Supplementary Information: Power for an epigenome-wide association study of atopic dermatitis

Contents

Introduction

This supplementary note details the analyses conducted to assess the power to detect associations at varying effect sizes in epigenome-wide association studies (EWAS) of atopic dermatitis (AD). To get an idea of the power to detect associations, data were simulated and EWAS were run on these simulated data. Our empirical EWAS results were then mapped to the simulated results to assess the power of this study to detect associations between DNA methylation and AD at different effect sizes. The details of how simulations were setup are in the **Methods** section and the simulation results and code can be found in the **Results** section.

Methods

Data were simulated as to perform a logistic regression analysis using the `glm()` function in R, with the model,

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X \tag{1}$$

where $p(X) = Pr(Y = 1|X)$. X is the methylation level and Y is the atopic dermatitis status of the individual (case or control).

Across the simulations, data were simulated whilst varying these parameters:

- mean methylation level
- standard deviation
- sample size
- proportion of cases
- difference in methylation between cases and controls

These values were chosen to try and reflect the actual data as best as possible. Under each scenario (one set of parameters), the `glm()` function was run to perform the logistic regression as above - an odds ratio and P value was extracted from the `glm()` output. The simulations were repeated 1000 times for each scenario.

Power was calculated like so,

$$power = \frac{\sum_{i=1}^{n} P_i < 1e-7}{n} \tag{2}$$

where n = the number of simulations (so 1000) and P_i = the P value from that simulation.

Results

Loading packages

```
library(tidyverse)
library(knitr)
library(bookdown)
library(kableExtra)
library(usefunc) # own package of useful functions - github.com/thomasbattram/usefunc
```

Study data

Before launching into the simulations, lets remind ourselves of the empirical results. **Table 1** gives a summary of the data and results.

Table 1: Meta-analyses results summary

trait	n	K	OR per SD median (range)	Lowest P
Childhood AD	5133	0.33	1.05 (1.02, 1.14)	4.3e-06
Early-onset AD	5846	0.26	1.06 (1.01, 1.14)	2.0e-07
Persistent AD	3189	0.12	1.09 (1.05, 1.14)	1.6e-06

For each trait there were three models, to summarise results here I took the top 30 hits from these models to calculate the median OR per SD increase in DNA methylation and minimum P value. n = sample size, K = proportion of cases

Simulations

Generate a function that will simulate data based on altering the variables listed in the Methods section and run logistic regression on those simulated data.

```
#' Calculate power to detect an EWAS association
#'
#' @param meth_mean mean methylation of CpG site
#' @param meth_diff methylation difference between cases and controls
#' @param meth_sd SD of methylation of CpG site
#' @param n sample size
#' @param case_prop proportion of the samples that are cases
#'
#' @return tibble of input parameters, OR and P from `glm()`
calc_power <- function(meth_mean, meth_diff, meth_sd, n, case_prop)
{
  ## Generate cases and controls and methylation levels for each
  n_case <- n * case_prop
  mean_m_case <- meth_mean + meth_diff/2
  mean_m_control <- meth_mean - meth_diff/2
  tab <- tibble(case = c(rep(1, n_case), rep(0, n - n_case)))
  m_case <- rnorm(n_case, mean=mean_m_case, sd=meth_sd)
  m_control <- rnorm(n - n_case, mean=mean_m_control, sd=meth_sd)
  tab$meth <- c(m_case, m_control)

  ## Run the logistic regression and extract OR and P
  fit <- glm(case ~ meth, data = tab, family = "binomial")
  or <- exp(coef(fit)[2])
  or_sd <- convert_units(or, meth_sd)
  p <- summary(fit)$coefficients[2,4]
  out <- tibble(meth_mean, meth_diff, meth_sd, n, case_prop, or, or_sd, p)
  return(out)
}

#' Convert OR to OR per SD increase
#'
#' @param OR numeric vector of odds ratios
#' @param sd numeric vector of corresponding SDs
#'
#' @return numeric vector of ORs per SDs
convert_units <- function(OR, sd)
{
  beta <- log(OR)
  return(exp(beta * sd))
}
```

Set the parameters under which the simulations will run.

```
## Note, we use data from the original study data to make sure we include the range of
## sample sizes and case proportions within our data
parameters <- expand.grid(meth_mean = seq(0.3, 0.7, 0.1),
                          meth_diff = seq(0.01, 0.05, 0.01),
                          meth_sd = seq(0.05, 0.2, 0.05),
                          n = seq(signif(min(study_dat$n), 1),
                                  signif(max(study_dat$n), 1),
                                  500),
                          case_prop = as.numeric(c(min(study_dat$K), max(study_dat$K))),
                          sim = 1:1000) %>% as_tibble()
```

Run the simulations. *Warning the simulations below take several hours to complete, so for the purposes of this supplementary note we have loaded in the results to make Figure 1*

```
## Run sims
out_res <- lapply(1:nrow(parameters), function(x) {
  set.seed(x)
  df <- parameters[x, ]
  res <- with(df, calc_power(meth_mean, meth_diff, meth_sd, n, case_prop))
  return(res)
})
fin_res <- bind_rows(out_res)
## Calculate power from sims
alpha_level <- 1e-7
power_levels <- fin_res %>%
  group_by(meth_diff, meth_sd, n, case_prop) %>%
  summarise(power = sum(p < alpha_level) / (max(fin_res$sim) * length(unique(fin_res$meth_mean))),
        avg_or = median(or), min_or = min(or), max_or = max(or)) %>%
  mutate(K = usefunc::comma(case_prop))
```

Figure 1 shows how power changes across sample size (n), proportion of cases (K) and different characteristics of DNA methylation sites. About 90% of CpG sites assayed with the Illumina Infinium® HumanMethylation450 (HM450) BeadChip have methylation levels with an SD of 0.1 or less. Thus, in simulation scenarios where meth_sd = 0.1, the power estimates are conservative for about 90% of probes.

```
power_levels$K <- as.factor(power_levels$K)
ggplot(power_levels, aes(x = n, y = power, colour = K, group = K)) +
  geom_point() +
  geom_line() +
  facet_grid(meth_diff ~ meth_sd,
         labeller = label_both) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90))
```

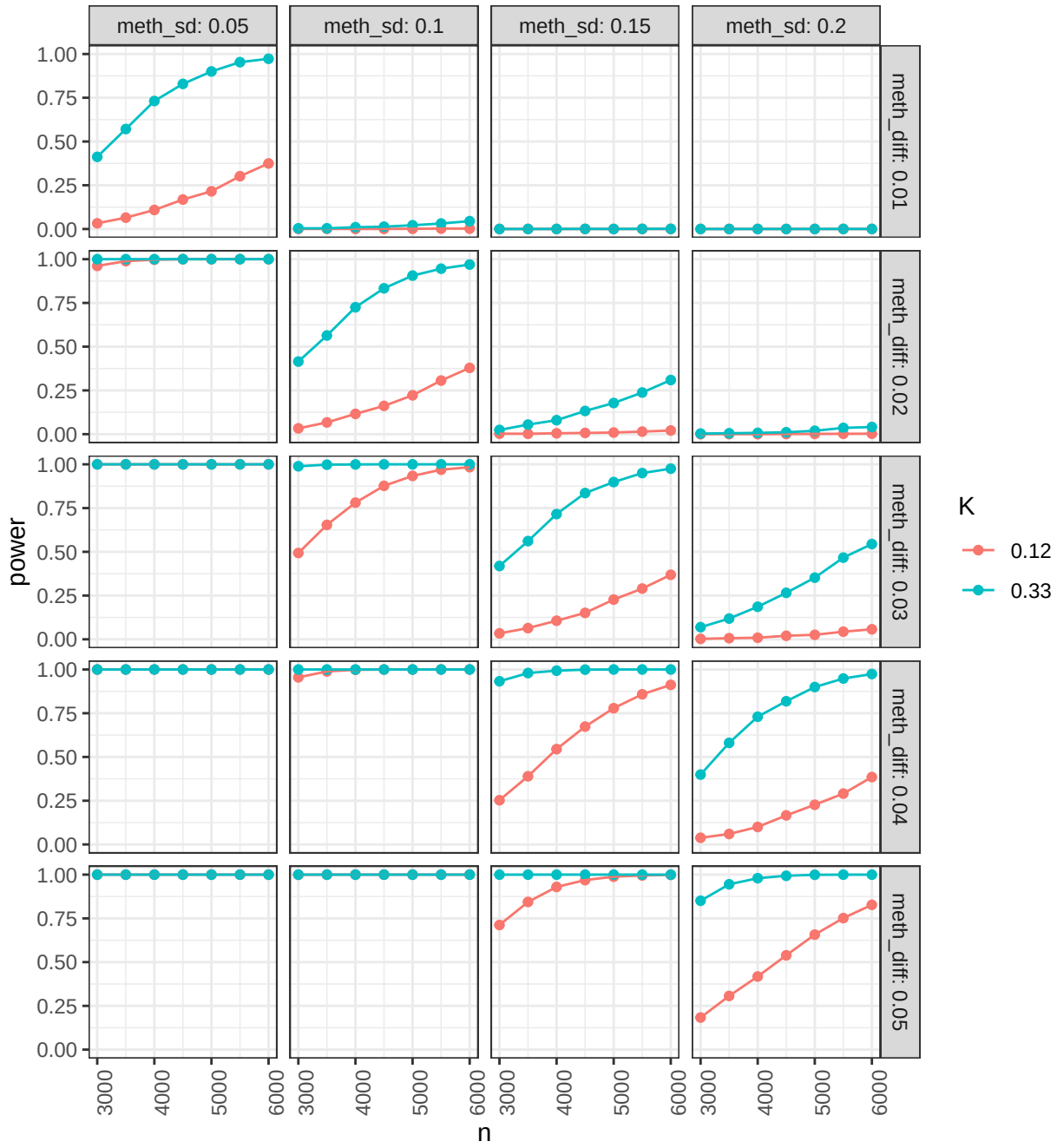

Figure 1: **EWAS power simulations**. n = sample size, K = proportion of cases, meth_sd = standard deviation of methylation levels, meth_diff = difference in methylation levels between cases and controls.

Conclusion

From our power calculations we estimate that for roughly 90% of CpG sites we have >99% power to detect DNA methylation differences of 0.03 between cases and controls and 80-90% power to detect differences of 0.02 between cases and controls for our Childhood AD EWAS (1694 cases and 3439 controls) and our Early-onset AD EWAS (1520 cases and 4326 controls). For our Persistent AD EWAS (383 cases and 2806 controls), we estimate, that for roughly 90% of CpG sites, we have 50% power to detect DNA methylation differences of 0.03 between cases and controls and >5% power to detect differences of 0.02 between cases and controls. For a DNA methylation difference of 0.02 and 0.03 between cases and controls, the OR per

SD increase in DNA methylation equivalent had a median and range of 1.22 (0.99-1.47) and 1.35 (1.11-1.73) respectively.