GenomicSEM of MDD symptoms sample counts

Mark Adams, Bradley Jermy, Jackson Thorp, Andrew Grotzinger, Michel Nivard

Make table of counts of sypmtom presence/absence used in final analyses

library(readxl)  
library(readr)  
library(dplyr)  
library(tidyr)  
library(stringr)

File for aligning sumstats filenames to cohort/symptom

cohort\_alignment <- read\_tsv('meta/cohort\_alignment.txt')

## Rows: 245 Columns: 3  
## ── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (3): filename, cohort, reference  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Symptom descriptions

mdd\_symptoms <- read\_tsv('dsm\_mdd.tsv')

## Rows: 15 Columns: 7  
## ── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (7): Ref, Reference, h, v, abbv, Symptom, Description  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

List of basic.xls files from Ricopili with sample counts

basic\_xls\_list <- list.files('meta/distribution', pattern='basic.\*.xls', full.names=TRUE, recursive=TRUE)  
  
names(basic\_xls\_list) <- str\_match(basename(basic\_xls\_list), "basic\\.([A-Za-z\_]+\\.MDD[0-9A-Za-z\_]+)")[,2]  
  
basics <- bind\_rows(lapply(basic\_xls\_list, read\_excel), .id='meta')  
  
basics\_datasets <- basics %>%  
 filter(Dataset != 'SUM') %>%  
 select(-meta) |>  
 distinct()  
   
basics\_meta <-  
basics %>%  
 filter(Dataset == 'SUM') %>%  
 mutate(Dataset=meta) %>%  
 select(-meta)

Merge and reformat

basics\_aligned <-   
cohort\_alignment %>%  
filter(!reference %in% c('MDD3', 'MDD4')) %>%  
mutate(Dataset=str\_replace(str\_remove(filename, pattern='daner\_'), 'gz', 'align.gz')) %>%  
left\_join(basics\_datasets, by='Dataset')  
  
basics\_summary <- basics\_aligned |>  
mutate(meta=case\_when(cohort %in% c('PGC', 'AGDS', 'GenScot', 'janpy') ~ "Clinical",  
 cohort %in% c('ALSPAC', 'EstBB', 'UKBB') ~ "Community",  
 TRUE ~ NA\_character\_)) |>  
group\_by(meta, reference) |>  
summarize(N\_cases=sum(N\_cases, na.rm=T), N\_controls=sum(N\_controls, na.rm=T), N\_eff\_half=sum(N\_eff\_half, na.rm=T)) |>  
ungroup()

## `summarise()` has grouped output by 'meta'. You can override using the  
## `.groups` argument.

presence\_absence <- basics\_summary |>  
 str\_glue\_data("{N\_eff\_half} ({round(100\*N\_cases/(N\_cases + N\_controls))}%)")  
  
basics\_formatted <-   
basics\_summary %>%  
mutate(PresenceAbsence=as.character(presence\_absence)) %>%  
select(meta, reference, PresenceAbsence) %>%  
pivot\_wider(names\_from=meta, values\_from=PresenceAbsence) |>  
left\_join(mdd\_symptoms, by = c('reference' = 'Reference')) |>  
mutate(Symptom = str\_glue("{Ref}. {Symptom}")) |>  
select(Symptom, Abbr. = abbv, Clinical, Community)  
  
knitr::kable(basics\_formatted)

| Symptom | Abbr. | Clinical | Community |
| --- | --- | --- | --- |
| 1. Depressed mood | Dep | 2471 (93%) | 102071 (52%) |
| 2. Anhedonia | Anh | 4494 (90%) | 98458 (39%) |
| 3a. Weight loss / decrease in appetite | AppDec | 10119 (39%) | 35837 (52%) |
| 3b. Weight gain / increase in appetite | AppInc | 9259 (38%) | 26681 (38%) |
| 4a. Insomnia | SleDec | 9418 (74%) | 28741 (79%) |
| 4b. Hypersomnia | SleInc | 10031 (49%) | 18377 (50%) |
| 5a. Psychomotor agitation | MotoInc | 10380 (46%) | 218 (3%) |
| 5b. Psychomotor slowing | MotoDec | 11130 (53%) | 543 (9%) |
| 6. Fatigue | Fatig | 3907 (91%) | 25790 (84%) |
| 7. Feelings of worthlessness / guilt | Guilt | 5503 (85%) | 46694 (59%) |
| 8. Diminished concentration | Conc | 3793 (91%) | 32827 (76%) |
| 9. Recurrent thoughts of death or suicide | Sui | 10545 (65%) | 50035 (44%) |

basics\_sum <-  
basics\_aligned %>%  
mutate(sample=if\_else(cohort %in% c('AGDS', 'PGC', 'GenScot', 'janpy'), true='enriched', false='unselected')) %>%  
group\_by(sample, reference) %>%  
summarise(N\_cases=sum(N\_cases), N\_controls=sum(N\_controls))

## `summarise()` has grouped output by 'sample'. You can override using the  
## `.groups` argument.

basics\_sum %>%  
group\_by(sample) %>%  
summarise(minCa=min(N\_cases, na.rm=T), maxCa=max(N\_cases, na.rm=T), minCo=min(N\_controls, na.rm=T), maxCo=max(N\_controls, na.rm=T))

## # A tibble: 2 × 5  
## sample minCa maxCa minCo maxCo  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 enriched 21681 21681 1748 1748  
## 2 unselected 113 107956 2995 126167

Sample prevalences

basics\_sum %>%   
transmute(sample, reference, P=N\_cases / (N\_controls + N\_cases)) %>%   
mutate(d50=abs(0.5-P)) %>%  
arrange(sample, d50) %>%  
print(n=Inf)

## # A tibble: 24 × 4  
## # Groups: sample [2]  
## sample reference P d50  
## <chr> <chr> <dbl> <dbl>  
## 1 enriched MDD1 0.925 0.425   
## 2 enriched MDD2 NA NA   
## 3 enriched MDD3a NA NA   
## 4 enriched MDD3b NA NA   
## 5 enriched MDD4a NA NA   
## 6 enriched MDD4b NA NA   
## 7 enriched MDD5a NA NA   
## 8 enriched MDD5b NA NA   
## 9 enriched MDD6 NA NA   
## 10 enriched MDD7 NA NA   
## 11 enriched MDD8 NA NA   
## 12 enriched MDD9 NA NA   
## 13 unselected MDD4b 0.501 0.000871  
## 14 unselected MDD3a 0.519 0.0195   
## 15 unselected MDD1 0.520 0.0204   
## 16 unselected MDD9 0.444 0.0562   
## 17 unselected MDD7 0.586 0.0863   
## 18 unselected MDD2 0.391 0.109   
## 19 unselected MDD3b 0.383 0.117   
## 20 unselected MDD8 0.763 0.263   
## 21 unselected MDD4a 0.787 0.287   
## 22 unselected MDD6 0.836 0.336   
## 23 unselected MDD5b 0.0908 0.409   
## 24 unselected MDD5a 0.0343 0.466

Datasets

cohort\_sumstats <- basics\_aligned |>  
 transmute(cohort, reference, dataset=str\_match(Dataset, "([A-Za-z0-9\_]+)")[,2],  
 N\_cases, N\_controls, `N\_eff\_half`, `LAMBDA-GC`, `N-SNPs`) |>  
 filter(!is.na(N\_cases))  
  
meta\_sumstats <-  
basics\_meta %>%  
 mutate(cohort=str\_match(Dataset, "([A-Za-z\_]+)")[,2],  
 reference=str\_match(Dataset, "(MDD[0-9a-b]+)")[,2]) |>  
 select(cohort, reference, dataset=Dataset,  
 N\_cases, N\_controls, `N\_eff\_half`,  
 `LAMBDA-GC`, `N-SNPs`)  
   
sumstats <- bind\_rows(cohort\_sumstats, meta\_sumstats)  
   
write\_csv(sumstats, "meta/meta.csv")

Totals

cohort\_sumstats |> filter(cohort %in% c('PGC', 'AGDS', 'GenScot', 'janpy')) |>  
 mutate(N=N\_cases+N\_controls, cohort=if\_else(cohort=='PGC', true=sapply(str\_split(dataset, "\_"), last), false=cohort)) |>  
 group\_by(cohort) |>  
 summarize(N=max(N)) |>  
 ungroup() |>  
 summarize(N=sum(N))

## # A tibble: 1 × 1  
## N  
## <dbl>  
## 1 30148

cohort\_sumstats |> filter(cohort %in% c('ALSPAC', 'EstBB', 'UKBB')) |>  
 mutate(N=N\_cases+N\_controls) |>  
 group\_by(cohort) |>  
 summarize(N=max(N)) |>  
 ungroup() |>  
 summarize(N=sum(N))

## # A tibble: 1 × 1  
## N  
## <dbl>  
## 1 207436