Supplemental Material

General workflow

The main goal of the meal response Genome-Wide Association Study (GWAS) is to identify genetic variants that modulate postprandial metabolite levels independently from fasting, baseline levels. In order to remove the baseline related variation from the postprandial levels, a prediction model of postprandial levels has to be generated and subsequently the residuals have to be calculated. As will be shown in the remainder of this section, two aspects have to be taken into account when constructing a predictor: the postprandial levels of some metabolites are linked in a nonlinear fashion to the baseline levels, and, both baseline and postprandial levels are measured with considerable amount of measurement noise. We show here that an orthogonal nonlinear least squares (OrNLS)[1] regression model accomodates both aspects and, consequently, provides the optimal framework for determining the postprandial response.

We used the workflow shown in Fig. S1 to compute the residuals for each metabolite. To make the OrNLS fit robust to outliers, during the least squares minimization process two types of outliers were removed from the dataset. Zero-outliers, which are postprandial and baseline measurements close to zero, were removed from both the regression procedure and the final residual computation, and SD-outliers, which are data points that lie further away from the regression line than 5 standard deviations, were eliminated from the regression procedure but were included in the final calculation of the residuals. In addition to these steps, for each trait four plots were generated showing: 1) the postprandial-baseline scatterplot and OrNLS regression line; 2) the scatterplot of the vertical residuals versus the estimated baseline levels; 3) a histogram of the raw, untransformed residuals; 4) a Gaussian QQ plot of the residuals. These plots were inspected as quality control to ensure that the OrNLS procedure did not get stuck in local minima. Finally, the association analyses were done on the rank-based inverse normal transformed TG and NMR data.

1. `Eliminate zero-outliers from the dataset, which are the baseline and postprandial measurements that are lower than 0.001 times the median`
2. `Define the SD-outliers as an empty set`
3. `Fit the OrNLS regression model on the data except the SD-outliers`
4. `Determine the SD-outliers as the data points that have orthogonal residuals larger than 5 times the standard deviation`
5. `If the set of SD-outliers has changed in the last step, go to step 3 and reperform the OrNLS regression`
6. `If the set of SD-outliers has not changed, calculate the vertical residuals of all data points except the zero-outliers`
7. `Perform a rank-based inverse normal transformation on the residuals`

Figure S1: Workflow for computing the postprandial response.

[1] In our discussion, we distinguish orthogonal from ordinary least squares regression in abbreviations by using 'Or' for orthogonal and 'O' for ordinary.

An errors-in-variables model for postprandial response

To determine the proper statistical model to investigate genetic determinants that are specific for the postprandial response, we first assume an errors-in-variables model in which the postprandial levels y_i are linearly proportional to the baseline levels x_i

$$y_i = \beta_0 + \xi_i\beta_1 + g_i + \varepsilon_i \quad (1)$$

$$x_i = \xi_i + \delta_i \quad (2)$$

In equation (1) and (2), ξ_i are the actual baseline levels without measurement noise, g_i the meal response specific genetic contribution to variation in postprandial levels, β_0 and β_1 the intercept and baseline reponse coefficient, and δ_i and ε_i the measurement noise on the baseline and postprandial data respectively.[2] In the GWAS setting, the genetic factor g_i is assumed to be composed of a linear combination of genetic variants v_{ij} with effect sizes γ_j

$$g_i = v_{i0}\gamma_0 + v_{i1}\gamma_1 + v_{i2}\gamma_2 + \cdots$$

In the remainder of this section, we assume that δ and ε are uncorrelated with g and ξ and that δ and ε are normally distributed with the same variance $\mathcal{N}(0, \sigma^2)$. Note that the goal of the postprandial response GWAS is to determine the genetic effects g_i. Genetic variants that only affect the baseline levels are assumed to be incorporated in the term ξ_i and are assumed to affect y_i in the same proportion β_1 as the non-genetically determined baseline levels. As a consequence, baseline variants that affect postprandial levels with an effect size that is different from β_1 are assumed to affect baseline and postprandial levels independently, and their effect will therefore be separated into a pure baseline and response portion.

For the purpose of regression it is important to realize that the actual baseline levels ξ_i cannot be observed, and that performing ordinary least-squares (OLS) regression of y_i on x_i gives biased estimates of the regression coefficients due to the measurement error δ

$$\begin{aligned} E\left(\hat{\beta}_1^{\mathrm{OLS}}\right) &= \frac{E(x \cdot y)}{E(x^2)} \\ &= \beta_1 \frac{E(\xi^2)}{E(\xi^2) + \sigma^2} + \frac{E(\xi \cdot g)}{E(\xi^2) + \sigma^2} \end{aligned}$$

where we simplified notation by assuming that, without loss of generality, all intercept terms are set to zero (i.e. ξ and g have zero mean and $\beta_0 = 0$). The term $E(\xi \cdot g)$ refers to the expected association between baseline levels and the genetic component that has independent effects on baseline and postprandial levels. This association will not be exactly zero but can be assumed to be negligable with respect to the total variance in baseline levels $E(x^2)$ – thus not contributing to the OLS estimate of β_1.

A consequence of the biased OLS estimate is that the residuals are correlated with the actual baseline levels ξ_i and therefore they do not give a clean measure of the contribution of meal-specific genetic effects g_i since now also the effects of genetic variants affecting ξ_i are included

$$\begin{aligned} e_i^{\mathrm{OLS}} &= y_i - x_i \cdot E\left(\hat{\beta}_1^{\mathrm{OLS}}\right) \\ &= g_i + \beta_1 \frac{\sigma^2}{E(\xi^2) + \sigma^2}\xi_i + \varepsilon_i - \beta_1 \frac{E(\xi^2)}{E(\xi^2) + \sigma^2}\delta_i \end{aligned}$$

[2] In this discussion, variables with subscript i are used to denote actual data whereas the same variable without subscript is used to denote the random variable. That is, y_i refers to the postprandial level of subject i, whereas y refers to the abstract random process that generated the values y_i and that is composed of the processes ξ, g and ε.

where again we assumed zero intercepts and $E(\xi \cdot g) \ll E(x^2)$.

The correct method to perform regression on variables with equal amount of measurement error is orthogonal least-squares (OrLS), which is a special case of Deming regression (Fuller, 1987). In contrast to OLS, which minimizes the vertical distance of each data point to the regression line, OrLS minimizes the perpendicular distance of each data point to the regression line – i.e. the distance to the closest point on the regression line – which gives an unbiased estimate of the regression coefficients in case of uncorrelated measurement noise with equal variances. Due to the genetic term in (1), the OrLS estimate contains a minor bias:

$$E\left(\hat{\beta}_1^{\mathrm{OrLS}}\right) = \frac{E(y^2) - E(x^2)}{2E(x \cdot y)} + \sqrt{\left(\frac{E(y^2) - E(x^2)}{2E(x \cdot y)}\right)^2 + 1}$$

with

$$\frac{E(y^2) - E(x^2)}{2E(x \cdot y)} = \frac{1}{2\beta_1}\left(\beta_1^2 - 1 + \frac{E(g^2)}{E(\xi^2)}\right)$$

where again we assumed zero intercepts and $E(\xi \cdot g) \ll E(x^2)$. Refactoring shows that the bias can be approximated as

$$\begin{aligned} E\left(\hat{\beta}_1^{\mathrm{OrLS}}\right) &= \frac{\beta_1}{2}\left(1 - \beta_1^{-2} + B + \beta_1^{-2}B\right) + \frac{\beta_1}{2}\sqrt{\left(1 + \beta_1^{-2} + B - \beta_1^{-2}B\right)^2 + 4\beta_1^{-2}B^2} \\ &\approx \beta_1(1 + B) \end{aligned}$$

assuming that $\left(1 + \beta_1^{-2} + B - \beta_1^{-2}B\right)^2 \gg 4\beta_1^{-2}B^2$, with B the bias in the OrLS estimate

$$B = \frac{E(g^2)}{E(\xi^2) + E((\beta_1\xi)^2)}$$

That is, the OrLS estimate is biased by the ratio of the genetic variance and the sum of the baseline variance and baseline related variance in postprandial levels $\frac{E(g^2)}{E(\xi^2)+E((\beta_1\xi)^2)}$. Importantly, the OrLS bias will typically be much smaller than the bias $\frac{\sigma^2}{E(\xi^2)+\sigma^2}$ in the OLS estimate, which depends on the variance of the measurement noise. The vertical OrLS residuals now become

$$\begin{aligned} e_i^{\mathrm{OrLS}} &= y_i - x_i \cdot E\left(\hat{\beta}_1^{\mathrm{OrLS}}\right) \\ &= g_i - \beta_1 \frac{E(g^2)}{E(\xi^2) + E((\beta_1\xi)^2)}\xi_i + \varepsilon_i - \beta_1\left(1 + \frac{E(g^2)}{E(\xi^2) + E((\beta_1\xi)^2)}\right)\delta_i \end{aligned}$$

which shows that the association between the vertical OrLS residuals and the baseline levels ξ_i is much lower than for the OLS residuals. Importantly, only the vertical OrLS residuals provide an unbiased measure of genetic effects g_i and not the perpendicular ones that are minimized by orthogonal least squares regression. In fact, it can be shown that the values of e_i^{perp} are shrunk by a constant factor with respect to e_i^{OrLS}, which would therefore result in biased estimates of the genetic effect sizes γ_j

$$\begin{aligned} e_i^{\mathrm{perp}} &= \pm\sqrt{\left(x_i - \hat{\xi}_i\right)^2 + \left(y_i - \hat{\beta}_1^{\mathrm{OrLS}}\hat{\xi}_i\right)^2} \\ &= \frac{1}{\sqrt{\left(\hat{\beta}_1^{\mathrm{OrLS}}\right)^2 + 1}}\left(y_i - x_i \cdot \hat{\beta}_1^{\mathrm{OrLS}}\right) \end{aligned}$$

where $\hat{\xi}_i$ are the estimated baseline levels, i.e. the observed baseline levels corrected for measurement noise

$$\hat{\xi}_i = \frac{1}{\left(\hat{\beta}_1^{\mathrm{OrLS}}\right)^2 + 1}\, x_i + \frac{\hat{\beta}_1^{\mathrm{OrLS}}}{\left(\hat{\beta}_1^{\mathrm{OrLS}}\right)^2 + 1}\, y_i$$

In order words, for the fitting of the regression line the summed squares of the orthogonal residuals are used, whereas for the association analyses the vertical (y-axis) residuals are used to estimate the effect of g_i.

At this point it should be mentioned that a more straightforward alternative for determining meal responses of metabolites is to calculate the difference (delta) between postprandial and baseline levels or to adjust postprandial levels for baseline levels as a covariate. From our discussion here it follows that although the delta approach is robust for cases in which only the intercept β_0 is affected by a meal but not the slope of the response (i.e. $\beta_1 \approx 1$), in cases in which the postprandial levels are proportional with a coefficient $\beta_1 \neq 1$ the delta measure is by definition associated with the baseline levels ξ_i. On the other hand, the covariate approach makes no assumptions about the size of β_1, but since the estimate is biased due to measurement noise this adjustment will still induce an association with the baseline levels. After inspection of the baseline-postprandial scatterplots of all NMR metabolites we saw that both cases occur in our dataset – i.e. metabolites with a coefficient β_1 different from 1 and metabolites that had substantial measurement error on baseline and postprandial values (Fig. S2A-C). Therefore, computing the vertical residuals of the orthogonal regression model is the best approach to cover all cases.

Importantly, we observed that for a number of metabolites, amongst which glucose and several lipids and ketones, the postprandial levels were related in a strongly nonlinear fashion to the baseline levels (Fig. S2D-F). For these metabolites equation (1) has to be generalized to the nonlinear case. We found that the most pronounced nonlinear responses showed saturation effects for high baseline levels, which could be adequately modelled by the formula

$$y_i = \beta_0 + \beta_1 \frac{\xi_i}{\xi_i \beta_2 + 1} + g_i + \varepsilon_i \quad (3)$$

$$x_i = \xi_i + \delta_i \quad (4)$$

Note that equation (3) has the linear relationship (1) as special case, namely, when $\beta_2 = 0$. Therefore, we applied the nonlinear errors-in-variables model (3) and (4) on all TG and NMR metabolites in order to use a single statistical framework and also to model the response of metabolites with more subtle nonlinear effects.

A possible source for the nonlinear baseline-response relationship of certain metabolites is that baseline levels are indicative of metabolic health and therefore also affect the time when the peak is reached in the postprandial response and the size of the peak. For instance, blood glucose levels in normal glycemic subjects quickly respond to a meal and return relatively fast to zero, whereas (pre)diabetic subjects have a much slower and more blunted glucose response. So where the postprandial time measurement of 150 min. will be relative to the peak in the total postprandial time curve will therefore determine in a nonlinear fashion what the value is of y_i.

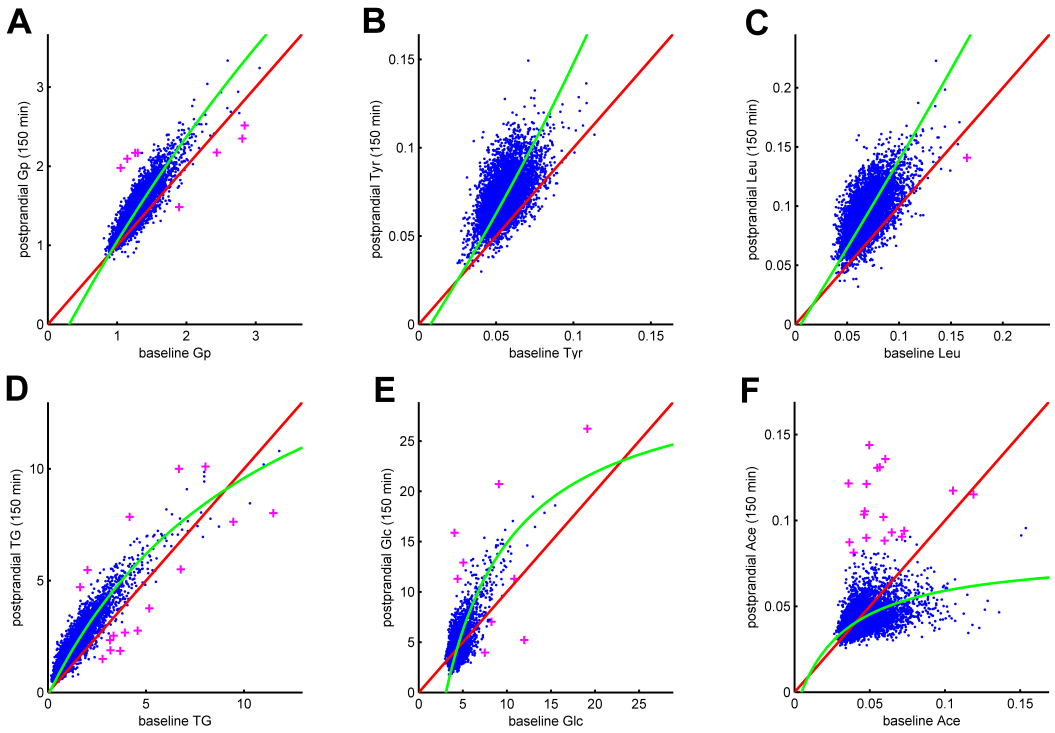

Figure S2: Scatterplots of postprandial versus baseline levels of glycoprotein acetyls (A), tyrosine (B), leucine (C), triglycerides (D), glucose (E) and acetate (F). The size of the postprandial responses of glycoprotein acetyls, tyrosine and leucine is proportional to the baseline levels, showing that for some NMR metabolites the regression coefficient β_1 is larger than one. The postprandial responses of triglycerides, glucose and acetate show pronounced nonlinear effects. The red line in the figure is the $y = x$ line, the green line is the OrNLS regression fit, and the magenta data points are the SD outliers.

References

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.