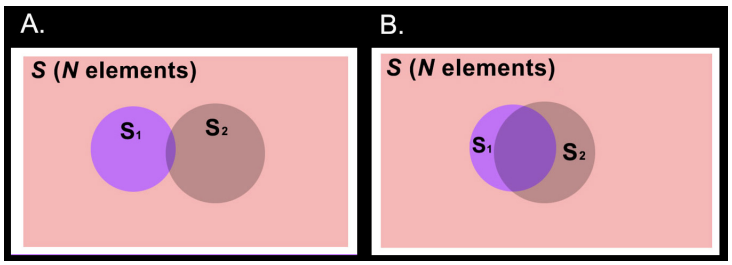
University of Pennsylvania

BIOL4536 Fall 2023

Professor: Gregory R. Grant

Final Exam (Practice #1) SOLUTIONS

Question 1. If we model pathway enrichment analysis with a hypergeometric distribution, then Which of the following has a smaller p-value?



ANSWER: B, because it has greater overlap.

Question 2. What causes us to need to correct pathway enrichment p-values for multiple testing?

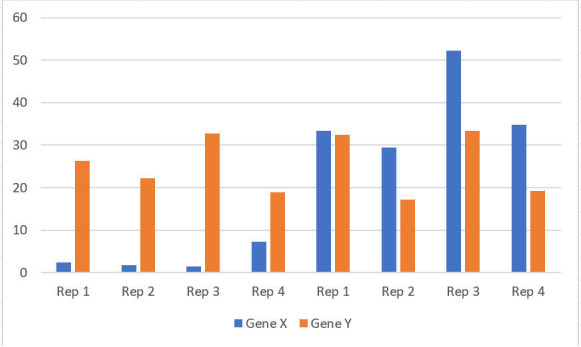
ANSWER: Because we get a p-vaule for each pathway, so a multiple testing problem.

Question 3. True or False. When doing a differential expression pathway enrichment analysis, raising the q-value cutoff used to produce the DE list will always lead to an equal or lower enrichment q-values.

ANSWER: False. It will lead to a bigger input list to the pathway analysis, but that doesn't mean there'll be a lower enrichment p-value or q-value.

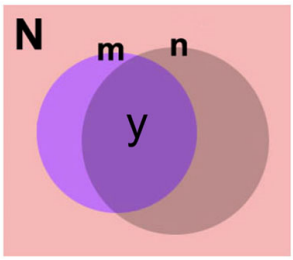
Question 4. This is a figure from the slides on GSEA (Gene Set Enrichment Analysis). Where do those numbers come from in the column all the way on the right with a question mark? In other words how are they defined? We're not looking for a mathematical formula, just their defining property.

	Condition 1				Condition 2				?
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 4	
Gene X	2.4	1.8	1.4	7.2	33.4	29.42	52.17	34.83	0.936986436
Gene Y	26.3	22.24	32.75	18.91	32.4	17.2	33.4	19.23	0.039800984
Class	0	0	0	0	1	1	1	1	



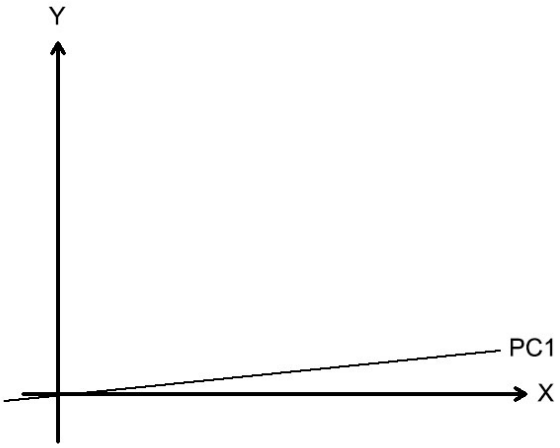
ANSWER: Those are the correlations between the data and the class labels.

Question 5. This is the diagram we saw in class illustrating the hypergeometric random variable. Which of the things m, n, N and y are random quantities? Circle all that apply.



ANSWER: Only y is random. Everything else is a parameter of the distribution.

Question 6. Consider the PC1 subspace in the figure. Circle the one correct answer.
(A) The X loading is higher than the Y loading. ⟵ **THIS ONE**
(B) The X loading is equal to the Y loading.
(C) The X loading is lower than the Y loading.
(D) None of the above can be said definitively.



Question 7. True or False. A latent variable can combine the information from multiple genes (dimensions) into one variable (dimension).

ANSWER: True.

Question 8. In Principle Components Analysis, to obtain PC1, we project onto the one-dimensional subspace. In this subspace the data has
(A) Strictly greater variance as it had in the original space.
(B) Strictly less variance as it had in the original space.
(C) Greater or equal variance as it had in the original space.
(D) Less or equal variance as it had in the original space. ⟵ **THIS ONE**

Question 9. True or False. Principle Components Analysis is a linear method and UMAP is non-linear.

ANSWER: True.

Question 10. The null hypothesis for a Mann-Whitney test is (circle one):
(A) Equal means
(B) Equal means and variances
(C) Equal medians
(D) Equal distributions ⟵ **THIS ONE**
(E) Equal standard deviations

Question 11. In your own words, what's the difference between technical and biological variation in a data set.

ANSWER: Technical variation is due to things like measurement error, batch effects, experimental errors. Biological variance is the variation inherent in the population of organisms the samples are drawn from.

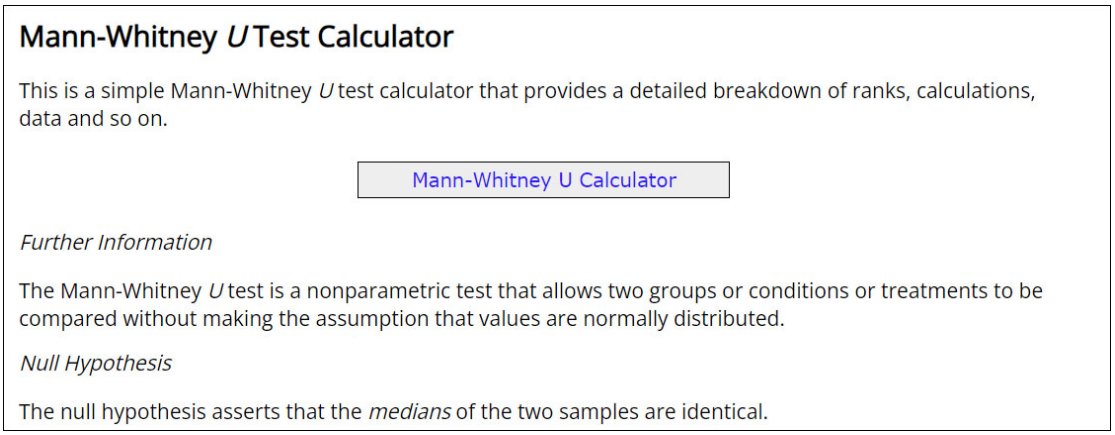
Question 12. Consider the two data sets shown below. True or False. The Mann-Whitney p-value is the same for both data sets.

Data Set 1	
condition1	condition2
5	8
4	7
3	6

Data Set 2	
condition1	condition2
5	800
4	7
3	600

ANSWER: True, because in both cases the smallest value in condition 2 is bigger than the largest in condition 1.

Question 13. This is a screen shot detail from a webpage with a Mann-Whitney calculator. The authors of this page made a grievous mistake, what is it?



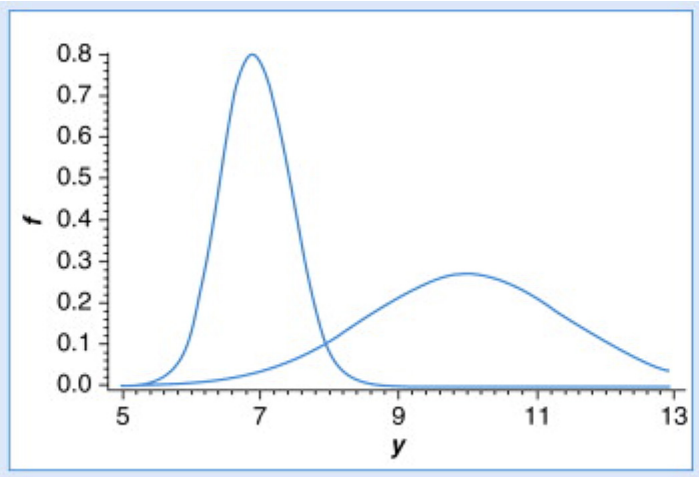
Not that it matters, but the website is: `https://www.socscistatistics.com/tests/mannwhitney`

ANSWER: They misstate the null hypothesis to be about medians when it should be equality of the entire distributions.

Question 14. True or False. Non-parametric tests tend to be blind to outliers.

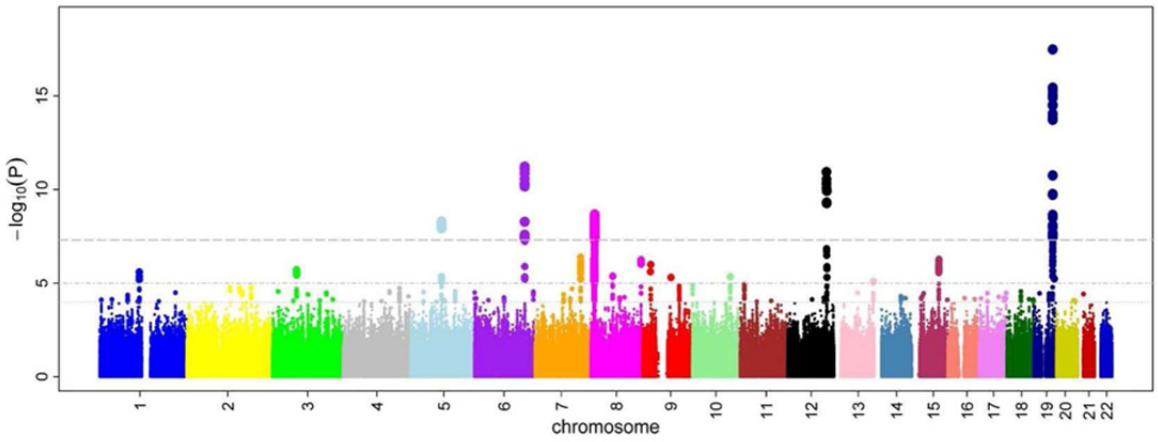
ANSWER: True, because they tend to be based on ranking.

Question 15. It would be legitimate to compare these two distributions using (circle all that apply)
(A) Mann-Whitney ⟵ **THIS ONE**
(B) A Permutation test ⟵ **THIS ONE**
(C) A parametric T-test



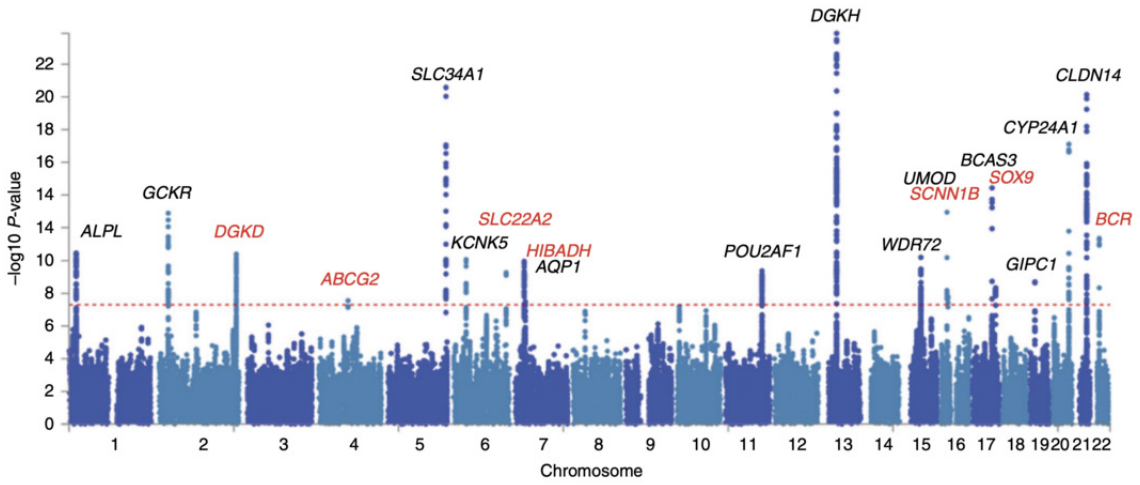
NOTE: *Cannot apply a parametric T-test unless the variances are equal between the two distributions.*

Question 16. True or False. It is called a Manhattan plot because multiple points are in the exact same genomic location.



ANSWER: False. No two points are in the exact same location, they just look that way on the plot because the genome is so large.

Question 17. True or False. At each locus in a Manhattan plot, the SNP with the smallest p-value (highest on the vertical axis) is the causative SNP.



ANSWER: False. Because of linkage disequilibrium, it could be any SNP in the locus.

Question 18. What is the multiple testing issue in GWAS?

(A) Multiple genes
(B) Multiple subjects
(C) Multiple phenotypes
(D) Multiple chromosomes
(E) Multiple SNPs ⟵ **THIS ONE**

Question 19. Is it necessary to run a test for significant association on the following data?

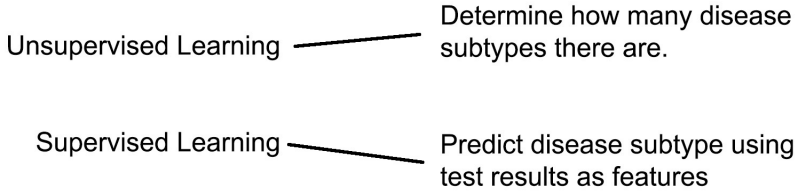
	Controls	Cases
A/A	100	1000
A/B	200	2000
B/B	900	9000

ANSWER: No, that would be pointless since every value in the cases column is exactly 10 times the value in controls. So there can be no interesting association.

Question 20. True or False. A polygenic risk score combines information from multiple SNPs into one score that assesses risk of one genetically associated condition.

ANSWER: True. That's basically the definition of a polygenic risk score.

Question 21. Connect the thing on the left with the correct thing on the right.

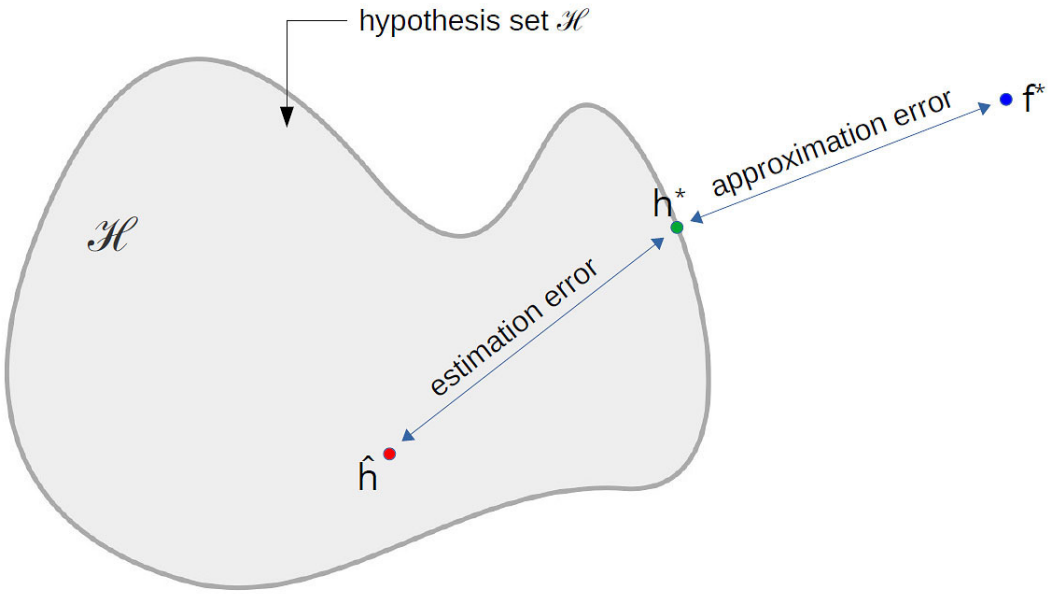


Question 22. In the figure, which functions are generally unknowable? Circle all that apply.

(A) $\hat{h}$

(B) h^* ⟵ **THIS ONE**

(C) f^* ⟵ **THIS ONE**



Question 23. True or False. For a continuous dependent variable, we use the absolute value loss function $|\hat{y} - y|$ instead of the quadratic loss function $(\hat{y} - y)^2$ because the former is differentiable.

ANSWER: False. It's the exact opposite, we use quadratic loss because the absolute value is not differentiable.

Question 24. True or False. An element of the hypothesis set is a function of the dependent variable.

ANSWER: False. It's a function of the indepdendent variables.

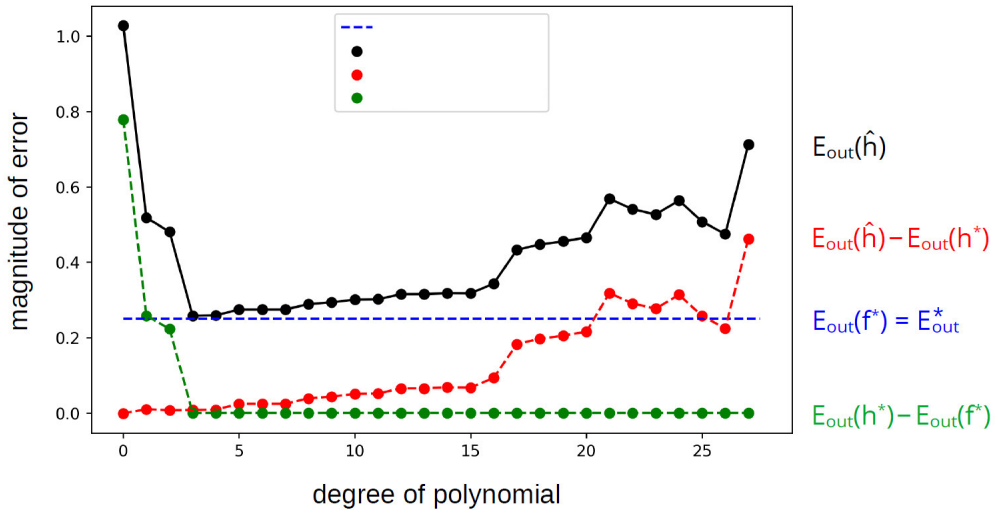
Question 25. In the figure, which color line represents the Bayes Risk?

(A) black

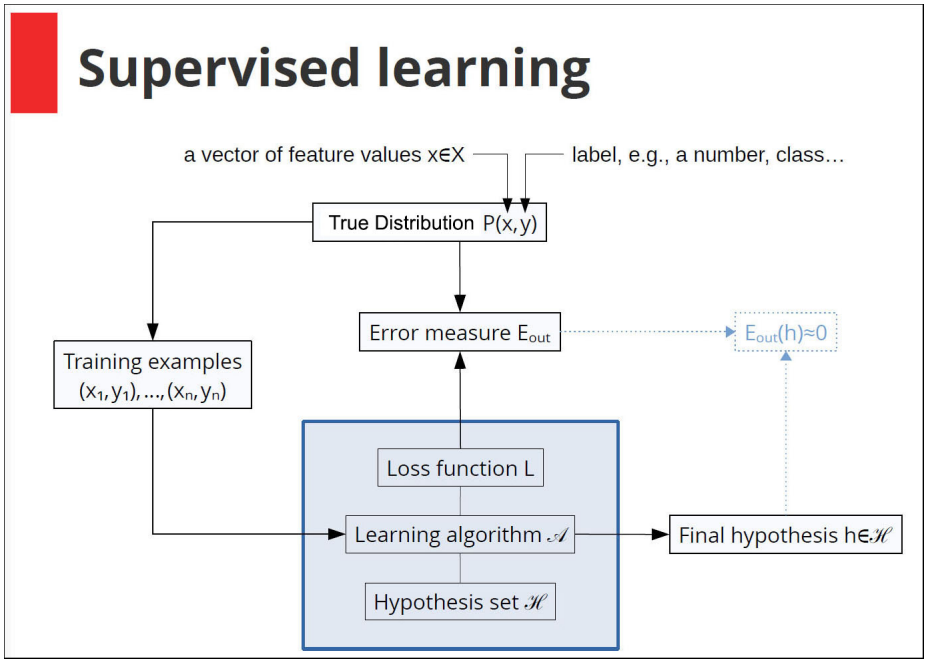
(B) blue ⟵ **THIS ONE**

(C) orange

(D) green



Question 26. Refer to the diagram below. True or False, the Final hypothesis $h \in \mathcal{H}$ depends on the choice of Loss function L.



ANSWER: True.

Question 27. Draw a line between each R data structure and the matching description

Data structure	Description
data frame	One-dimensional, where each element can have a different type
list	One-dimensional, where each element must have the same data type
matrix	Two-dimensional table, where all elements must have the same data type
vector	Two-dimensional table, where each column can have a different data type

Question 28. Here are the first six lines from a text file:

```
species,island,bill_length_mm,year
Adelie,Torgersen,39.1,2007
Adelie,Torgersen,39.5,2007
Adelie,Torgersen,40.3,2007
Adelie,Torgersen,NA,2007
Adelie,Torgersen,36.7,2007
```

Which R function would you use to read this file into R?

(A) read_xlsx()
(B) read_csv() ⟵ **THIS ONE**
(C) read_tsv()
(D) read_RDS()

Question 29. Consider the following function definition:

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

If we call the function with the following code, what value is assigned to the 'q_value_cutoff' argument?

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

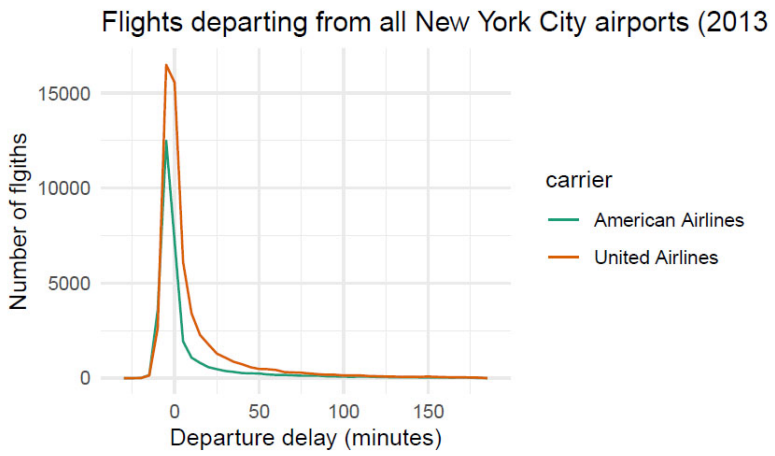
(A) contents of 'intput_data'
(B) 2
(C) 0.05 ⟵ **THIS ONE**
(D) "limma"
(E) 1

Question 30. True or False. SummarizedExperiement objects are designed to store the results of an experimental assay, along with metadata describing the experiment?

ANSWER: True.

Question 31. Consider this table and graph

```
# A tibble: 6 x 6
  flight carrier          sched_dep_time dep_delay sched_arr_time arr_delay
   <int> <chr>                     <int>     <dbl>          <int>     <dbl>
1   1545 United Airlines             515         2            819        11
2   1714 United Airlines             529         4            830        20
3   1141 American Airlines           540         2            850        33
4   1696 United Airlines             558        -4            728        12
5    301 American Airlines           600        -2            745         8
6    194 United Airlines             600        -2            917         7
```



To draw this graph, which aesthetic(s) would you need to map? (Circle all that apply)
(A) x ⟵ **THIS ONE**
(B) y
(C) shape
(D) color ⟵ **THIS ONE**

Question 32. True or False. These two R expressions are equivalent.

```r
penguins |> mutate(flipper_length_cm = flipper_length_mm / 10)

mutate(flipper_length_cm = flipper_length_mm / 10, penguins)
```

ANSWER: False.

Question 33. Consider these two tibbles:

Tibble A:

```
# A tibble: 24 x 3
  gene_name sample_id   read_counts
  <chr>     <chr>             <int>
1 Lcn2      Saline_9574          63
2 Lcn2      Saline_9575          41
3 Lcn2      IL1B_9577         39976
4 Lcn2      IL1B_9578         44056
5 Ido2      Saline_9574        1734
6 Ido2      Saline_9575        1129
# i 18 more rows
```

Tibble B:

```
# A tibble: 6 x 5
  gene_name Saline_9574 Saline_9575 IL1B_9577 IL1B_9578
  <chr>           <int>       <int>     <int>     <int>
1 Lcn2               63          41     39976     44056
2 Ido2             1734        1129       280       230
3 Fam83a              6           5        94       210
# i 3 more rows
```

Which R function would you use to reshape Tibble A into Tibble B?

(A) `left_join()`
(B) `pivot_longer()`
(C) `pivot_wider()` ⟵ **THIS ONE**
(D) `bind_rows()`