All Weighted Plot Results

Alex Stern

2022-06-29

We used mappings from https://github.com/EBISPOT/EFO-UKB-mappings to organize our data based on the EFO ontology. First we filtered out any data in our original table that was rank-normalized (irnt) to prevent multiple rows with the same code. We then joined the ICD10 codes in our data frame to the codes in the EFO master file to associate them with the proper zooma query name. Finally, we joined the zooma query names of our updated data frame to the original zooma table which listed the type of trait for every trait in our table (disease, measurement, etc.). We then separated the traits into three different tables: diseases, quantitative traits, and others. For each of these tables, we plotted the mean increasing allele frequency for each phenotype in a histogram.

Stringent p-value (5e-8)

```
df <- fread("../output/results/all_weighted_5e-8_0.01.txt.gz")
master_file <- fread("../data/EFO/UK_Biobank_master_file.txt", header = TRUE)
zooma <- fread("../data/EFO/ukbiobank_zooma.txt", header = TRUE)
df <- separate(df, col=filename, into=c('code', 'rest', 'bothsexes', 'tsv'), sep='\\.') %>% select("cod
```

```
## Warning: Expected 4 pieces. Additional pieces discarded in 4539 rows [1, 2, 3,
## 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

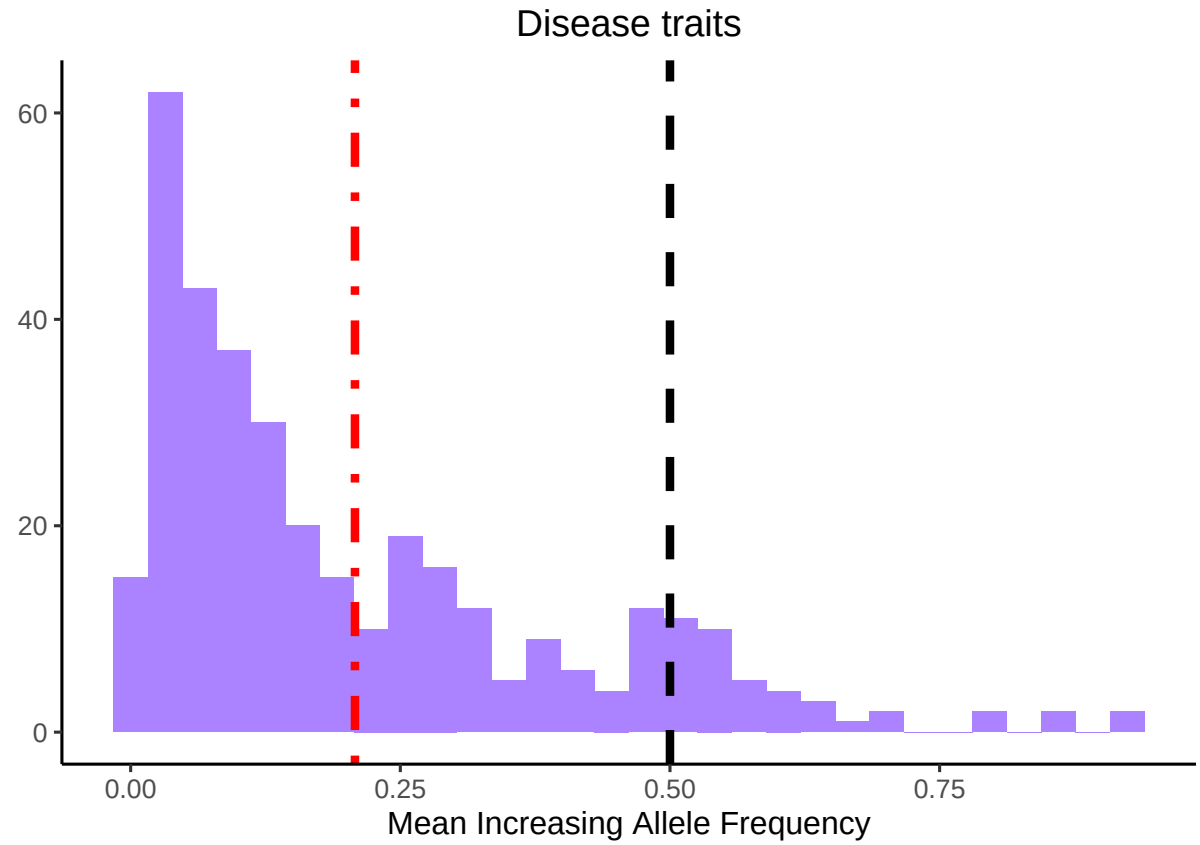
```
# drop measurement type irnt
df$measurement_type <- "N"
for(i in 1:nrow(df)) {
  id <- df[i,1]
  id_parts <- str_split(id, pattern="_")[[1]]
  if (length(id_parts) > 1) {
      if (id_parts[2] == 'irnt' | id_parts[2] == 'raw') {
        df[i,1] <- id_parts[1]
        df[i,'measurement_type'] <-  id_parts[2]
    }
  }
}

df <- filter(df, measurement_type != "irnt")

# merge with masterfile
df_master <- inner_join(df, master_file, by = c("code" = "ICD10_CODE/SELF_REPORTED_TRAIT_FIELD_CODE"))

# merge with zooma
df_merged <- inner_join(df_master, zooma, by = c("ZOOMA QUERY" = "PROPERTY_VALUE")) %>% distinct(code,

# seperate traits based on type
diseases <- filter(df_merged, PROPERTY_TYPE == "disease")
```

```
quantitative <- filter(df_merged, PROPERTY_TYPE == "measurement")

other_phenotypes <- filter(df_merged, (PROPERTY_TYPE != "disease") & (PROPERTY_TYPE != "measurement"))
```

```
# function to plot histogram
plot_data <- function(data, color_name, trait_type) {
  pl <- ggplot(data = data, mapping = aes(x = mean_iaf)) + geom_histogram(fill = color_name) + theme_cl
  return(pl)
}
```

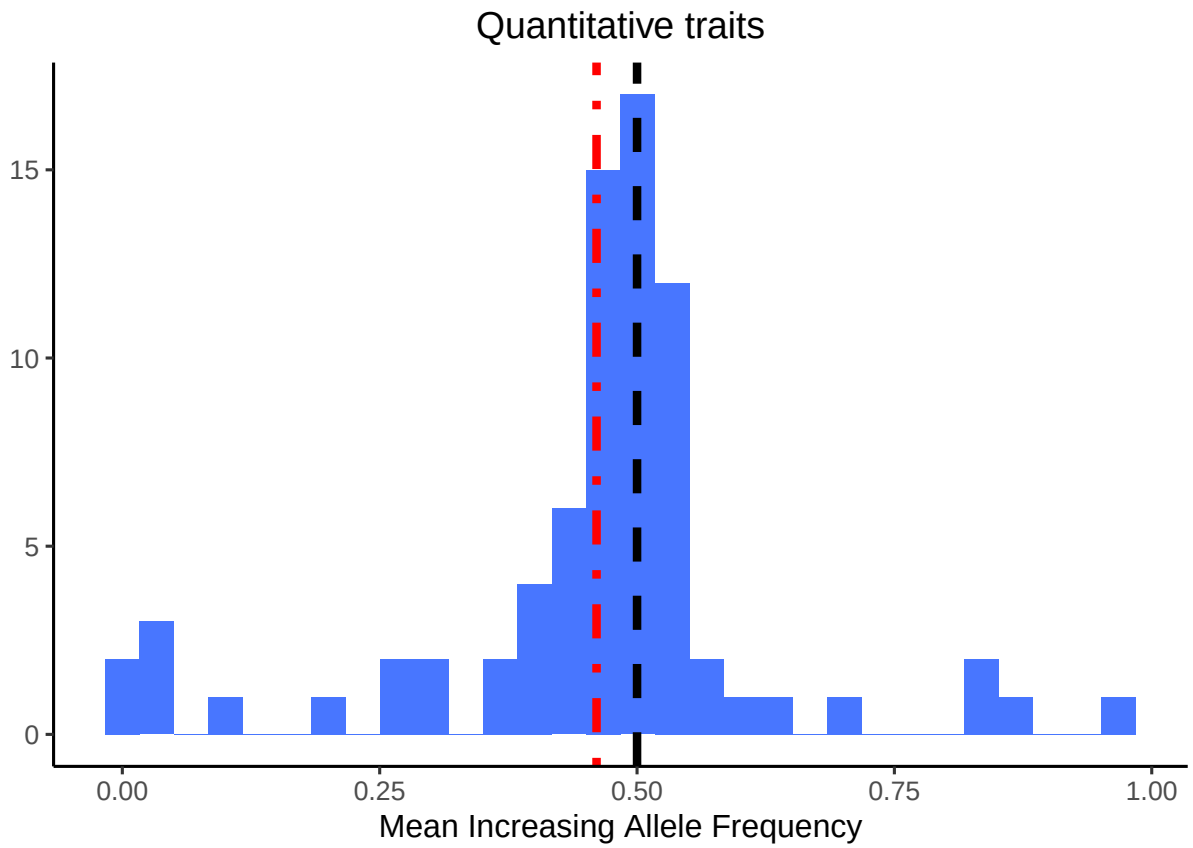
```
plot_data(diseases, "mediumpurple1", "Disease")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



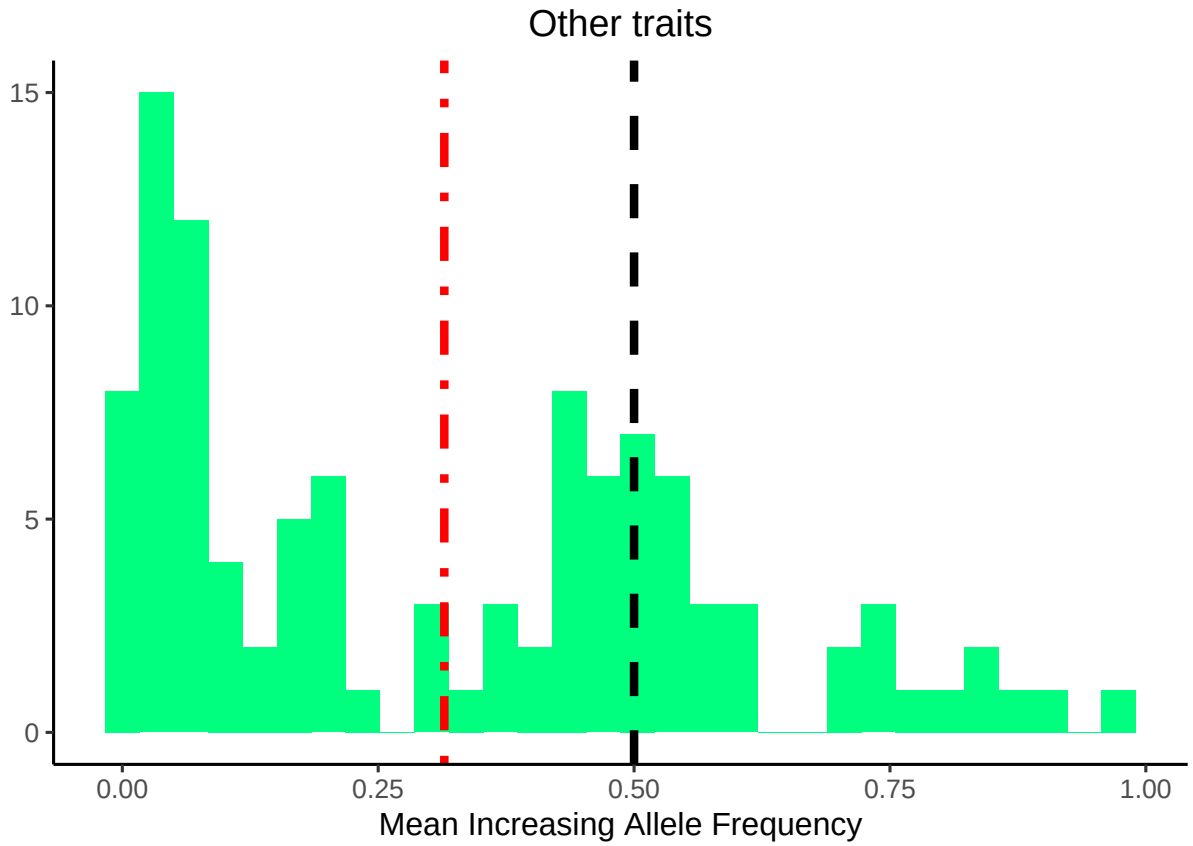

```
plot_data(quantitative, "royalblue1", "Quantitative")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
plot_data(other_phenotypes, "springgreen", "Other")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For the more stringent p-values, which are believed to have fewer spurious SNP's, the disease traits show a strong signal of mutational bias and the mean increasing allele frequency is less than 0.5. The quantitative traits in comparison have an average mean increasing allele frequency near 0.5. The miscellaneous traits show clustering around 0 and 0.5, suggesting a mix of disease and quantitative traits.

Lenient p-value threshold (1e-5)

```
getwd()
```

```
## [1] "/Users/alexstern/Desktop/ukbb_ss_pipeline/analysis"
```

```
df <- fread("../output/results/all_weighted_1e-5_0.01.txt.gz")
master_file <- fread("../data/EFO/UK_Biobank_master_file.txt", header = TRUE)
zooma <- fread("../data/EFO/ukbiobank_zooma.txt", header = TRUE)
df <- separate(df, col=filename, into=c('code', 'rest', 'bothsexes', 'tsv'), sep='\\.') %>% select("cod
```

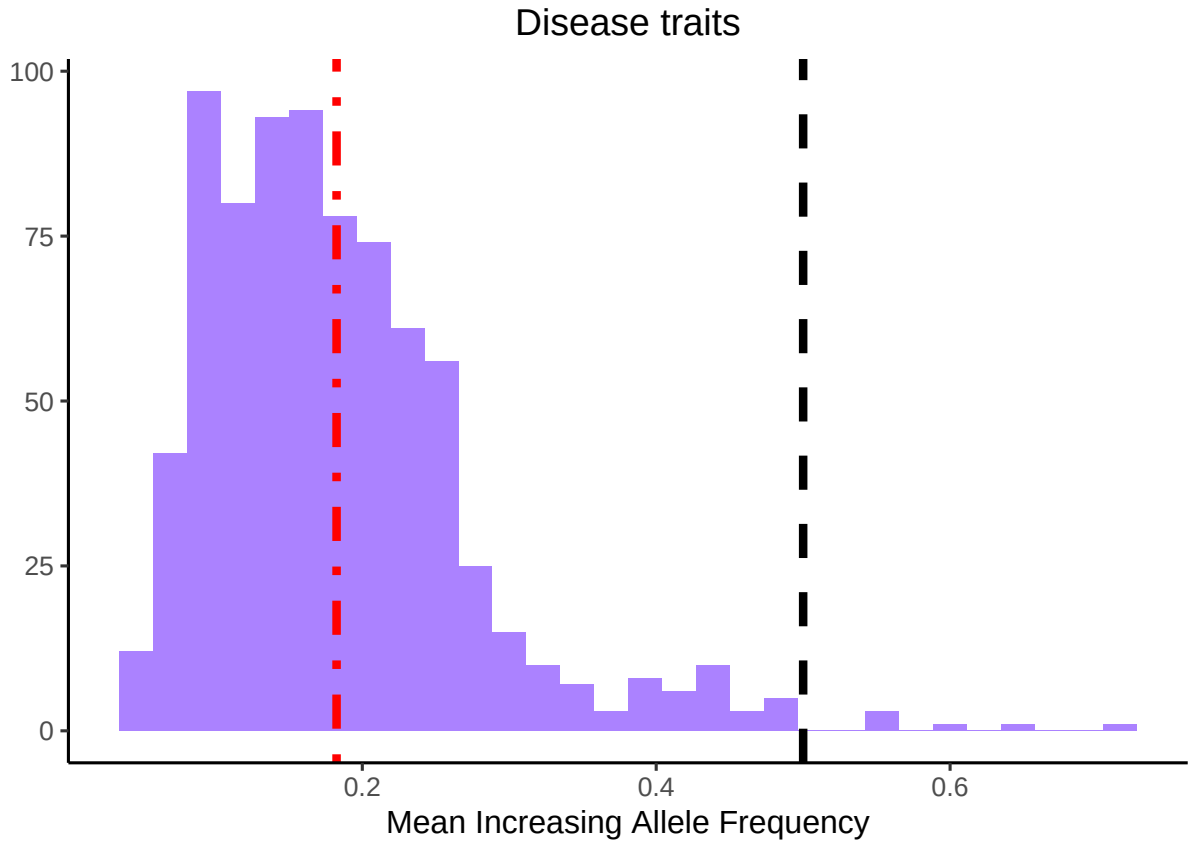
```
## Warning: Expected 4 pieces. Additional pieces discarded in 4539 rows [1, 2, 3,
## 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
# drop measurement type irnt
df$measurement_type <- "N"
for(i in 1:nrow(df)) {
  id <- df[i,1]
  id_parts <- str_split(id, pattern="_")[[1]]
  if (length(id_parts) > 1) {
      if (id_parts[2] == 'irnt' | id_parts[2] == 'raw') {
        df[i,1] <- id_parts[1]
        df[i,'measurement_type'] <-  id_parts[2]
```

```
    }
  }
}

df <- filter(df, measurement_type != "irnt")

# merge with masterfile
df_master <- inner_join(df, master_file, by = c("code" = "ICD10_CODE/SELF_REPORTED_TRAIT_FIELD_CODE"))

# merge with zooma
df_merged <- inner_join(df_master, zooma, by = c("ZOOMA QUERY" = "PROPERTY_VALUE")) %>% distinct(code,

# seperate traits based on type
diseases <- filter(df_merged, PROPERTY_TYPE == "disease")

quantitative <- filter(df_merged, PROPERTY_TYPE == "measurement")

other_phenotypes <- filter(df_merged, (PROPERTY_TYPE != "disease") & (PROPERTY_TYPE != "measurement"))
```

```
plot_data(diseases, "mediumpurple1", "Disease")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

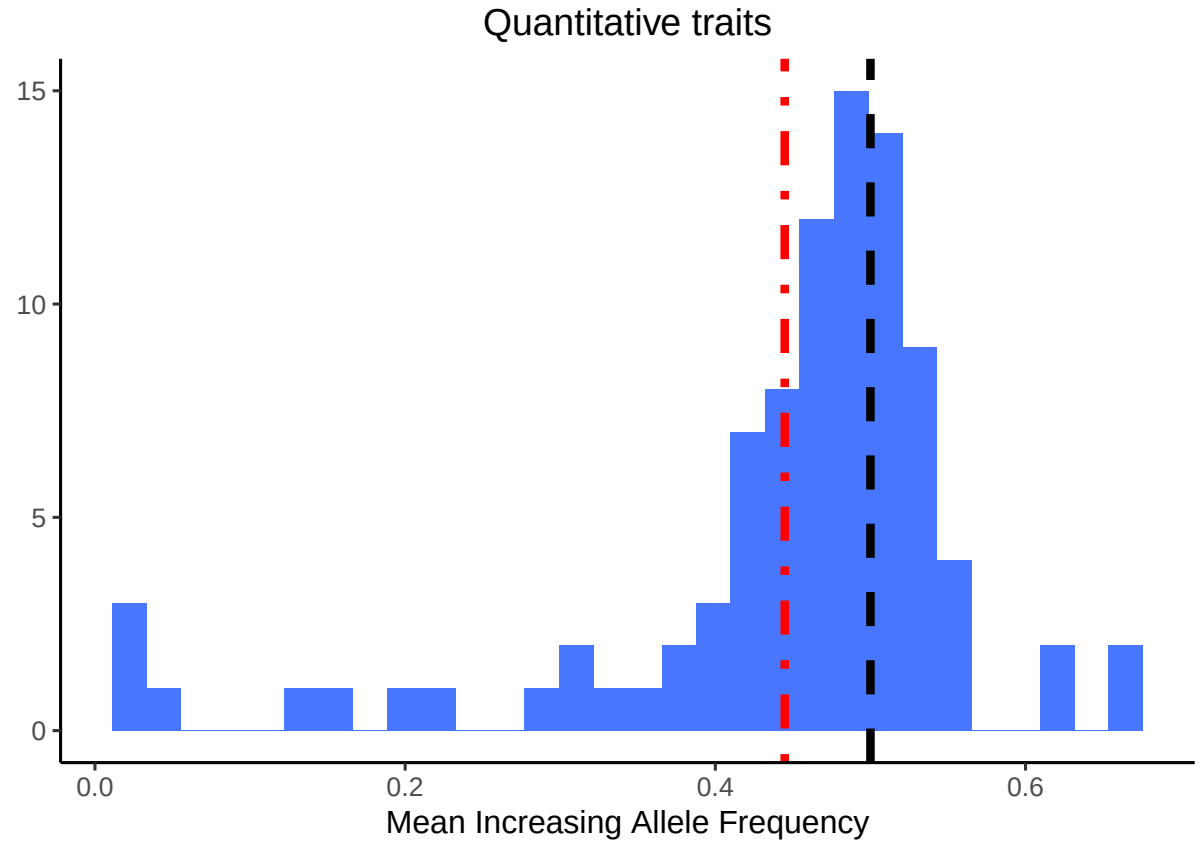



```
print(mean(diseases$mean_iaf))
```

```
## [1] 0.1824837
```

```
plot_data(quantitative, "royalblue1", "Quantitative")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

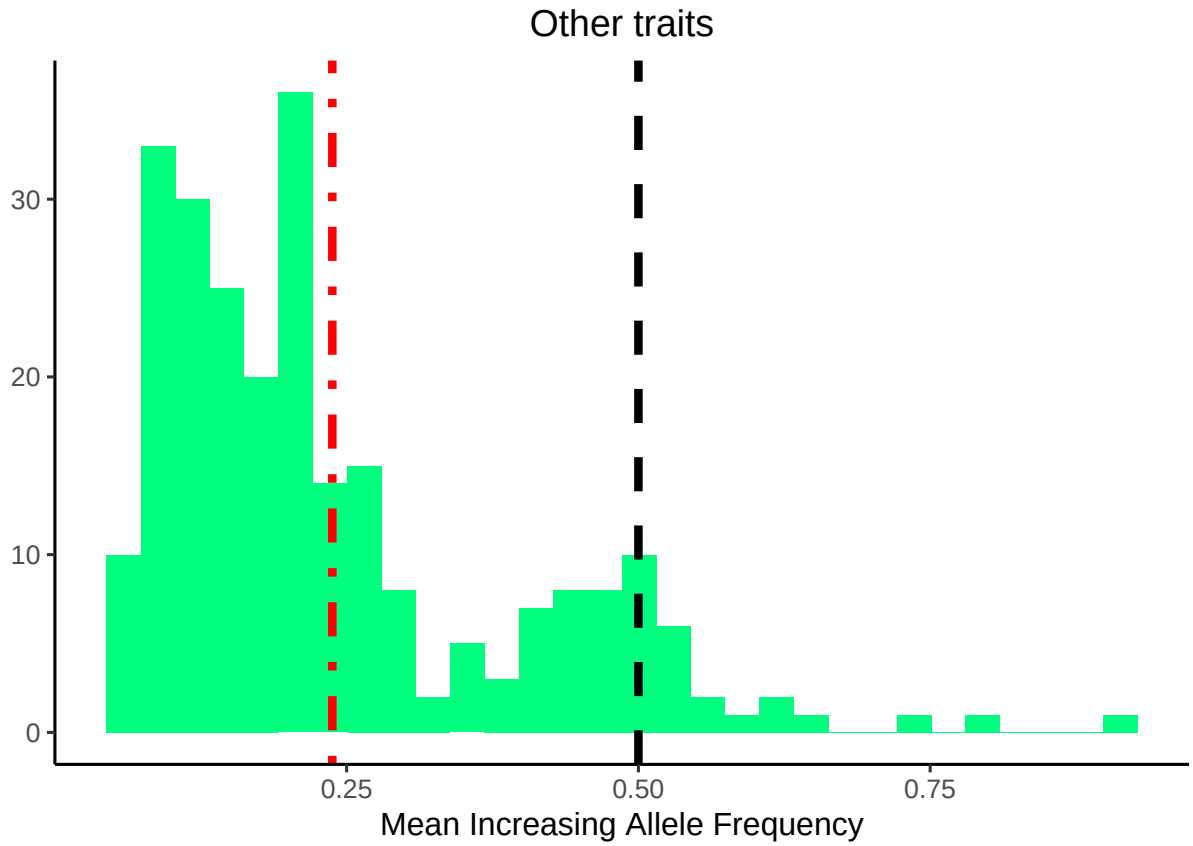



```
print(mean(quantitative$mean_iaf))
```

```
## [1] 0.4447198
```

```
plot_data(other_phenotypes, "springgreen", "Other")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For the more lenient p-value threshold, the average mean increasing allele frequency of the disease traits again seems to be much less than 0.5, but with a spike centered around 0.2. This may have to do with the increase of spurious associations and case-control bias. The quantitative traits still seem to have an average mean increasing allele frequency near 0.5.

Based on the data from both p-values and the presence of a spike in mean increasing allele frequency near 0.2 for disease traits which is not present in quantitative, there is evidence to suggest the case-control power imbalance is responsible for the spike at the lenient p-value threshold.