariates if $\hat{\delta}$, the estimated effect of X on the candidate covariate is within the *unconditional* inclusion interval (orange), the conditional inclusion interval (blue), or both (i.e. matches either inclusion criteria, red). For simplicity, and to avoid the issue of outcome-covariate effect estimation, in the presence of multicollinearity we used principal component of covariates instead of their raw values. Note that we do not apply this transformation in the final version of the algorithm.

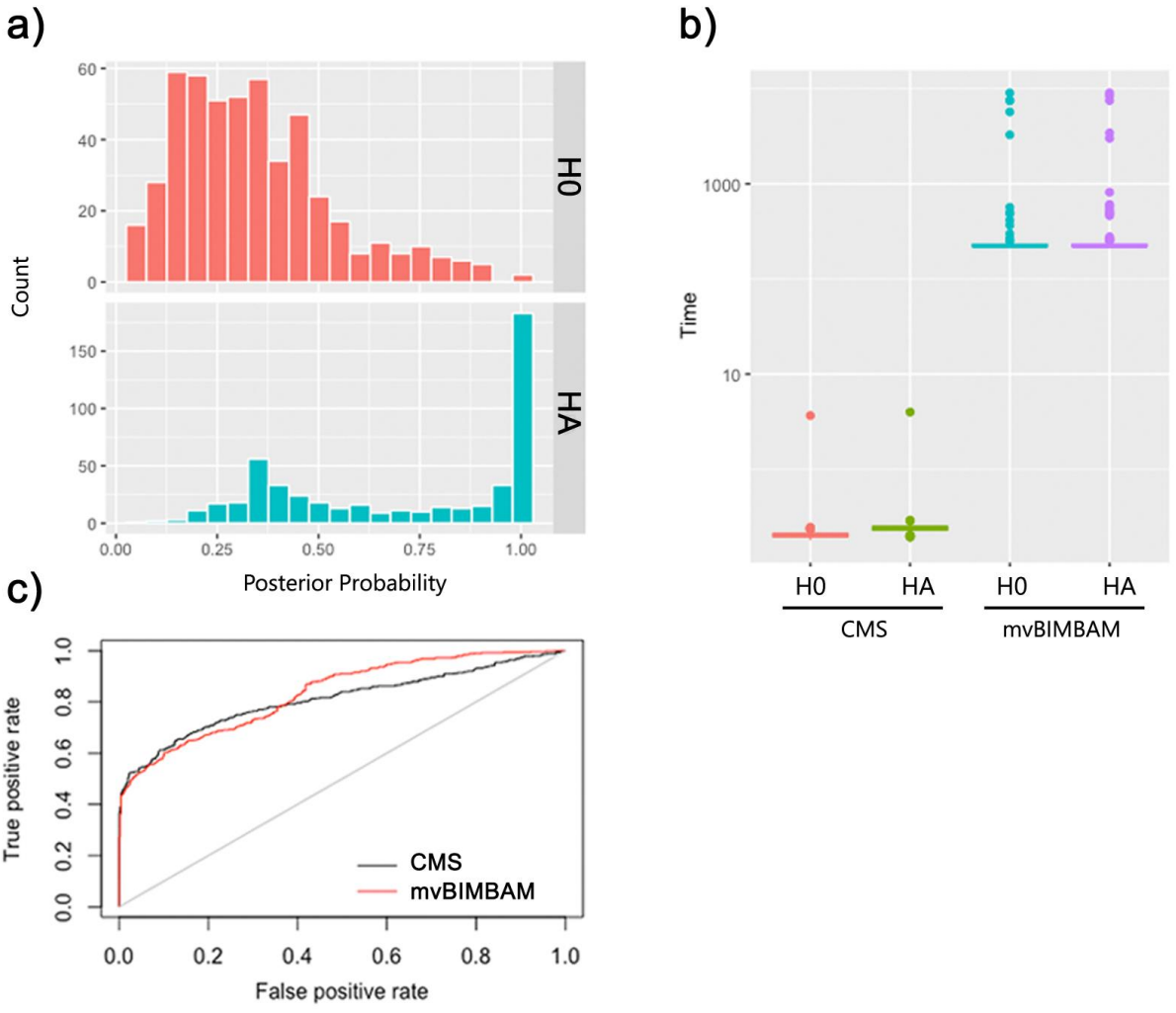


Supplementary Figure 5. Example of collinearity bias in multivariate analysis

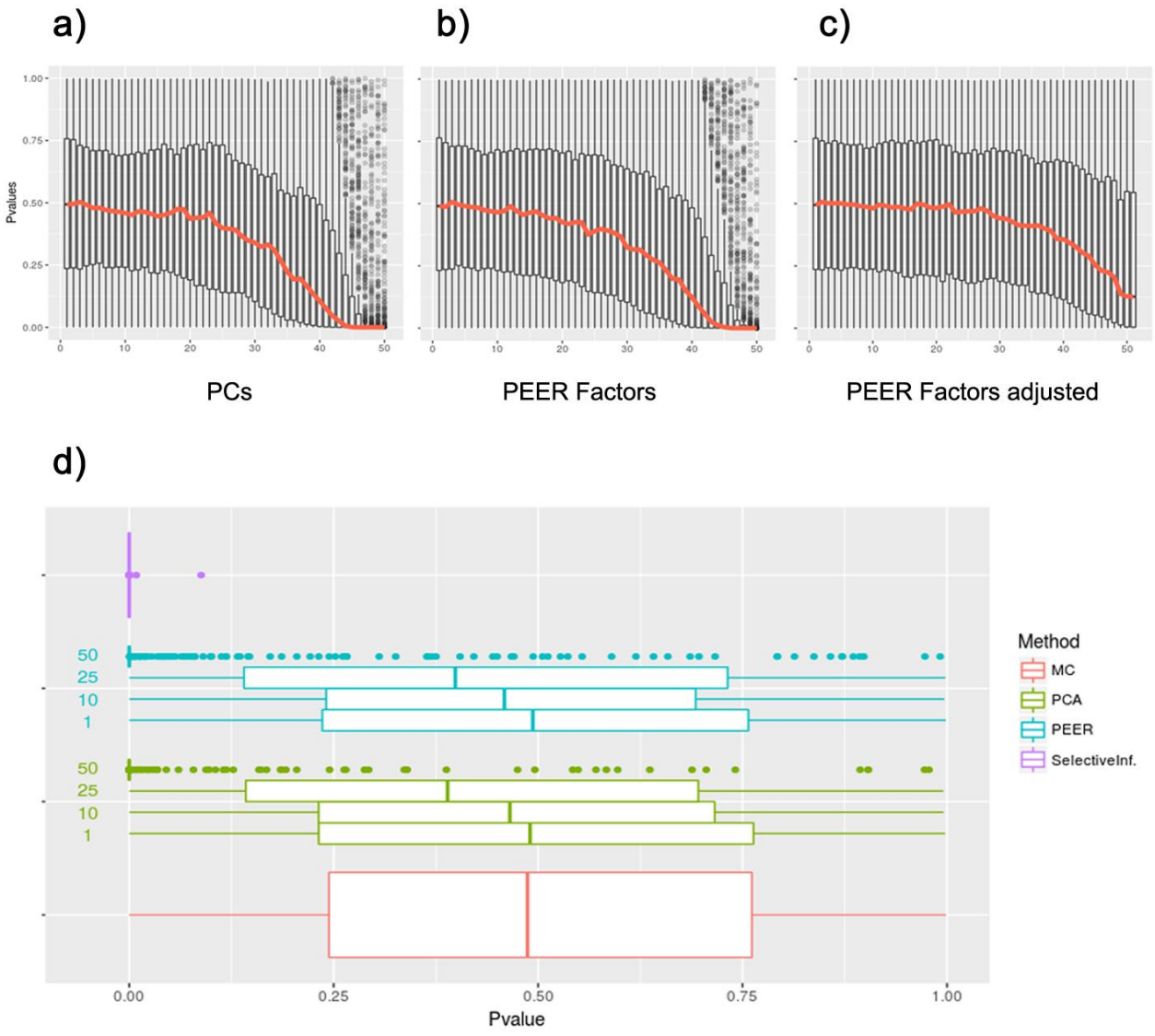
We simulated 1,000 replicates including each 30 correlated variables, one primary outcome and 29 secondary outcomes, and a one predictor across 1,000 individuals. For simplicity, both the outcomes and the predictor were normally distributed with mean 0 and variance 1. We considered a scenario where, on average, 75% of the variance of the primary outcome can be explained by the secondary outcomes. The predictor was associated with 35% of the secondary outcome but not associated with the primary outcome. We performed three four test of association: a standard (unadjusted) linear regression (LR), the *CMS* approach, a standard regression adjusting for all other (29) simulated variables (ADJall) and a so-called *Reverse regression* (REV), where the predictor was treated as the outcome and all outcomes (i.e. primary and secondary) are treated as predictors. For each test we estimated $\hat{\beta}$, the effect of the predictor on the primary outcome and the associated *p*-value. The upper and lower panels shows the distribution of $\hat{\beta}$ and the type I error rate at $\alpha = 5\%$ for MR, *CMS*, REV, and ADJall, respectively.



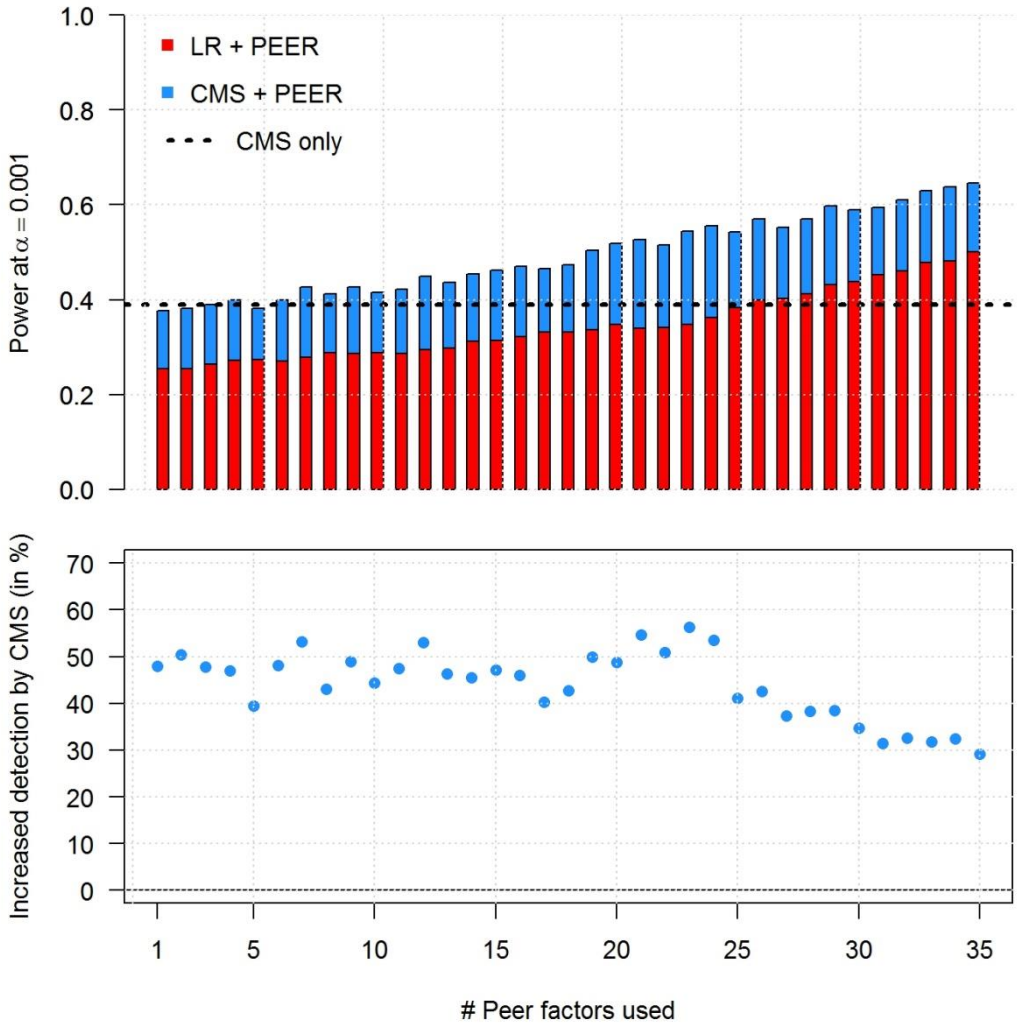
Supplementary Figure 6. Comparison between CMS and mvBIMBAM.
We simulated 500 series of datasets including 10,000 individuals. For each individual, we generated jointly an outcome and 10 covariates from a multivariate normal distribution so that pairwise correlation varies in [-0.6, 0.6], and a genetic variant with minor allele frequency drawn uniformly in [0.05, 0.95]. For half of the simulations we added a genetic effect to a random subset of the covariates but not to the outcome (H0), while in the second half, we added a genetic effect to the outcome but not to the covariates (HA). We performed a test of association between the outcome and the genetic variants using *CMS* and derived the posterior probability of the genetic variant being associated directly or indirectly with the outcome from *mvBIMBAM*. Panel a) shows the posterior probabilities for H0 and HA from *mvBIMBAM*, while panel b) shows the computation time in second for both approach in each scenario. Panel c) shows the ROC curves derived from the p-value for association for *CMS* and from the posterior probabilities for *mvBIMBAM.*



Supplementary Figure 7. Limitation of PEER factors and PC adjusted analysis.
We simulated series of datasets including 10,000 individuals. For each individual, we generated jointly an outcome and 49 covariates from a multivariate normal distribution so that pairwise correlation varies in [-0.4, 0.4], and a genetic variant with minor allele frequency drawn uniformly in [0.05, 0.95]. We added an effect of the genetic variant on some of the covariates but not with the outcome. We performed a test of association between the outcome and the genetic variants using four approaches: i) *CMS*, ii) standard linear regression adjusted for principal components (PCs), iii) standard linear regression adjusted for PEER factors, and iv) a LASSO regression method (Selective Inference). Panel a) and b) show the p-value for association as a function of the number of principal components (PC) and PEER factor added to the model, respectively. Panel c) shows the boxplot of p-values for each of the four approaches. In this simple scenario, only *CMS* has a correct uniform p-value distribution, while alternative approaches show deflation of the median p-values.

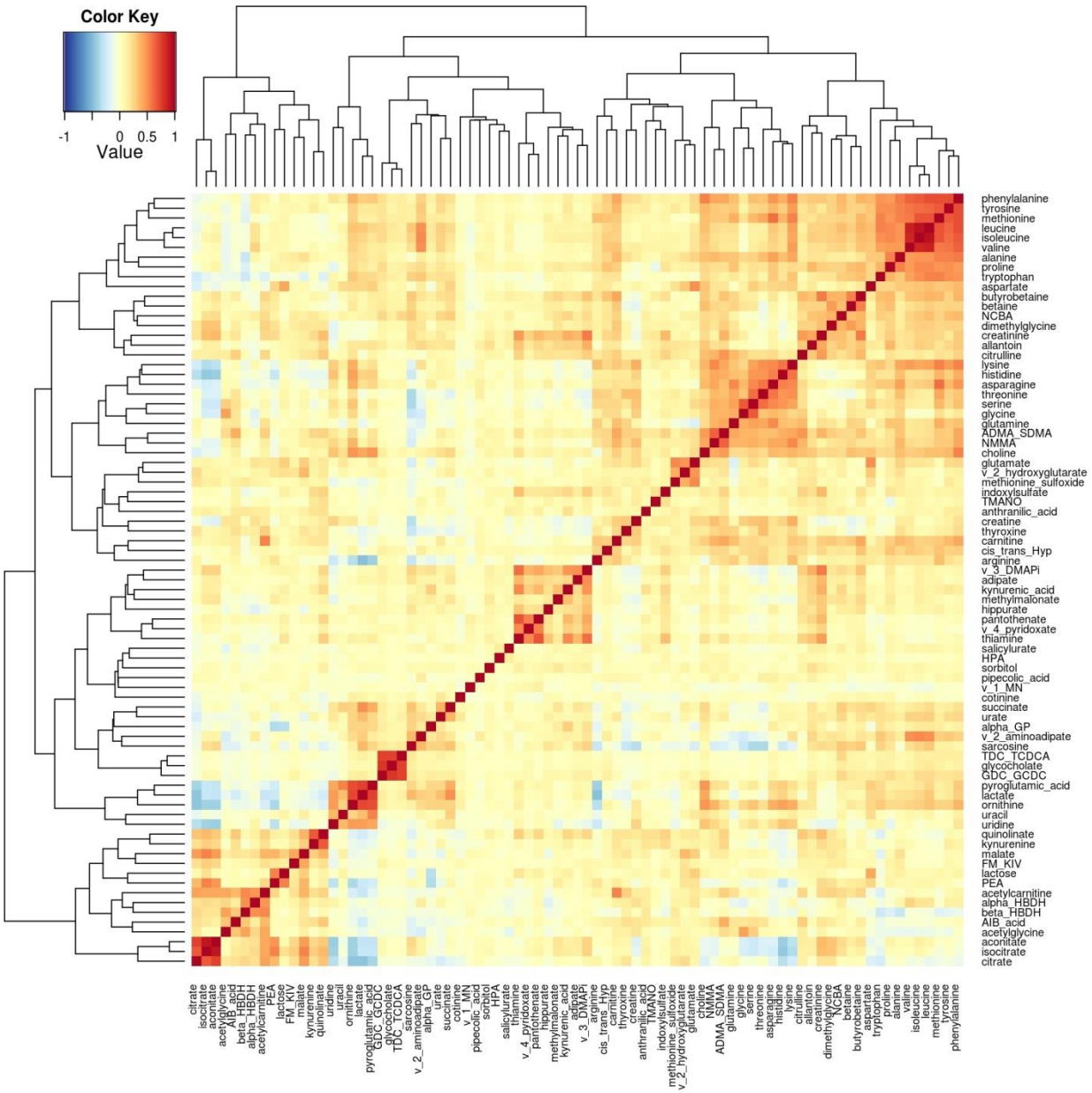


Supplementary Figure 8. Power of CMS when used in conjunction with PEER.
We simulated series of datasets including 10,000 individuals. For each individual, we generated jointly a primary outcome and 49 secondary outcomes from a multivariate normal distribution so that pairwise correlation varies in [-0.4, 0.4], and a genetic variant with minor allele frequency drawn uniformly in [0.05, 0.95]. We added an effect of the genetic variant on the primary outcome but not with the secondary outcomes. We then derived the PEER factors from all outcomes, and derived the residuals of each outcome after adjusting from 1 to 35 PEER factors, the largest number we could include without inducing substantial bias in Figure S11. We performed a test of association between the residual of the primary outcome and the genetic variants using either standard linear regression (LR, red bars) or *CMS* (blue bars). The upper panel shows the power for alpha of 0.001 for both approaches while increasing the number of PEER factors used for the adjustment. The bottom panel shows the corresponding increase in detection by *CMS* over LR.



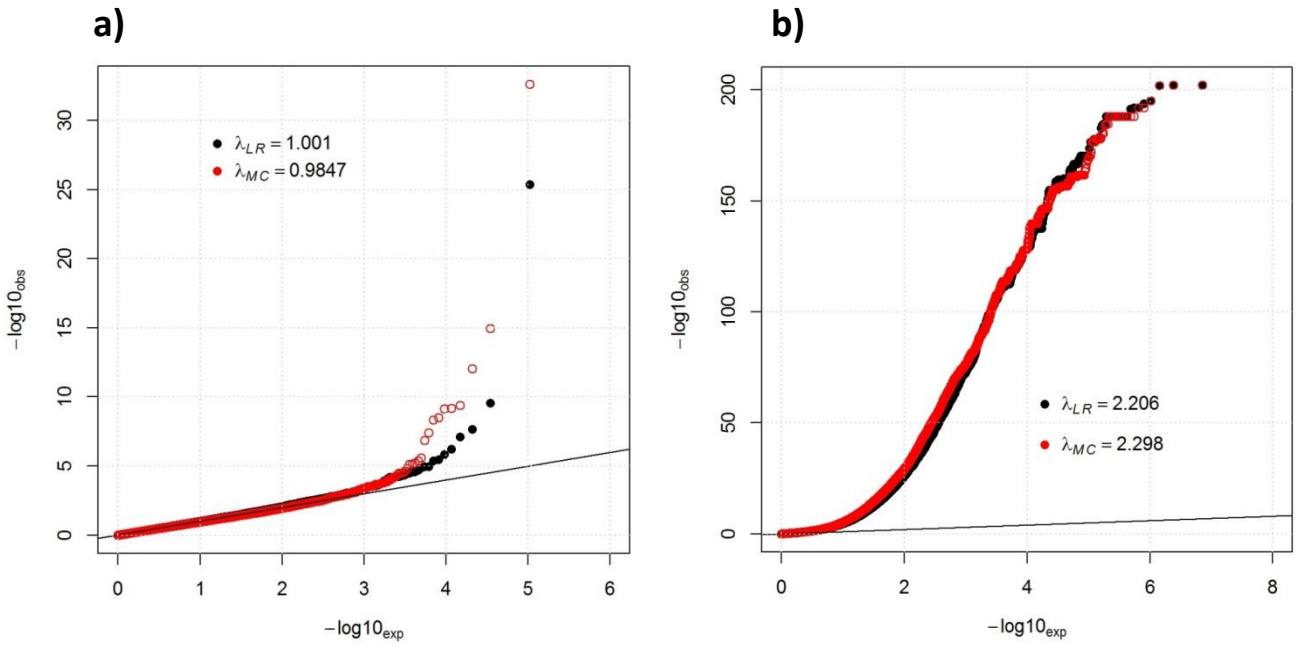
Supplementary Figure 9. Correlation matrix of metabolites.

Pairwise Pearson correlation between the 79 metabolites collected as part of the NHS, HPFS, PHS and WHI studies. Positive and negative correlations are highlighted in red and blue, respectively. Dendrogram were draw based on hierarchical clustering derived from the *hclustfun()* R function.



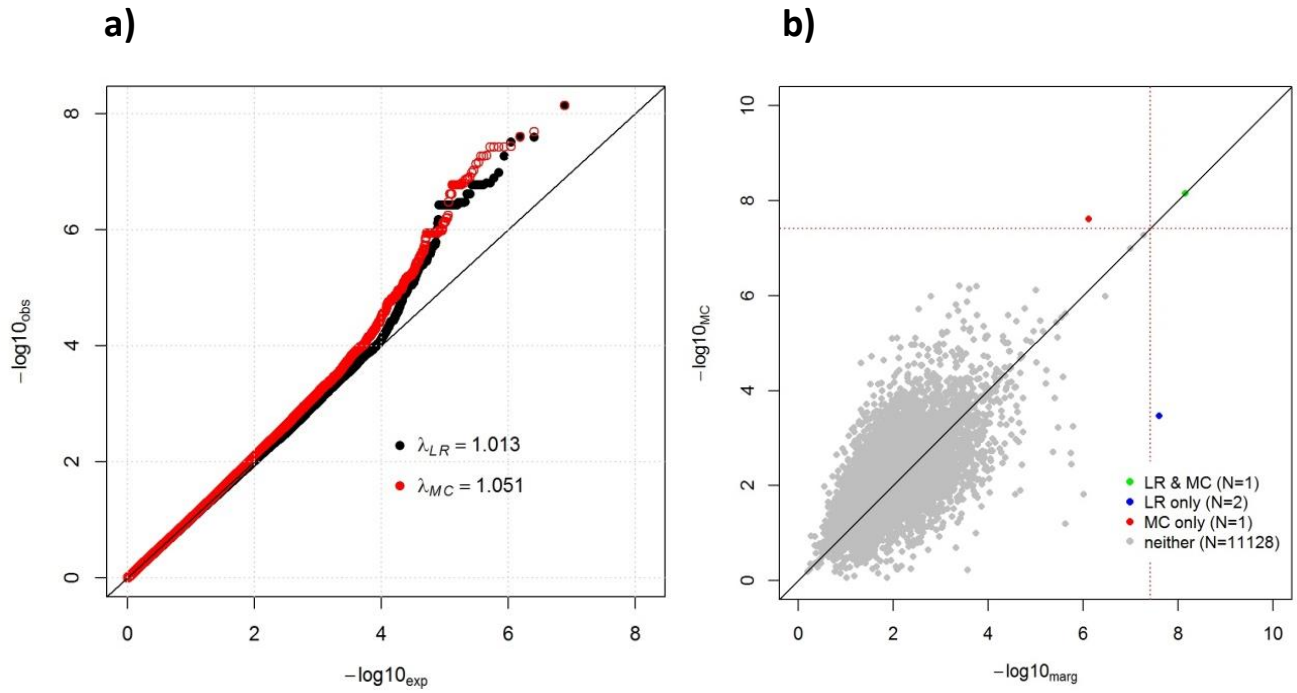
Supplementary Figure 10. QQplot from real data analyses.

Panel a) shows the QQplot from the association screening between 79 metabolites and 668 SNPs using 1,192 individuals from the PanScan study. Observed -log10(*p*-value) of this screening are plotted against an expected uniform *p*-value distribution. Panel b) present the genome-wide cis-eQTL screening in the gEUVADIS data. 11,694 genes were tested for association with genetic variants in close physical proximity for a total of 3.4 million tests. Observed -log10(*p*-value) of this screening are plotted against an expected uniform *p*-value distribution.



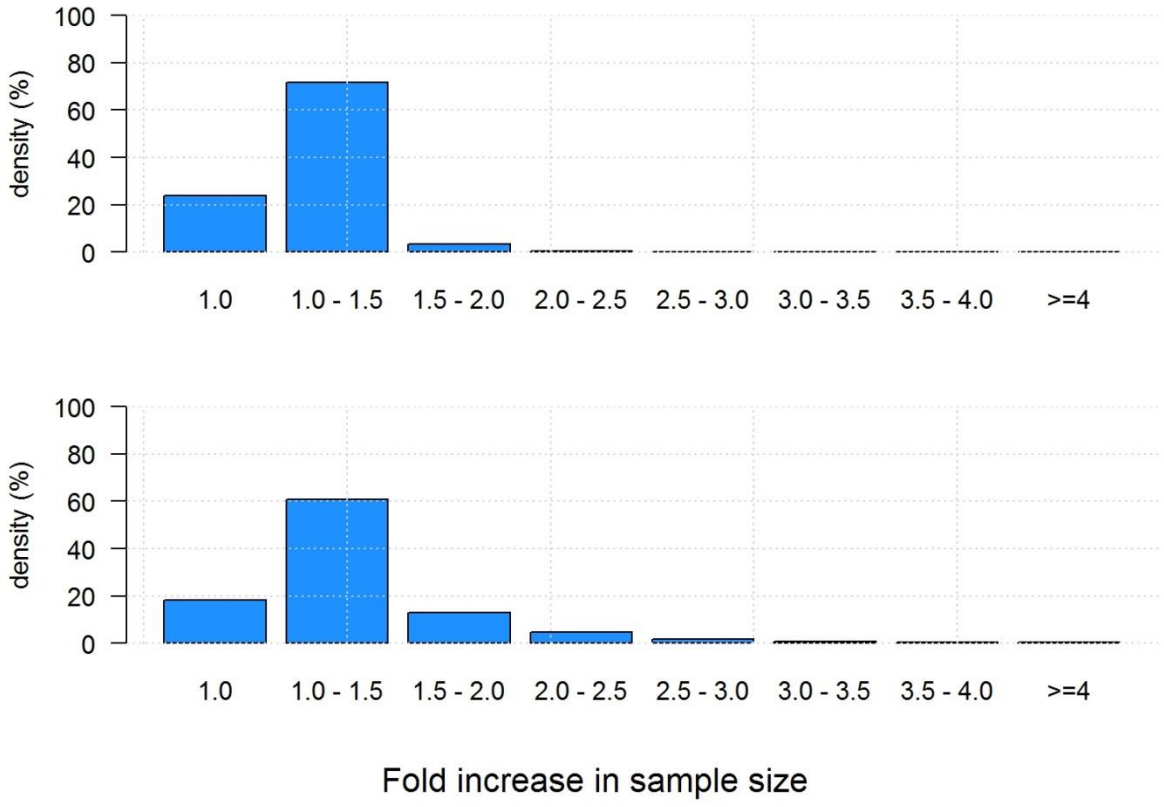
Supplementary Figure 11. Comparison of the standard approach and CMS in a quasi-null experiment

We aimed at mimicking the original genome-wide *cis*-eQTL mapping performed from in the gEUVADIS study, but under a quasi-null model of no association. To do so we kept all parameters of the real data analysis similar but we selected SNPs tested for *cis*-effects on a different chromosome that the targeted gene. Most of the tests are expected to be under the null, although some *trans* effects might be captured in this experiment. Analysis was performed using standard linear regression (LR, black) and the *CMS* (red) approach. Both consisted in running a linear regression adjusted for 10 PEER factors, while the *CMS* analysis also included 0 to 50 additional covariates per SNP/gene pair tested. We compared the $-\log_{10}$(*p*-value) of the two approaches against an expected uniform *p*-value distribution (panel a) and again each other's (panel b).

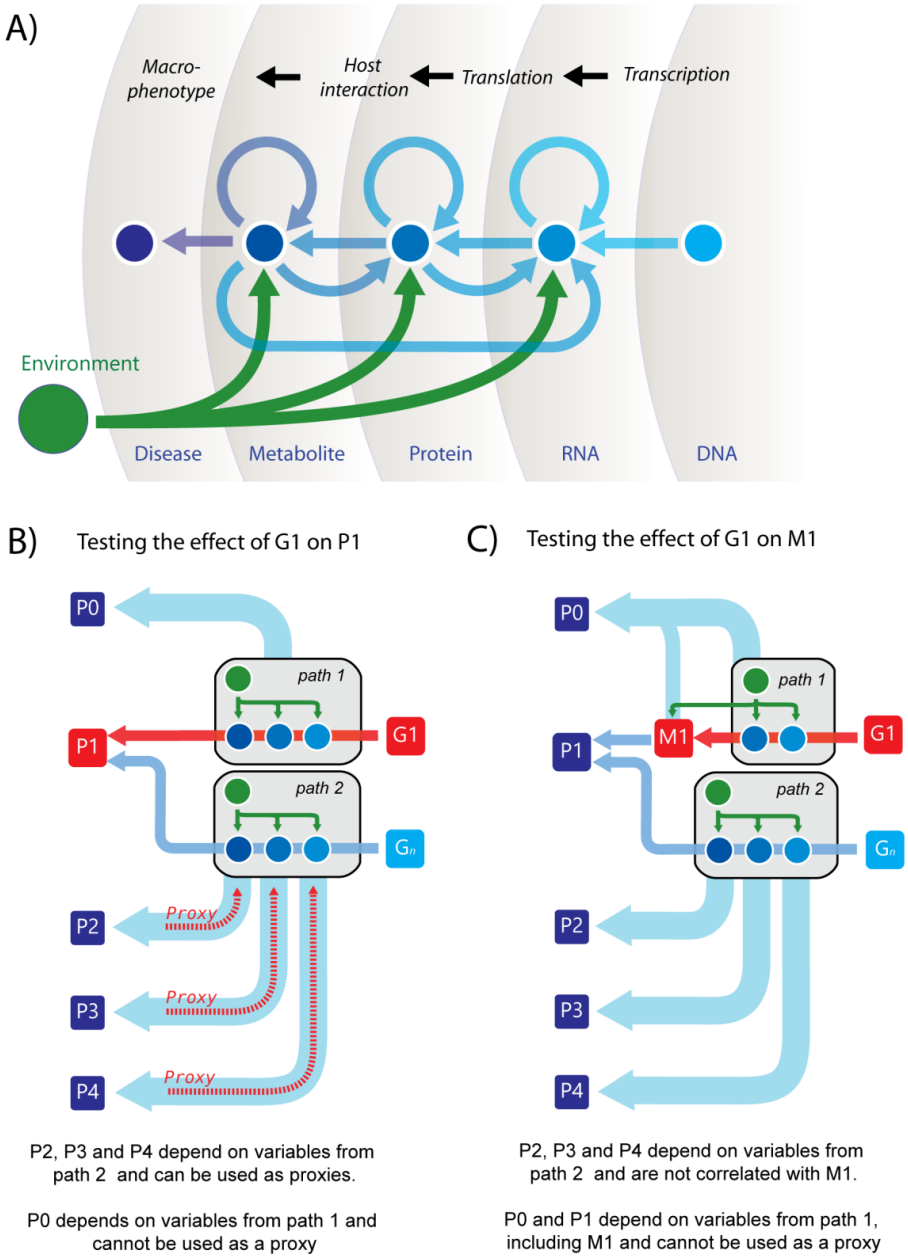


Supplementary Figure 12. Observed effective fold-increase in sample size

For both real data analysis, we estimated the increase in variance of outcomes explained after adding covariates selected by the *CMS* approach. This increase was measured as the difference in adjusted r-squared between the model including only the SNP tested and confounding factors for the metabolites analysis, and the PEER factors for the gene expression analysis. We used those estimates to derive the distribution of equivalent fold increase in sample size for in the gene expression (upper panel) and the metabolites (lower panel) data.

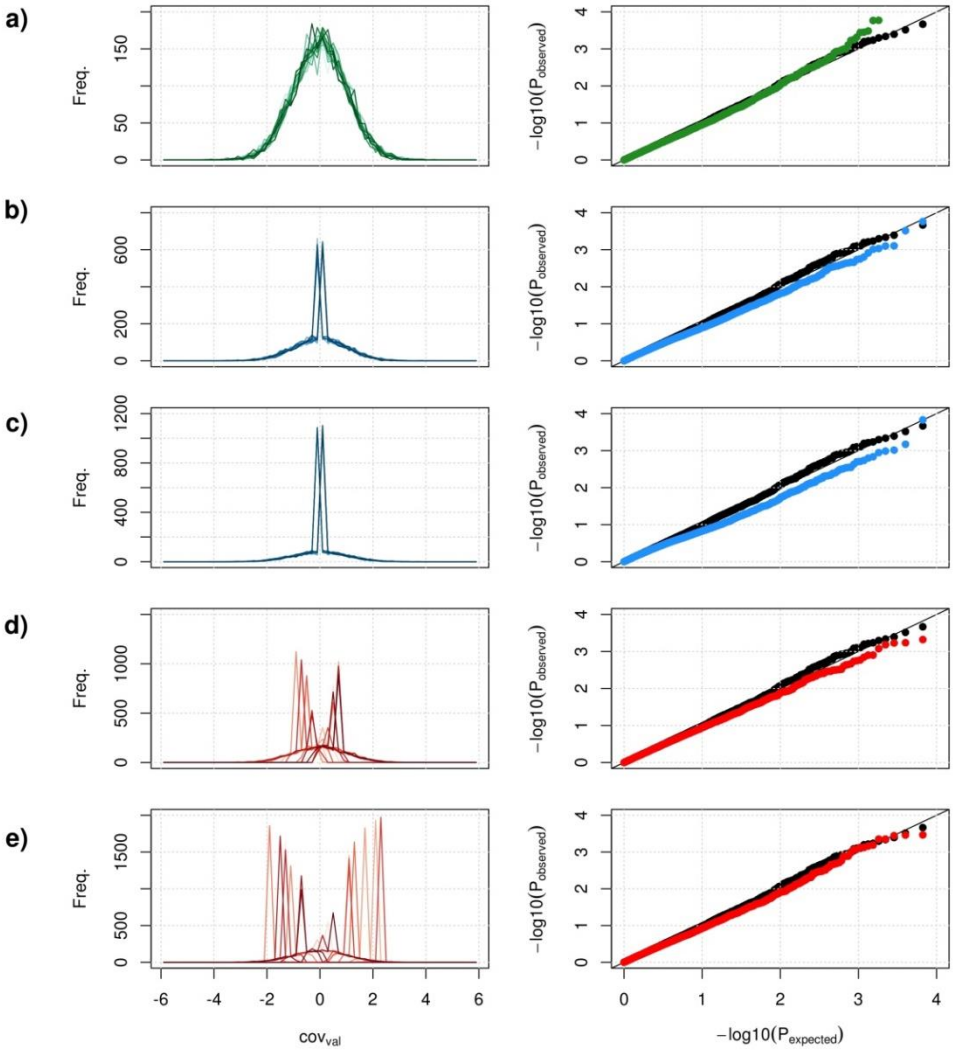


Supplementary Figure 13. Rationale for applying CMS – the example of genomic data.
In genomics data the underlying causal pathway is partially understood (A). Such information can be used to pre-select candidate covariates of interest. In particular, variables from the same structural level can be leverage to detect association with variables upstream in the causal pathway. For example in (B), one can test the association between a genotype G1 and an outcome P1 while leveraging other available phenotypes not on the pathway from G1 to P1 (P2, P3, P4). Conversely it is more difficult to leverage variables downstream the outcome considered as correlation might be due to a causal relationship and not by shared risk factors. For example in (C), intermediate variable such as M1, might be on the path from G1 to P1, and therefore should be excluded *a priori*, when applying the *CMS* approach.

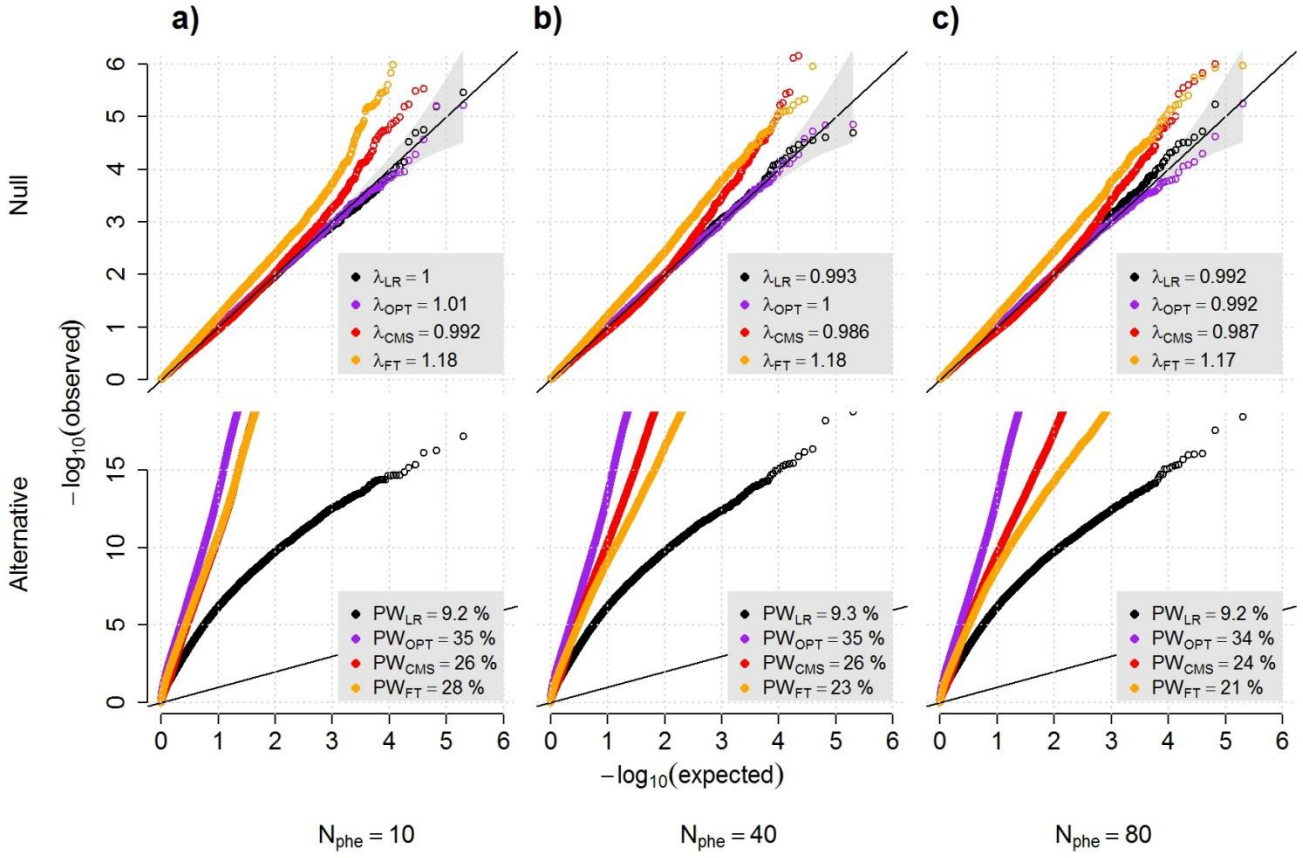


Supplementary Figure 14. Inference of missing values

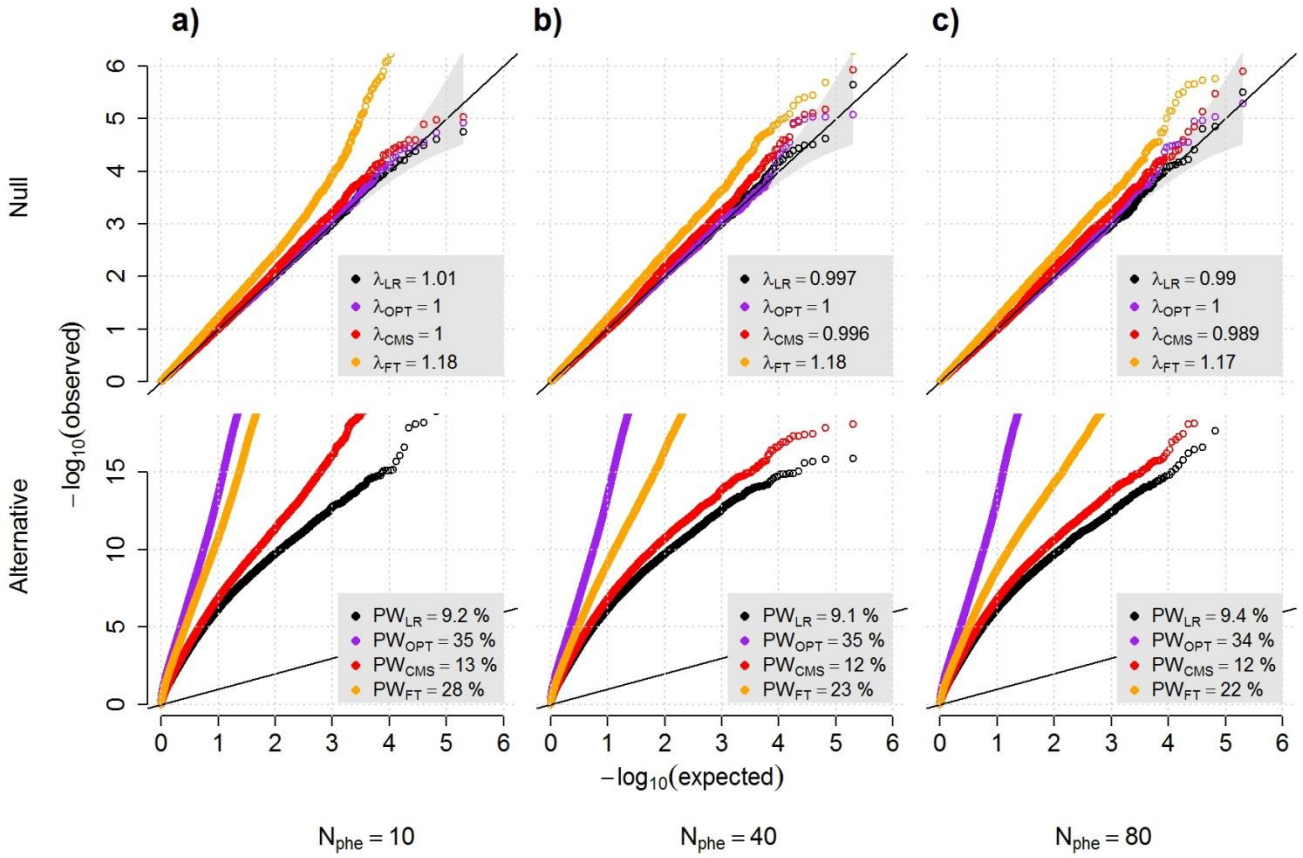
We generated a series of 10,000 replicates including 2,000 individuals. For each individual we generated independently a single predictor X, and 50 variables from a multivariate normal distribution with covariance matrix defined so that the pairwise correlation varies in [-0.4, 0.4]. We randomly defined one of those variables as the primary outcome and tested it for association with the predictor, and treated the remaining variables as candidate covariates. To explore the impact of missing values on *CMS*, we set either 0% (a), 25% (b and d), or 50% (c and e) of the covariates values as missing and performed a mean-imputation for the missing values. We considered unstructured missing values (b and c), or structured missing values, respectively. For the unstructured missingness, we randomly choose missing values, while for the structure missingness; we arbitrarily define a threshold and set as missing all values being either above or below the threshold. The left panel shows an example of the resulting distributions after imputation. The right panels show the QQplot from the *CMS* approach applied on the imputed data.



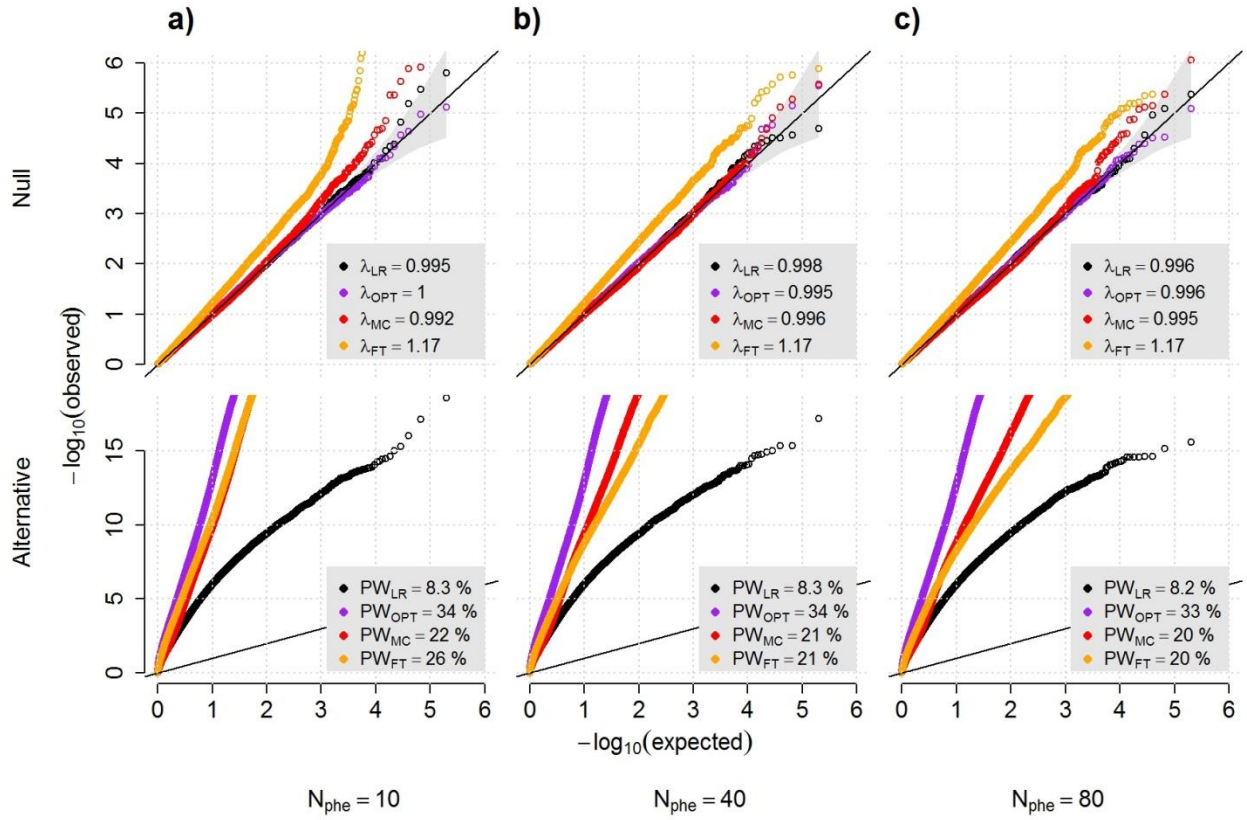
Supplementary Figure 15. Impact of decreasing the transition point in the CMS algorithm.
We simulated data similarly to Figure 4, except we modified *CMS* so that the transition point for which we start down-weighting the conditional δ interval is two-fold smaller. We simulated series of 100,000 datasets including 10 (a), 40 (b) and 80 (c) outcomes under a null model (upper panels), where a predictor of interest is not associated with a primary outcome but is associated with either 0%, 15% or 35% of the other outcomes with probability 0.75, 0.2 and 0.05 respectively, and under the alternative (lower panels), where the predictor is associated with the primary outcome only. The variance of the primary outcome that can be explained by the other outcomes was randomly chosen from [25%, 50%, 75%] with equal probability. In each replicate we applied four tests of association between the primary outcome and the predictor: a standard marginal univariate test (LR); the optimally adjusted test (OPT) that includes as covariates only the outcomes not associated with the predictor ; the *CMS* approach ; and a univariate test that include as covariate all outcomes with a *p*-value for association with the predictor above 0.1 (FT). For the null models we derived the genomic inflation factor λ_{GC}, while for the alternative model we estimated power at an α threshold of 5×10^{-7}, to correct for 100,000 tests.



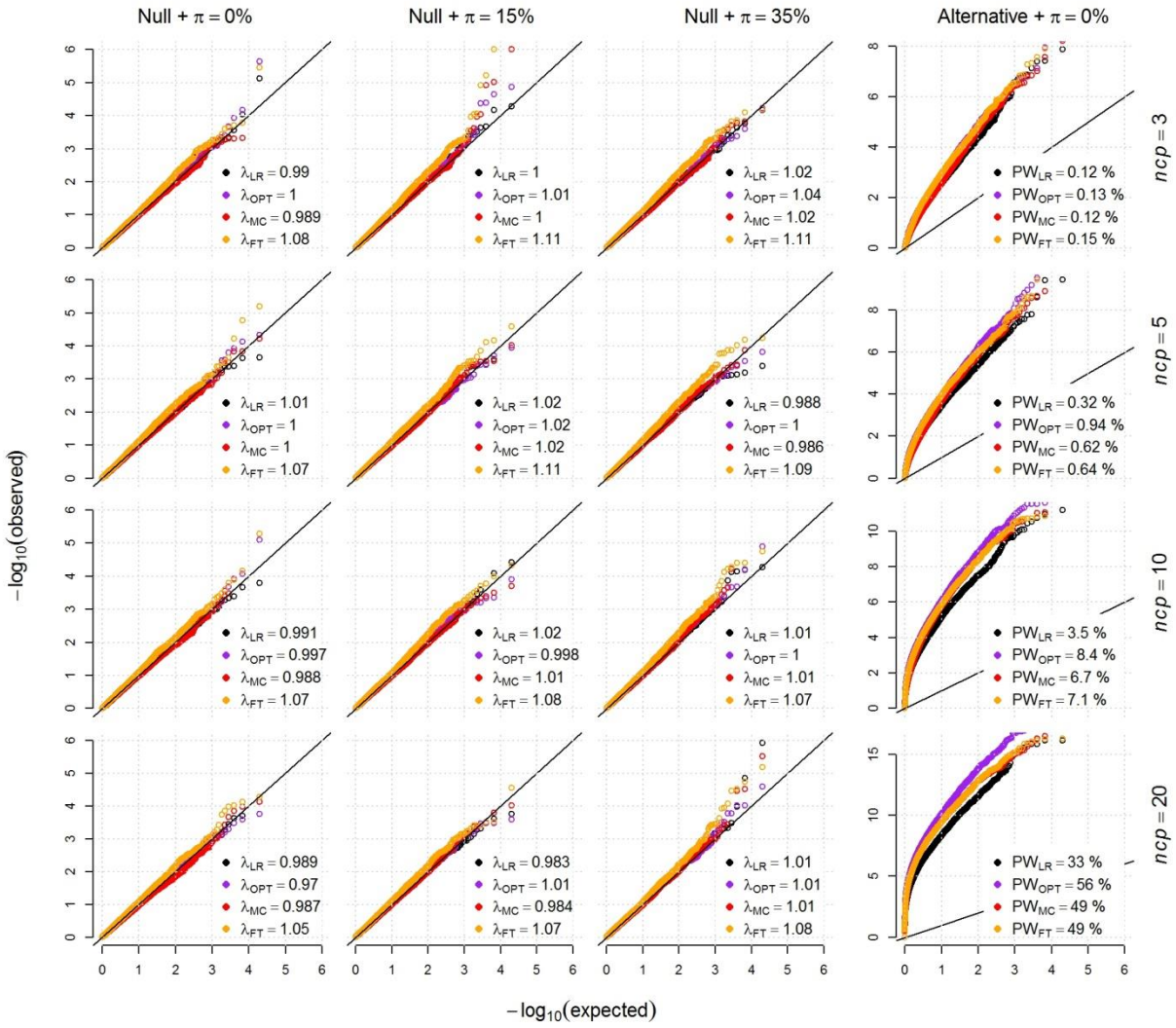
Supplementary Figure 16. Impact of shrinking the inclusion area.
We simulated data similarly to Figure 4, except we modified *CMS* so that the maximum conditional and unconditional δ interval equals $\sigma/2$ instead of our suggestion of 2σ, where σ corresponds to the standard error of either the conditional or unconditional distribution. We simulated series of 100,000 datasets including 10 (a), 40 (b) and 80 (c) outcomes under a null model (upper panels), where a predictor of interest is not associated with a primary outcome but is associated with either 0%, 15% or 35% of the other outcomes with probability 0.75, 0.2 and 0.05 respectively, and under the alternative (lower panels), where the predictor is associated with the primary outcome only. The variance of the primary outcome that can be explained by the other outcomes was randomly chosen from [25%, 50%, 75%] with equal probability. In each replicate we applied four tests of association between the primary outcome and the predictor: a standard marginal univariate test (LR); the optimally adjusted test (OPT) that includes as covariates only the outcomes not associated with the predictor ; the *CMS* approach ; and a univariate test that include as covariate all outcomes with a *p*-value for association with the predictor above 0.1 (FT). For the null models we derived the genomic inflation factor λ_{GC}, while for the alternative model we estimated power at an α threshold of $5\text{x}10^{-7}$, to correct for 100,000 tests.



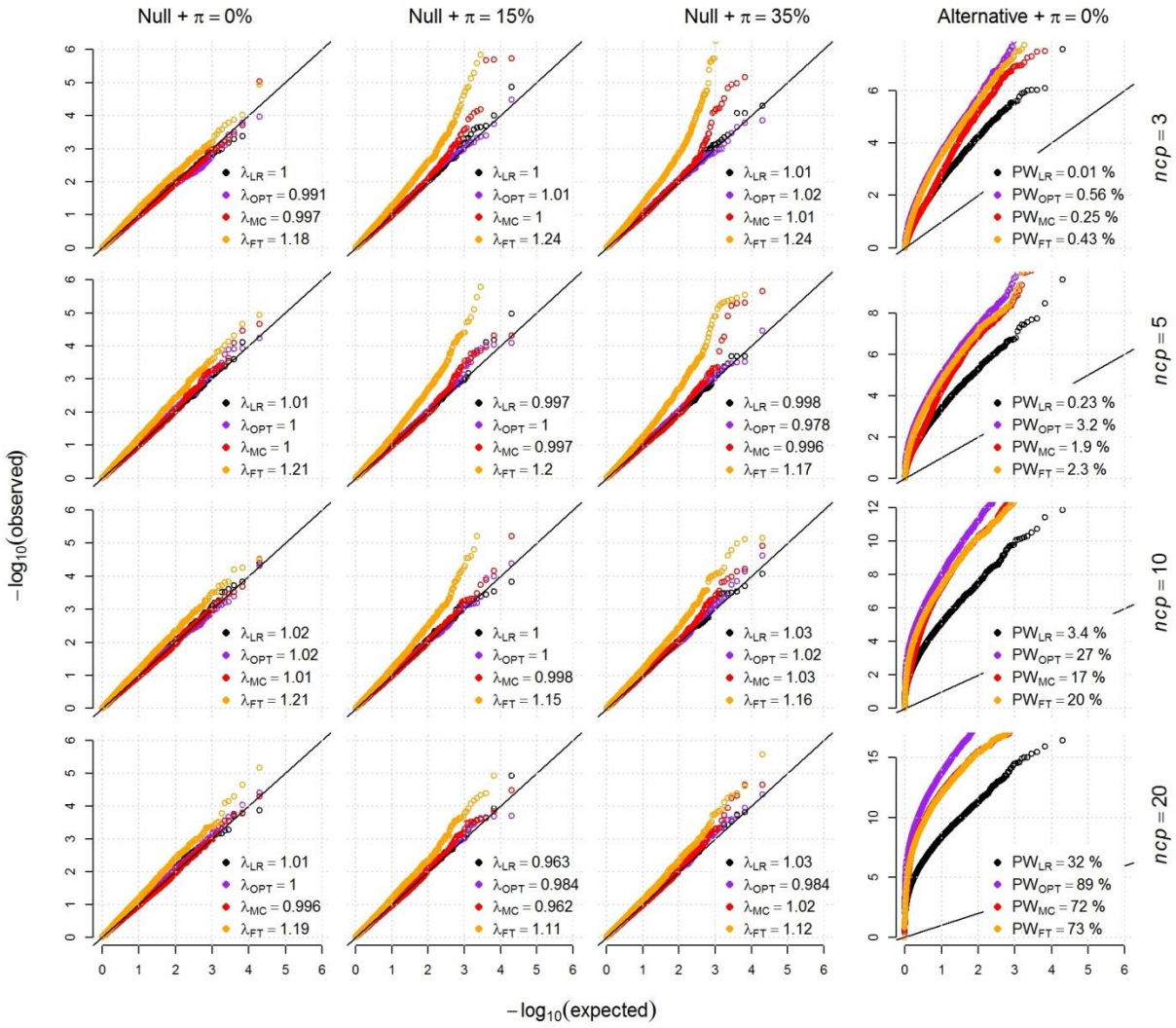
Supplementary Figure 17. Impact increasing the multivariate test threshold parameter.
We simulated data similarly to Figure 4, except we modified *CMS* so that t_{MUL}, the p-value threshold from the multivariate test equals 0.2, instead of 0.05, as in the final version of *CMS*. We simulated series of 100,000 datasets including 10 (a), 40 (b) and 80 (c) outcomes under a null model (upper panels), where a predictor of interest is not associated with a primary outcome but is associated with either 0%, 15% or 35% of the other outcomes with probability 0.75, 0.2 and 0.05 respectively, and under the alternative (lower panels), where the predictor is associated with the primary outcome only. The variance of the primary outcome that can be explained by the other outcomes was randomly chosen from [25%, 50%, 75%] with equal probability. In each replicate we applied four tests of association between the primary outcome and the predictor: a standard marginal univariate test (LR); the optimally adjusted test (OPT) that includes as covariates only the outcomes not associated with the predictor ; the *CMS* approach ; and a univariate test that include as covariate all outcomes with a *p*-value for association with the predictor above 0.1 (FT). For the null models we derived the genomic inflation factor λ_{GC}, while for the alternative model we estimated power at an α threshold of 5×10^{-7}, to correct for 100,000 tests.



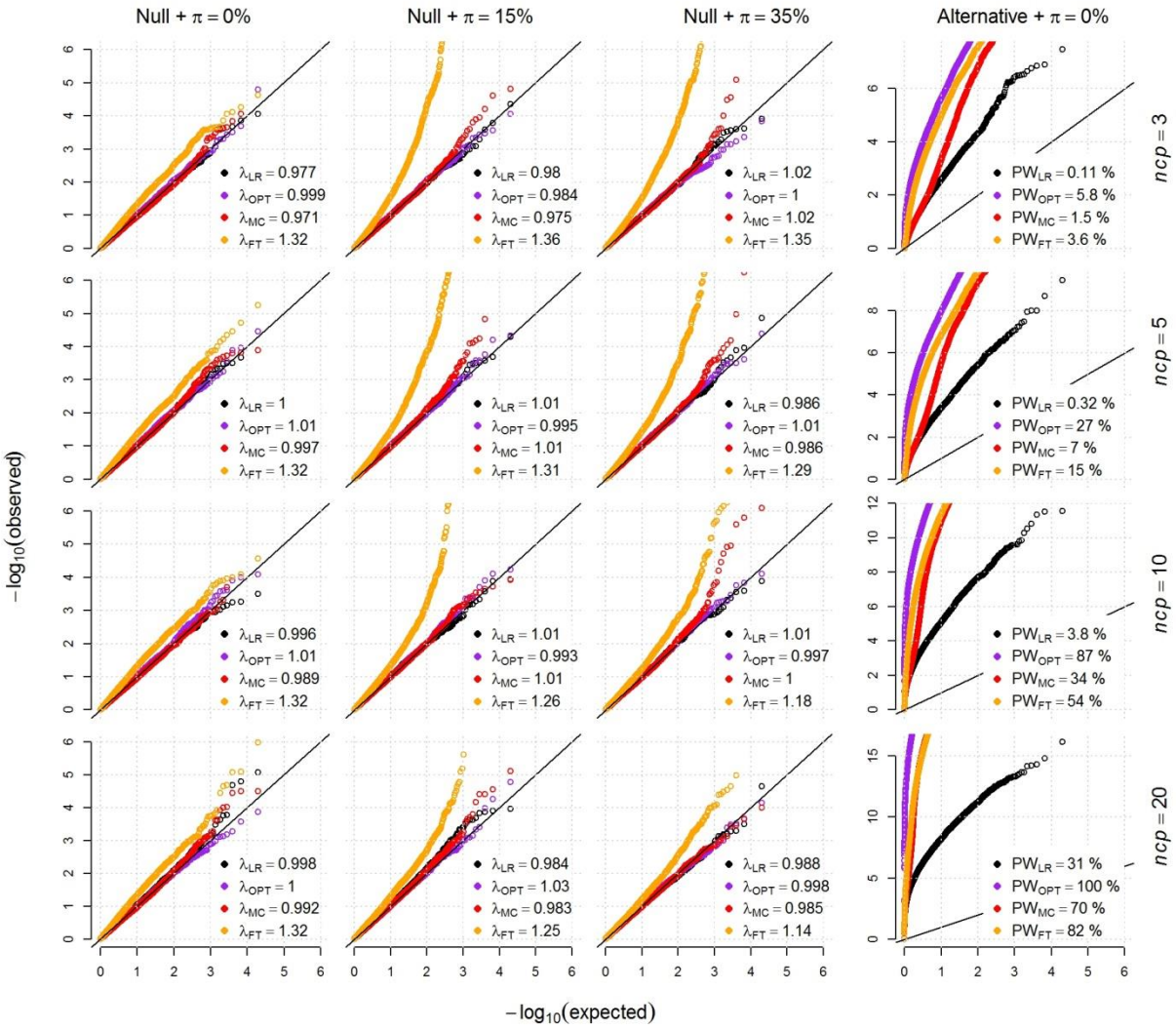
Supplementary Figure 18. ***CMS*** **QQplots for 10 phenotypes, 300 individuals and 25% of outcome variance explained**



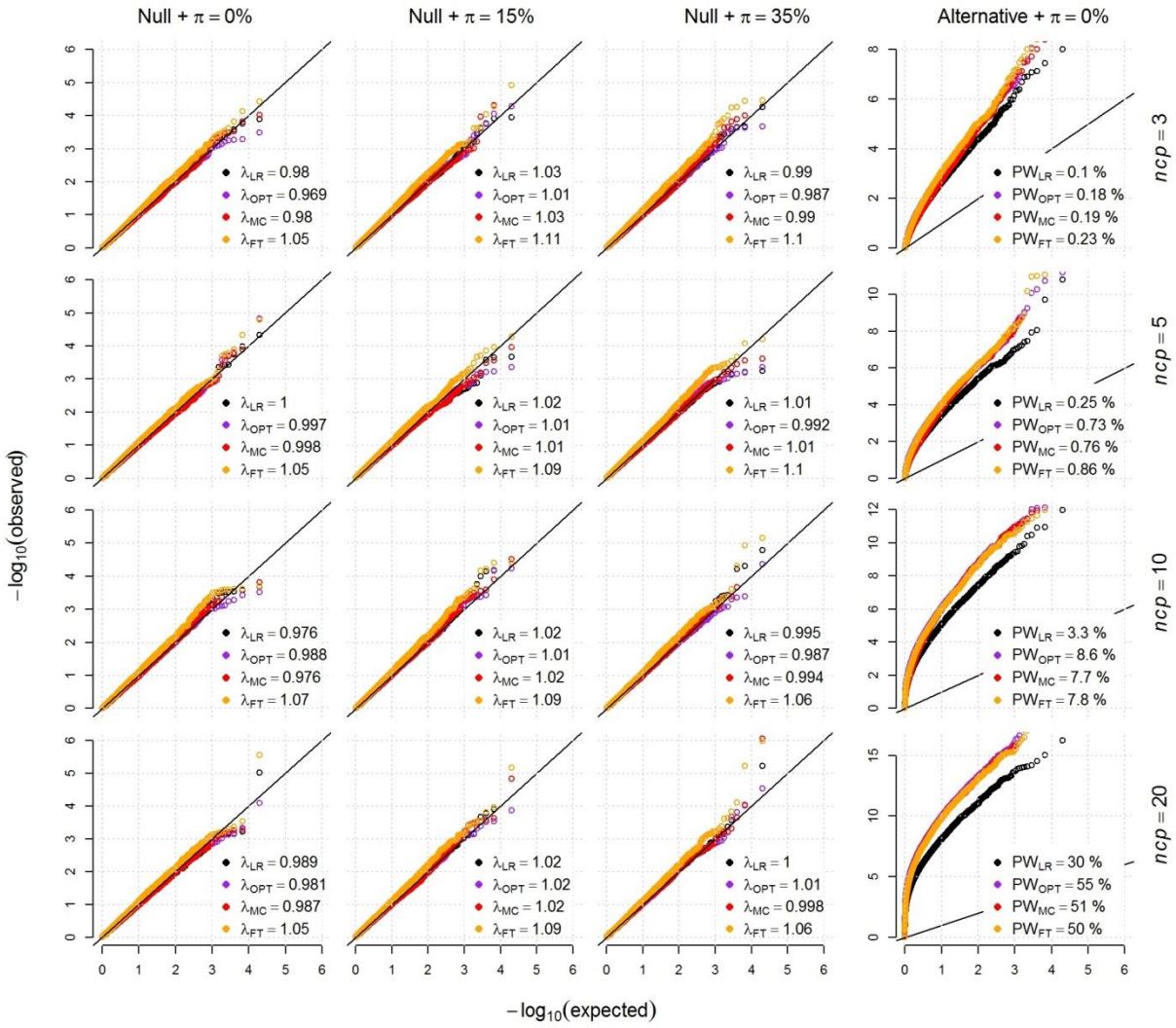
Supplementary Figure 19. ***CMS*** **QQplots for 10 phenotypes, 300 individuals and 50% of outcome variance explained**



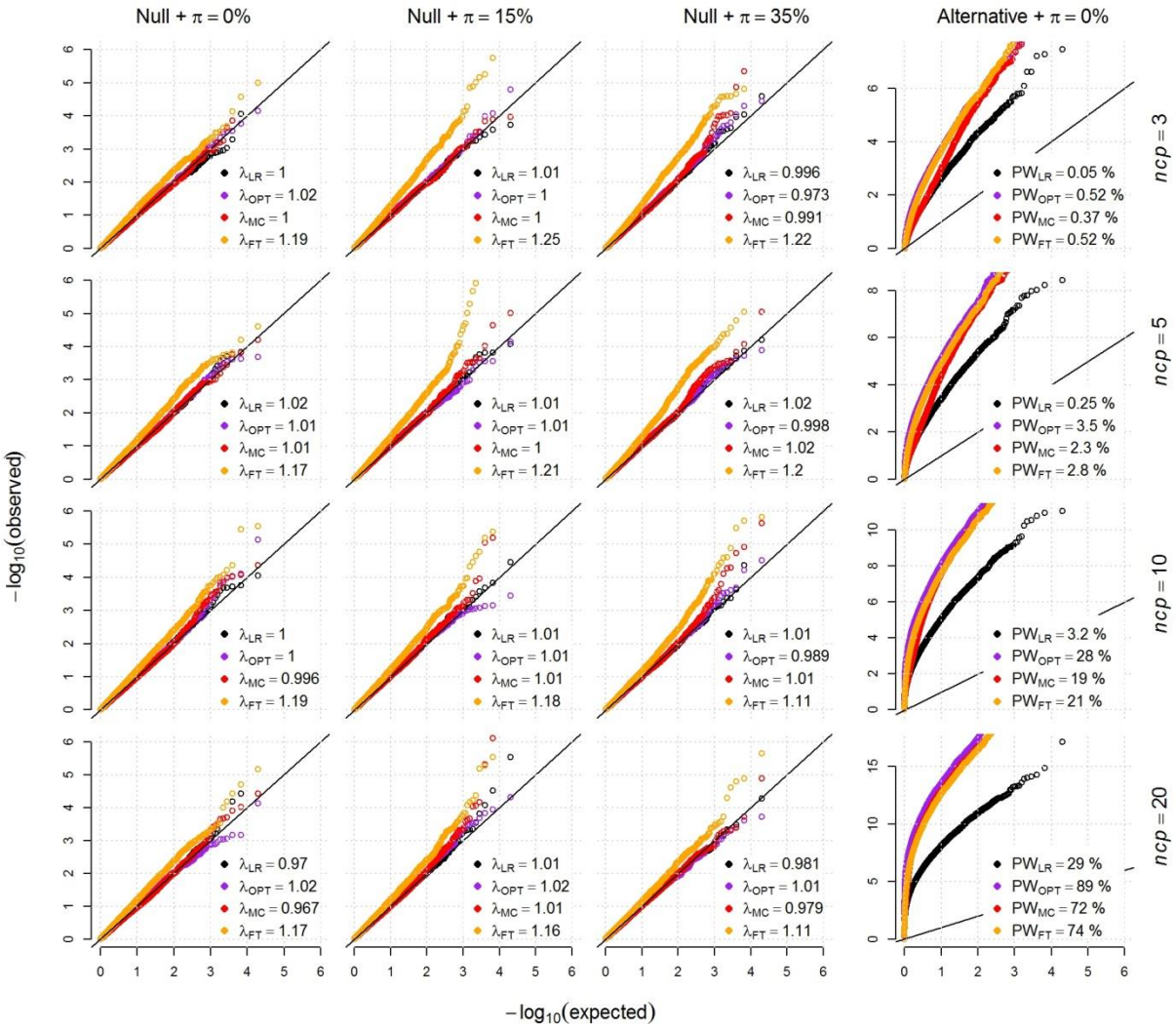
Supplementary Figure 20. ***CMS*** **QQplots for 10 phenotypes, 300 individuals and 75% of outcome variance explained**



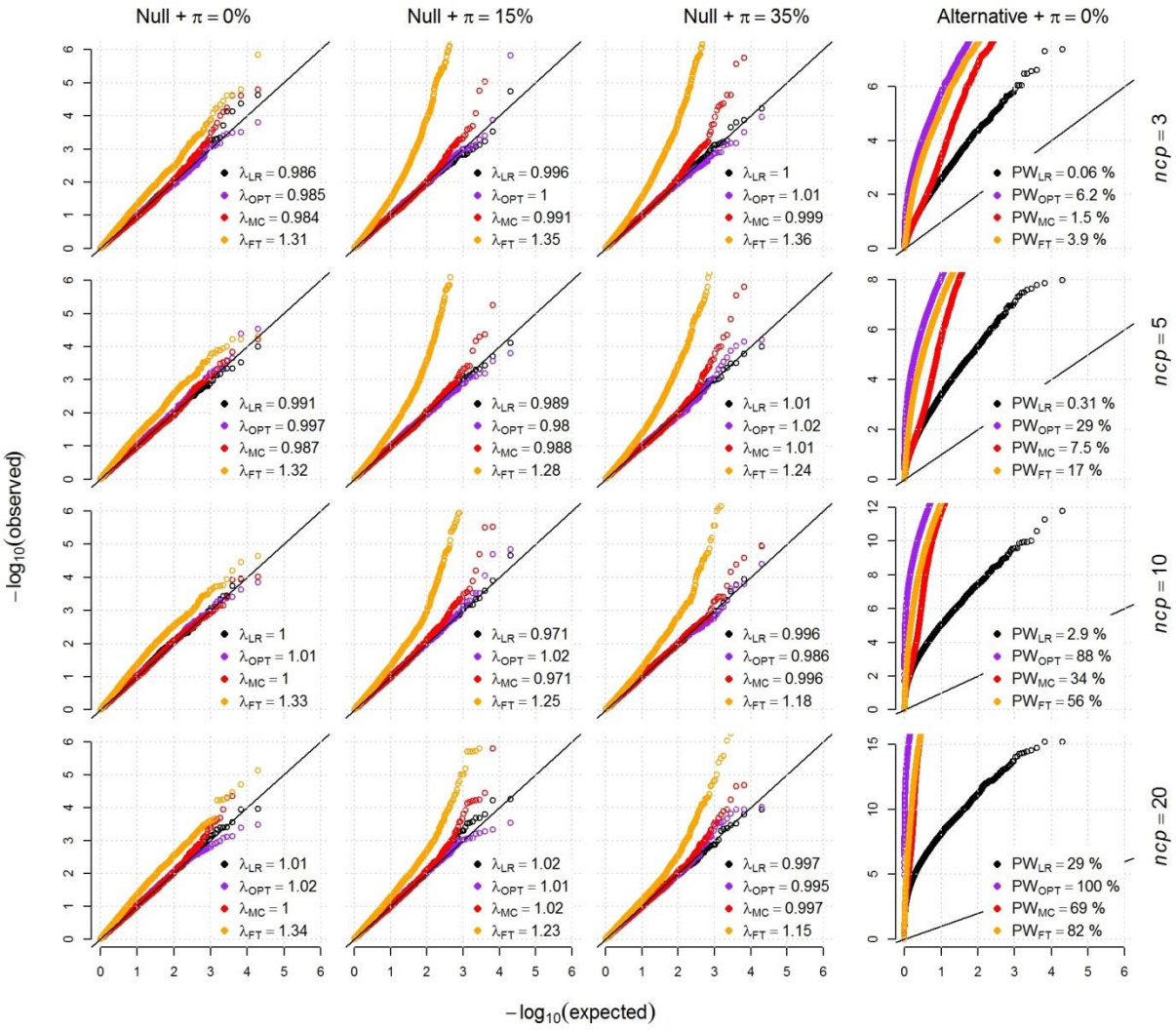
Supplementary Figure 21. ***CMS*** **QQplots for 10 phenotypes, 2000 individuals and 25% of outcome variance explained**



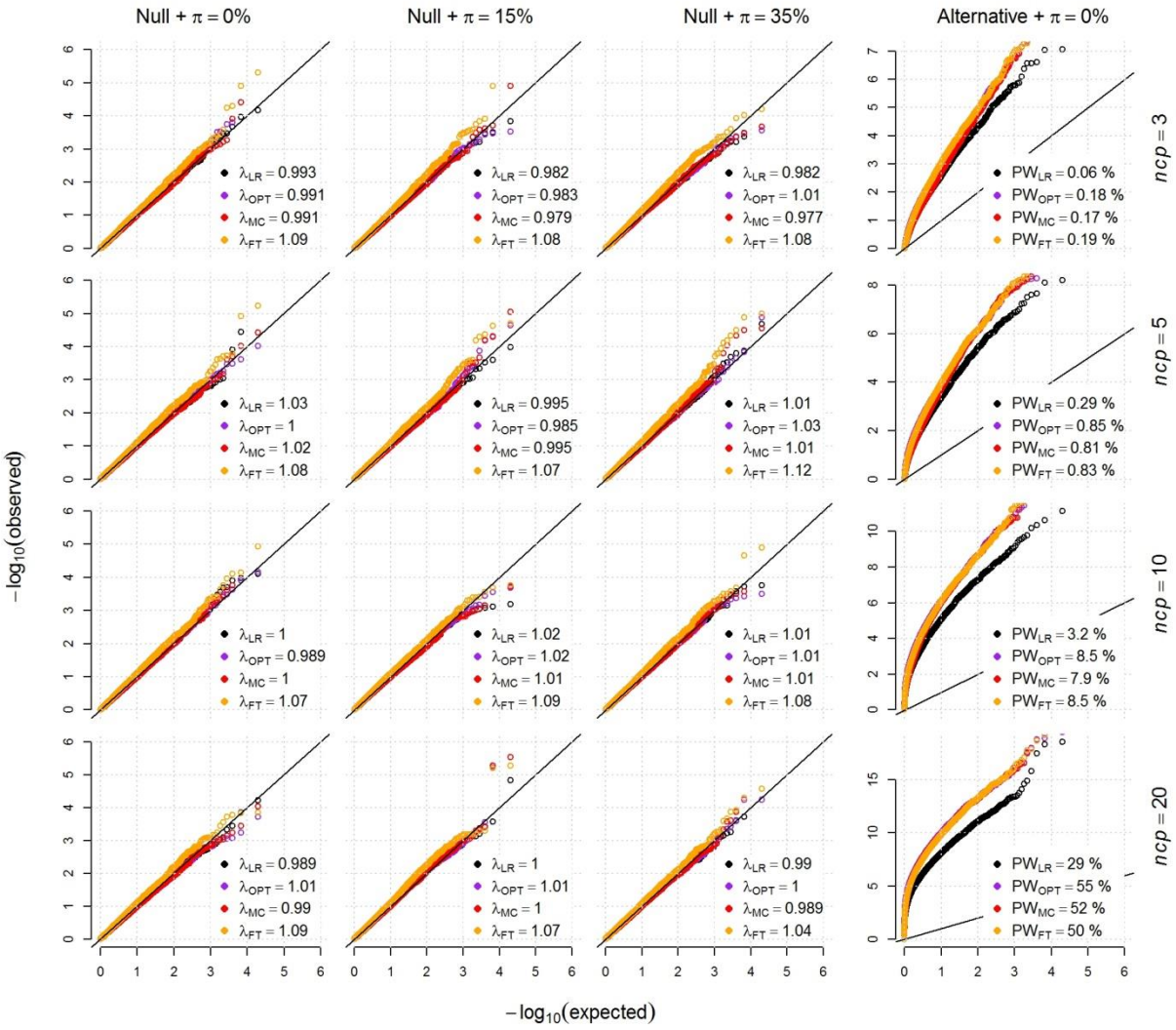
Supplementary Figure 22. ***CMS*** **QQplots for 10 phenotypes, 2000 individuals and 50% of outcome variance explained**



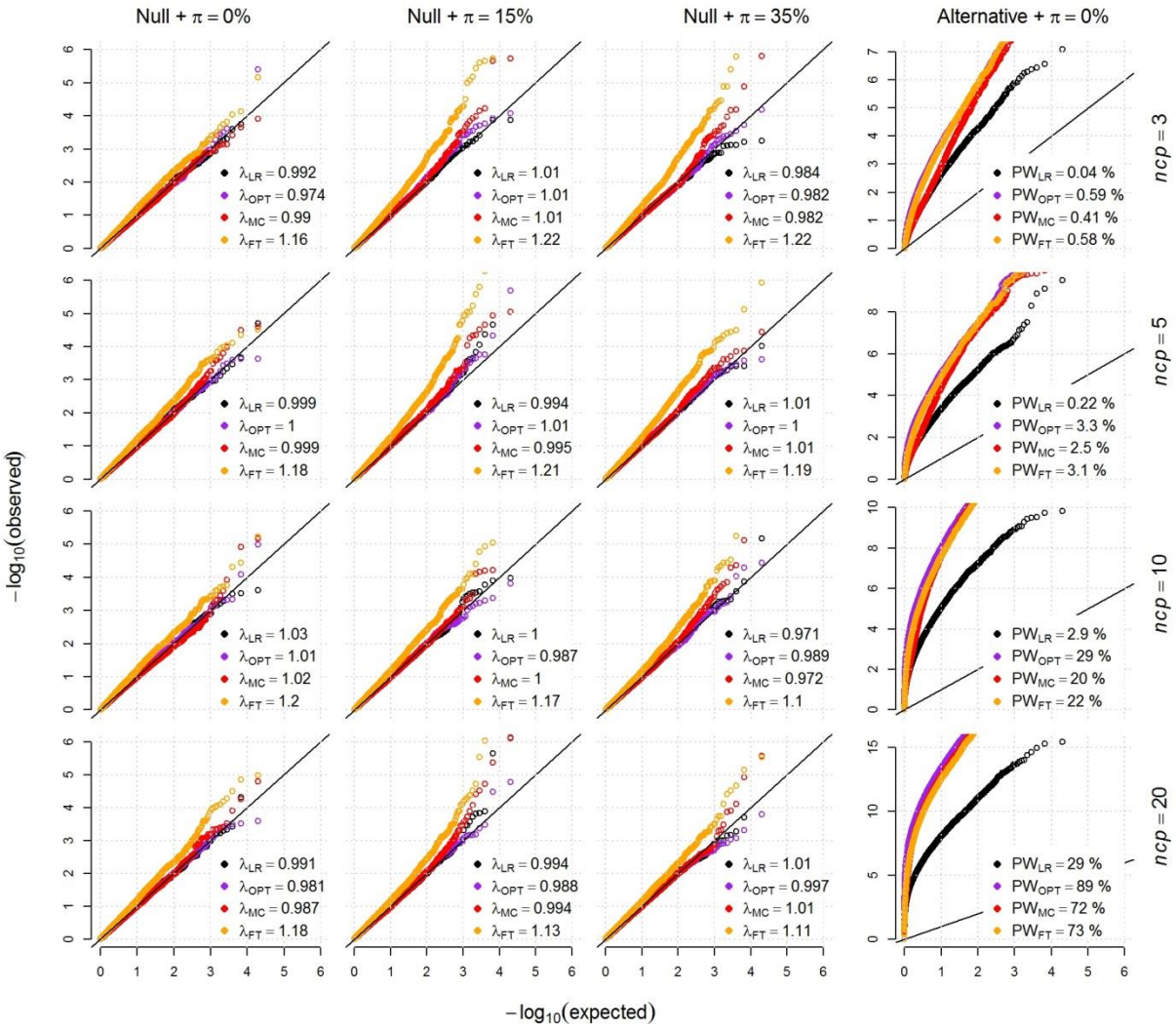
Supplementary Figure 23. ***CMS*** **QQplots for 10 phenotypes, 2000 individuals and 75% of outcome variance explained**



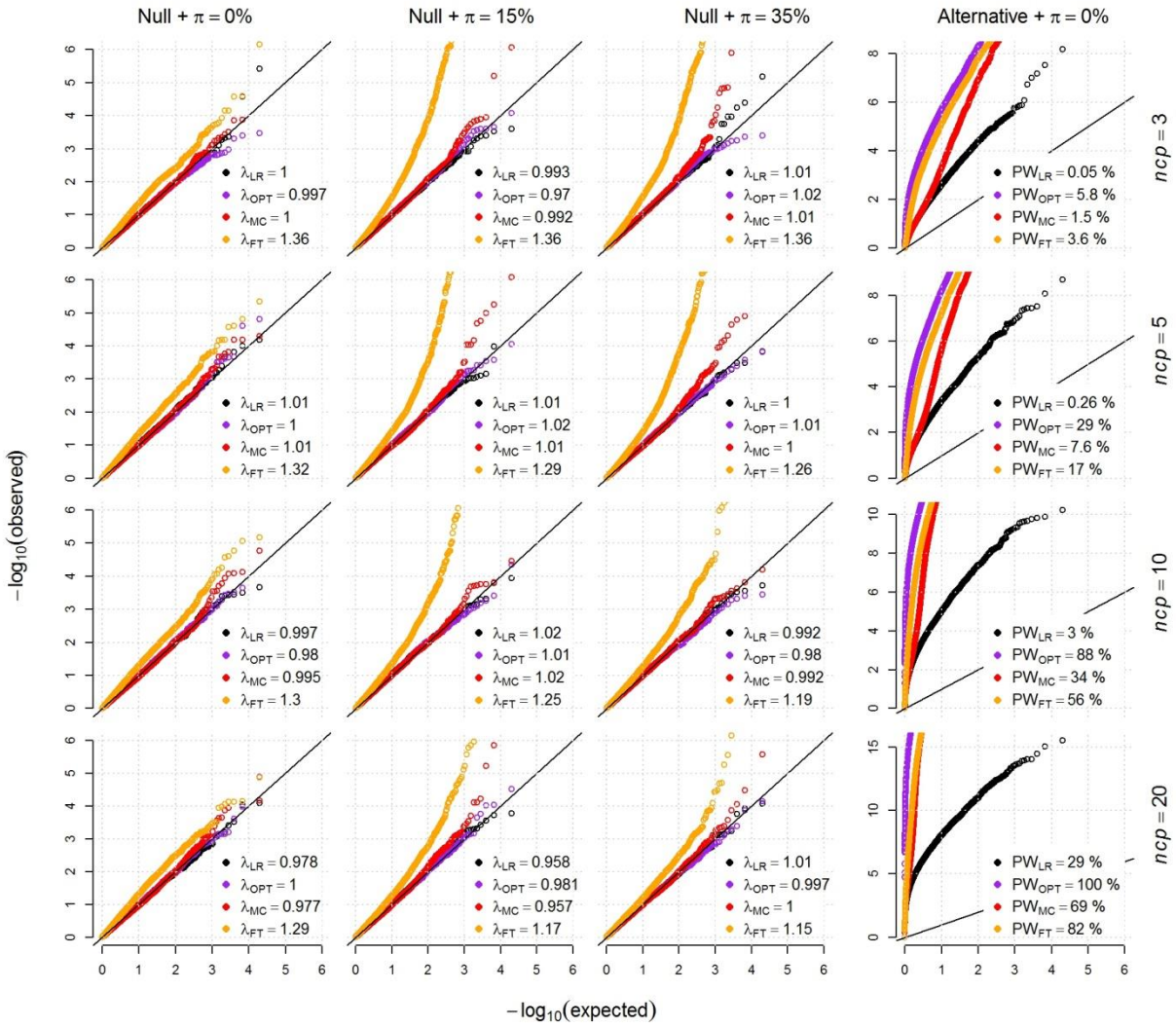
Supplementary Figure 24. ***CMS*** **QQplots for 10 phenotypes, 6000 individuals and 25% of outcome variance explained**



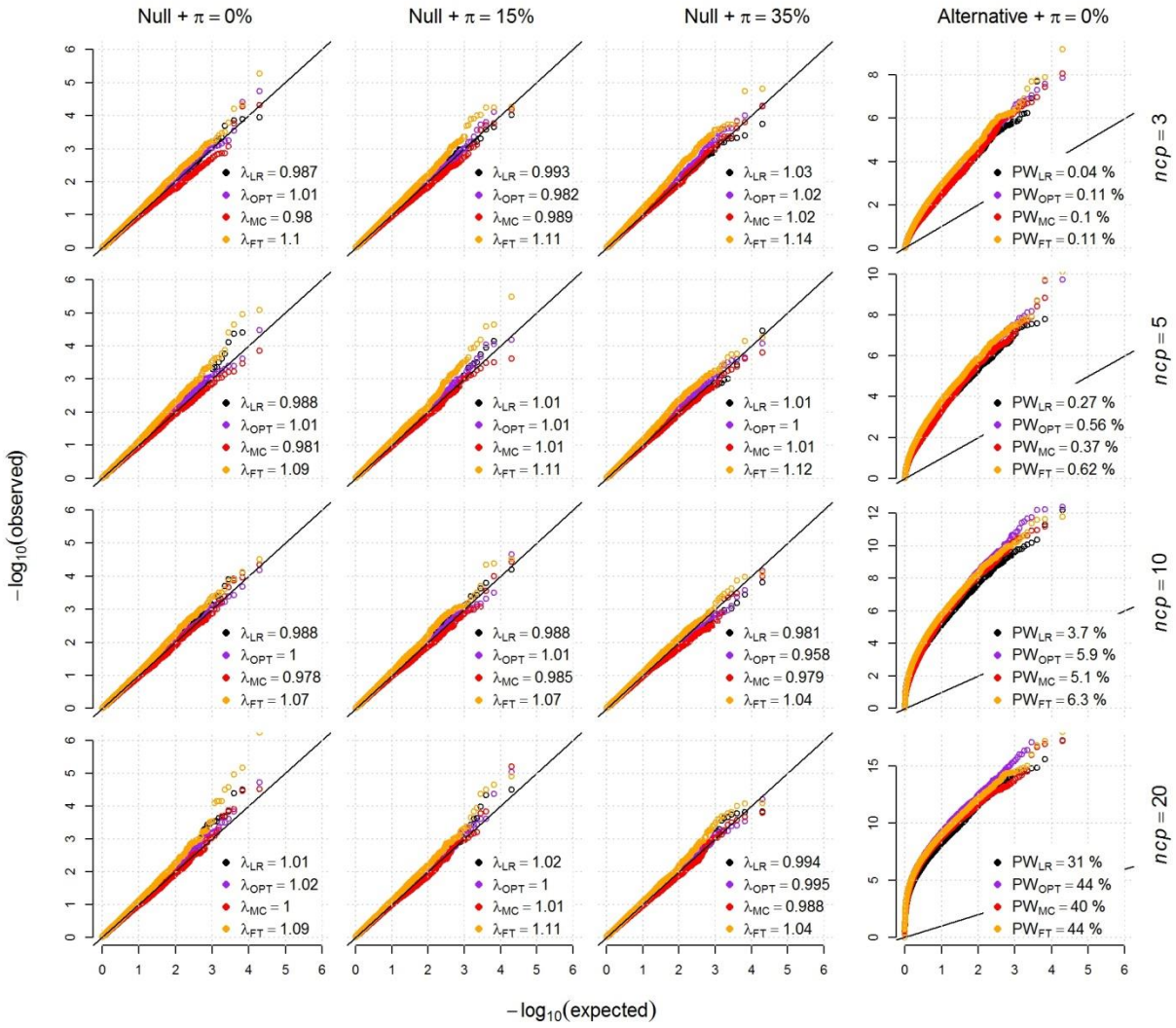
Supplementary Figure 25. ***CMS*** **QQplots for 10 phenotypes, 6000 individuals and 50% of outcome variance explained**



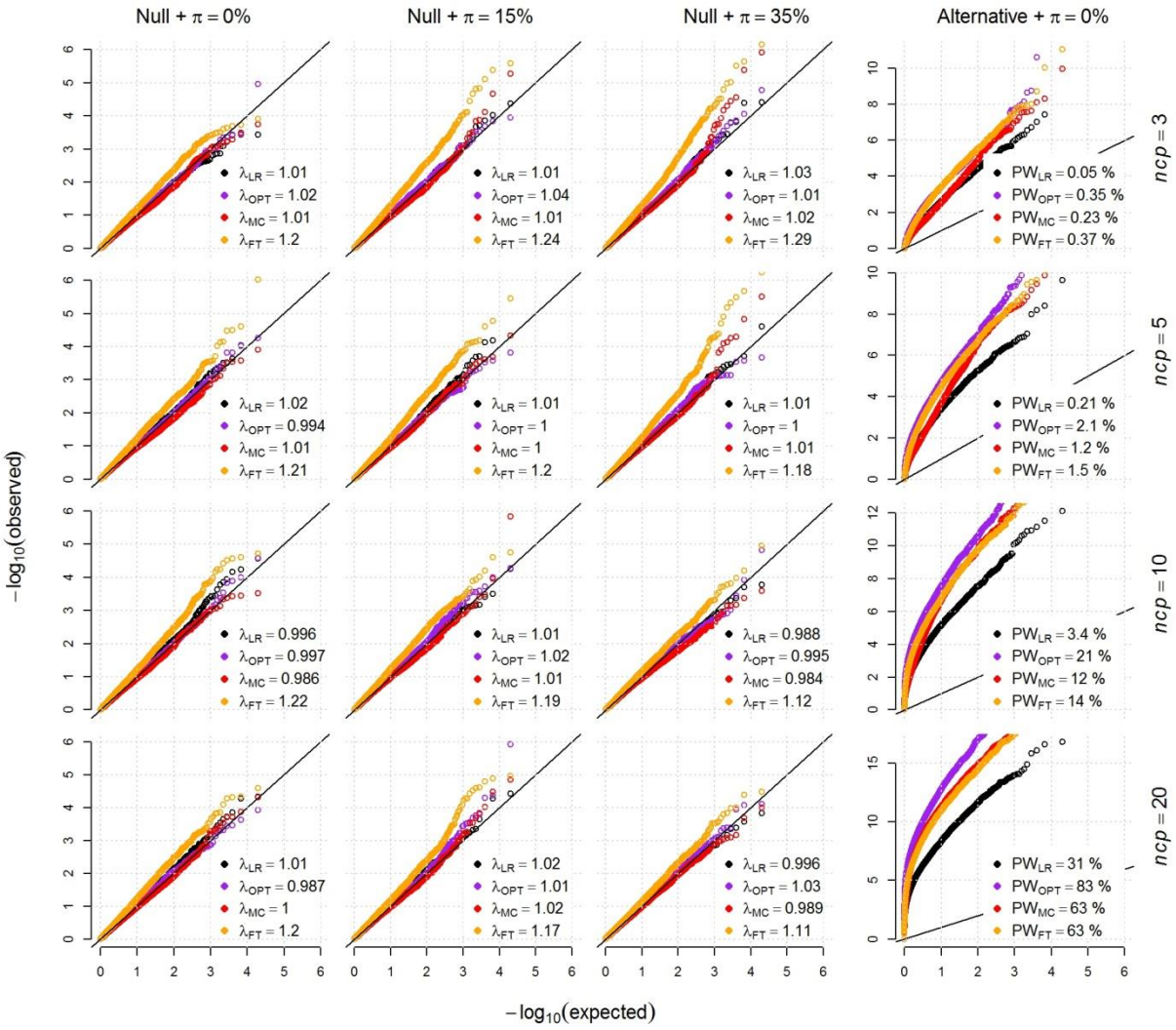
Supplementary Figure 26. ***CMS*** **QQplots for 10 phenotypes, 6000 individuals and 75% of outcome variance explained**



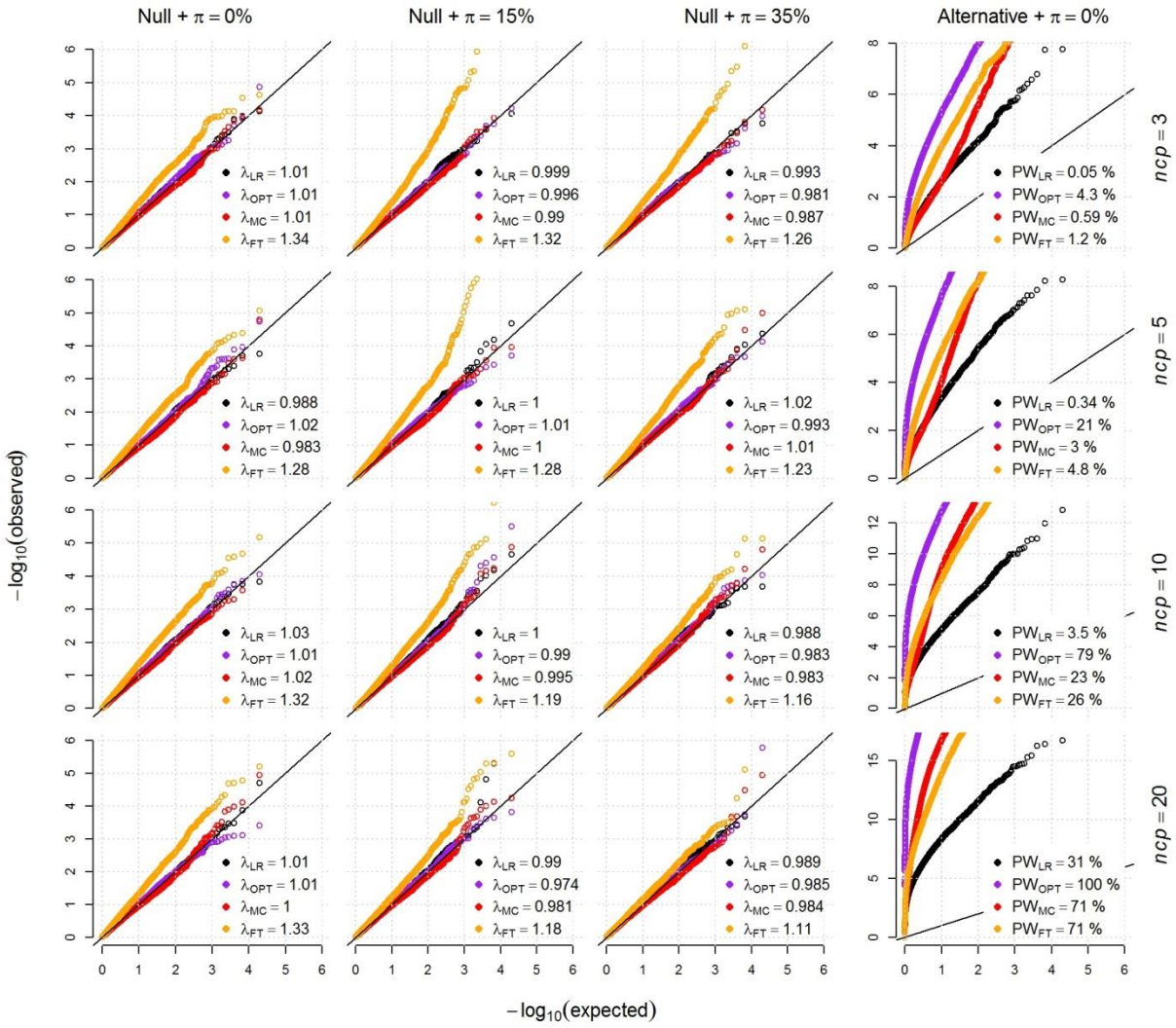
Supplementary Figure 27. ***CMS*** **QQplots for 40 phenotypes, 300 individuals and 25% of outcome variance explained**



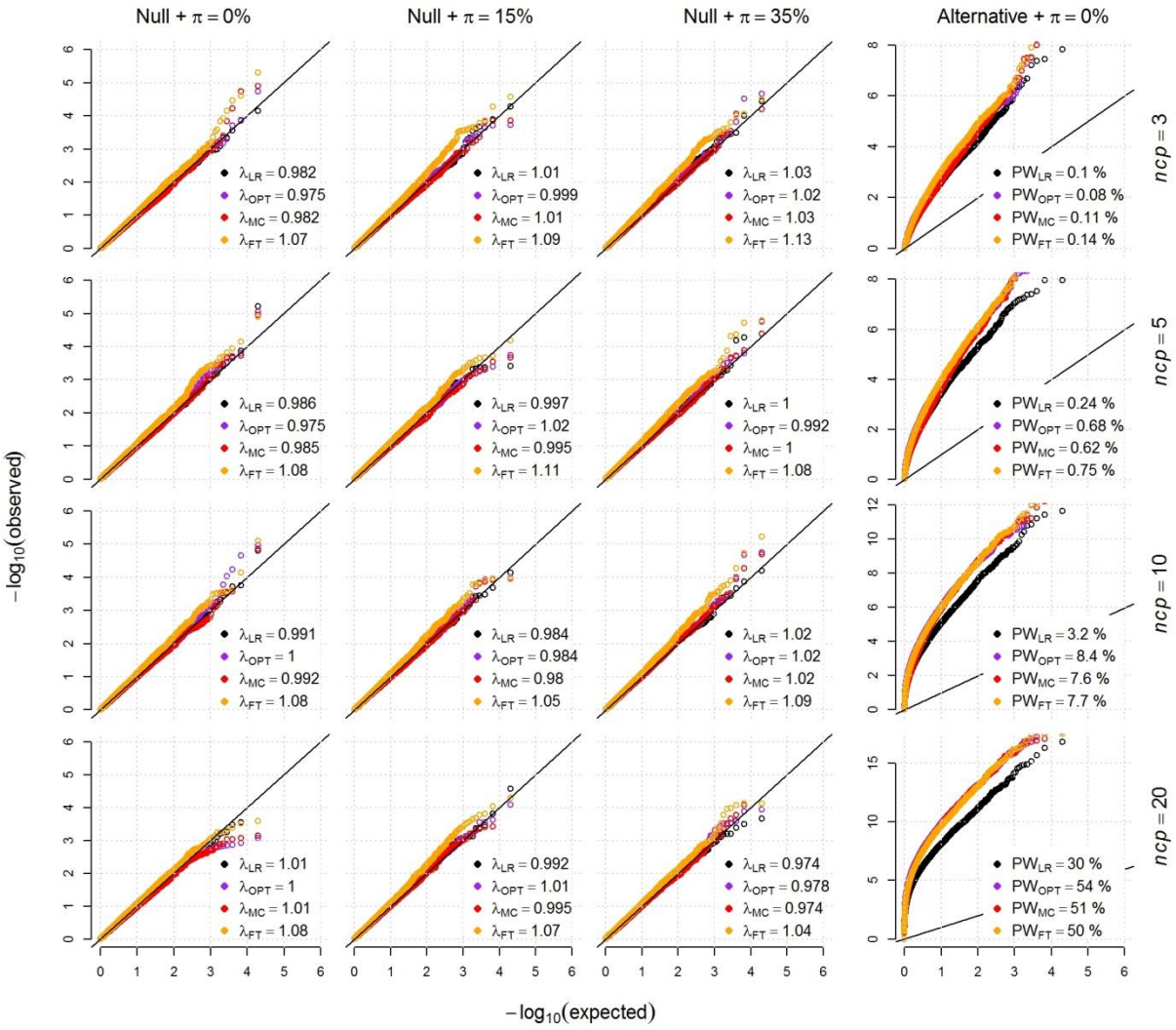
Supplementary Figure 28. ***CMS*** **QQplots for 40 phenotypes, 300 individuals and 50% of outcome variance explained**



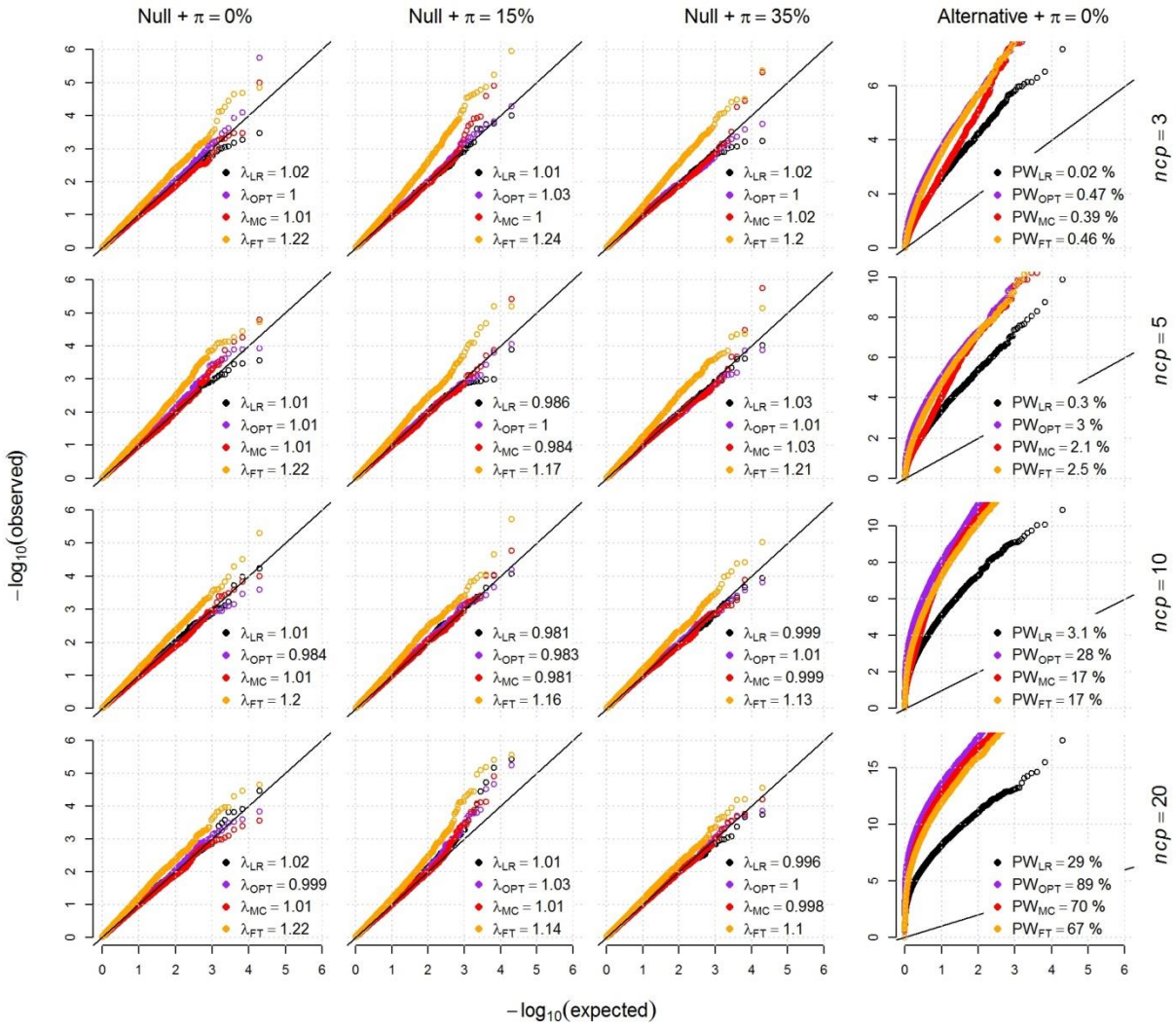
Supplementary Figure 29. ***CMS*** **QQplots for 40 phenotypes, 300 individuals and 75% of outcome variance explained**



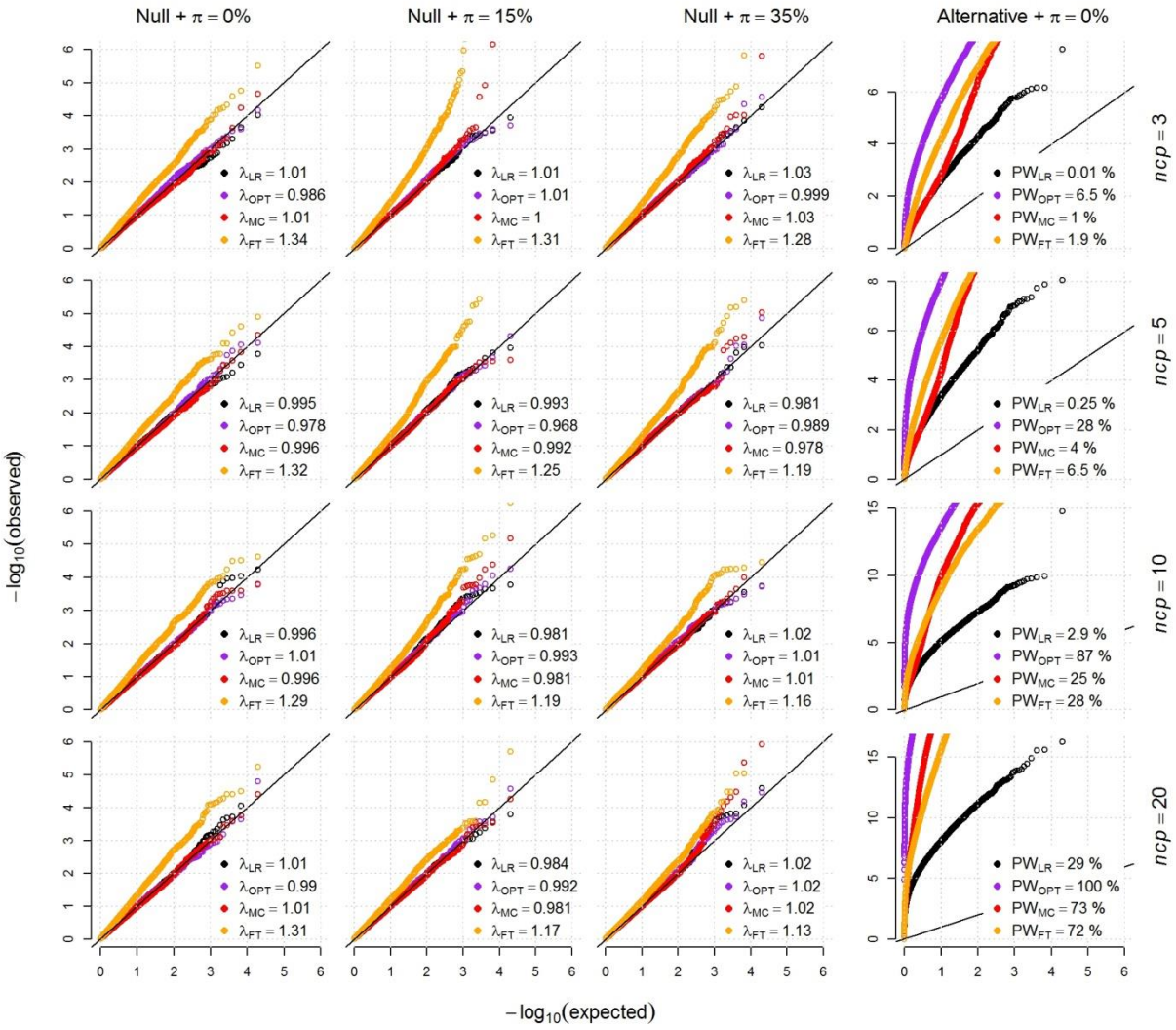
Supplementary Figure 30. ***CMS*** **QQplots for 40 phenotypes, 2000 individuals and 25% of outcome variance explained**



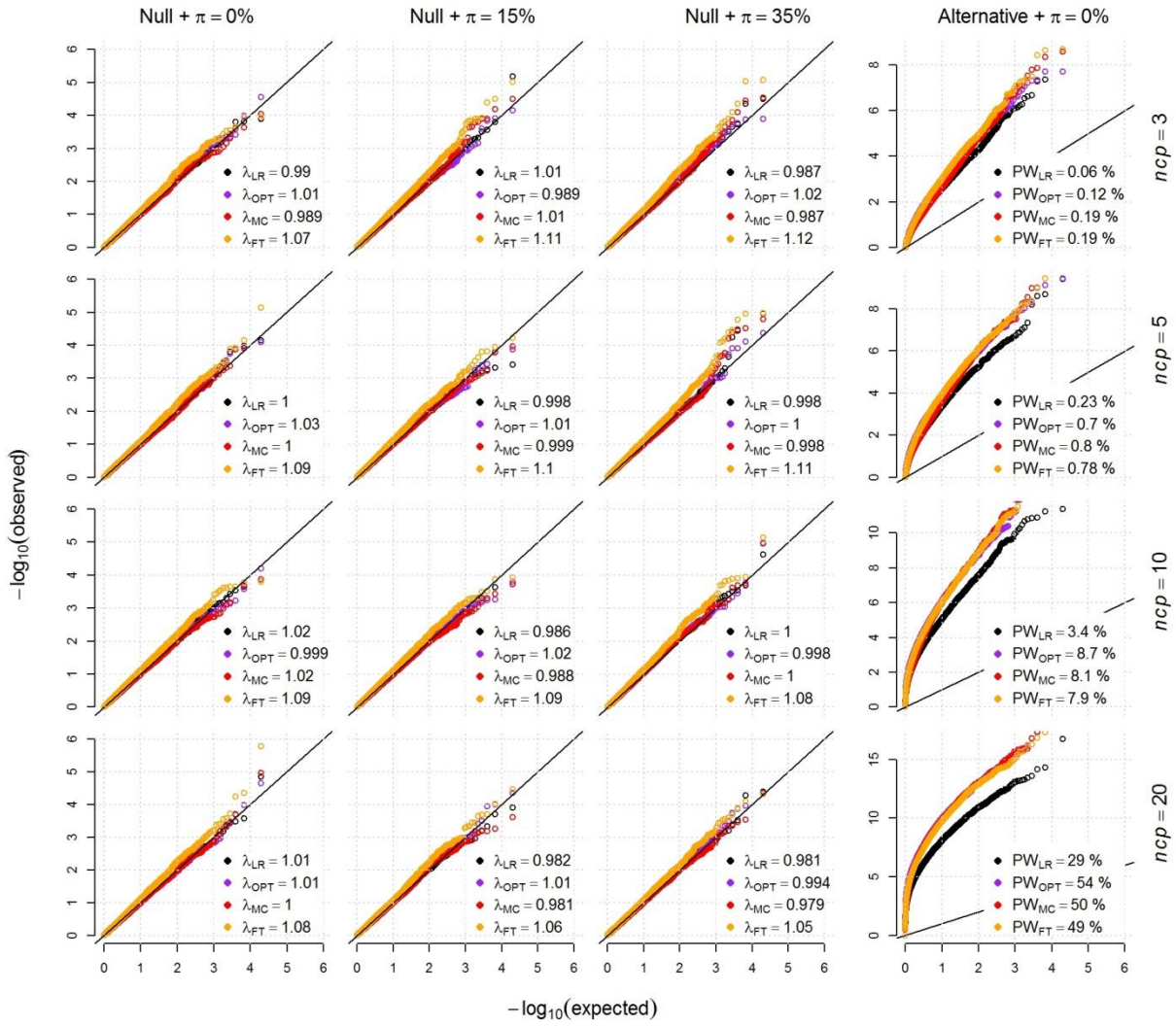
Supplementary Figure 31. ***CMS*** **QQplots for 40 phenotypes, 2000 individuals and 50% of outcome variance explained**



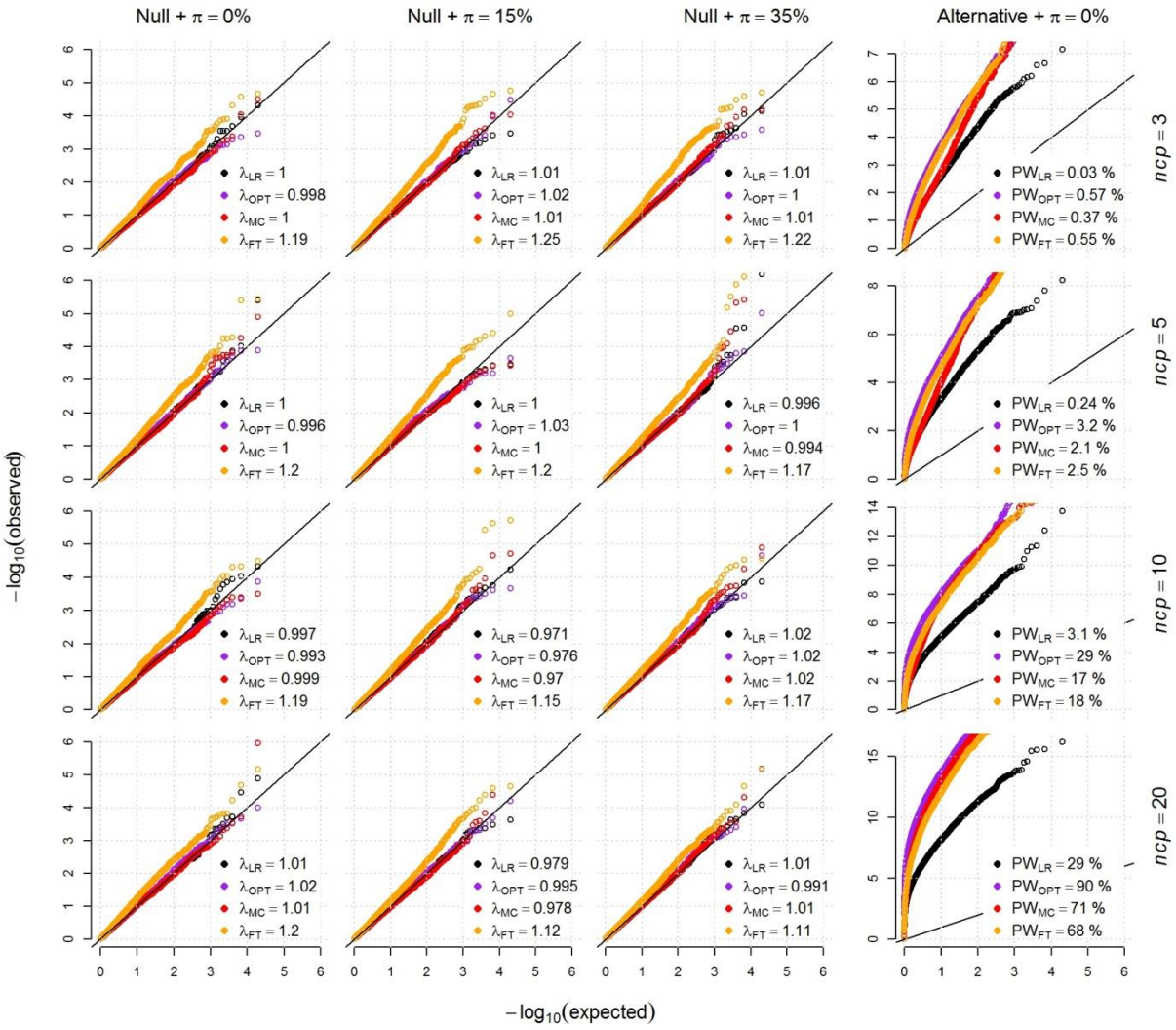
Supplementary Figure 32. ***CMS*** **QQplots for 40 phenotypes, 2000 individuals and 75% of outcome variance explained**



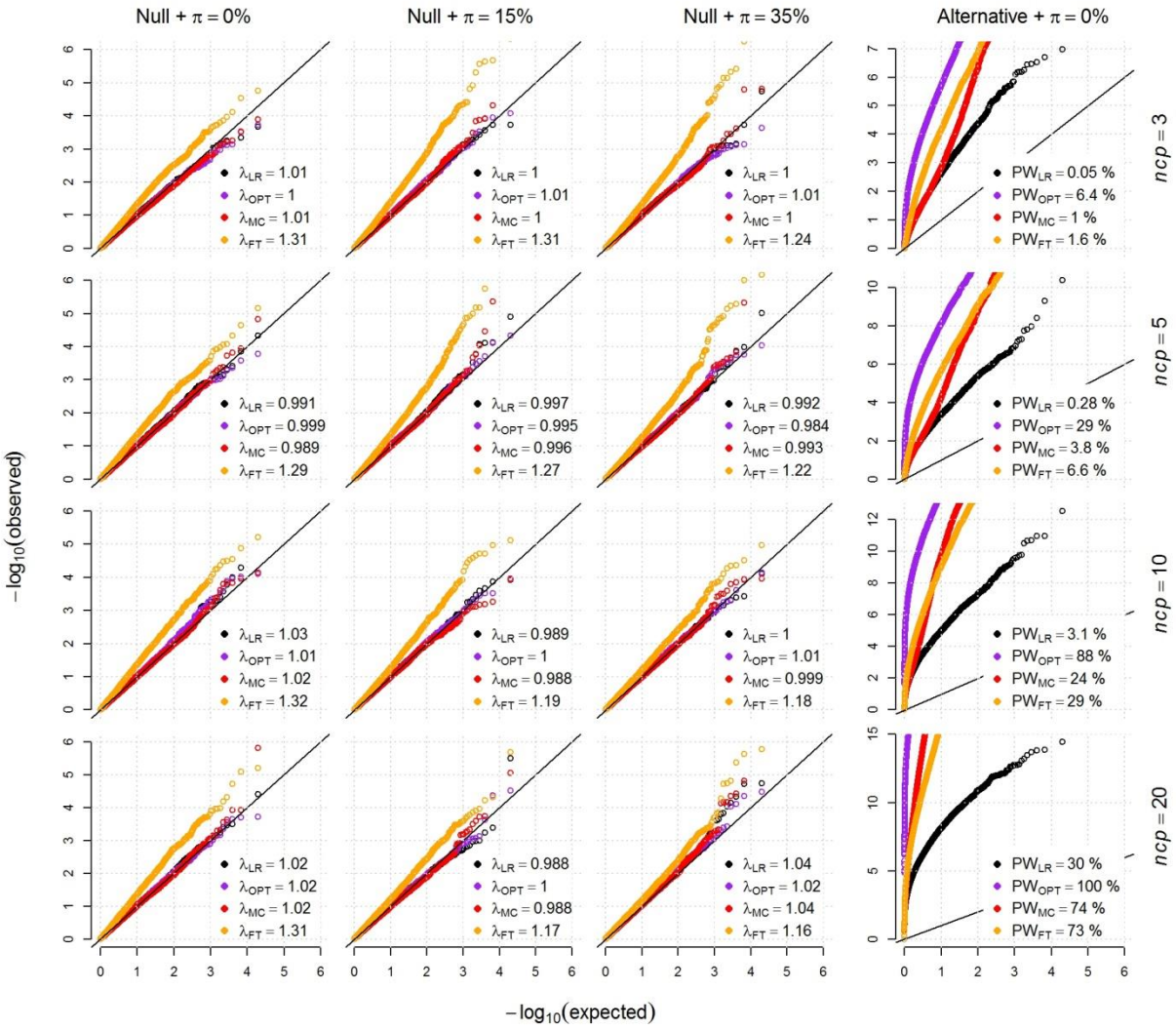
Supplementary Figure 33. ***CMS*** **QQplots for 40 phenotypes, 6000 individuals and 25% of outcome variance explained**



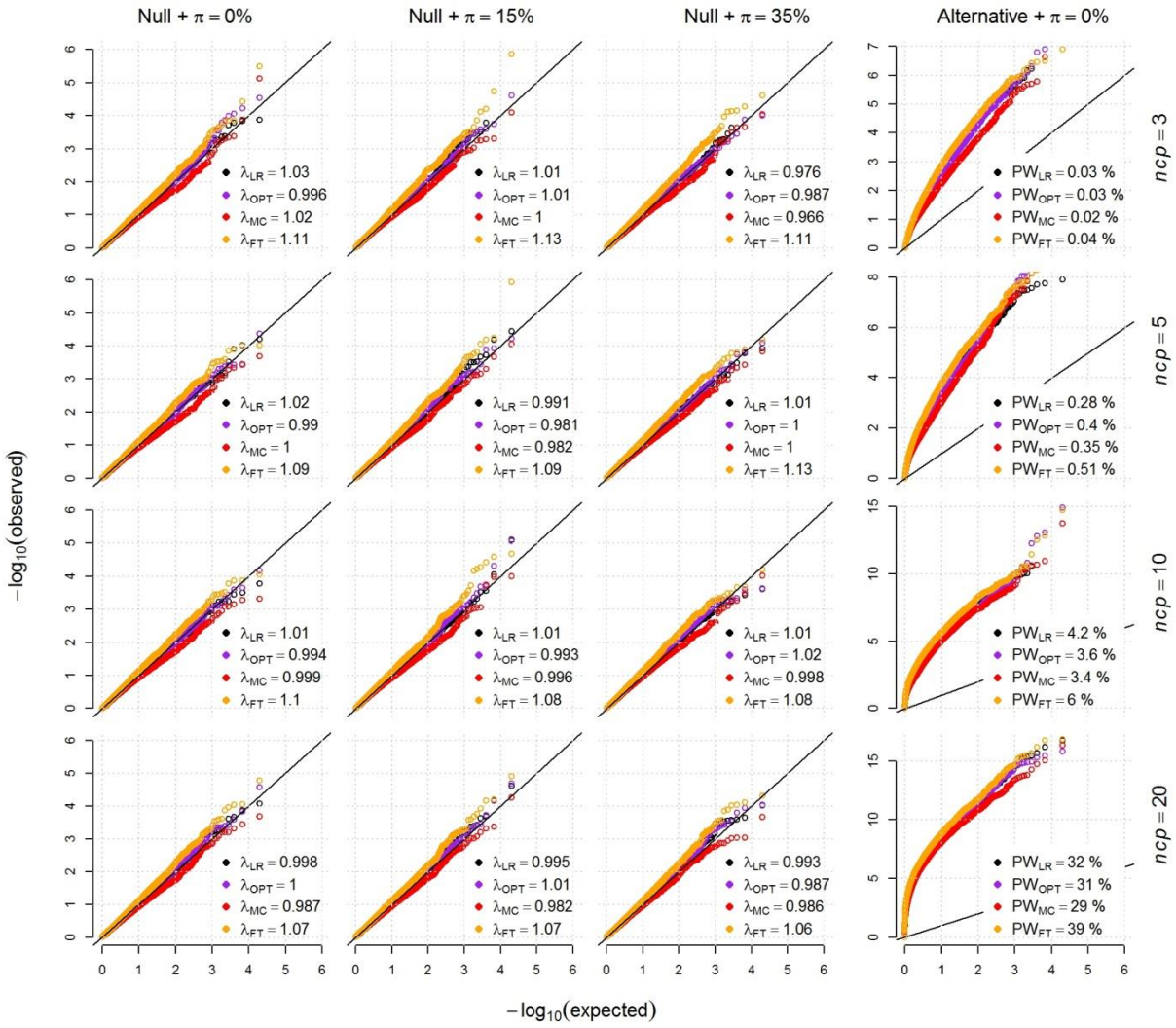
Supplementary Figure 34. ***CMS*** **QQplots for 40 phenotypes, 6000 individuals and 50% of outcome variance explained**



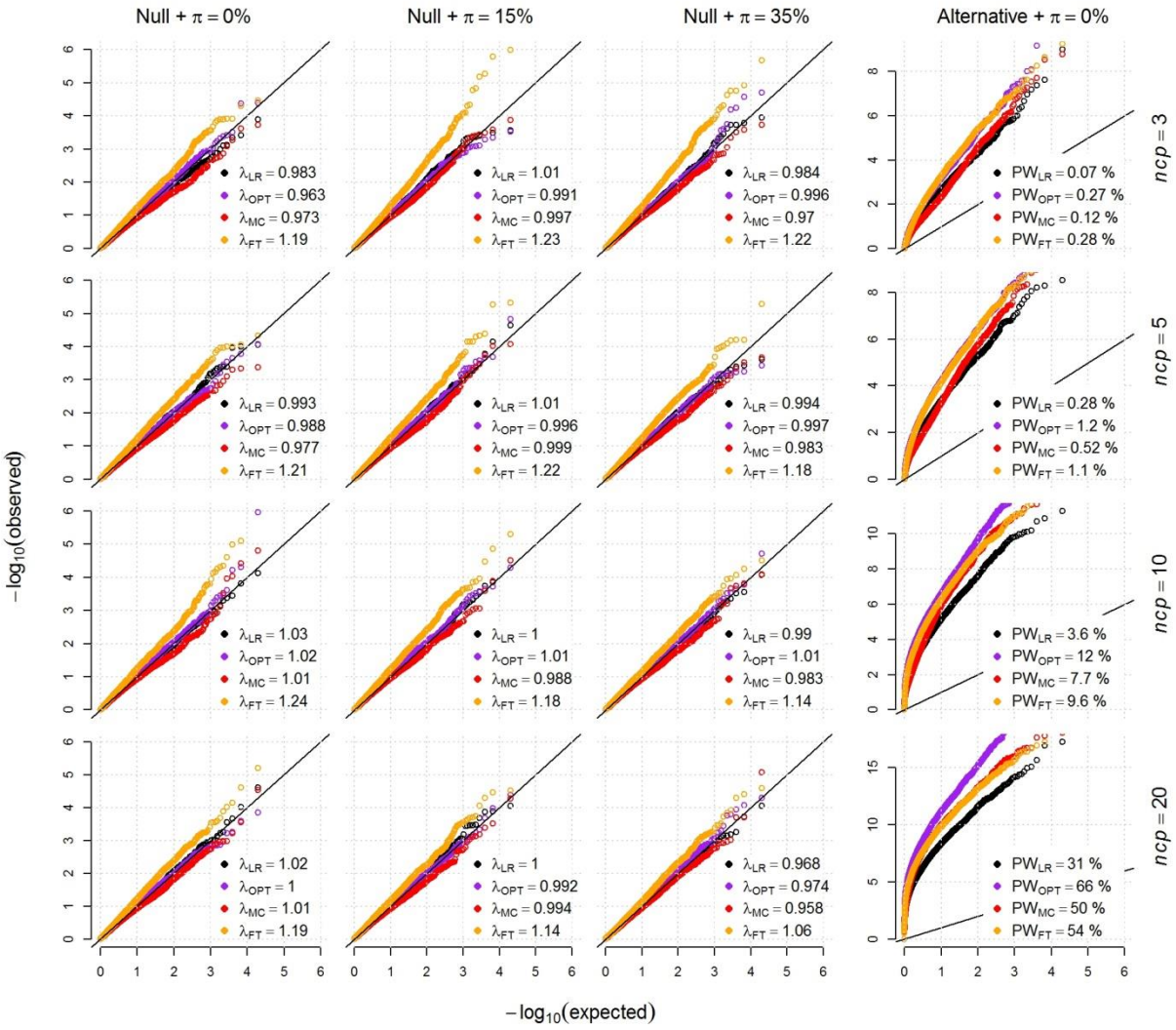
Supplementary Figure 35. ***CMS*** **QQplots for 40 phenotypes, 6000 individuals and 75% of outcome variance explained**



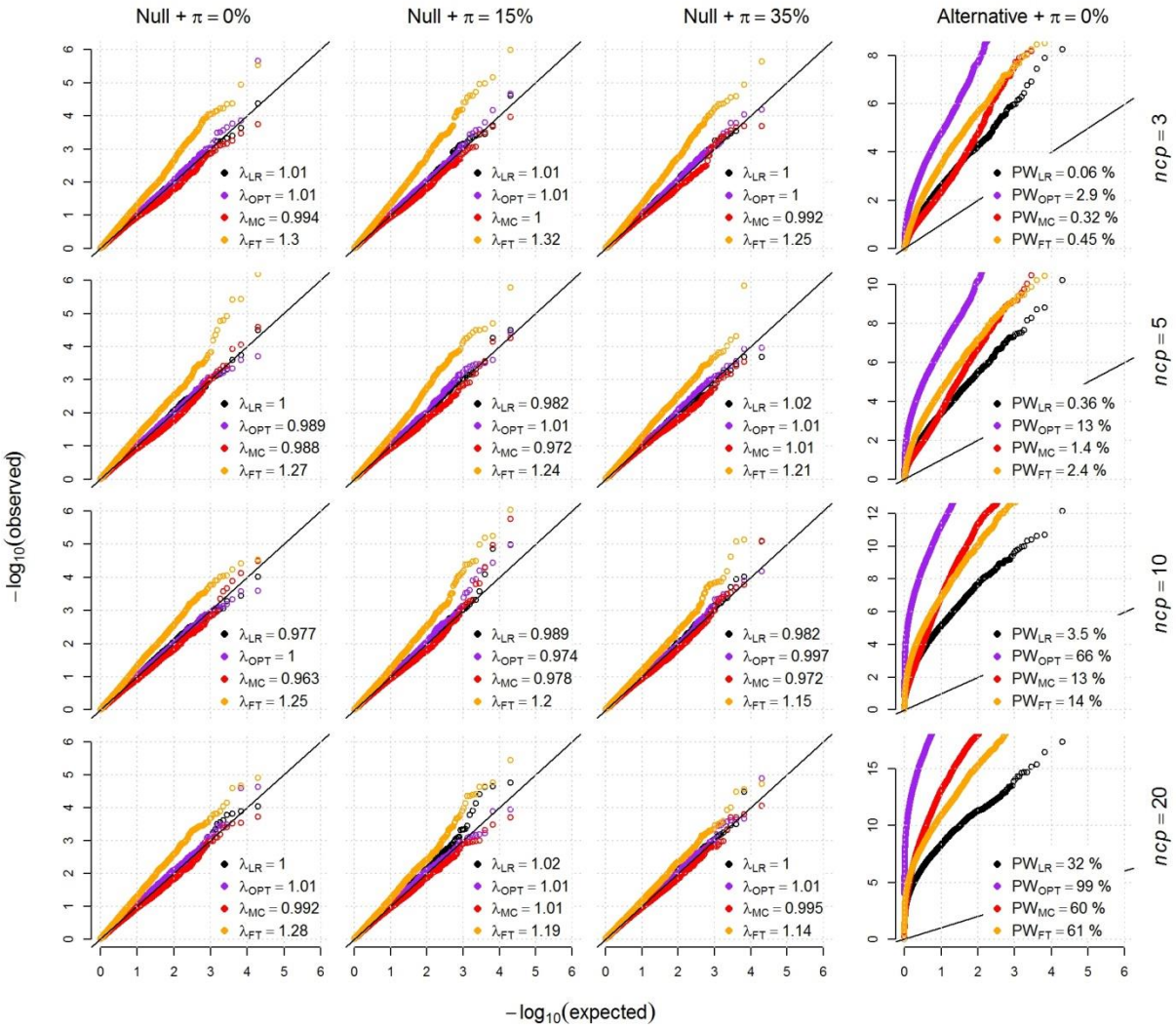
Supplementary Figure 36. ***CMS*** **QQplots for 80 phenotypes, 300 individuals and 25% of outcome variance explained**



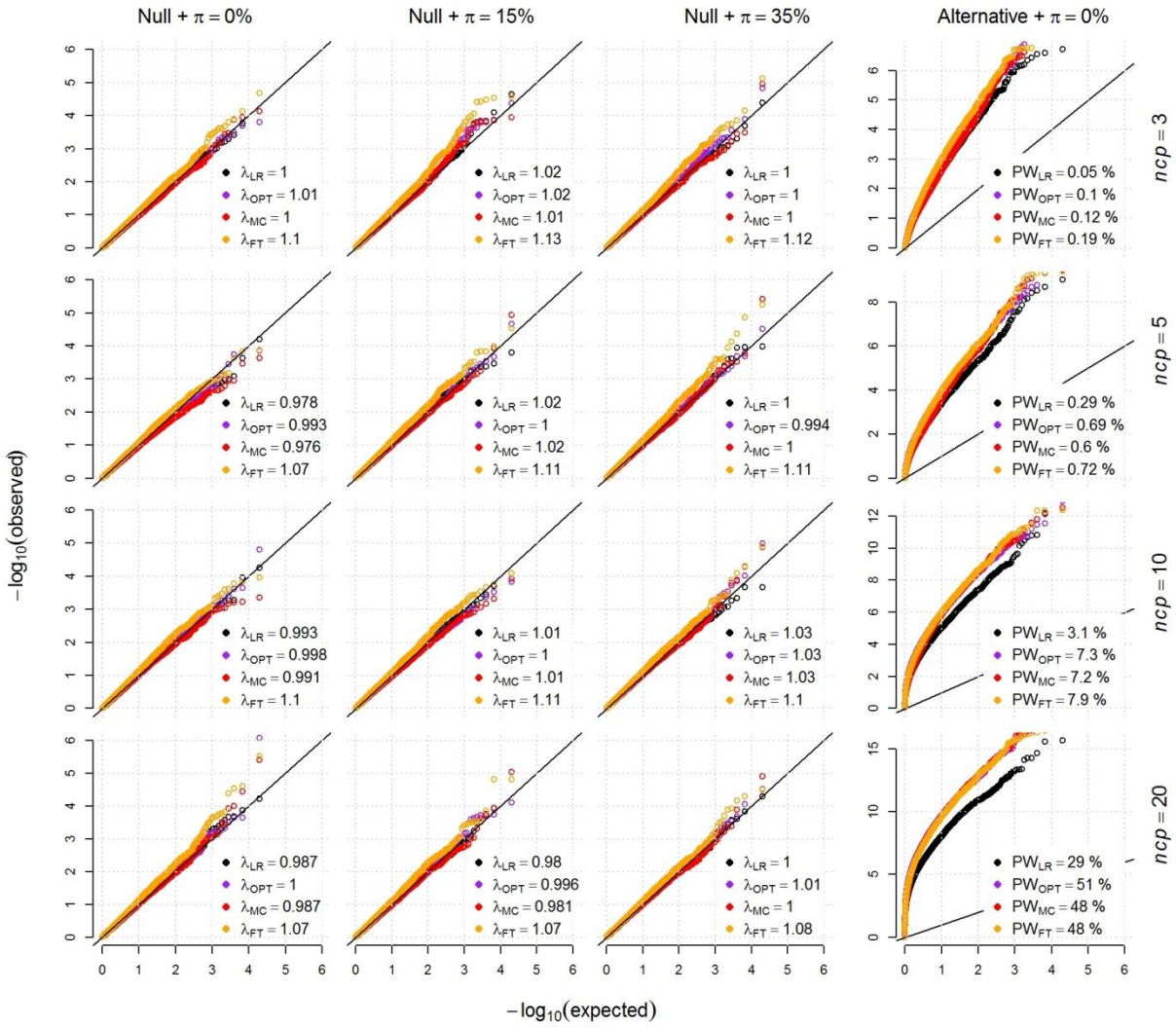
Supplementary Figure 37. ***CMS*** **QQplots for 80 phenotypes, 300 individuals and 50% of outcome variance explained**



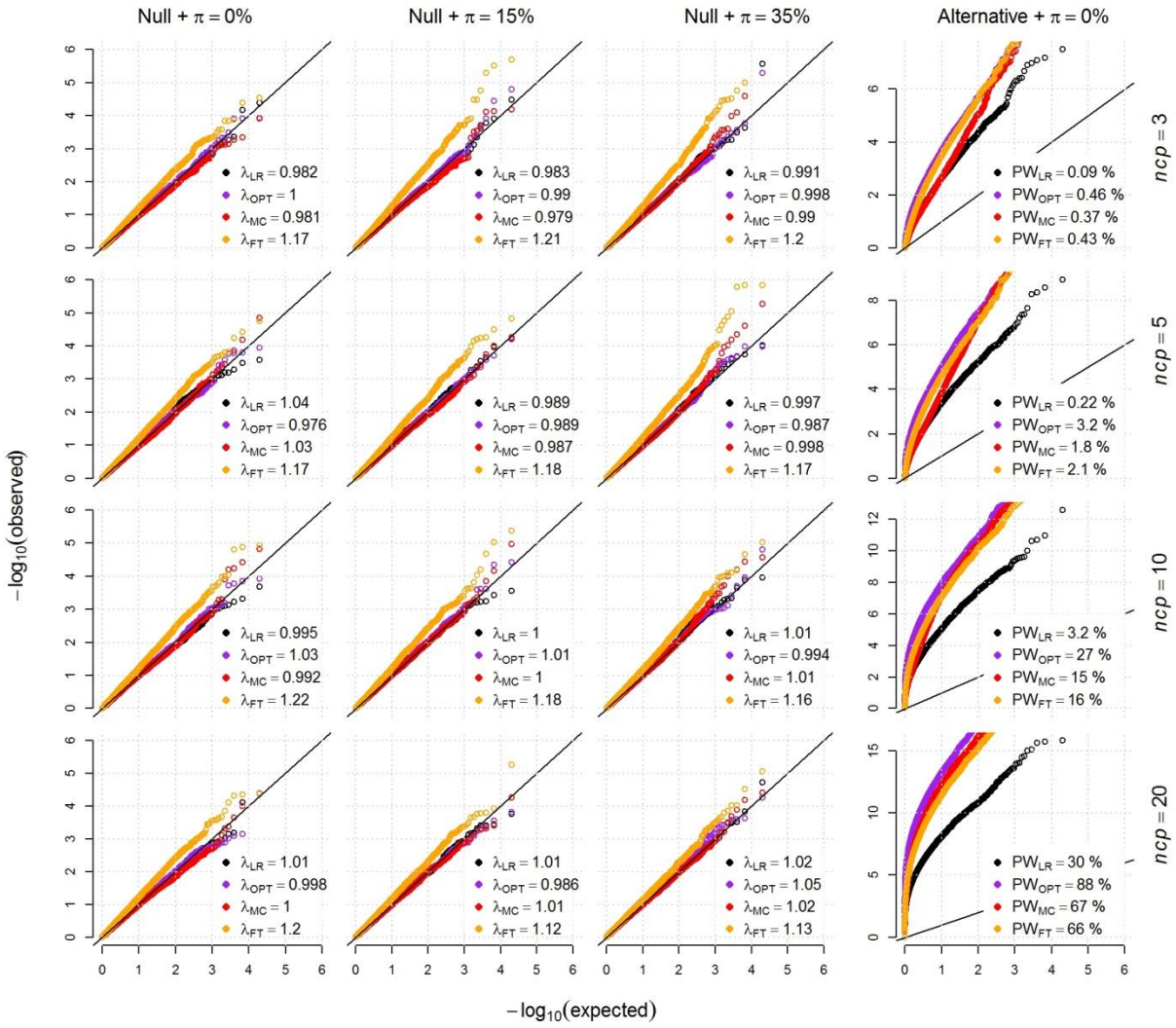
Supplementary Figure 38. ***CMS*** **QQplots for 80 phenotypes, 300 individuals and 75% of outcome variance explained**



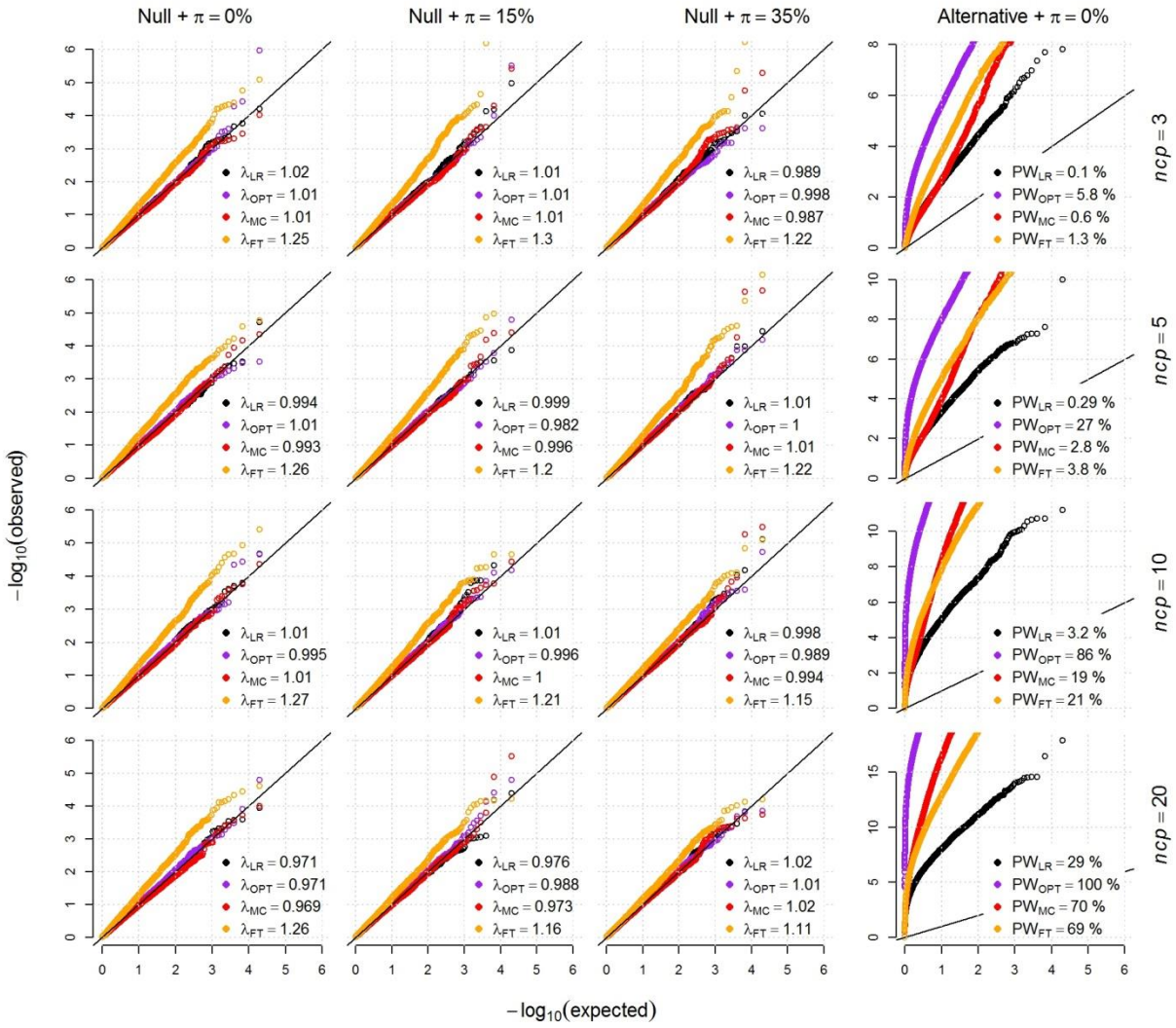
Supplementary Figure 39. ***CMS*** **QQplots for 80 phenotypes, 2000 individuals and 25% of outcome variance explained**



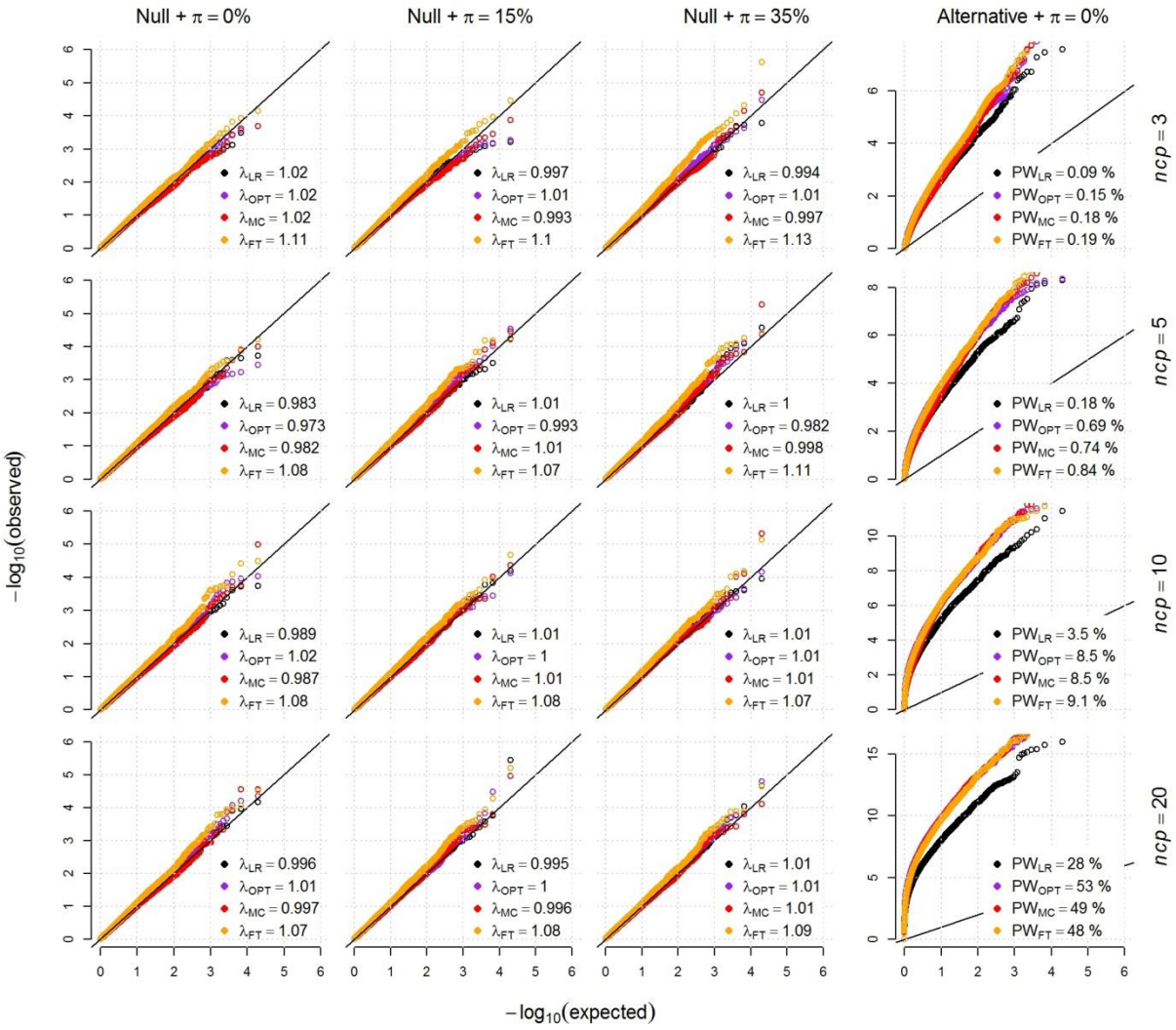
Supplementary Figure 40. ***CMS*** **QQplots for 80 phenotypes, 2000 individuals and 50% of outcome variance explained**



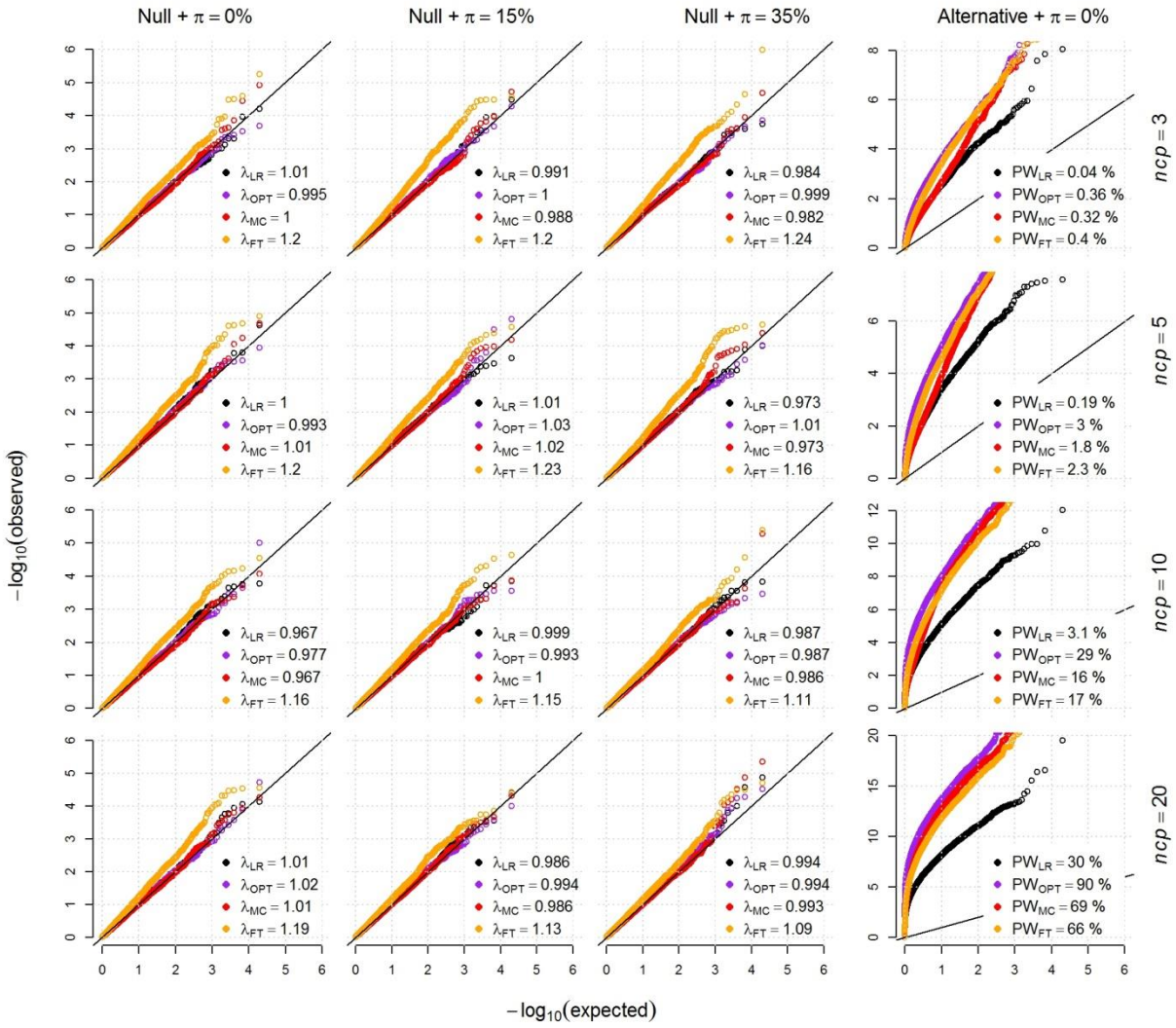
Supplementary Figure 41. ***CMS*** **QQplots for 80 phenotypes, 2000 individuals and 75% of outcome variance explained**



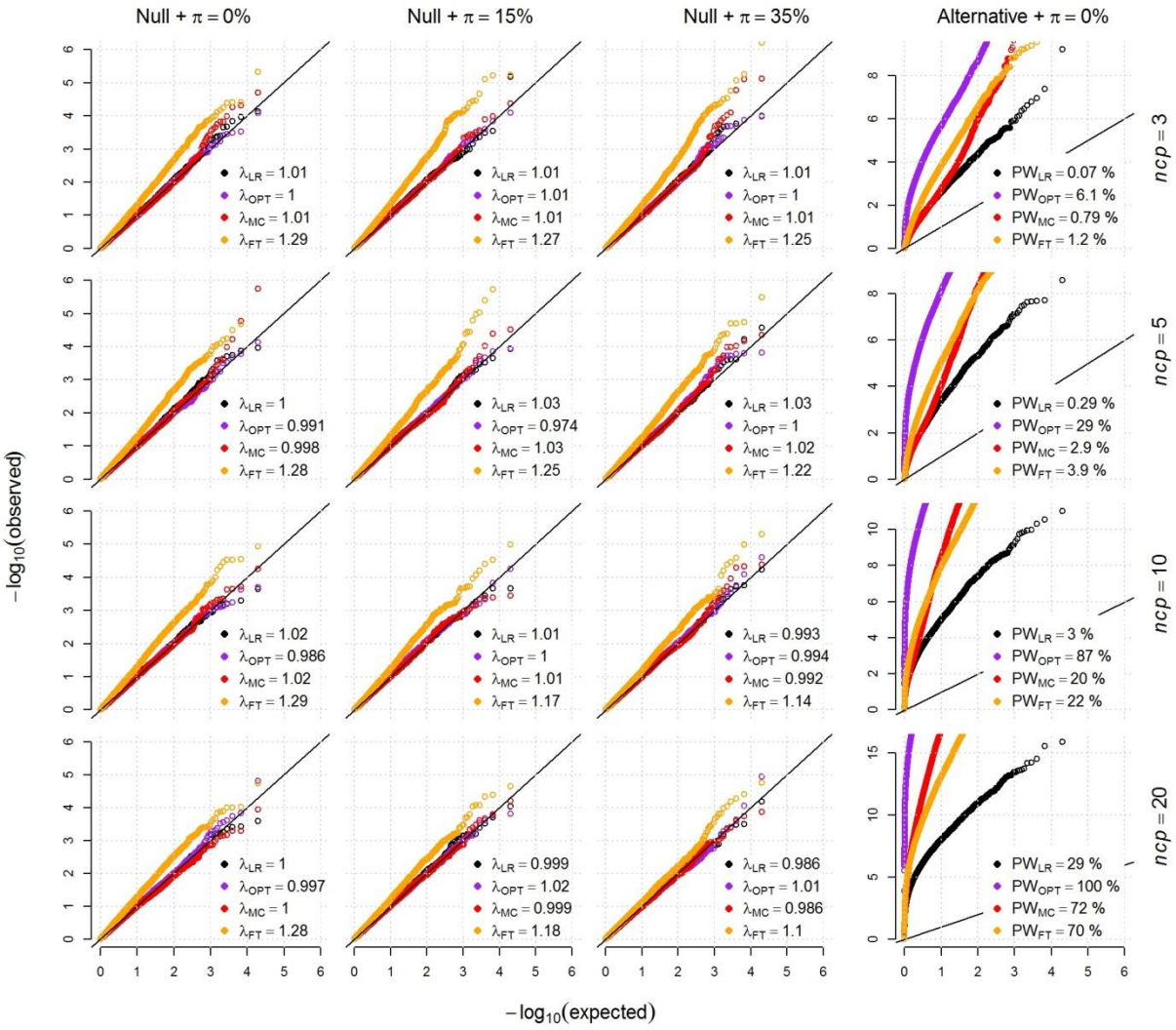
Supplementary Figure 42. ***CMS*** **QQplots for 80 phenotypes, 6000 individuals and 25% of outcome variance explained**



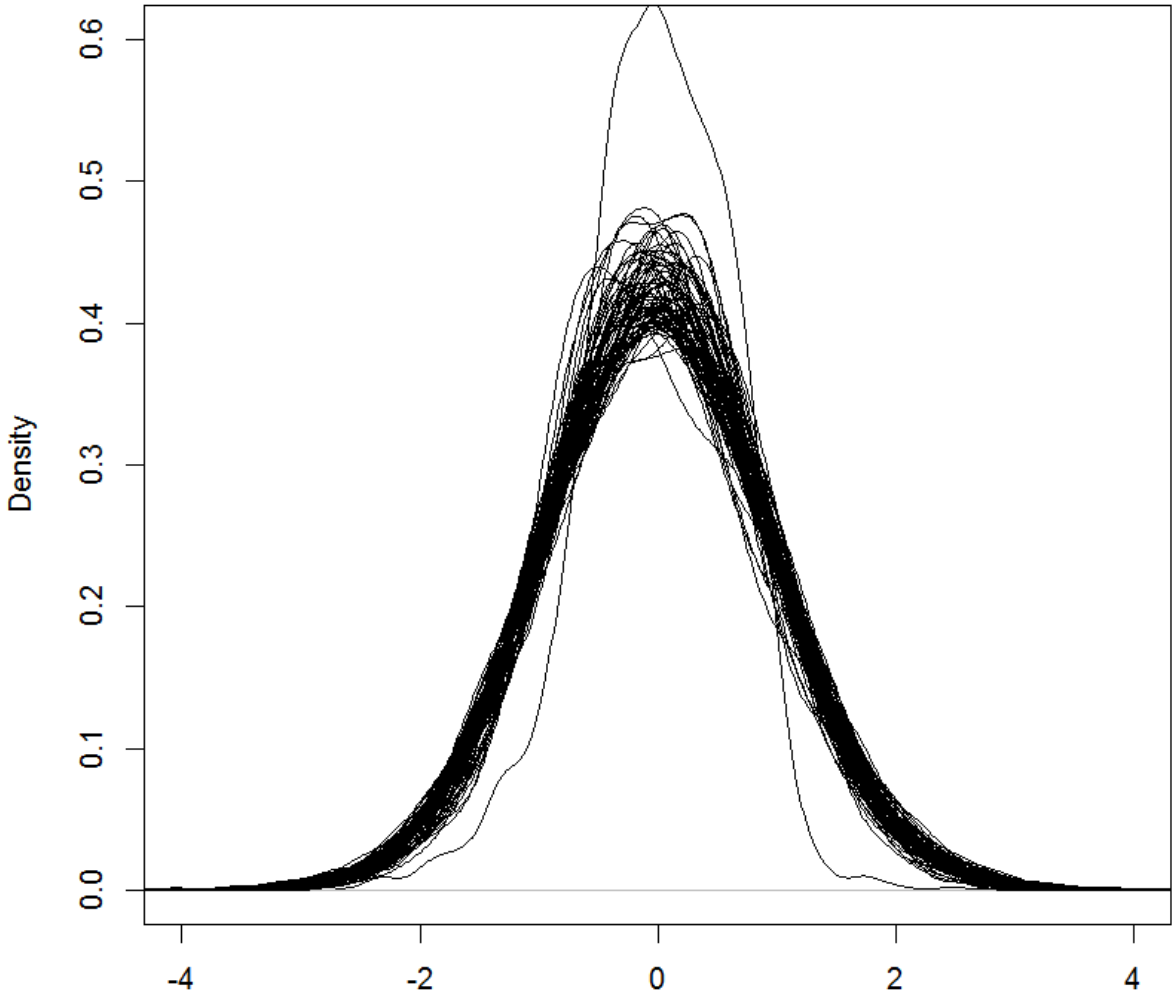
Supplementary Figure 43. ***CMS*** **QQplots for 80 phenotypes, 6000 individuals and 50% of outcome variance explained**



Supplementary Figure 44. ***CMS*** **QQplots for 80 phenotypes, 6000 individuals and 75% of outcome variance explained**



Supplementary Figure 45. Density plot for the metabolite levels.
Distribution of the 79 metabolites after adjusting for pancreatic cancer case-control status, age at blood draw, fasting status, self-reported race, and gender, and standardization (mean centered and scaled by standard deviation).



References

1. Aschard, H., Vilhjalmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet* **96**, 329-39 (2015).
2. Marchand, E. Computing the moments of a truncated noncentral chi-square distribution. *Journal of Statistical Computation and Simulation* **54**, 387-391 (1996).
3. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
4. Lee, J.D., Sun, D.L., Sun, Y. & Taylor, J.E. Exact post-selection inference, with application to the lasso. 907-927 (2016).
5. Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PLoS One* **8**, e65245 (2013).
6. Dahl, A., Guillemot, V., Mefford, J., Aschard, H. & Zaitlen, N. Adjusting For Principal Components Of Molecular Phenotypes Induces Replicating False Positives. *bioRxiv* (2017).
7. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770 (2010).
8. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837-845 (1988).
9. Karr, J.R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389-401 (2012).
10. Shin, S.Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-50 (2014).
11. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581 (2014).
12. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-7 (2014).
13. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269-76 (2012).
14. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54-60 (2011).
15. Rhee, E.P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab* **18**, 130-43 (2013).
16. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
17. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-7 (2007).
18. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-50 (2009).
19. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-7 (2010).
20. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-72 (2010).
21. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).
22. Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res* **23**, 716-26 (2013).

23. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).