Supplementary Information: Epigenome-wide association study architecture and power

Contents

Introduction

Epigenome-wide association study (EWAS) associations in which DNA methylation levels correlate with complex trait variation can arise due to (1) forward causality (whereby DNA methylation changes cause trait variation), (2) reverse causality (whereby trait variation causes change in DNA methylation levels) or (3) confounding. Each of these models is subject to different constraints under a multi-factorial model of complex traits. Suppose that our complex trait y has P_y factors and a methylation site m_j is influenced by P_{m_j} factors. Though methylation sites are known to be multi-factorial themselves, in general we would consider that $P_y \gg P_{m_j}$.

If you have P factors that influence a value y then they have the following constraint

$$R^2_T = \sum_{j=1}^{P} R^2_{m_j y} \leq 1$$

where R^2_T = the total variance of a trait and $R^2_{m_j y}$ = the variance explained by methylation site j.

In principle the variance not captured by $R^2_{m_j y}$ is basically stochastic noise. Therefore the larger the value of P, the smaller the value that $R^2_{m_j y}$ can take, and the lower the power to detect any factor for y. Suppose that $Var(y) = 1$ and $Var(m_j) = 1$ then $\beta_{m_j y} \sim N(0, R^2_T/P)$.

Below we discuss the implication for discovery of CpG sites in EWAS when associations arise due to forward causality, reverse causality, and confounding.

Models

Forward causal

If all CpG-trait associations are due to CpGs being causal then they are amongst the P_y causal factors for y then the effects are constrained to be drawn from $\beta_{m_j y} \sim N(0, R^2_{T_y}/P_y)$. So $R^2_{m_j y}(1) = \beta^2_{m_j y}$.

Reverse causal

If all CpG-trait associations are reverse causal, then each CpG is independently influenced by y with an effect of $\beta_{y m_j} \sim N(0, R^2_{T_{m_j}}/P_{m_j})$. So $R^2_{m_j y}(2) = \beta^2_{y m_j}$

Confounded

If all CpG-trait associations are confounded then each CpG is independently influenced by some confounder u that also influences y. So the effect of u on m_j is subject to the constraint in (2) and the effect of u on y is subject to the constraint in (1).

$$\begin{aligned} R^2_{m_j y}(3) &= \frac{Cov(m_j, y)^2}{Var(m_j)Var(y)} \\ &= Cov(\beta_{u m_j} u, \beta_{uy} u)^2 \\ &= Var(u)^2 \beta^2_{u m_j} \beta^2_{uy} \\ &= \beta^2_{u m_j} \beta^2_{uy} \end{aligned}$$

Assume the variance of u is 1, overall the expected association will be $R^2_{m_j y}(3) = \beta^2_{uy} \beta^2_{u m_j}$.

Expected number of associations in EWAS

Statistical power of EWAS is mostly related to variance in the trait explained by the CpG ($R^2_{m_j y}$) and the sample size (N). For P_y DNA methylation sites that relate to y the number expected to be associated is the sum of the power across all sites.

Using the models above as a guide, below are simulations demonstrating how power to detect forward causal, reverse causal, and confounded associations in EWAS differ under various scenarios.

Simulations

Generate a function that will estimate the expected number of associations for each of the three models

```
suppressMessages(suppressPackageStartupMessages({
library(pwr)
library(dplyr)
library(ggplot2)
}))
#' Calculate power to detect an EWAS association under models of forward causality, reverse causality,
#'
#' @param Pm causal factors influencing DNAm
#' @param Py causal factors influencing the trait (Y)
#' @param R2m variance explained in DNAm by the trait
#' @param R2y variance explained in the trait by DNAm
#' @param N sample size
#' @param thresh P value threshold for a "significant" association
#'
#' @return tibble of input parameters and number of expected associations
calc_power <- function(Pm, Py, R2m, R2y, N, thresh)
{
  # Model 1 (forward causal)
  b1 <- rnorm(Py, mean=0, sd=sqrt(R2y/Py))
  pow1 <- pwr.r.test(N, b1, thresh)$power
  nsig1 <- sum(pow1)

  # Model 2 (reverse causal)
  b2 <- rnorm(Py, mean=0, sd=sqrt(R2m/Pm))
  pow2 <- pwr.r.test(N, b2, thresh)$power
  nsig2 <- sum(pow2)

  # Model 3 (confounded)
  b3 <- b1 * b2
  pow3 <- pwr.r.test(n=N, r=b3, sig.level=thresh)$power
  nsig3 <- sum(pow3)

  return(tibble(Pm=Pm, Py=Py, R2m=R2m, R2y=R2y, N=N, thresh=thresh,
                model=c("Forward causal", "Reverse causal", "Confounded"),
                nsig=c(nsig1, nsig2, nsig3)
  ))
}
```

Set the parameters across which the simulations will run

```
param <- expand.grid(
  Pm = c(5, 50, 500),
  Py = seq(500, 10000, by=500),
  R2m = c(0.02, 0.2, 0.8),
  R2y = c(0.3, 0.5, 0.7),
  N = c(1000, 10000, 100000),
  thresh=5e-7
)
res <- lapply(1:nrow(param), function(i) do.call(calc_power, param[i,])) %>% bind_rows()
```

Visualise the expected yield of associations from each model

```
subset(res, R2m == 0.2 & R2y == 0.5) %>%
ggplot(., aes(x=Py, y=nsig)) +
  geom_line(aes(colour=as.factor(model))) +
  facet_grid(N ~ Pm, labeller=label_both) +
  scale_y_log10() +
  geom_hline(yintercept=1, linetype="dotted") +
  scale_colour_brewer(type="qual") +
  labs(x="Number of causal factors for Y (Py)",
       y="Expected number of associations",
       colour="Model") +
  theme_bw() +
  theme(legend.position="bottom")
```

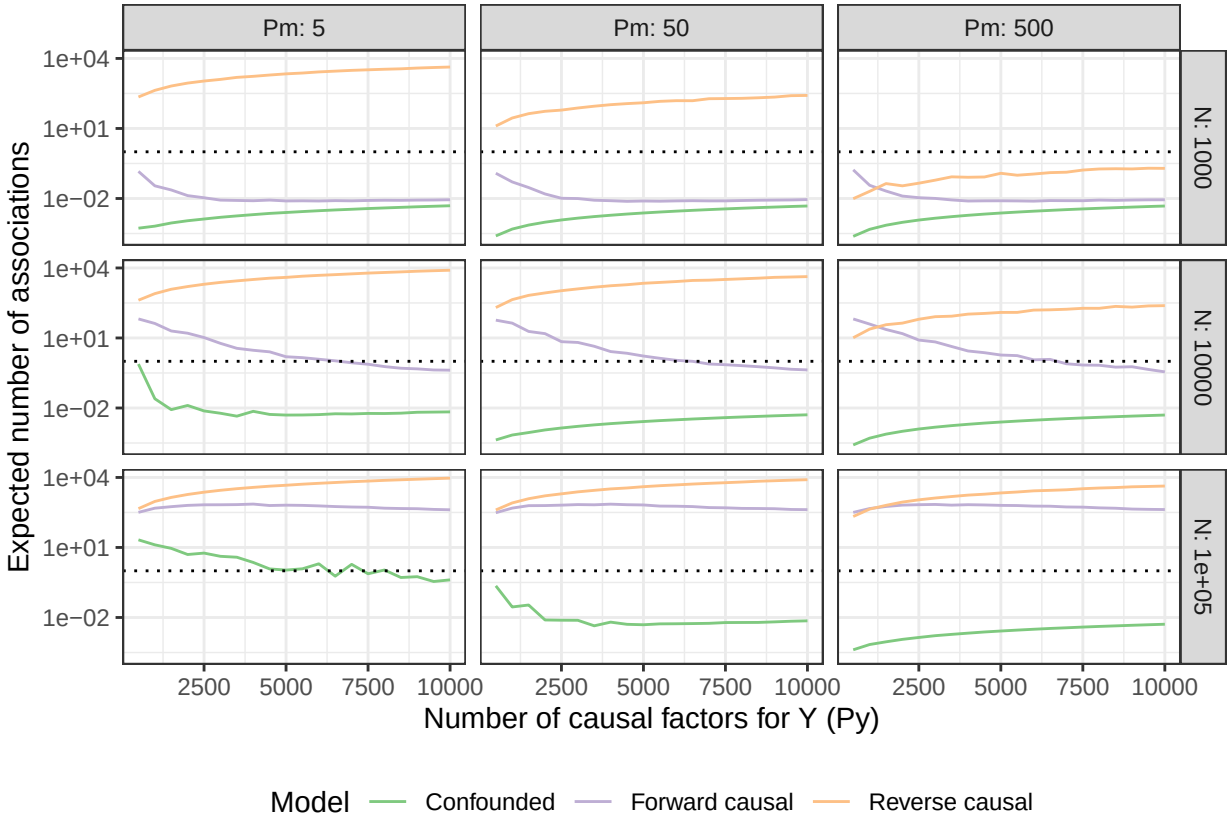

Figure 1: Each box represents the expected number of EWAS associations (y-axis, note the log scale) against how multi-factorial the trait is (x-axis) for the given parameters under each of the three causal models. Rows of boxes represent different sample sizes, and columns of boxes represent how multi-factorial the DNAm sites are. The black dotted line depicts the location on the y-axis for detecting a single causal variant, drawn for convenience.

Summary

In general, the power to detect EWAS associations whereby the complex trait of interest influences DNA methylation at a given site (reverse causal), greatly exceeds the power to detect DNA methylation changes that effect the trait (forward causal) when the trait is substantially more multi-factorial than each of the DNA methylation sites. EWAS typically have the least power to detect associations arising due to confounding, though this assumes that each confounder has a small effect on the trait (y).