# Máster interuniversitario de Bioestadística y Bioinformática
# Análisis de datos Ómicos (M0-157)
# Primera prueba de evaluación continua.

**Fecha publicación del enunciado: 25-05-2020**

**Fecha límite de entrega de la solución: 14-06-2020**

## Presentación

Esta PEC consta de ejercicios similares a los discutidos en los debates con los que podréis contrastar vuestra asimilación de los conceptos y métodos presentados en la primera parte del curso.

## Objetivos

El objetivo de esta PEC es ilustrar el proceso de análisis de datos de ultrasecuenciación mediante la realización de un estudio, de principio a fin, tal como se llevará a cabo en una situación real.

## Descripción de la PEC

La PEC se basará en los datos de un estudio que os proporcionaremos y del que deberéis extraer una muestra aleatoria con el fin de garantizar que cada conjunto de datos es distinto. A partir de dicho conjunto y de la información sobre el problema deberéis: (i) Plantear las cuestiones que deseáis responder (ii) Realizar los análisis necesarios y (iii) Elaborar un informe explicando problemas, métodos, resultados y discusión. Recordad que tan importante como el resultado es el razonamiento y el proceso que os leva a ello, es decir el consultor debe poder ver no tan sólo donde habéis llegado sino también como y porque habéis llegado hasta allí.

## Recursos

Los recursos para la solución de la PEC son los que se han proporcionado en el aula para las primeras unidades, es decir los materiales del curso y casos de estudio.

## Criterios de valoración

Tal como se indica en el plan docente la PEC vale el 40% de la nota.

## Código de honor

Cuando presentáis ejercicios individuales os adherís al código de honor de la UOC, con el que os comprometéis a no compartir vuestro trabajo con otros compañeros o a solicitar de su parte que ellos lo hagan. Asimismo aceptáis que, de proceder así, es decir, en caso de copia probada, la calificación total de la PEC será de cero, independientemente del papel (copiado o copiador) o la cantidad (un ejercicio o todos) de copia detectada.

## Formato

Para hacer la entrega se tiene que enviar un mensaje al buzón de entregas del aula. En este mensaje debéis adjuntar **únicamente** un fichero pdf o html obtenido a partir de vuestro archivo Rmarkdown. **Se penalizará la entrega de archivos adicionales o de un archivo** comprimido. El nombre del fichero debe ser la composición de vuestro apellido y vuestro nombre seguido de "_ADO_PEC1.doc" (por ejemplo: si vuestro nombre es "Jordi Pujol", el fichero debe llamarse "pujol_jordi_ADO_PEC1.pdf").

Además del archivo pdf podéis presentar vuestro estudio en formato reproducible mediante un repositorio de github cuya dirección deberá encontrarse en la primera página del informe.

**No olvidéis de poner vuestro nombre y apellidos en el informe!!!**

# Enunciado

El objetivo de esta práctica es doble:

- Partiendo de un problema y unos datos seleccionados como se iindica a continuación deberéis
    - Decidir un pipeline de análisis apropiado, con la herramienta que consideréis adecuada (R/Bioconductor o Galaxy)
    - Realizar el análisis siguiendo las pautas presentadas en los materiales.
- Una vez obtenidos los resultados deberéis redactar un informe con la estructura tradicional de un informe científico técnico (ver "*Guías para el informe*").

## *Selección de los datos*

El archivo **targets_and_counts.xls** contiene la información de las muestras de un estudio obtenido del repositorio (GTEx[1]). Este repositorio contiene datos de múltiples tipos en un total de 54 tejidos. Nosotros nos centraremos en los datos de expresión (RNA-seq) pertenecientes a un análisis del tiroides en donde se compara tres tipos de infiltración medido en un total de 292 muestras pertenecientes a tres grupos:

- *Not infiltrated tissues* (NIT): 236 samples
- *Small focal infiltrates* (SFI): 42 samples
- *Extensive lymphoid infiltrates* (ELI): 14 samples.

En este ejercicio no os pedimos que busquéis un estudio para analizar sino que ya os proporcionamos los datos preprocesados en una tabla de contajes y os pedimos que seleccionéis 30 muestras aleatoriamente, 10 de cada grupo. Para ello deberéis

- Leer los datos de las dos tablas "targets.csv" y "counts.csv" (o las dos pestañas del archivo targets_and_counts.xls) a R
- Escribir un pequeño script que extraiga 10 muestras del grupo 1 (NIT), 10 del grupo 2 (SFI) y 10 del grupo 3 (ELI). Tenéis esta información en la columna "Groups" del archivo targets.csv (pestaña "targets" del archivo excel).
- Con la información de las filas escogidas tenéis que "subsetear" las columnas escogidas en el archivo "counts.csv" (o en la pestaña counts del archivo excel)

NOTA: Podéis hacer esta extracción aleatoria en la forma que prefiráis. Únicamente aseguraros de explicarlo con detalle en el informe.

Una vez hayáis preparado los datos podéis proceder a realizar un análisis de expresión diferencial. Puesto que hay tres grupos podéis hacer tres comparaciones SFI-NIST, ELI-NIST y ELI-SFI.

NOTA: Si esto os parece demasiado complicado podéis hacer una sóla comparación, quie os dará menos trabajop pero también una calificación menor (como máximo un 8 en vez de un 10).

[1]he Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

# “Pipeline” de análisis

El pipeline da análisis de datos de RNA-seq es parecido al de microarrays salvo algunos pasos. Básicamente tendréis que proceder como en el caso anterior

1. Definición de los datos tal como se ha descrito en el párrafo anterior
2. Preprocesado de los datos: filtraje y normalización
3. Identificación de genes diferencialmente expresados
4. Anotación de los resultados
5. Busca de patrones de expresión y agrupación de las muestras (comparación entre las distintas comparaciones).
6. Análisis de significación biológica (“Gene Enrichment Analysis”)

## *Software para el análisis*

En esta segunda PEC, en la que ya sois más expert@s os voy a dejar más libertad. Podéis hacer los análisis en R o en Galaxy. Podéis usar el ejemplo del caso de uso con el paquete Deseq2 o bien utilizar otras alternativas como edgeR o limma. O incluso podéis proponer otros medios, siempre que los describáis y justifiquéis.

# Informe del análisis

Una vez realizado el análisis debéis redactar un informe exponiendo qué habéis hecho, como lo habéis hecho y qué resultados habéis obtenido.
Como cualquier informe científico-técnico vuestro informe tiene que tener las partes siguientes:

1. **Abstract**, con un resumen breve de no más de cinco líneas.
2. **Objetivos**: Que se pretende con este estudio
3. **Materiales y Métodos**
    1. Naturaleza de los datos, tipo de experimento, diseño experimental,
    2. Métodos y **herramientas** que habéis utilizado en el análisis:
        1. Procedimiento general de análisis (pasos, “workflow” o “pipeline” que habéis seguido)
        2. Software que habéis utilizado
    3. Que habéis hecho en cada paso (NO ES PRECISO entrar en el detalle de los métodos, más bien hacer una descripción cualitativa indicando porque se ha llevado a cabo cada paso, y cual ha sido el “input” suministrado al procedimiento y el “output” obtenido.
4. **Resultados**
    1. Que se obtiene como resultado del análisis
5. **Discusión**
    1. Que limitaciones consideramos que pueden haber en el estudio (si consideramos que hay alguna…)
6. **Conclusión**: NO HACE FALTA. Vuestro “rol” aquí es técnico. Como bioinformáticos se os presupondrá la capacidad de manejar la información biológica mediante los programas adecuados, pero ello no implica que debáis tener los conocimientos específicos que puede requerir la interpretación biológica de los resultados.
7. **Apéndice**: Podéis poner el código de R que hayáis utilizado en un apéndice con comentarios.

## *Algunos comentarios sobre el formato de entrega*

- La estructura indicada no es más que una propuesta. Podéis modificarla o adaptarla según vuestro propio criterio.
- Procurad facilitar la revisión
    - Tabla de contenidos
    - Secciones y subsecciones bien organizadas.

- Gráficos bien centrados, preferiblemente con número y pie
- Código o salida en formato courier y bien justificado
- Páginas numeradas
- Referencia bibliográficas completas.

Una cosa importante: El informe NO DEBE SER una colección de salidas de R como en algunos de los scripts y markdown de ejemplo que os he ido facilitando. Podéis poner algún fragmento de R si lo consideráis interesante pero tenéis que separar el informemdel código.
Si, como es de esperar, trabajáis con Rmarkdown os será muy sencillo ocultar las salidas de código que no deseáis mostrar utilizando las opciones del paquete knitr.

Observad especialmente que el objetivo de la práctica no es que generéis un "tocho" con un montón de información cogida de todas partes (que luego yo deberé leer) sino que realicéis un trabajo de síntesis que ilustre, de forma general, el proceso que va desde que el investigador se presenta delante vuestro diciendo "tengo unos datos que me gustaría que analicéis" hasta que le presentáis un informe con un "esto es lo que ha salido".

## Reproducibilidad del estudio

Una habilidad que debéis adquirir como bioinformatic@s es aseguraros de que vuestro trabajo sea reproducible. Una forma de conseguirlo es crear un proyecto de Rstudio y ponerlo bajo control de versión en github, tal como hemos discutido en los debates.

Cread un repositorio en github y poned en él vuestro proyecto de forma que se pueda clonar en otro ordenador y reproducir vuestro trabajo. Debéis indicar la url de vuestro repositorio en el informe.

Recordad, como resumen de la explicación anterior: Debéis producir dos cosas:

- **Un documento** único que contenga el informe del análisis en formato HTML o PDF. **NO ACEPTAREMOS entregas que contengan más de un archivo.**
- Un repositorio de github con todo lo necesario para reproducir vuestro análisis. Si hacéis el análisis en Galaxy considerar de proporcionar un workflow que permita reproducir vuestro anaálisis.