Question 1. Suppose somebody gives you a list of gene ID’s for a pathway enrichment analysis. True or False: you may have to convert those ID’s before you can input them to a pathway enrichment tool.

ANSWER: True.

Question 2. True or False. If we have a list of gene identifiers and we’re just testing one pathway for enrichment, then we still have a multiple testing problem because the pathway has multiple genes.

ANSWER: False.

Question 3. True or False. When we perform a GO enrichment analysis, the gene sets are disjoint, meaning they have (pairwise) empty intersections.

ANSWER: False.

Question 4. Consider the following pathway enrichment results table. Suppose we use a $q$-value significance cutoff of 0.1. How many false positives do we expect?

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Gene</th>
<th>Count</th>
<th>$%)$</th>
<th>P-Value</th>
<th>Benjamin</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>cytosol</td>
<td>R1</td>
<td>64</td>
<td>32.5</td>
<td>3.8E-6</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>nucleus</td>
<td>R1</td>
<td>82</td>
<td>41.6</td>
<td>7.3E-5</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>aminoacyl-tRNA_ligase_activity</td>
<td>R1</td>
<td>5</td>
<td>2.5</td>
<td>4.6E-4</td>
<td>8.7E-2</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>protein kinase binding</td>
<td>R1</td>
<td>15</td>
<td>7.6</td>
<td>4.8E-4</td>
<td>6.7E-2</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>translation</td>
<td>R1</td>
<td>11</td>
<td>5.6</td>
<td>6.2E-4</td>
<td>7.2E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>cytosolic riboside</td>
<td>R1</td>
<td>6</td>
<td>3.0</td>
<td>7.3E-4</td>
<td>7.0E-2</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>cytochrome</td>
<td>R1</td>
<td>25</td>
<td>12.7</td>
<td>9.6E-4</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>protein binding</td>
<td>R1</td>
<td>67</td>
<td>34.0</td>
<td>9.7E-4</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>RNA_ribonucleotide_mRNA_translation</td>
<td>R1</td>
<td>4</td>
<td>2.0</td>
<td>3.6E-3</td>
<td>1.0E-0</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>pentose phosphate metabolism</td>
<td>R1</td>
<td>5</td>
<td>2.5</td>
<td>5.0E-3</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>nucleoside</td>
<td>R1</td>
<td>45</td>
<td>22.6</td>
<td>5.4E-3</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>aminoacyl-tRNA_adenylation_activity</td>
<td>R1</td>
<td>3</td>
<td>1.5</td>
<td>5.6E-3</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>metabolic process</td>
<td>R1</td>
<td>7</td>
<td>3.6</td>
<td>6.3E-3</td>
<td>1.0E-0</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>Golgi apparatus</td>
<td>R1</td>
<td>23</td>
<td>11.7</td>
<td>6.6E-3</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>histone deacetylase_complex</td>
<td>R1</td>
<td>4</td>
<td>2.0</td>
<td>6.6E-3</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>ribosome</td>
<td>R1</td>
<td>7</td>
<td>3.6</td>
<td>7.4E-3</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>membrane</td>
<td>R1</td>
<td>76</td>
<td>38.6</td>
<td>7.4E-3</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>polysome</td>
<td>R1</td>
<td>4</td>
<td>2.0</td>
<td>8.6E-3</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>cytosolic translation</td>
<td>R1</td>
<td>5</td>
<td>2.5</td>
<td>8.3E-3</td>
<td>1.0E-0</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>cytosolic small ribosomal subunit</td>
<td>R1</td>
<td>4</td>
<td>2.0</td>
<td>8.9E-3</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>nuclear organizing center</td>
<td>R1</td>
<td>6</td>
<td>3.0</td>
<td>9.7E-3</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>regulation of translation</td>
<td>R1</td>
<td>6</td>
<td>3.0</td>
<td>1.1E-2</td>
<td>1.0E-0</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>cellular response to epithelial growth factor stimulus</td>
<td>R1</td>
<td>4</td>
<td>2.0</td>
<td>1.2E-2</td>
<td>1.0E-0</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>trans-cap network</td>
<td>R1</td>
<td>7</td>
<td>3.6</td>
<td>1.3E-2</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>GOTERM_BP_DIRECT</td>
<td>cytoplasmic metabolic process</td>
<td>R1</td>
<td>7</td>
<td>3.6</td>
<td>1.3E-2</td>
<td>1.0E-0</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>cell projection</td>
<td>R1</td>
<td>19</td>
<td>9.6</td>
<td>1.4E-2</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>endoplasmic reticulum</td>
<td>R1</td>
<td>24</td>
<td>12.2</td>
<td>1.5E-2</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>HPO:0000103</td>
<td>Zinc_finger</td>
<td>R1</td>
<td>23</td>
<td>11.7</td>
<td>1.7E-2</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>RNA_binding</td>
<td>R1</td>
<td>16</td>
<td>8.1</td>
<td>1.7E-2</td>
<td>6.3E-1</td>
</tr>
<tr>
<td>GOTERM_MF_DIRECT</td>
<td>valine-tRNA_ligase_activity</td>
<td>R1</td>
<td>2</td>
<td>1.0</td>
<td>1.8E-2</td>
<td>8.3E-1</td>
</tr>
</tbody>
</table>
A 0.1 cutoff results in six pathways of which we expect $6 \times 0.1 = 0.6$ false positives.

**Question 5.** In the following GSEA diagram that follows a DE analysis, the horizontal axis goes up to 14,000 and change. This number represents:

(A) The number of genes in the pathway.
(B) The non-DE genes.
(C) The DE genes.
(D) The number of genes tested for DE.  ← **THIS ONE**

![GSEA Diagram](image-url)

**Question 6.** How many one-dimensional subspaces are there in the two-dimensional plane?

**ANSWER:** Infinitely many.

**Question 7.** True or False. In Principle Components Analysis, it’s possible that neither the first, nor the second, principle components PC1 and PC2 are driven by the biological variation or batch effects.

**ANSWER:** True.

**Question 8.** On the following graph, draw in (approximately) the line corresponding to the first principle component subspace PC1.
Question 9. Suppose you have RNA-Seq data from three experimental conditions WT, Mutant 1 and Mutant 2 and you get the following PCA plot. What appears to be true?
(A) Both PC1 and PC2 are driven by genotype (variation across conditions)
(B) Both PC1 and PC2 are driven biological variation (variation within condition.)
(C) PC1 is driven by genotype and PC2 is driven by biological variation within condition. ← THIS ONE
(D) PC2 is driven by genotype and PC1 is driven by biological variation within condition.

Question 10. What is the smallest $p$-value one can obtain from a permutation test with 3 replicates in one group and 2 in the other?
**Question 11.** True or False. In a permutation test, the unpermuted data counts as one permutation.

**ANSWER:** True.

**Question 12.** Suppose a permutation $p$-value is calculated for paired (repeated measures) data from $n$ subjects. What’s the smallest $n$ can be for there to be any possibility of achieving a significant $p$-value? *note: consider the largest $p$-value cutoff for significance to be 0.05*

**ANSWER:** The number of (paired) permutations is $\frac{1}{n^2}$ which is only smaller than 0.05 when $n \geq 5$.

**Question 13.** We have $n$ independent observations from two conditions. What are the two conditions required to run a parametric $T$-test?

**ANSWER:** Observations are normally distributed and both conditions have the same variance.

**Question 14.** In the following single cell RNA-Seq tSNE plot each cell has different

(A) genome

(B) transcriptome

(C) species

(D) size

(E) gene expression ← THIS ONE

**Question 15.** Suppose we are doing a Mann-Whitney test for 3 vs. 2 replicates. Write down two different rankings of the five values that give the same value of the statistic $R$.

**ANSWER:** Number them 1, 2, 3, 4, 5. If the first group has 1,3,4 then $R = 8$ and if the first group has 1,2,5 then also $R = 8$.

**Question 16.** True or False. SNP calling done by microarray limits the analysis to a set of pre-determined single nucleotide variants.
**ANSWER:** True.

**Question 17.** True or False. In a Manhattan plot there are as many points in the graph as there are genome positions on the horizontal axis.

**ANSWER:** False.

**Question 18.** True or False. In each peak of SNP’s in a GWAS Manhattan plot that rises above the significance cutoff, there is exactly one causitive SNP, although it might not be the most significant one.

**ANSWER:** False, there could be more than one causative SNP in a given peak.

**Question 19.** If a point on a Manhattan plot is one unit higher than another point, then the significance level of the first point is

(A) the significance level of the second point plus 0.1
(B) the significance level of the second point times 0.1
(C) the significance level of the second point minus 0.1
(D) the significance level of the second point divided by 0.1

**Question 20.** True or False. Suppose there are five SNPs associated with a particular disease, then the five SNPs will result in five polygenic risk scores for the disease.

**ANSWER:** False, they will combined into one score.

**Question 21.** Connect the things on the left to the relevant things on the right.
Question 22. In supervised learning regression, suppose the hypothesis set consists of all cubics with leading coefficient equal to one. How many parameters are there in the model?

(A) One
(B) Two
(C) Three ← THIS ONE
(D) four

Question 23. In the Learning Inequality, if we raise the number of observations by one, how does that affect the
right hand side?
(A) It doubles it.
(B) It divides it by two.
(C) It multiplies it by \( e^{-2\varepsilon^2} \). ← THIS ONE
(D) It increases it by \( 2|\mathcal{H}| \).

Question 24. True or False. If we increase the size of \( \mathcal{H} \) indefinitely, it could be possible that \( \hat{h} = h^* = f^* \)?

\[
P \left( |E_{in}(h) - E_{out}(h)| > \varepsilon \right) \leq 2 \cdot |\mathcal{H}| \cdot e^{-2n\varepsilon^2}
\]

**ANSWER:** True.

**Question 25.** Assume we’re doing supervised machine learning and \( f^* \) is the true model of the data (as usual). Assume \( f^* = \hat{h} \). True or False, \( E_{out}(f^*) = E_{out}(h^*) \)

**ANSWER:** True, because \( f^* = \hat{h} \) means \( f^* = h^* = \hat{h} \) so \( f^* \) and \( h^* \) are actually equal, so must have equal out-of-sample error.

**Question 26.** In the \( k \)-means clustering diagram, what changes in the next iteration, the location of the points in the plane or the location of the centroids in the plane?

**ANSWER:** The centroids.
Question 27. You have a data table containing columns with text and numeric data. Would it be best to store these data in R as a data frame, a matrix, or a vector?

**ANSWER:** data frame

Question 28. You have the following data frame:

```r
# A tibble: 14,567 x 4
Gene_id log2_fold_change p_value q_value
<chr>        <dbl>       <dbl>     <dbl>
1 ENSMUSG00000032861 -1.16       0.000000993 0.0145
2 ENSMUSG00000025554 -0.541      0.00000282  0.0206
3 ENSMUSG00000025227  0.0185      0.00000677  0.0328
4 ENSMUSG00000006561  0.657       0.0000130   0.0464
5 ENSMUSG00000025252  0.317       0.0000159   0.0464
6 ENSMUSG00000024663  0.514       0.0000300   0.0617
7 ENSMUSG00000031677 -0.270      0.0000306   0.0617
8 ENSMUSG00000027860  0.573       0.0000339   0.0617
9 ENSMUSG00000039846  0.301       0.0000425   0.0681
10 ENSMUSG00000030623  0.172       0.0000484   0.0681
# ... with 14,557 more rows
```

You’d like to perform a series of transformations on this data frame. Match the `dplyr` function with the desired transformation:

- `select` Calculate \(-\log_{10}(\text{q\_value})\), and store the result in a new column
- `filter` Remove the `log2\_fold\_change` column
- `mutate` Sort rows according to the `p\_value` column
- `arrange` Find all rows with a value < 0.05 in the `p\_value` column

Question 29. which of these is the relational operator testing for equality between two values?

(A) \(\mid\rangle\)
(B) \(\langle\rangle\)
(C) \(\==\) ← **THIS ONE**
(D) \(\langle\langle\rangle\rangle\)
The following three questions are based on the Palmer Penguins dataset. Here is the formatted table of penguin data:

```
# A tibble: 344 x 6
   species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
     <fct>        <dbl>        <dbl>            <int>       <int> <fct>
 1     Adelie       39.1        18.7             181        3750  male
 2     Adelie       39.5        17.4             186        3800  female
 3     Adelie       40.3          18             195        3250  female
```

**Question 30.** Fill in the blanks to make this R code calculate the mean body mass and bill length within each species of penguin.

```
penguin_data |>
  group_by(species) |>
  summarise(mean_body_mass = mean(body_mass_g),
             mean_bill_length = mean(bill_length_mm))
```

**Question 31.** Fill in the three aesthetic mappings you would need to create the following `ggplot2` graph:

```
aes(x = bill_depth_mm,
    y = flipper_length_mm,
    color = species)
```

**Question 32.** The following plot displays the distribution of bill lengths across all Adelie penguins. Which `ggplot2 geom_` function would you use to create this graph?

(A) geom_point()
(B) geom_bar()
(C) geom_col()
(D) geom_histogram() ← THIS ONE
(E) geom_boxplot()
Question 33. Using the following plot for reference:

connect each ggplot2 aesthetic to the corresponding visual property it controls:
<table>
<thead>
<tr>
<th>x</th>
<th>Position for each point on vertical axis, and variable for calculating boxplot distribution statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>color</td>
<td>Symbol (circle, triangle, square) for each point</td>
</tr>
<tr>
<td>shape</td>
<td>Color of the boxplots’ interior space</td>
</tr>
<tr>
<td>y</td>
<td>Color of the points and boxplot outlines</td>
</tr>
<tr>
<td>fill</td>
<td>Position for each point and boxplot on the horizontal axis</td>
</tr>
</tbody>
</table>