University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Final Exam (Practice #2) SOLUTIONS

Question 1. Pathway analysis where the input is the list of 'siginificant' genes is based on
(A) The Binomial distribution
(B) The Negative Binomial distribution
(C) The Normal distribution
(D) The Hypergeometric distribution ⟵ **THIS ONE**
(E) The Geometric distribution

Question 2. Which of the following are pathway enrichment online servers? Circle all that apply.
(A) DAVID ⟵ **THIS ONE**
(B) STRING ⟵ **THIS ONE**
(C) IGV
(D) ENRICHR ⟵ **THIS ONE**
(E) ENSEMBL

Question 3. We model pathway enrichment with the hypergeometric test. What assumptios are we making implicitly? Circle all that apply. (A) Genes are expressed independently ⟵ **THIS ONE**
(B) Pathways are disjoint sets
(C) All genes are in some pathway.
(D) There are enough genes on the list to justify a normal approximation.

Question 4. What does "GO" stand for in "GO Analysis"? (A) Gene Ontology ⟵ **THIS ONE**
(B) Gene Overrepresentation
(C) Genetic Overlay
(D) Gamma Oligo
(E) Gamut Ontology

Question 5. True or False. Gene Set Enrichment Analysis (GSEA) performs a different random walk for each gene set.

ANSWER: True.

Question 6. True or False. The "kernel trick" of dimensionality reduction involves first embedding the data into an even higher dimensional space, before projecting.

ANSWER: True.

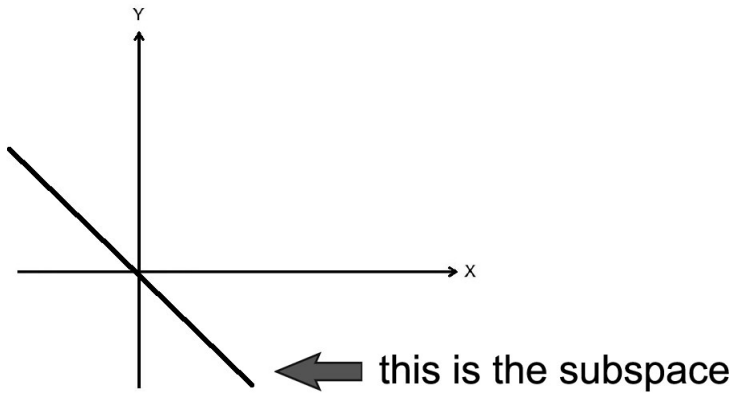
Question 7. True or False. For single cell RNA-Seq the standard methods of dimensionality reduction are (circle all that apply).

(A) UMAP ⟵ **THIS ONE**
(B) STAR
(C) t-SNE ⟵ **THIS ONE**
(D) PCA
(E) Phasing

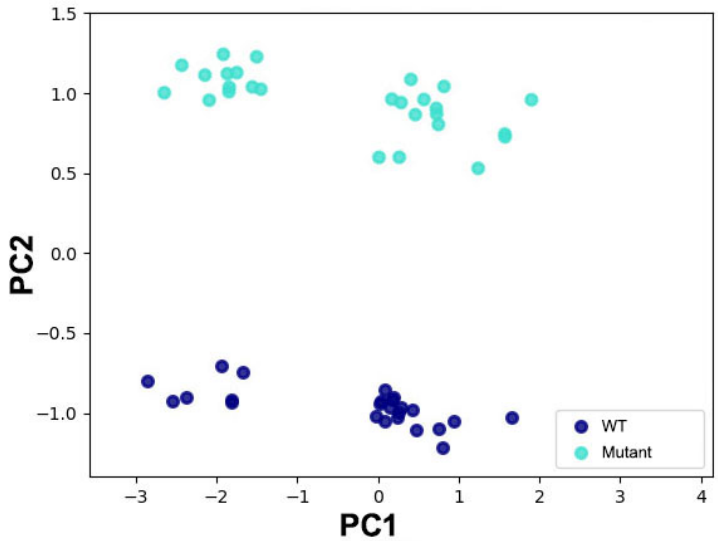
Question 8. True or False. In Principle Components Analyis PC2 is perpendicular to PC1 in the original high-dimensional space.

ANSWER: True.

Question 9. Draw the subspace that corresonds to loadings of $+1$ for X and -1 for Y.



Question 10. Suppose you have RNA-Seq data from WT and Mutant and you get the following PCA plot. Withuot having any more information, speculate on what could be driving the separation observed in PC2.



ANSWER: PC2 is pretty clearly driven by the genotype. So the biology.

Question 11. True or False. A Permutation test cannot declare significane of a 2-versus-2 comparison no matter what the data.

ANSWER: True. There are not enough permutations p for $1/p$ to be less than 0.05

Question 12. True or False. The Mann-Whitney test is for independent observations and the Wilcoxon Signed-Rank test is for repeated measures.

ANSWER: True.

Question 13. The data shows 7 observations of Condition 1 and 9 observations of condition 2. Is there a permutation that would raise the T-statistic higher in absolute value than it is on the unpermuted data? Explain.

Condition 1	Condition 2
23.2	112.91
14.13	197.12
82.29	99.81
47.01	132.94
22.82	217.17
32.31	189.94
39.97	101.4
	57.13
	416.27

ANSWER: Yes, swap the highest one in column 1 82.29 with 51.13 in the second column. That would have to change the ranking, in fact after doing that swap the highest value in column one is lower than the lowest in column two. Therefore, no further swapping could increase it further.R

Question 14. Suppose you are considering whether or not to do a non-parametric test in place of a parametric T-test, when there are 20 replicates per group.

(A) It's more important to consider the normality assumption than equal variance assumption.
(B) It's more important to consider the equal variance assumption than the normality assumption. ← **THIS ONE**
(C) It's equally important to consider the normality assumption as the equal variance assumption.
(D) The decision is not based on normality or variance.

ANSWER: It's (B) because once we have enough replicates for a T-test, we don't have to worry about normality. But we always have to worry about variance.

Question 15. The table below shows all 20 possible rankings of a 3-versus-3 comparision for a Mann-Whitney analysis. The table is split over two rows since it was too wide to display on one. Each ranking is equally likely, so each has probability $1/20 = 0.05$. What is the *two-sided* p-value for an observed value of $R = 7$? (*Note: Make sure to calculate the two-sided and not one-sided p-value*)

Cond.1	1,2,3	1,2,4	1,2,5	1,2,6	1,3,4	1,3,5	1,3,6	1,4,5	1,4,6	1,5,6
Cond. 2	4,5,6	3,5,6	3,4,6	3,4,5	2,5,6	2,4,6	2,4,5	2,3,6	2,3,5	2,3,4
R	6	7	8	9	8	9	10	10	11	12

Cond.1	2,3,4	2,3,5	2,3,6	2,4,5	2,4,6	2,5,6	3,4,5	3,4,6	3,5,6	4,5,6
Cond. 2	1,5,6	1,4,6	1,4,5	1,3,6	1,3,5	1,3,4	1,2,6	1,2,5	1,2,4	1,2,3
R	9	10	11	11	12	13	12	13	14	15

ANSWER: The p-value for $R = 7$ is the two-sided tail probability, so we have to add up the probabilities for $R = 6$ and $R = 7$ for the left tail and $R = 14$ and $R = 15$ for the right tail. Each has probability $1/20$. So $\frac{1}{20} + \frac{1}{20} + \frac{1}{20} + \frac{1}{20} = \frac{1}{5}$

Question 16. The term 'eQTL' stands for:

(A) Expanded Quality Template Locus
(B) Expected Quantitative Time Line
(C) Expression Quantitative Trait Locus ⟵ **ANSWER:**
(D) Expressed Query To Locus
(E) Evaluation of Quantitized Traits in Longtitude

Question 17. True or False. A Fisher Exact test is based on the hypergeometric distribution.

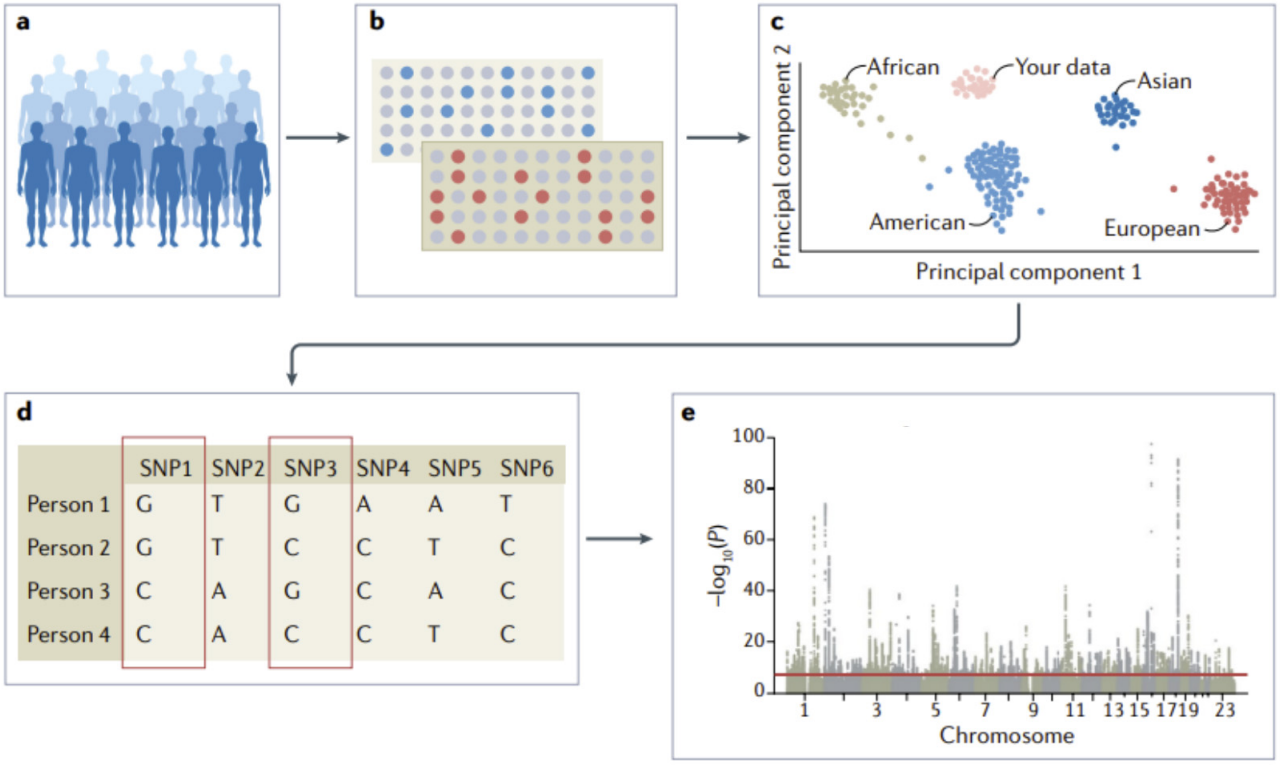
ANSWER: True.

Question 18. Give one mechanistic reason a SNP might be highly associated to a phenotype besides changing an amino acid.

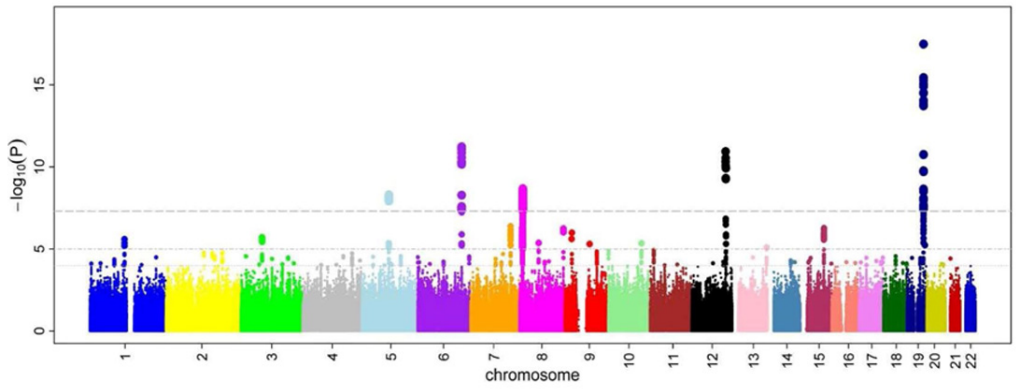
ANSWER: It could alter gene regulation.

Question 19. In the GWAS workflow shown below, the names of the five steps are listed. Write the letter (a through e) of the step from the diagram next to its corresponding item on the list.

- Imputation **D**
- Association Testing **E**
- Data Collection **A**
- Genotyping **B**
- Quality Control **C**

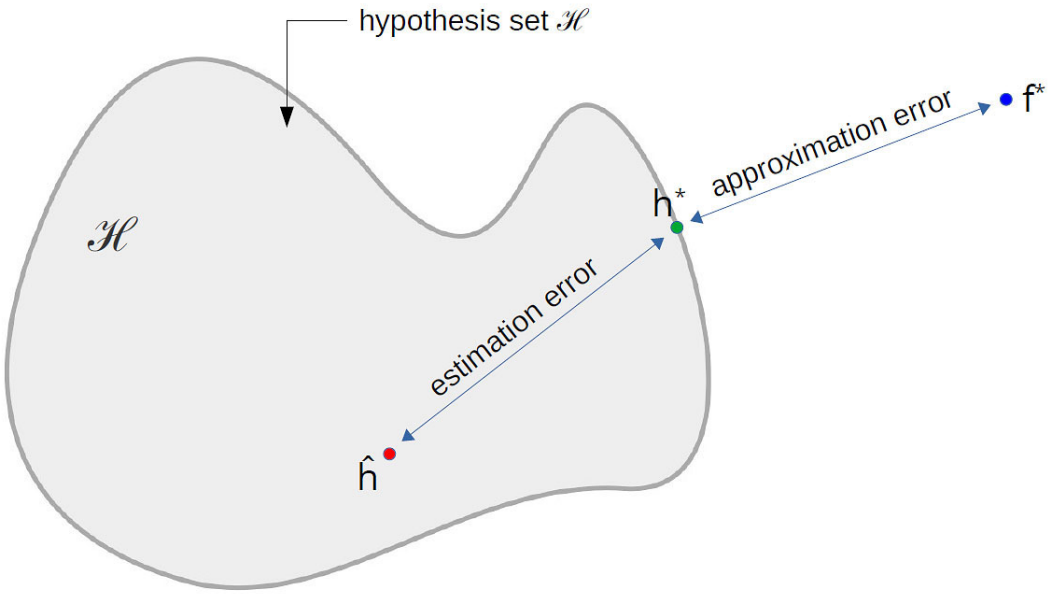


Question 20. In a Manhattan plot such as this, what do the dashed lines represent?



ANSWER: The dashed lines represent significance cutoffs. In other words thresholds for p-values or q-values to be called significant.

Question 21. For $\hat{h}$, h^* and f^* as in the figure, which out-of-sample error is the smallest?
(A) $E_{\text{out}}(\hat{h})$
(B) $E_{\text{out}}(h^*)$
(C) $E_{\text{out}}(f*)$ ⟵ **THIS ONE**



ANSWER: The out of sample error on f^* is lowest, since it's the true error of the the true model.

Question 22. In the left-hand side of the Learning Inequality, what part is random? Circle all that apply.
(A) $E_{\text{in}}(h)$ ⟵ **ANSWER:**
(B) $E_{\text{out}}(h)$
(C) ϵ

$$P\left(\left|E_{in}(h)-E_{out}(h)\right|>\varepsilon\right) \leqslant 2\left|\mathscr{H}\right|\cdot e^{-2n\varepsilon^2}$$

Question 23. In unsupervised clustering, we do not have training or test data where the truth is known. True or False, we can still evaluate the results using external sources of information.

ANSWER: True. We can use pathway enrichment for example.

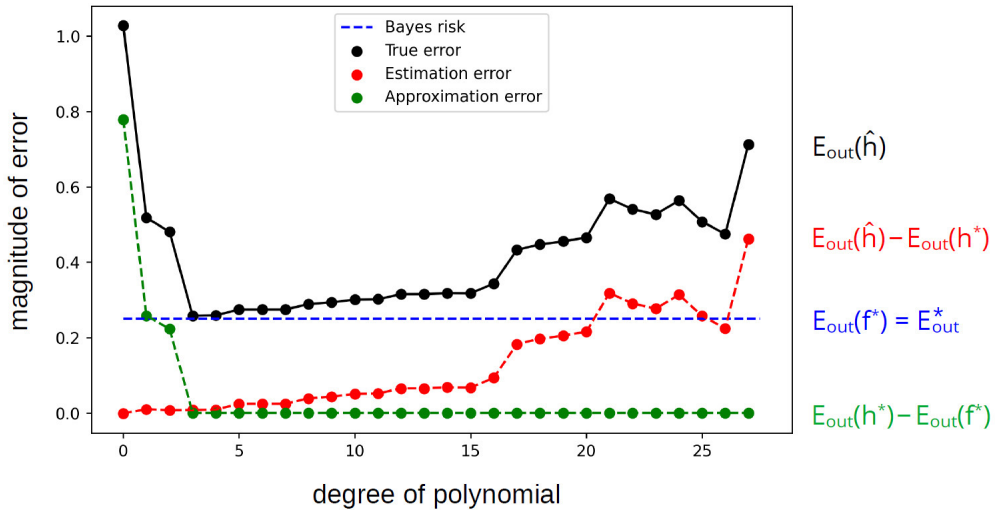
Question 24. Put the correct item in the correct place in the legend

(A) Estimation Error

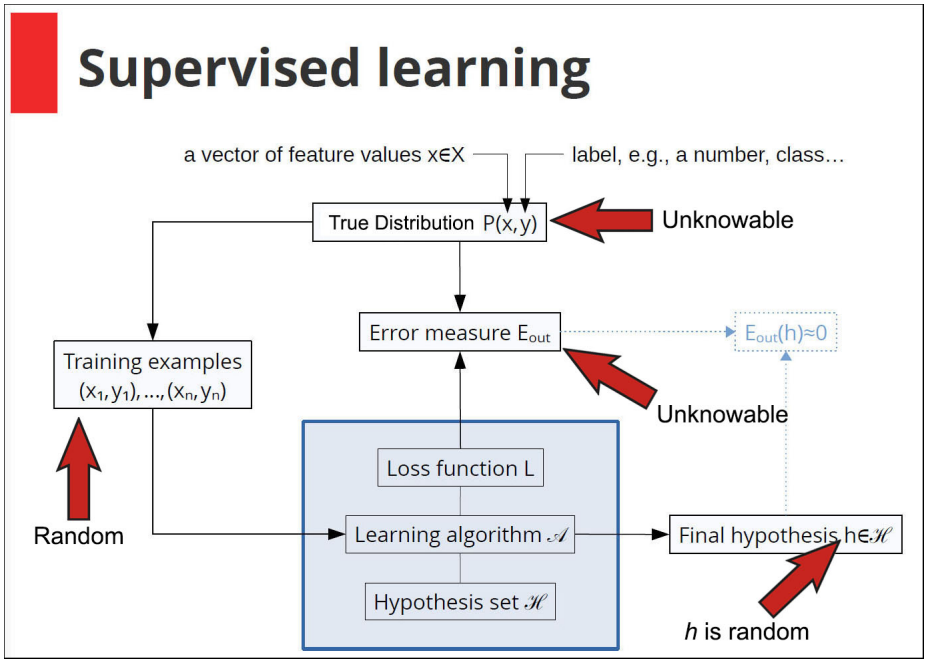
(B) Approximation Error

(C) True Error

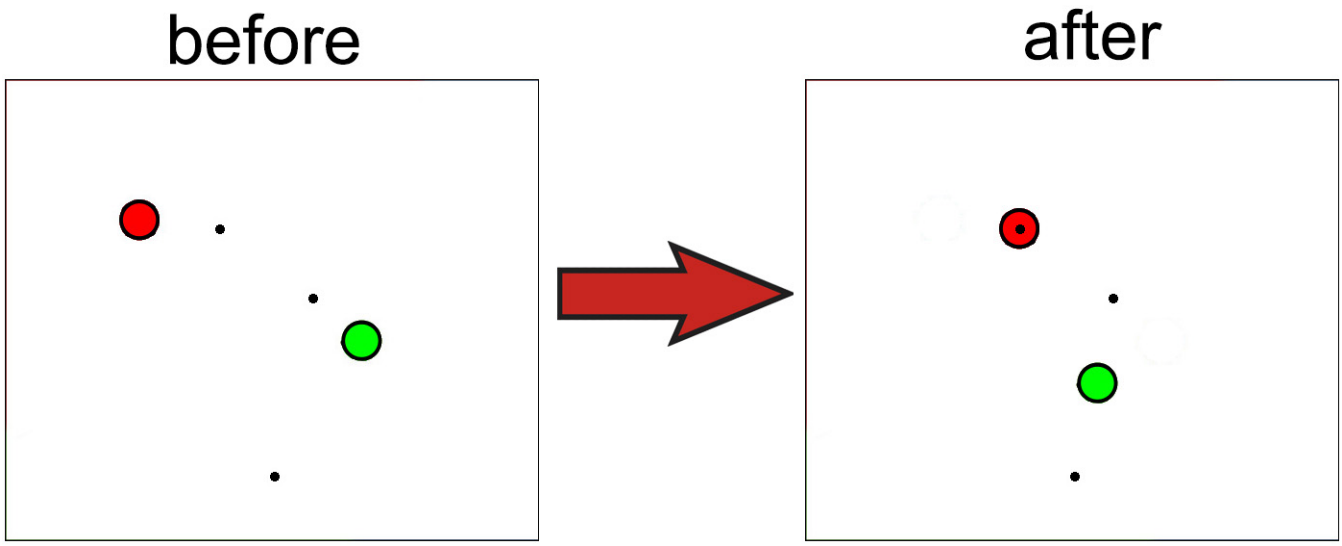
(D) Bayes' Risk



Question 25. On the diagram of supervised learning below, indicate one thing that is unknowable and one thing that is random.



Question 26. In the figure, there are three data points and two centroids undergoing k-means clustering (where $k = 2$). Indicate where the two centroids will move to in the next iteration.



Question 27. Which R command displays the help documenation for the `summary()` function?
(A) `?summary` ⟵ **ANSWER:**
(B) `base::summary()`
(C) `summary(help)`
(D) `help |> summary()`

Question 28. Which of the following `geom_` functions are designed to dsiplay the distributions of continous data (circle all that apply)?
(A) `geom_boxplot()` ⟵ **ANSWER:**
(B) `geom_bar()`
(C) `geom_violin()` ⟵ **ANSWER:**
(D) `geom_histogram()` ⟵ **ANSWER:**

Question 29. This code uses the `left_join` function to merge the 'airline' tibble to the 'flights' tibble:

```
flights |>
    left_join(airlines,
              by = join_by(carrier))
```

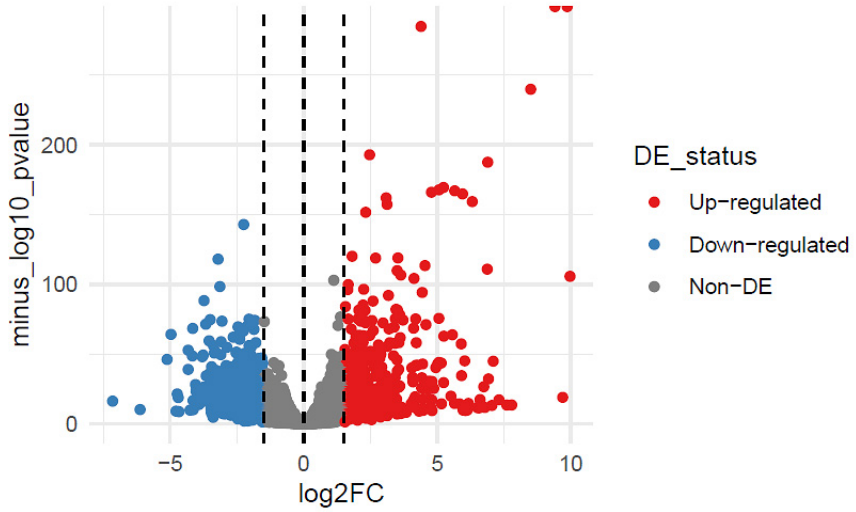
What is the name of the column used to line up data from both tibbles?

ANSWER: *carrier*

Question 30. Consider the 'de_results' data frame of differential expression results:

```
# A tibble: 6 x 6
  gene_id             log2FC   pvalue     padj minus_log10_pvalue DE_status
  <chr>                <dbl>    <dbl>    <dbl>              <dbl> <fct>
1 ENSMUSG00000058006  1.12   1.90e-11 1.63e-10            10.7    Non-DE
2 ENSMUSG00000021336 -1.32   1.34e- 8 7.94e- 8             7.87   Non-DE
3 ENSMUSG00000011158 -0.199  1.30e- 1 1.93e- 1             0.885  Non-DE
4 ENSMUSG00000032085  0.246  1.47e- 1 2.14e- 1             0.833  Non-DE
5 ENSMUSG00000004364  0.0278 8.19e- 1 8.64e- 1             0.0866 Non-DE
6 ENSMUSG00000113428  0.0823 8.22e- 1 8.66e- 1             0.0853 Non-DE
```

Here's a volcano plot made from the 'de_results' data frame:



What is the data type of the values stored in the 'DE_status' column of the 'de_results' data frame?

ANSWER: *factor*

Question 31. Draw a line between each R function and the task it is best suited to perform. *Note: this is a 1-to-1 correspondence*

Function	Task
t.test()	Perform a Bonferroni correction for multiple testing
wilcox.test()	Test for significant associations between two continuous variables
tidy()	Test if means of two groups are different, assuming data from both groups are normally distributed
p.adjust()	Convert output from a function performing a statistical test to data frame format
lm()	Test if distributions of two groups are different, without assuming the data have a specific distribution

Question 32. Consider the following function definition:

```
get_de_genes <- function(de_results,
                         de_method,
                         q_value_cutoff = 0.05,
                         log2_fc_cutoff = 1) {
    # Body of the function
}
```

Suppose we call the function with this code:

```
input_data |>
    get_de_genes(log2_fc_cutoff = 2,
                 "limma")
```

What value is assigned to the 'de_results' argument?
(A) contents of 'input_data' ⟵ **ANSWER:**
(B) 2
(C) 0.05
(D) "limma"
(E) 1

Question 33. Draw a line connecting the R function to the type of data file it is designed to read. *Note: this is a 1-to-1 correspondence*

Function	Data file
readRDS()	Binary file of tabular data created by Excel
read_csv()	Binary, R data serialized file
read_xlsx()	Tabular raw text file where comma characters separate data from each column
read_tsv()	Tabular raw text file where tab characters separate data from each column