Ejercicio tipo informe.

Lo resolveremos en clase el próximo miércoles 19 de noviembre.

Vamos a elaborar un informe reproducible en R Markdown utilizando los datos publicados en Perez-Rial et al. (2024).

El objetivo final será generar un documento HTML que incluya texto formateado, código, resultados, figuras, tablas y citas bibliográficas, respondiendo a una hipótesis concreta.

Antes de comenzar, guarda los siguientes archivos en la misma carpeta donde vayas a guardar el archivo R Markdown (los puedes encontrar en Moodle):

- libraries.bib (contiene bibliografía para los paquetes de R más utilizados)
- zotero.bib (contiene bibliografía de artículos científicos de interés)
- 12870_2024_5411_MOESM1_ESM.xlsx

1. **Encabezado YAML. Modifica el encabezado del archivo .Rmd para que:**
   - En el html final aparezca como título "Informe NILs" y autor "Nombre Apellidos y vuestro correo@uco.es."
   - La fecha sea aumática.
   - La salida en html **tenga índice y secciones numeradas**
   - En la bibliografía incluye a "zotero.bib" y "libraries.bib". Activa los links de las citas.

2. Incluye el **LOGO de la UCO** y el de la **Facultad de ciencias,** con un tamaño **width = 15%, en el encabezado.**

3. Elimina los apartados ## R Markdown y ## Including Plots para dejar el archivo limpio.

4. Copia el siguiente texto y aplica el formato indicado. Debes introducir las citas correctamente mediante @.

```
# Introducción

## El trabajo seleccionado

Vamos a trabajar con los datos de Perez-Rial et al. 2024. Este trabajo describe genes candidatos de la floración en garbanzo (*Cicer arietinum*) como ***MED16***, ***BBX24*** o ***ELF3***, este último también reportado por otros estudios como Ridge et al. 2017. Los mecanismos básicos de floración estudiados en la planta modelo *Arabidopsis thaliana* están conservados en las leguminosas (Weller et al. 2019, Hecht et al. 2005, Weller et al. 2015), exceptuando ciertos genes como ***CONSTANS*** (Wong et al. 2014).
```

5. En el siguiente apartado de la introducción del informe (## Los datos) debemos importar los datos y podemos mostrar una primera vista. Para ello, debemos:
    a. **Importar los datos del archivo** 12870_2024_5411_MOESM1_ESM.xlsx **y asignarlos a un objeto llamado** Data.

    b. **Incluir el link donde se puede encontrar el trabajo original. Debes proporcionarle un nombre al link** https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-024-05411-y

    c. **Primera vista de los datos.**
    Muestra una tabla con formato html de los primeros 8 elementos de los datos, alineada en el centro y que se llame "Tabla 1: Ocho primeras filas"

    d. **Añade texto explicativo e inserta código R en línea.** Escribe el código necesario para incorporar el número de filas y columnas. Por ejemplo:
    Los datos tienen `r ___(Data)` filas y `r ____(Data)` columnas.
    Cada fila se corresponde con un SNP, por tanto, hay `r ___(Data)` SNPs.

6. En el mismo apartado anterior (# Introducción, ##Los datos) podemos empezar a trabajar con los datos para seleccionar aquellos que nos interesan. Para ello, primero debemos **cargar las librerías necesarias (dplyr y ggplot2). Recuerda NO mostrar pasos innecesarios en el output mediante opciones de chunk, ni warnings nin message**. A continuación, podemos:

    a. Filtrar los datos para quedarte únicamente con los SNPs detectados en cromosomas (sus códigos empiezan por NC_) y los sigues almacenando sobre el mismo objeto "Data" del siguiente modo:

    ```
    Data <- Data %>% filter(CHROM %in% c("NC_021160.1","NC_021161.1",
                            "NC_021162.1", "NC_021163.1", "NC_021164.1",
                            "NC_021165.1","NC_021166.1", "NC_021167.1")
    ```

    b. Volver a escribir **Código R en línea. Por ejemplo:**

    Del total de SNPs detectados, `r ___(Data)` se corresponden con SNPs situados en cromosomas.

    c. Calcula el número de SNPs por cromosoma detectados con la función `count`. Guarda los datos ordenados de forma descendiente en un nuevo objeto "Cromosoma".

    d. Utiliza la variable nueva "Cromosoma" para hacer un gráfico de barras incluyendo el pie de figura "Figura 1. Número de SNPs por cromosomas".
    Posición centrada y tamaño reducido al 60% del ancho del texto.
    Recuerda darle un nombre a los ejes.

7. El último apartado de la introducción puede ser el de proponer una hipótesis (## Hipótesis).
    Redacta una hipótesis en el documento. Por ejemplo:
    "Los cromosomas con mayor número de SNPs presentan mayor % de SNPs que no pasan el filtro de calidad"

8. En el apartado de **Materiales y Métodos (# Materiales y Métodos)** debemos incluir las citas correctas de los paquetes de R utilizados. Debes introducir las citas correctamente mediante @. Por ejemplo:

   Para llevar a cabo el análisis usamos R [@____] con las librerias dplyr [@____] y ggplot [@____ ; @____]

9. **En resultados (# Resultados),** realizamos las operaciones necesarias para responder a la hipótesis planteada ("Los cromosomas con mayor número de SNPs presentan mayor % de SNPs que no pasan el filtro de calidad"). Para ello, primero seleccionamos únicamente las columnas que nos interesan del dataset "Data". Selecciona del dataset "Data" únicamente las columnas CHROM, POS y FILTER mediante "select". Guarda el resultado en un nuevo objeto llamado "datos_filtrados".

   Recuerda NO mostrar este paso en el output mediante las opciones de chunk.

   Puedes describir las columnas con las que te quedas. Por ejemplo: En CHROM aparecen los diferentes cromosomas del garbanzo, en POS aparecen las posiciones de los SNPs detectados y en FILTER aparece el filtro de calidad. Los SNPs que pasan el filtro aparecen descritos como PASS.

10. En el mismo apartado de resultados, crea una nueva variable en la que todo valor distinto de PASS se transforme en NOT_PASS. Puedes usar `mutate()` junto con `ifelse()`.
    Este paso NO debe aparecer en el output.

11. **## Cálculo de porcentajes por cromosoma**
    Agrupa los datos por cromosoma (CHROM) mediante y calcula:
    a. el número total de SNPs
    b. el número total de SNPs que no pasan el filtro
    c. porcentaje de SNPs que no pasan el filtro. (redondeado a un decimal).

    Puedes usar las funciones group_by y summarise.

12. Muestra el resultado como una tabla con `knitr::kable()`. Nombre de la tabla: "Tabla 2. Porcentaje de SNPs que pasan y no pasan el filtro por cromosoma".

13. **## Representación gráfica**
    Haz un gráfico de dispersión mediante ggplot2 que muestre la relación entre número total de SNPs (x) y porcentaje de SNPs no filtrados (y).
    Añade una línea de tendencia con geom_smooth(method="lm") y un título descriptivo. Recuerda darle nombre a los ejes.

14. **#Conclusiones.** Interpreta los resultados: Indica si los datos apoyan o no la hipótesis inicial y explica brevemente por qué.

El índice que debéis obtener es el siguiente:

Introducción
  El trabajo seleccionado
  Los datos
  Hipótesis
Materiales y Methods
Resultados
  Cálculo de porcentajes por cromosoma
  Representación gráfica
Conclusiones

NOTA:

Modificar las opciones de chunk necesarias para que aparezca o no el resultado del bloque de código en el output.

Añadir todos los comentarios explicativos sobre lo que vais haciendo a lo largo del ejercicio. Tenéis que explicar el código en vuestro informe.