Plot Results

Alex Stern

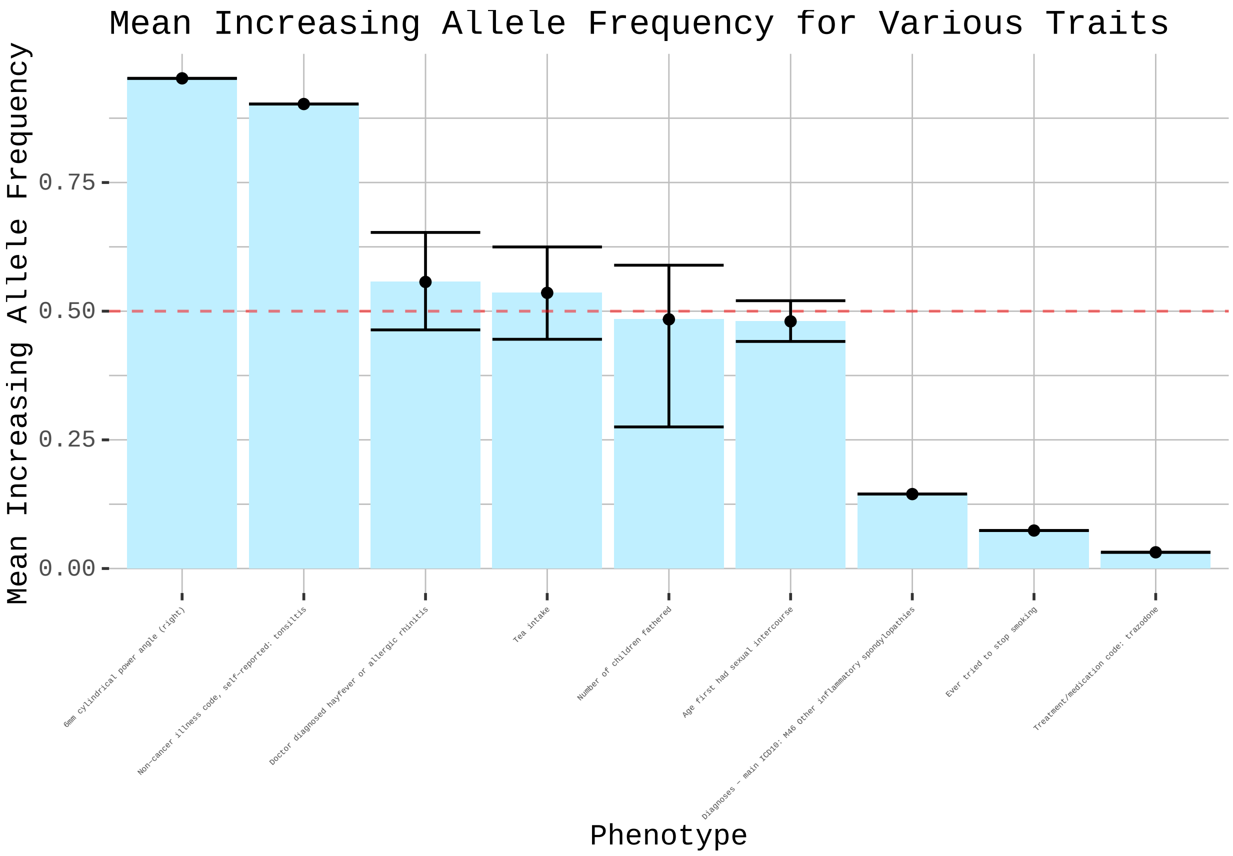
2022-04-28

Increasing Allele Frequency for Various Traits

This code builds off of a previous pipeline which takes GWAS data and returns the average increasing allele frequency (with confidence intervals) for various phenotypes. Using these output files, this code creates a bar graph that compares the mean increasing allele frequency for a select few of these traits. A p-value threshold of 5e-8 is used throughout this report to create a cutoff for which SNP's are considered significant.

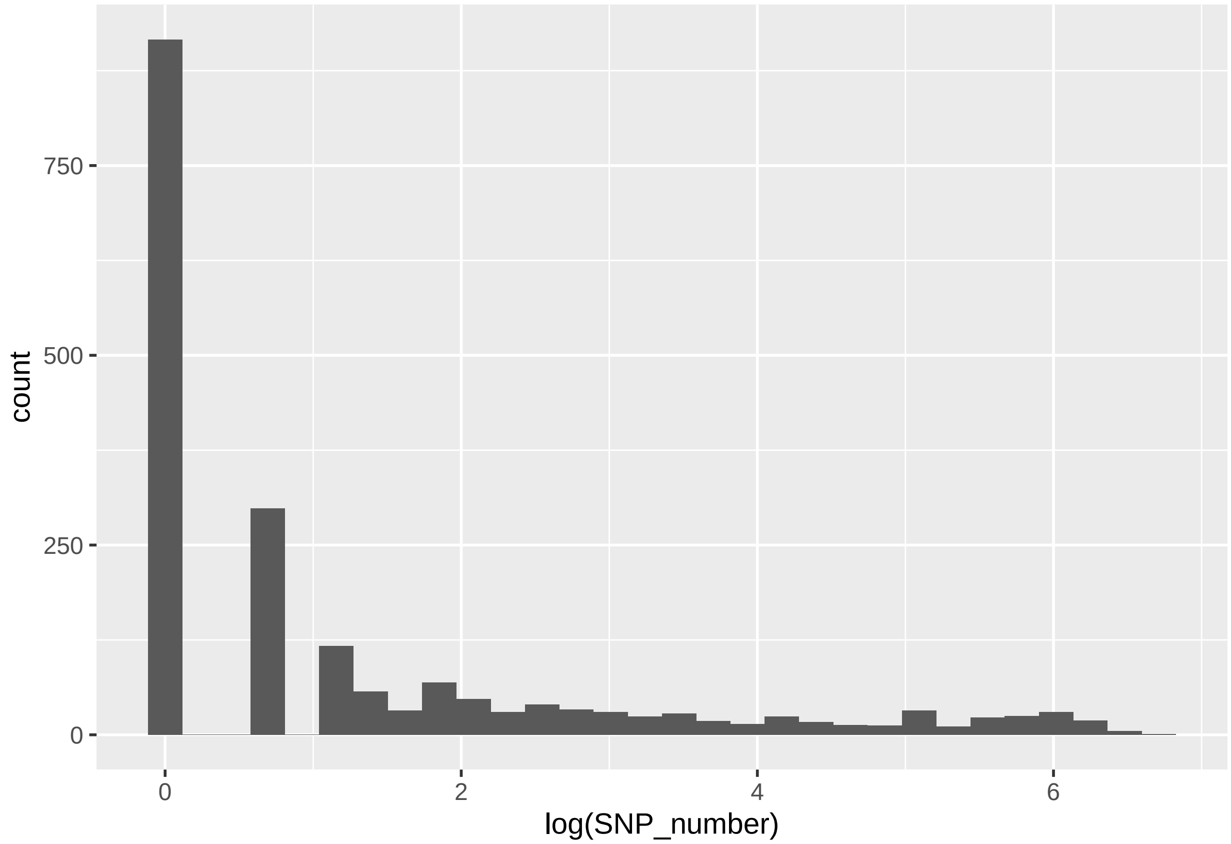
We first ran the pipeline on a small set of 10 phenotypes to demonstrate desired plots that will come from the broader dataset. The mean increasing allele frequency of the traits are plotted on the y-axis and each bar represents one phenotype.

```
# Read in data
df <- fread("output/results/small_run.txt")

# Remove rows with N/A
df <- df %>%
    drop_na(mean_iaf)

# Create a bar graph with IAF
# data
ggplot(data = df, mapping = aes(x = reorder(phenotype_name,
    -mean_iaf), y = mean_iaf)) +
    geom_col(fill = "lightblue1") +
    theme(axis.text.x = element_text(angle = 45,
        hjust = 1, size = 3)) +
    geom_errorbar(aes(y = mean_iaf,
        ymax = upper_ci, ymin = lower_ci)) +
    geom_point(aes(y = mean_iaf)) +
    geom_hline(yintercept = 0.5,
        alpha = 0.5, linetype = "dashed",
        color = "red") + ggtitle("Mean Increasing Allele Frequency for Various Traits") +
    xlab("Phenotype") + ylab("Mean Increasing Allele Frequency") +
    theme(plot.title = element_text(family = "mono"),
        axis.title.x = element_text(family = "mono"),
        axis.title.y = element_text(family = "mono"),
        axis.text.x = element_text(family = "mono"),
        axis.text.y = element_text(family = "mono"),
        panel.background = element_rect(fill = "white",
            colour = "white", size = 0.5,
            linetype = "solid"),
        panel.grid.major = element_line(size = 0.25,
            linetype = "solid",
            colour = "grey"), panel.grid.minor = element_line(size = 0.25,
            linetype = "solid",
            colour = "grey"))
```



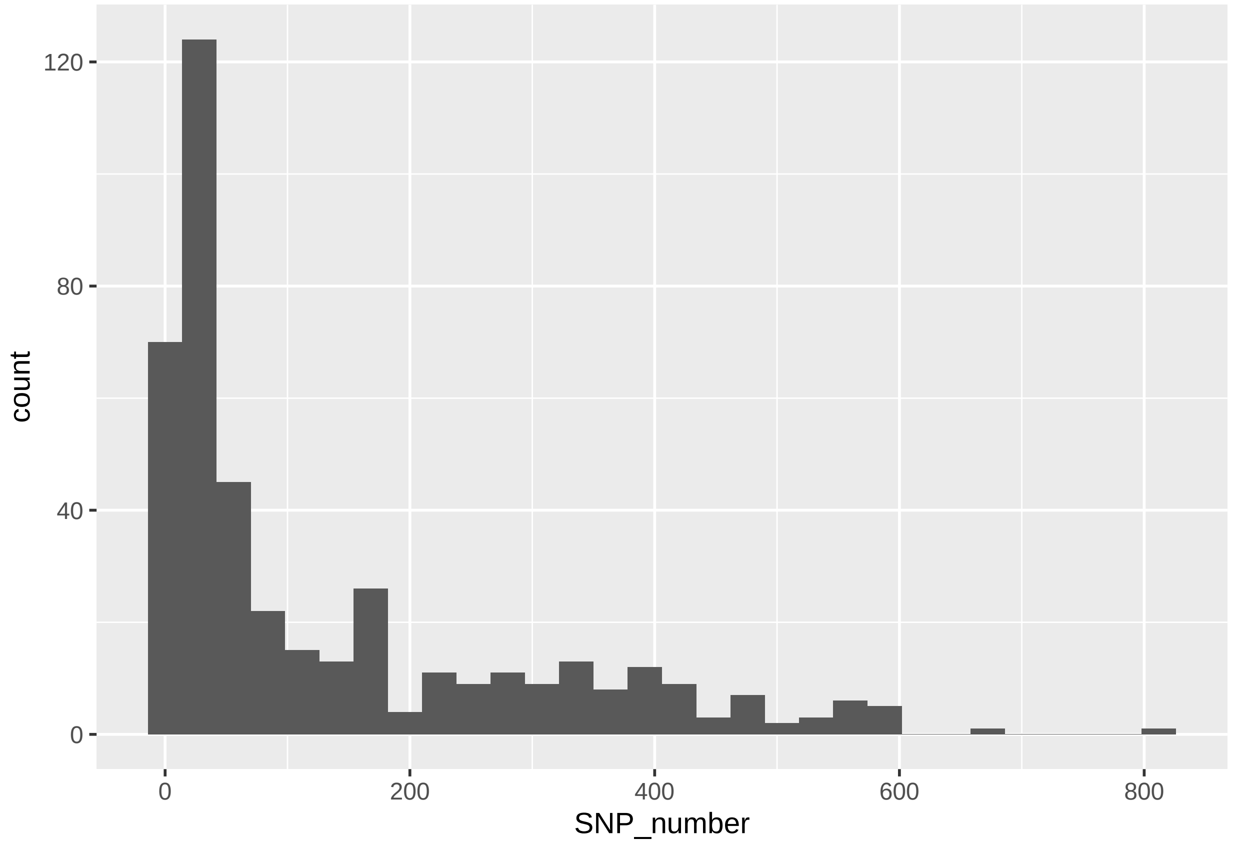
We then ran the pipeline on our larger data set from all of the UK Biobank data. Traits with no significant SNPs and therefore no data for mean increasing allele frequency were removed from the data set. We then plotted the number of significant SNPs associated with each trait on a logarithmic scale. As shown in the graph, most traits only had one significant SNP, meaning the increasing allele frequency data was not very informative due to the small sample size.

```
df_all <- fread("output/results/all_5e-8_0.01.txt.gz")
df_all <- drop_na(df_all, SNP_number)
df_all <- distinct(df_all, phenotype_name,
    .keep_all = TRUE)

ggplot(data = df_all, mapping = aes(x = log(SNP_number))) +
    geom_histogram()
```



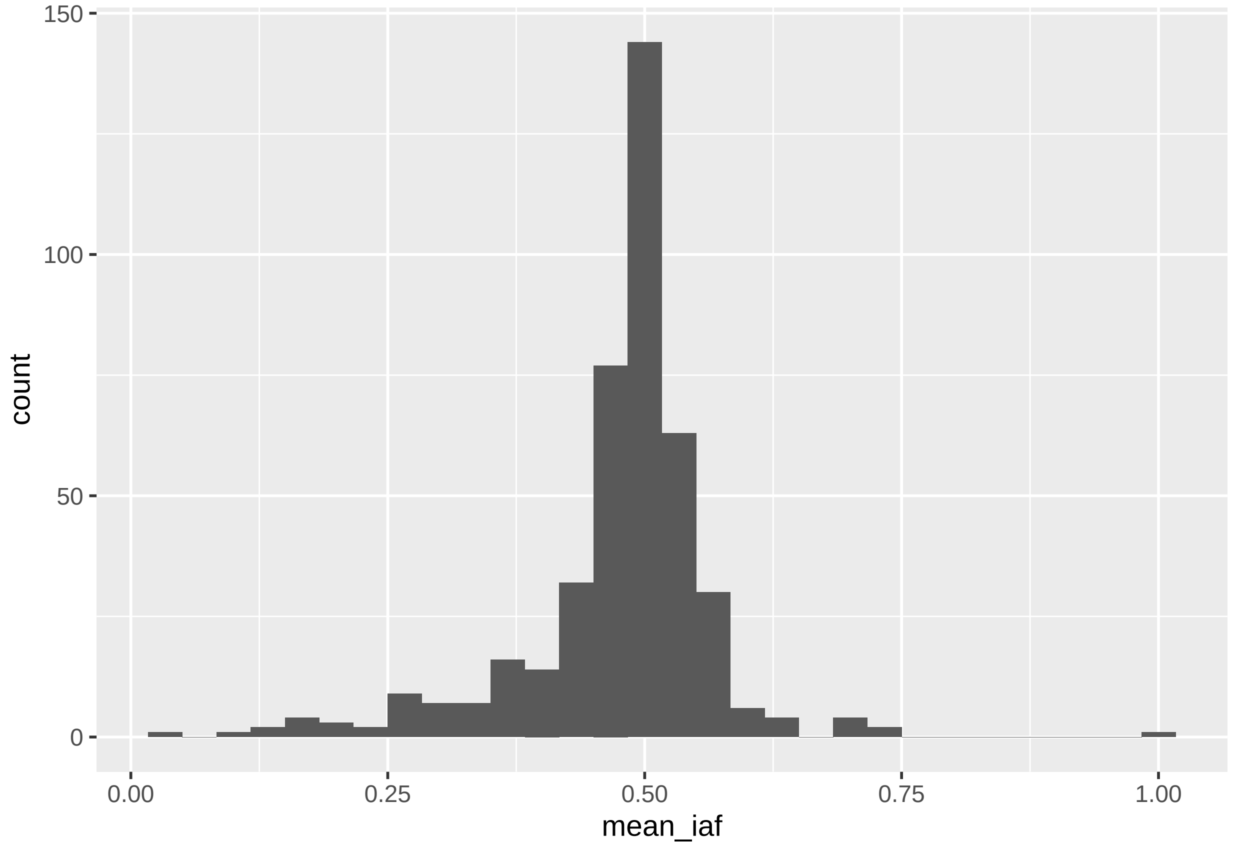
We then filtered out all traits with less than 10 significant SNPs. The SNP numbers were again plotted in a histogram, not on a logarithmic scale this time.

```
df_10 <- filter(df_all, SNP_number >=
    10)
ggplot(data = df_10, mapping = aes(x = SNP_number)) +
    geom_histogram()
```



To get an idea of the distribution of mean increasing allele frequencies, we plotted each traits mean iaf in a histogram. As the plot shows, most traits have an average increasing allele frequency of 0.5.

```
ggplot(data = df_10, mapping = aes(x = mean_iaf)) +
    geom_histogram()
```



We calculated the average of the mean iaf's for all of the traits with more than 10 SNP's.

```
mean_of_mean_iaf <- mean(df_10[,
    mean_iaf])
mean_of_mean_iaf
```

```
## [1] 0.4761936
```

The average is below 0.5, suggesting mutational bias towards increasing alleles.

We then filtered our data into two groups. The first group was of traits with a mean increasing allele frequency statistically significantly higher than 0.5. The second was of traits significantly lower than 0.5. Traits whose confidence intervals included 0.5 we removed entirely.

```
df_sig_dif <- filter(df_10, lower_ci >
    0.5 | 0.5 > upper_ci)
df_greater_0.5 <- filter(df_sig_dif,
    lower_ci > 0.5)
df_less_0.5 <- filter(df_sig_dif,
    0.5 > upper_ci)
```

```
num_traits <- nrow(df_10)
traits_greater <- nrow(df_greater_0.5)
percent_greater <- traits_greater/num_traits
traits_less <- nrow(df_less_0.5)
percent_less <- traits_less/num_traits
```

```
cat("There were ", traits_greater,
    " traits with mean iaf significantly higher than 0.5.
This is ",
    100 * round(percent_greater,
        4), "% of traits with 10 or more significant SNPs.")
```

```
## There were  32  traits with mean iaf significantly higher than 0.5.
## This is  7.46 % of traits with 10 or more significant SNPs.
```

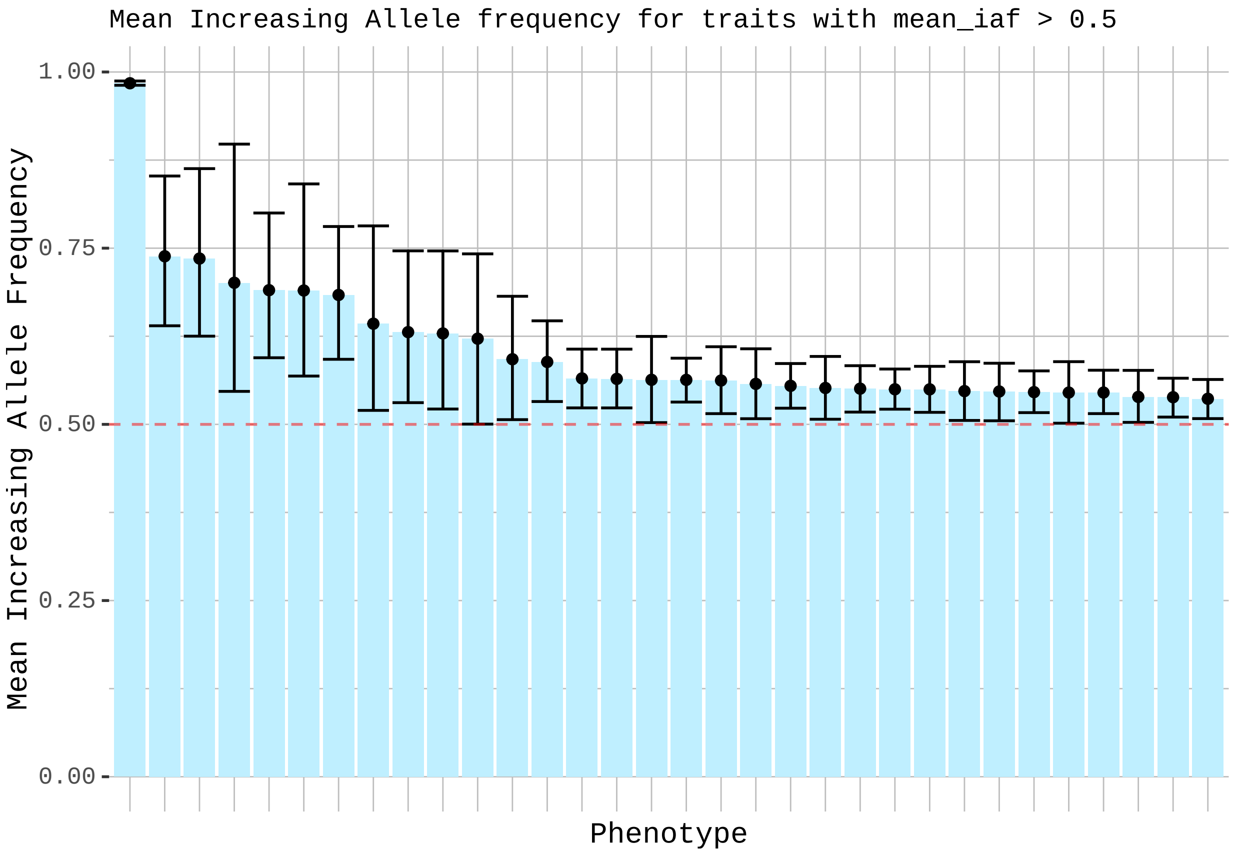
```
cat("There were ", traits_less,
    " traits with mean iaf significantly lower than 0.5.
This is ",
    100 * round(percent_less, 4),
    "% of traits with 10 or more significant SNPs.")
```

```
## There were  68  traits with mean iaf significantly lower than 0.5.
## This is  15.85 % of traits with 10 or more significant SNPs.
```

```
plot_data <- function(dat, title) {
    pl <- ggplot(data = dat, mapping = aes(x = reorder(phenotype_name,
        -mean_iaf), y = mean_iaf)) +
        geom_col(fill = "lightblue1") +
        theme(axis.text.x = element_text(angle = 45,
            hjust = 1, size = 3)) +
        geom_errorbar(aes(y = mean_iaf,
            ymax = upper_ci, ymin = lower_ci)) +
        geom_point(aes(y = mean_iaf)) +
        geom_hline(yintercept = 0.5,
            alpha = 0.5, linetype = "dashed",
            color = "red") + ggtitle(title) +
        xlab("Phenotype") + ylab("Mean Increasing Allele Frequency") +
        theme(plot.title = element_text(family = "mono",
            size = 10), axis.title.x = element_text(family = "mono"),
            axis.title.y = element_text(family = "mono"),
            axis.text.x = element_blank(),
            axis.ticks.x = element_blank(),
            axis.text.y = element_text(family = "mono"),
            panel.background = element_rect(fill = "white",
                colour = "white",
                size = 0.5, linetype = "solid"),
            panel.grid.major = element_line(size = 0.25,
                linetype = "solid",
                colour = "grey"),
            panel.grid.minor = element_line(size = 0.25,
                linetype = "solid",
                colour = "grey"))
    return(pl)
}
```

We then plotted the mean increasing allele frequency with error for all of the traits with mean iaf significantly greater than 0.5.

```
plot_data(df_greater_0.5, "Mean Increasing Allele frequency for traits with mean_iaf > 0.5")
```



The following table contains the phenotype names for all of the traits included in the above graph. They are sorted in the order they appear from left to right.

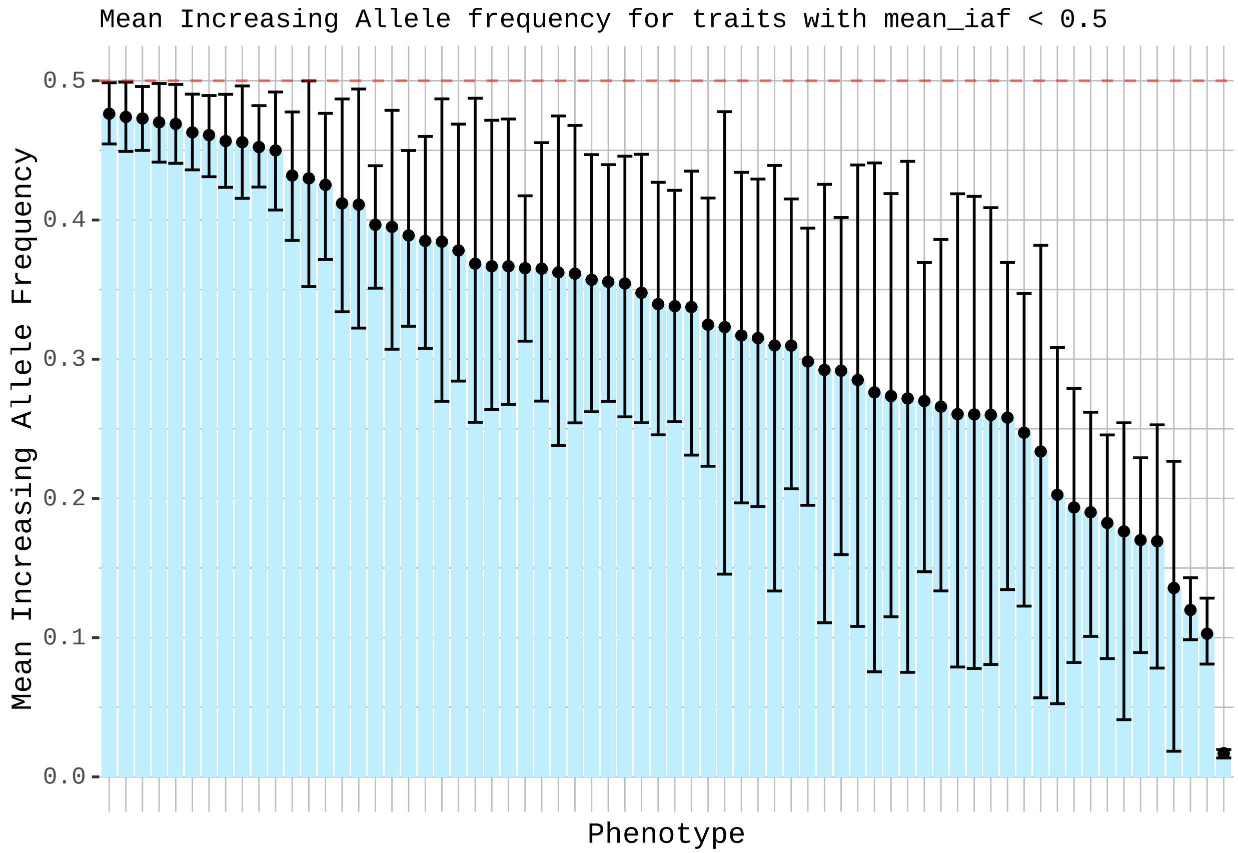
```
greater_table <- select(df_greater_0.5,
    c("mean_iaf", "SNP_number",
        "phenotype_name")) %>%
    arrange(desc(mean_iaf))
kable(greater_table)
```

mean_iaf	SNP_number	phenotype_name
0.9840106	14	Both eyes present: Yes
0.7384865	10	Non-oily fish intake
0.7352980	10	Types of physical activity in last 4 weeks: Light DIY (eg: pruning, watering the lawn)
0.7008749	12	Diagnoses - main ICD10: R31 Unspecified haematuria
0.6904252	17	Number of days/week of moderate physical activity 10+ minutes
0.6899265	13	Seen a psychiatrist for nerves, anxiety, tension or depression
0.6835833	18	Potassium in urine
0.6427946	15	Mouth/teeth dental problems: None of the above
0.6308838	17	Salad / raw vegetable intake
0.6289810	20	Diagnoses - main ICD10: N40 Hyperplasia of prostate
0.6215750	12	3mm cylindrical power (right)
0.5924121	21	Pain type(s) experienced in last month: None of the above
0.5885436	61	Hot drink temperature
0.5652180	182	Cholesterol (mmol/L)
0.5645054	180	Cholesterol (quantile)
0.5631863	51	C-reactive protein (mg/L)

mean_iaf	SNP_number	phenotype_name
0.5631169	290	HDL cholesterol (quantile)
0.5622314	133	Hair colour (natural, before greying): Dark brown
0.5574454	111	Glucose (quantile)
0.5546298	284	HDL cholesterol (mmol/L)
0.5516103	165	LDL direct (quantile)
0.5506903	268	Apoliprotein A (quantile)
0.5497812	329	Eosinophill count
0.5496864	260	Apoliprotein A (g/L)
0.5471828	117	Spherical power (right)
0.5466209	121	Spherical power (left)
0.5457667	303	Total protein (quantile)
0.5451159	170	LDL direct (mmol/L)
0.5450958	293	Total protein (g/L)
0.5389423	172	Hand grip strength (right)
0.5385134	372	Eosinophill percentage
0.5362948	382	Glycated haemoglobin (quantile)

Finally, we plotted the mean increasing allele frequency for traits with mean iaf significantly less than 0.5.

```
plot_data(df_less_0.5, "Mean Increasing Allele frequency for traits with mean_iaf < 0.5")
```



A table containing the data represented in the bar graph is included below.

```
less_table <- select(df_less_0.5,
    c("mean_iaf", "SNP_number",
        "phenotype_name")) %>%
   arrange(desc(mean_iaf))
kable(less_table)
```

mean_iaf	SNP_number	phenotype_name
0.4762992	436	Whole body fat mass
0.4740094	420	Red blood cell (erythrocyte) count
0.4729073	429	Arm fat mass (left)
0.4701164	336	High light scatter reticulocyte count
0.4689953	346	High light scatter reticulocyte percentage
0.4629142	343	Haemoglobin concentration
0.4609835	321	Reticulocyte percentage
0.4566873	214	Comparative body size at age 10
0.4558521	171	3mm strong meridian (right)
0.4523605	325	Reticulocyte count
0.4499477	95	Smoking status: Never
0.4319116	154	Hair colour (natural, before greying): Blonde
0.4298697	12	Number of operations, self-reported
0.4251808	103	Mean corpuscular haemoglobin concentration
0.4119039	37	Mouth/teeth dental problems: Dentures
0.4110226	20	Smoking status: Current
0.3965677	112	Non-cancer illness code, self-reported: hypothyroidism/myxoedema
0.3950779	50	Direct bilirubin (umol/L)
0.3888699	41	Irritability
0.3849221	27	Medication for pain relief, constipation, heartburn: None of the above
0.3843881	11	Leisure/social activities: Sports club or gym
0.3780944	34	Palmar fascial fibromatosis [Dupuytren]
0.3685937	10	Diseases of the musculoskeletal system and connective tissue
0.3667529	16	Coxarthrosis arthrosis of hip
0.3667529	16	Diagnoses - main ICD10: M16 Coxarthrosis [arthrosis of hip]
0.3653615	73	Treatment/medication code: levothyroxine sodium
0.3649800	31	Fibroblastic disorders
0.3623553	18	Cereal type: Other (e.g. Cornflakes, Frosties)
0.3614701	15	Diagnoses - main ICD10: C50 Malignant neoplasm of breast
0.3570304	17	Cancer code, self-reported: breast cancer
0.3555771	31	Frequency of tiredness / lethargy in last 2 weeks
0.3543538	30	Diagnoses - main ICD10: M72 Fibroblastic disorders
0.3476914	14	Medication related adverse effects
0.3395423	24	Diagnoses - main ICD10: I48 Atrial fibrillation and flutter
0.3380975	19	Disorders of lens
0.3374726	15	Non-cancer illness code, self-reported: migraine
0.3247638	15	Treatment/medication code: thyroxine product
0.3230537	12	Number of fluid intelligence questions attempted within time limit
0.3170245	20	Hair colour (natural, before greying): Red
0.3151430	14	Non-cancer illness code, self-reported: atrial fibrillation
0.3098694	10	Number of self-reported cancers
0.3097428	13	Medication related adverse effects (Asthma/COPD)
0.2983464	19	Cardiac arrhytmias, COPD co-morbidities
0.2922681	10	Cancer diagnosed by doctor
0.2917110	13	Cancer code, self-reported: basal cell carcinoma
0.2850355	12	Venous thromboembolism
0.2761649	10	Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor: Blood clot in the leg (DVT)
0.2735533	13	DVT of lower extremities and pulmonary embolism

mean_iaf	SNP_number	phenotype_name
0.2717945	10	Non-cancer illness code, self-reported: deep venous thrombosis (dvt)
0.2699631	14	Treatment/medication code: insulin product
0.2659029	20	Non-cancer illness code, self-reported: psoriasis
0.2605741	11	Other ILD-related CVD-co-morbidities
0.2603050	11	Diagnoses - main ICD10: I26 Pulmonary embolism
0.2599982	10	DVT of lower extremities
0.2580085	12	Started insulin within one year diagnosis of diabetes
0.2472024	13	Medication for cholesterol, blood pressure or diabetes: Insulin
0.2336803	12	Treatment/medication code: rosuvastatin
0.2025516	11	Endocrine, nutritional and metabolic diseases
0.1935038	12	Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Insulin
0.1900540	21	Non-cancer illness code, self-reported: malabsorption/coeliac disease
0.1823798	10	Non-cancer illness code, self-reported: sarcoidosis
0.1763187	11	Diagnoses - main ICD10: K90 Intestinal malabsorption
0.1700973	13	Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis
0.1691988	10	Treatment/medication code: warfarin
0.1356605	11	Non-cancer illness code, self-reported: rheumatoid arthritis
0.1198611	10	Never eat eggs, dairy, wheat, sugar: Wheat products
0.1028306	11	Coeliac disease
0.0169908	34	Distance between home and job workplace

Based on the data, it seems that traits are approximately slightly biased towards a mean_iaf of below 0.5. This matches theoretical expectations regarding mutational biases towards traits with lower mean_iaf.