SSGAC data task

kindle zhang

April 2025

1 Summary

This is a data task from SSGAC. I used `R` as the main tool, with an `.Rmd` file (`ssgac_data_task.rmd`) as the code script, and Overleaf with LaTeX syntax to generate the PDF for this task.

The assignment consists of two questions in total. Below are my answers; for detailed code, please refer to `ssgac_data_task.rmd`.

The PDF file is designed to answer the questions using concise language and clear figures, while the R Markdown file provides additional details and implementation steps.

2 Question 1

2.1 clean up data set

1. SNPs with nucleotide values `(A, T, G, or C)` mistakenly recorded in the position column are corrected by shifting the values into their appropriate columns.
2. For SNPs where the `beta_hat` field contains chromosome counts instead of effect size estimates, we correct the misaligned columns by shifting the values into their appropriate positions.
3. The `A1` and `A2` alleles should each be one of `"A"`, `"C"`, `"G"`, or `"T"`. The minor allele frequency (MAF) should lie between 0 and 0.5. In addition, `A1` should be different from `A2`.
4. Identify SNPs with a large sample size ($N > 70{,}000$), statistical significance at the 95% confidence level ($p < 0.05$), and an accuracy greater than 90% (e.g., INFO score > 0.90).
5. Check SNPs where the sign of the Z-score and the beta coefficient are opposite. If the signs are inconsistent, a flip correction should be applied.
6. Check for SNPs where the chromosome is not 22 or where SNPs are duplicated. For duplicated SNPs, retain only the entry with the highest INFO score.

Here is a table showing how many snps are changed or deleted in each steps:

No. Step	A	B
1	701	701
2	345	345
3	2152	2070
4	7023	7498
5	0	0
6	1	0

Table 1: number of snps changed in each steps

Finally, we have 825 snps left in Trait A and 432 snps left in Trait B.

2.2 find common snp in Trait A and Trait B

Now we get two data sets cleaned, and we can find the common SNPs in two traits.

There were a total of 48 SNPs shared between the two traits. However, two of them—`rs12072405` and `rs4662139`—had inconsistent alleles (`A1` and `A2`) between the two traits, which could not be explained by strand flips or reference/alternate allele swaps. To be cautious, we excluded these two SNPs from further analysis.

As a result, we retained 46 SNPs that are common between the two datasets. Among them, the SNP with the largest absolute Z-score in dataset A is `rs214342`.

3 Question 2

3.1 produce the Q-Q plot

3.2 answer the question

1. **interpret the Q-Q plot:**

   The 45-degree line in the QQ plot represents the distribution of p-values we would expect under the global null hypothesis—i.e., if none of the SNPs are associated with the trait. In this analysis, deviations from the line suggest inflation or enrichment of small p-values.

   Specifically, the p-values for Trait B show a marked upward deviation from the 45-degree line, indicating strong evidence of association between SNPs and the trait. In contrast, Trait A follows the null line more closely, suggesting fewer or weaker signals of association.

2. **why it should be monotonic**

   Q plots are inherently monotonic because they plot quantiles of two distributions, both of which are sorted in ascending order. As a result, each

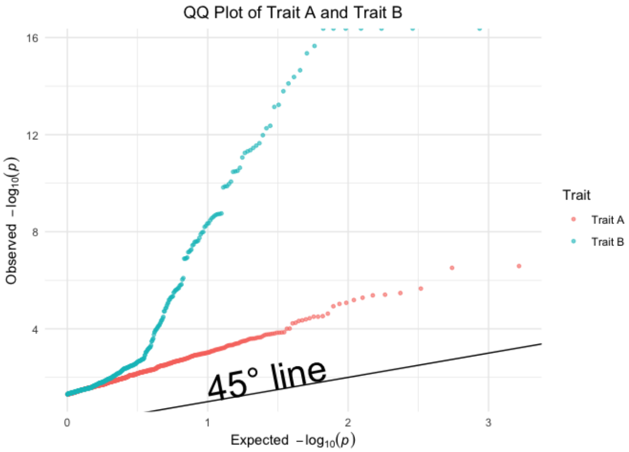

Figure 1: Q-Q plot between CDF from empirical GWAS and null hypothesis

successive point must have equal or higher x and y values than the previous one, making the plot monotonically increasing regardless of the specific data being compared.

3. **describe the trend of different traits**

   In the QQ plot, the data for Trait A closely follows the 45-degree line, indicating that most SNPs do not deviate from the null distribution. This suggests few, if any, SNPs are significantly associated with Trait A.

   In contrast, Trait B shows a substantial upward deviation from the 45-degree line, especially in the tail, which indicates strong enrichment of small p-values. This suggests that many SNPs are likely associated with Trait B, and the genetic signal is stronger compared to Trait A.

4 Discussion

1. In this analysis, the p-values used for plotting were derived from the dataset after an initial round of cleaning, rather than from the original raw data. Notably, one of the cleaning steps involved filtering for SNPs that were already significant. As a result, the distribution of p-values is biased toward smaller values, which may explain why the Q–Q plot does not align closely with the 45-degree reference line in its early (non-significant) range. This is a limitation of the current approach and could be improved in future analyses by avoiding selection based on significance prior to generating the Q–Q plot.

2. The data cleaning procedures can be encapsulated in a separate script and bundled into a package to enhance productivity and reproducibility.