HW3

Submitted by: Assaf Lovton & Adi Falach

ID: 209844414 & 323859231

Question 1:

1.1:

A research paper is based on original research.The kind of research may vary depending on your field or the topic (experiments, survey, interview, questionnaire, etc.), but authors need to collect and analyze raw data and conduct an original study. On the contrary, a review paper (also called review article) is based on other published articles. Review articles generally summarize the existing literature on a topic in an attempt to explain the current state of understanding on the topic.

1.2:

The impact factor (IF) is a measure of the frequency with which the average article in a journal has been cited in a particular year. It is used to measure the importance or rank of a journal by calculating the times it's articles are cited. The calculation is based on a two-year period and involves dividing the number of times articles were cited by the number of articles that are citable.

1.3:

A review paper:

> Five years of GWAS discovery. Visscher PM, Brown MA, McCarthy MI, Yang J. Am J Hum Genet.2012 Jan 13.

A research paper:

> GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. Yano K, Morinaka Y, Wang F, Huang P, Takehara S, Hirai T, Ito A, Koketsu E, Kawamura M, Kotake K, Yoshida S, Endo M, Tamiya G, Kitano H, Ueguchi-Tanaka M, Hirano K, Matsuoka M. Proc Natl Acad Sci US 2019 Oct 15;

1.4.1:

> A.am j hum genet impact factor-10.502
>
> B.Proc Natl Acad Sci U S A impact factor-9.412

1.4.2:

A.There is one author's affiliation.

B.There are five author's affiliations.

1.5.1:

The trait that has been studied was the "Rice architecture". Rice architecture is a complex trait affected by plant height, tillering, and panicle morphology. In further detail in this study PCA was performed on 8 typical traits related to plant architecture, including days-to-heading, culm length, panicle number, and 5 panicle-related traits (panicle length, rachis length, primary branch number per panicle, secondary branch number per panicle, and spikelet number per panicle).

1.5.2:

According to the SI appendix- " Two japonica rice panels comprising 169 and 133 varieties were collected from various sites in Japan (1, 2), and grown in Togo Field, Field Science Center, Nagoya University (Dataset S1 and S3). DNA preparation and genotyping were conducted as previously described (3). In total, 381,007 SNPs and 58,886 INDELs were identified in the 169 set, and 1,417,019 SNPs and 231,064 INDELs were found in the 133 set. The sequence data was deposited in the DDBJ Sequence Read Archive (DRA) under accession numbers DRA004358 and DRA008452."

From references:

1. Z. Hashimoto et al., Genetic diversity and phylogeny of Japanese sake-brewing rice as revealed by AFLP and nuclear and chloroplast SSR markers. Theor Appl Genet. 109, 1586–1596 (2004).
2. K. Ebana, Y. Kojima, S. Fukuoka, T. Nagamine, M. Kawase, Development of mini core collection of Japanese rice landrace. Breed Sci. 58, 281–291 (2008).
3. K. Yano et al., Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 48, 927–934 (2016).

1.5.3:

The results were positive, the researched trait was a "complex trait" and researched by taking into consideration 8 typical traits related to plant architecture, the results were that PCA can be good indicators for plant architecture and heading date.

From the results section:

"For results in 2015, PC1 explained 62% of the trait variance (Fig. 1A). **Except for the traits of days-to-heading and panicle number, the other 6 traits showed high positive loadings on PC1 (0.80–0.92)** , while panicle number showed negative loading (−0.54) (Fig. 1 A and B). This result suggested that **plants with high PC1 scores exhibited long culms, large panicle sizes, and small panicle numbers, and vice versa.** This corresponds to a trade-off relationship between panicle number and panicle weight. PC2 explained 16% of the total variance, and the loading on PC2 was high for days-to-heading

(0.83) (Fig. 1A), suggesting that **PC2 is representative of days-to-heading.** This component was also loaded (0.55) with panicle number, which is **consistent** with the observation that prolonged vegetative growth due to late heading increases the number of panicles per plant. The PCA results for traits measured in 2014 were consistent with 2015 results (SI Appendix, Fig. S3), suggesting that PC1 and PC2 **can be used as quantitative indices to characterize plant architecture and heading date**, respectively."

From the results section under GWAS for PC Scores:

"These results **supported our hypothesis that PC1 and PC2 could be good indicators for plant architecture and heading date,** respectively."

From the results section under GA Signaling Regulates Rice Plant Architecture paragraph:

"Further analyses revealed that the revertant had a single nucleotide polymorphism (SNP) designated C863T in SLR1, where A288 was substituted with V in the LHR1 domain (SI Appendix, Fig. S17). **This polymorphism suppressed the function of intact SLR1 in a semidominant manner, resulting in enhanced GA responsiveness and growth rate (SI Appendix, Fig. S18)."**

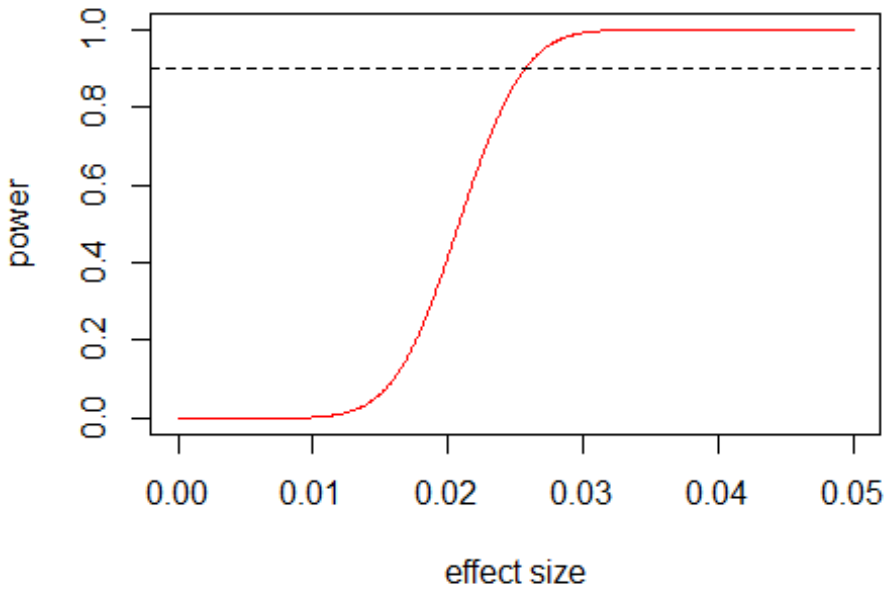
The final citation is found under the discussion section:

"In this study, PCA on 8 architectural traits revealed that PC1 captured 62% of variations for most traits, whereas PC2 captured 16% of variations that primarily impacted days-to-heading (Fig. 1 A and B); thus, **PC1 is a good indicator for plant architecture.** Using the PC scores for GWAS, **we identified significant peaks associated with PCs;** these included genes previously reported for regulating plant architecture along with other peaks that were considered novel. The peak with the strongest effect on PC1, the most important index for plant height and panicle structure (Fig. 1C), was further investigated. **Genetic studies confirmed that OsSPY is a causal gene for this peak and responsible for plant architecture.** OsSPY functions as a negative regulator in GA signaling by enhancing the suppressive function of DELLA proteins (13, 15). Thus, we considered GA signaling to be a major mechanism regulating plant architecture. This was confirmed by studies using 9 isogenic plant types showing different levels of GA signaling. In general, **PCA for these isogenic plants was very similar to the GWAS panel except for the days-to-heading trait;** this might have been caused by allelic variation in heading genes that are only present in the GWAS panel."

1.5.4:

```
f = 0.1 #MAF
b.alt= seq(0, 0.05 ,0.0001)
sigma = sqrt(1-2*f*(1-f)*b.alt^2)
ns = 381207  #candidate values for n
ses= sigma/sqrt(ns*2*f*(1-f)) #SEs corresponding to each candidate n
q.thresh= qchisq(5e-8, df = 1, ncp= 0, lower = F) #chi-sqrthreshold correspal
pha=5e-8
pwr= pchisq(q.thresh, df = 1, ncp=(b.alt/ses)^2, lower=F) #power at alpha=5e-
8 for VECTOR of SE values
```

```
plot(b.alt, pwr, col ="red", xlab="effect size", ylab = "power", t="l", lwd=
1.5)
abline(h = 0.9, lty= 2)
```




```
minimal_effect_size<-b.alt[min(which(pwr>=0.9))]
minimal_effect_size
```

```
## [1] 0.0258
```

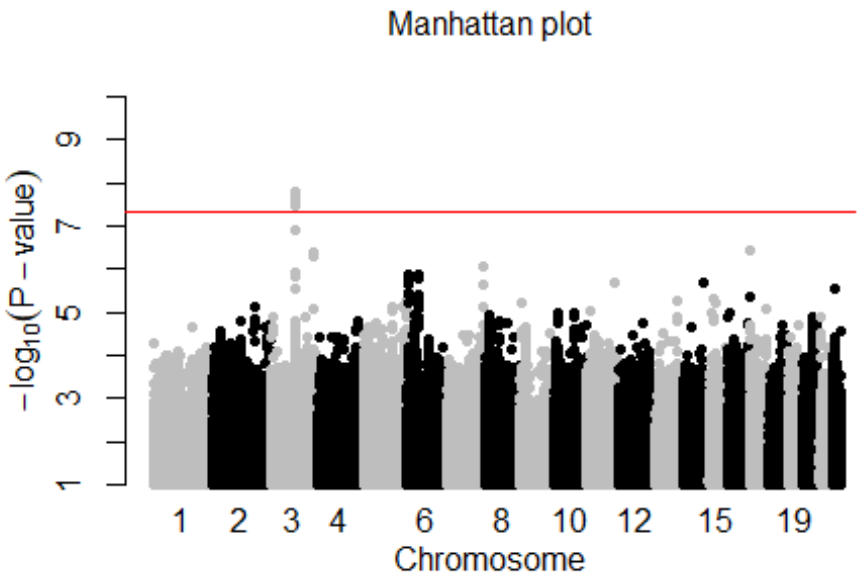
##Question 2:

2.1:

The name of the paper is “Meta-analysis of genome-wide association studies of anxiety disorders”.

2.2:

```
library(data.table)
library(fastman)
data<-fread("anxiety.meta.full.cc.tbl.gz")
#summary(data)
#Genomic locations were based on NCBI build 37/UCSC hg 19 data.
fastman(data, chr = "CHR", ps = "BP",p = "P.value", main = "Manhattan plot",
  suggest_line = FALSE, gws_line = -log10(5e-08),
  color = c("grey", "black"), chr_build = "GRCh37", yscale = NA,
  xlab_all = F, turbo = TRUE)
```



2.2.1:

The most significant SNPs are located in chromosome 3 as we can see in the Manhattan plot.

2.2.2:

Yes, according to the definition of LD we saw in the totorial-LD is a measure of non-random association between alleles at different location at the same chromosome in a given population. We know that SNPs are in LD when the frequency of association of their alleles is higher than expected under random assortment. We can see in the Manhattan plot that these SNPs (the significant ones-those above the red line) are located closely to each other and they are all on the same line parallel to the y axis, therefore we can assume that they are in LD.

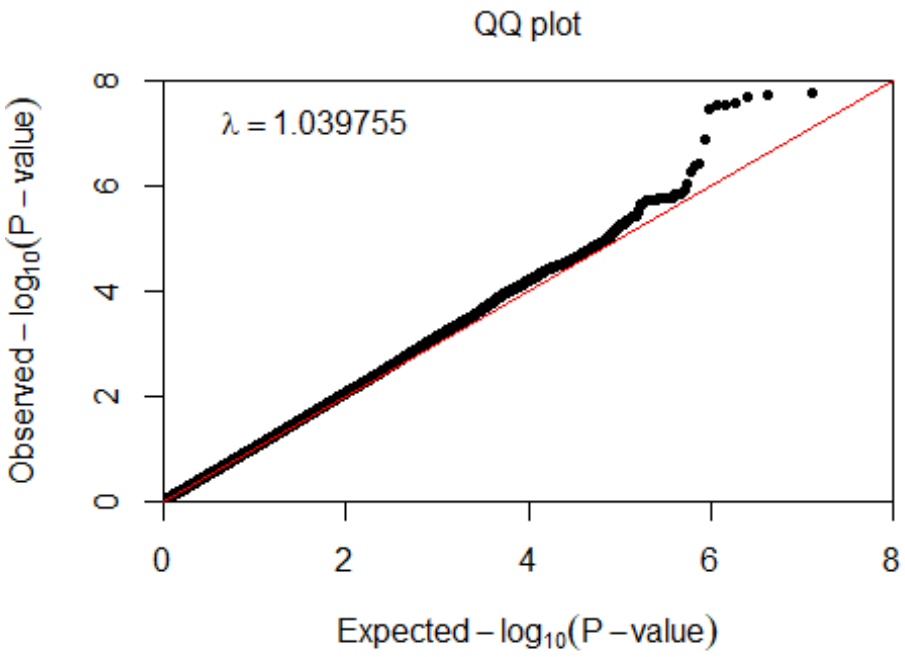
2.2.3:

rs1709393- we can see in our data that the most significant SNP is rs1709393 because he has the lowest p-Value. rs1709393 is located in the gene “LINC02085”. The gene is intron variant, therefore it is a non protein-coding gene.

2.3:

We think the results show that the analysis is good, since only part of the snp’s are in association with the trait (those at the right side of the plot). If we have gotten a curve that is all the way above the line therefore we would go and perform the additional QC.

```
fastqq(data,p="P.value", main = "QQ plot")
```



2.2.4:

supported in the zip file.