Table of Contents

Manual de Genómica

Este manual esta dirigido a estudiantes con conocimientos basicos de programación y de biología molecular. El objetivo es generar una guía sencilla y didáctica para mejorar la comprensión de los estudiantes que posean interes en programación y genómica.

Se presentan laboratorios, actividades y resolucion de problemas y dudas que van surgiendo a medida que se van realizando las actividades de interés.

Introducción a la Programación en R

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Es una implementación de software libre. Se trata de uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la mineria de datos, investigación biomedica, bioinformáticca y matemáticas financieras.

R es parte del sistema GNU y se distribuye bajo la licencia de GNU GPL. Esta disponible para los sistemas operativo Windows, IOs, Unix, GNU/Linux. En este caso, nos vamos a basar en la instalación y posterior introducción con el sistema operativo de Linux.

Descargando R

Dependiendo del sistema operativo se realiza la descarga del paquete que proporciona los elementos necesarios para la instalación de R. La pagina de descarga es: R

En el caso particular de Linux no es necesario realizar ninguna descarga ya que este paquete viene incorporado por defecto en el sistema operativo.

Instalando R

Antes de comenzar la instalacion de R, se recomienda instalar RStudio.

RStudio es una herramienta de visualización y ejecución de codigos, todo en uno. Para instalar RStudio es necesario dirigirse a la pagina: RStudio. Una vez descargada la versión acorde a su sistema operativo procederemos a la instalación de RStudio mediante terminal.

En la terminal de Linux escribir el siguiente comando para la instalacion del paquete de RStudio (tips: es importante posicionarse en la terminal en donde este nuestro archivo descargado, si no el sistema no lo reconocerá y enviara un mensaje de error).

El siguiente comando desempaquetará e instalará RStudio: `sudo dpkg -i rstudio.deb`

Interactuando con R

Una vez instalado R y RStudio vamos a comenzar a conocer algunos comandos basicos para ocuparlo. Para esto es necesario abrir RStudio. Una vez abierto el programa podemos observar que esta se divide en 4 pequeñas pantallas. Tenemos una pantalla que simula una especie de terminal o consola en la cual podemos observar los procesos que vamos

reaizando. Otra donde podemos buscar archivos o carpetas e importar trabajos. Una pantalla esta disponible para mostrar el historial de las actividades que se van realizando y finalmente la ultima pantalla que se utiliza para programar en R.

A continuacion se adjunta una página en la que se explican los comandos básicos para comenzar el trabajo con R: Manuales R. A estas páginas también puedes recurrir: 1, 2 y 3

Tutorial en R

Si desea adquirir mas conocimientos con respecto a R, recomiento realizar este tutorial.

Tips: Se recomienda realizar el tutorial desde el comienzo y sin saltarse ningún paso, ya que es escencial que domine las funciones básicas para luego abordar problemas mas complejos.

Introducción a Bioconductor para análisis de secuencias

Bioconductor es una fuente libre, abierta y de desarrollo de software abierto para el análisis y comprensión de los datos genómicos obtenidos mediante los experimentos de laboratorio húmedo de la biología molecular. Bioconductor proporciona herramientas para el análisis y comprensión de los datos genómicos de alto rendimiento. Utiliza el lenguaje de programación estadística R.

Introducción a Bioconductor

Bioconductor permite el análisis y la comprensión de los datos genómicos de alto rendimiento. Tiene un gran número de paquetes que permiten el análisis estadístico riguroso de datos de gran tamaño. Bioconductor ayuda a los usuarios que colocan sus resultados analíticos en su contexto biológico, con grandes oportunidades para la visualización. La reproducibilidad es un objetivo importante que Bioconductor analiza. Los diferentes tipos de análisis que se pueden llevar a cabo son:

- **Secuenciación:** RNASeq, ChIPSeq, variantes, el número de copias.
- **Microarrays:** Expresión, SNP.
- **Análisis específico de dominio:** La citometría de flujo, proteómica.

Para estos análisis, uno normalmente importa y trabaja con diversos tipos de archivos relacionadas con secuencias, incluyendo los archivos FASTA, FASTQ, BAM, gtf, bed, wig files, entre otros.

Los siguientes paquetes ilustran la diversidad de funcionalidades disponibles; todos se encuentran en la versión de lanzamiento de Bioconductor que puedes encontrar en la pagina web. Adicionalmente, adjunto el link para que se acceda directamente al workflows donde se podran encontrar los paquetes antes mencionados y las descripciones del funcionamiento de cada uno de estos.

Instalando Bioconductor

La versión actual de Bioconductor es la versión 3.3; funciona con R versión 3.3.0. Los usuarios con versiones anteriores de R y Bioconductor deben actualizar los programas para aprovechar de mejor forma esta herramienta. Para instalar la ultima version de Bioconductor debe abrir R e ingresar los siguientes comandos en la consola de RStudio:

```
biocLite()
```

Este comando permitira descargar los paquetes incorporados por defecto a esta herramienta. En algun minuto del proceso le preguntara si actualiza todos, algunos o ninguno de los paquetes. Es recomendable que los actualice todos de una vez, asi el proceso sera completamente automático y no debera hacer cambios de forma manual. Si desea esta opcion debe presionar la letra a, en caso de preferir algunos la letra s y si no desea ninguno la letra n.

En una segunda ocacion le preguntara si desea utilizar una biblioteca personal. En caso de que usted posea una biblioteca y desee utilizarla presione la letra y, en caso contrario la letra n.

TIPS: La instalacion de Bioconductor puede tomar un largo tiempo, sin embargo no deberia presentar errores. Al finalizar la instalacion, el programa le mostrara en la consola la ruta en la que sus paquetes fueron instalados.

Rangos de infraestructura

Muchos paquetes de Bioconductor dependen en gran medida de la infraestructura IRanges / GenomicRanges. Por lo tanto vamos a comenzar con una breve introducción a estos y luego cubrir diferentes tipos de archivos.

El paquete GenomicRanges nos permite asociar una serie de cromosomas coordinado con un nombre de secuencia (por ejemplo, cromosoma) y una hebra. Tales intervalos genómicos son muy útiles para describir tanto los datos (por ejemplo, las coordenadas del Alineados lee, llamados picos de chip, SNPs, o el número de copias variantes) y anotaciones (por ejemplo, modelos de genes, elementos reguladores Hoja de Ruta Epigenomics, conocidas variantes clínicamente relevantes de dbSNP) . Granges es un objeto que representa un vector de localizacion genómicas y anotaciones asociadas. Cada elemento en el vector está constituido por un nombre de secuencia, un rango, un filamento, y los metadatos opcionales (por ejemplo, puntuación, contenido de GC, etc.).

A medida que vamos desarrollando los practicos, se iran instalando paquetes que seran requeridos para el desarrollo de los tutriales. En los links que se describen en cada uno de los pasos se mostrara el codigo necesario para la instalacion del cada paquete.

Para instalar el paquete de Genomic/Range es necesario escribir el siguiente comando en la consola de RStudio.

```
library(GenomicRanges)
GRanges(seqnames=Rle(c('chr1', 'chr2', 'chr3'), c(3, 3, 4)),
      IRanges(1:10, width=5), strand='-',
      score=101:110, GC = runif(10))
```

Usted deberia optener una sallida de este tipo:

GRanges object with 10 ranges and 2 metadata columns: seqnames ranges strand | score GC | [1] chr1 [1, 5] - | 101 0.934405585518107 [2] chr1 [2, 6] - | 102 0.2653759396635 [3] chr1 [3, 7] - | 103 0.182115187868476 [4] chr2 [4, 8] - | 104 0.47889164602384 [5] chr2 [5, 9] - | 105 0.902476062532514 [6] chr2 [6, 10] - | 106 0.104757135966793 [7] chr3 [7, 11] - | 107 0.760009138146415 [8] chr3 [8, 12] - | 108 0.00772674381732941 [9] chr3 [9, 13] - | 109 0.42318176664412 [10] chr3 [10, 14] - | 110 0.196629931451753

seqinfo: 3 sequences from an unspecified genome; no seqlengths

Si necesita mas informacion sobre el paquete GenomicRange digite este comando en la consola de RStudio

```
vignette(package="GenomicRanges")
```

DNA/ Secuencia de aminoacidos desde archivos FASTA

Biostrings se utilizan para representar secuencias de ADN o de aminoácidos. En el siguiente ejemplo vamos a crear una cadena de ADN y mostrar algunas manipulaciones.

A continuacion, vamos a descargar todas las secuencias de ADNc Homo sapiens desde el archivo FASTA 'Homo_sapiens.GRCh38.cdna.all.fat' de Ensembl usando AnnotationHub.

Para esto es necesario instalar el paquete AnnotationHub con el siguiente comando:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R") biocLite("AnnotationHub")
```

TIPS: La instalación de este recurso puede tomar algunos minutos.

Este archivo se descarga como un archivo en formato Fasta () del paquete ShortRead. Vamos a abrir el archivo y obtener las secuencias y las anchuras de los registros en el archivo de la FASTA utilizando RSamtools.

Una vez abierto el archivo deberiamos optener una salida como la siguiente:

GRanges object with 170893 ranges and 0 metadata columns: seqnames ranges strand [1] ENST00000434970 [1, 9]
[2] ENST00000415118 [1, 8]
[3] ENST00000448914 [1, 13]
[4] ENST00000431870 [1, 16]
[5] ENST00000414852 [1, 16]
... ... [170889] ENST00000444082 [1, 3808]
[170890] ENST00000615390 [1, 859]
[170891] ENST00000512197 [1, 115]
[170892] ENST00000414573 [1, 204]
[170893] ENST00000428912 [1, 797]

seqinfo: 170893 sequences from an unspecified genome

La interpretación de estos resultados puede ser complicada al comienzo, sin embargo, todas las salidas tienen el mismo estilo, lo que a la larga facilitará el entendimiento de estos. Si quieres conocer la utilizacion de otros formatos te recomiendo que ingreses a sequencing. Aquí podras encontrar todos los pasos detallados como se mostraron anteriormente pero en otros formatos.

Genética de Poblaciones

La genética de poblaciones es la rama de la genética cuyo objetivo es describir la variación y distribución de la frecuencia alélica para explicar los fenómenos evolutivos.

Para ello, se define a una población como un grupo de individuos de la misma especie que están aislados reproductivamente de otros grupos afines, en otras palabras es un grupo de organismos que comparten el mismo hábitat y se reproducen entre ellos. Estas poblaciones, están sujetas a cambios evolutivos en los que subyacen cambios genéticos, los que a su vez están influidos por factores como la selección natural, la deriva genética, el flujo genético, la mutación y la recombinación genética.

Existen cuatro procesos que son fundamentales para entender el concepto de Genetica de Poblaciones.

Selección Natural

La selección natural es el proceso mediante el cual ciertas características de un individuo hacen que sea más probable sus supervivencia y reproducción. La selección natural actúa sobre fenotipos, o las características observables de organismos, pero la base genética hereditaria de cualquier fenotipo que da una ventaja reproductiva se hará más común en la población.

Deriva Génica

La deriva genética es el cambio en la frecuencia alélica de las especies como efecto estocástico del muestreo aleatorio en la reproducción y la pérdida de alelos por azar. Los alelos de los hijos son un muestreo aleatorio de los alelos de los padres. Los cambios en la deriva genética no son a consecuencia de selección natural, y pueden ser beneficiosos, neutrales o negativos para la reproducción y supervivencia.

Mutación

Las mutaciones son las principales fuentes de variabilidad genética en la forma de nuevos alelos. Pueden dar lugar a varios tipos de cambios en las secuencias del ADN; estos cambios pueden tener efectos neutros, positivos (alterando el producto génico) o negativos (impidiendo que funciona el gen).

Flujo Génico

El flujo genético o migración es la transferencia de alelos de genes de una población a otra, gracias a diferentes factores como la movilidad. Usualmente se da entre la misma especie, formándose híbridos cuando se da el caso contrario (al proceso de flujo genético entre especies se le denomina transferencia horizontal).

Una vez aclarados estos conceptos, pasaremos a la parte práctica de la Genética de Poblaciones.

En este breve tutorial, se explicará como analizar estructuras de poblaciones y se reproduciran los resultados con programas informáticos mediante el lenguaje de programación en R. El método funciona para cualquier sistema operativo, y que no requiere la instalación de una estructura o programas adicionales. El programa permite ejecutar en R los algoritmos de inferencia, la elección del número de racimos y mostrar coeficientes de mezcla.

Lo primero que vamos a hacer es instalar los paquetes de R que utilizaremos en este tutorial. Para eso, es necesario que en la consola de RStudio ingrese los siguientes comandos:

```
source("http://bioconductor.org/biocLite.R")
biocLite("LEA")
```

Una vez realizada la instalación de los paquetes procedemos a importar los archivos con los que trabajaremos mediante el siguiente comando:

```
struct2geno(file = input.file, TESS = FALSE, diploid = TRUE, FORMAT = 2, extra.row = 0, extra.col = 0, output = "./genotype.geno")
```

Ejemplo 1

Los datos se compone de 60 muestras de población de 10 individuos diploides que se genotipó en 100 marcadores múltiples alélicos. Los datos pueden ser descargados y se convierten de la siguiente manera.

```
input.file = "http://membres-timc.imag.fr/Olivier.Francois/secondary_contact.str"
struct2geno(file = input.file, TESS = TRUE, diploid = TRUE, FORMAT = 2, extra.row = 0, extra.col = 0, output = "secondary_contact.geno")
```

Dado que los datos también contenía la información geográfica para las muestras, se utilizó el TESS = TRUE flag. Tenga en cuenta que el script de conversión exporta las coordenadas geográficas (longitud, latitud) en un archivo .coord. Para comenzar, se realizará un análisis de estructura de la población que asume K = 3 grupos. Esto se puede hacer mediante el uso de la función *snmf* del paquete LEA. Para esto, es necesario ingresar el diguiente comando:

```
obj.snmf = snmf("secondary_contact.geno", K = 3, alpha = 100,
project = "new")
qmatrix = Q(obj.snmf, K = 3)
```

Una vez realizado, deberá optener una salida como esta:

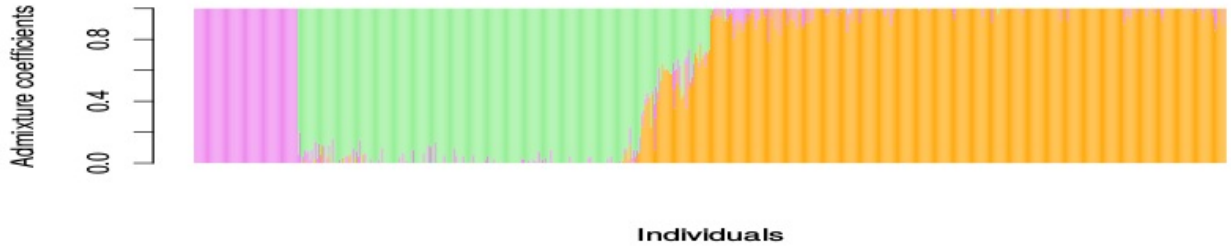
The project is saved into : tact.snmfProject

To load the project, use: project = load.snmfProject("tact.snmfProject")

To remove the project, use: remove.snmfProject("tact.snmfProject")

[1] " *sNMF K = 3 repetition 1* "

El algoritmo converge muy rápidamente. Al final de la carrera, el objeto qMatriz contiene la matriz de coeficientes de ascendencia para cada individuo y para k = 3 grupos. La matriz Q tiene 600 filas y 3 columnas, y se muestra tradicionalmente utilizando una representación barplot. Para esta representación, sólo tiene que utilizar la función barplot de R (Figura 1).



En la figura 1 se muestra una simulación de Contacto secundario. Diagrama de caja de coeficientes ascendencia de 600 individuos (datos alélicos).

Ahora se mostrarán las estimaciones de población de la mezcla mediante la superposición de gráficos circulares en un mapa geográfico de Europa. Para este objetivo, es necesario leer las coordenadas geográficas de muestras y crear identificadores de muestra. Hay 60 muestras de población de 10 individuos en cada población. Esto puede hacerse de la siguiente manera.

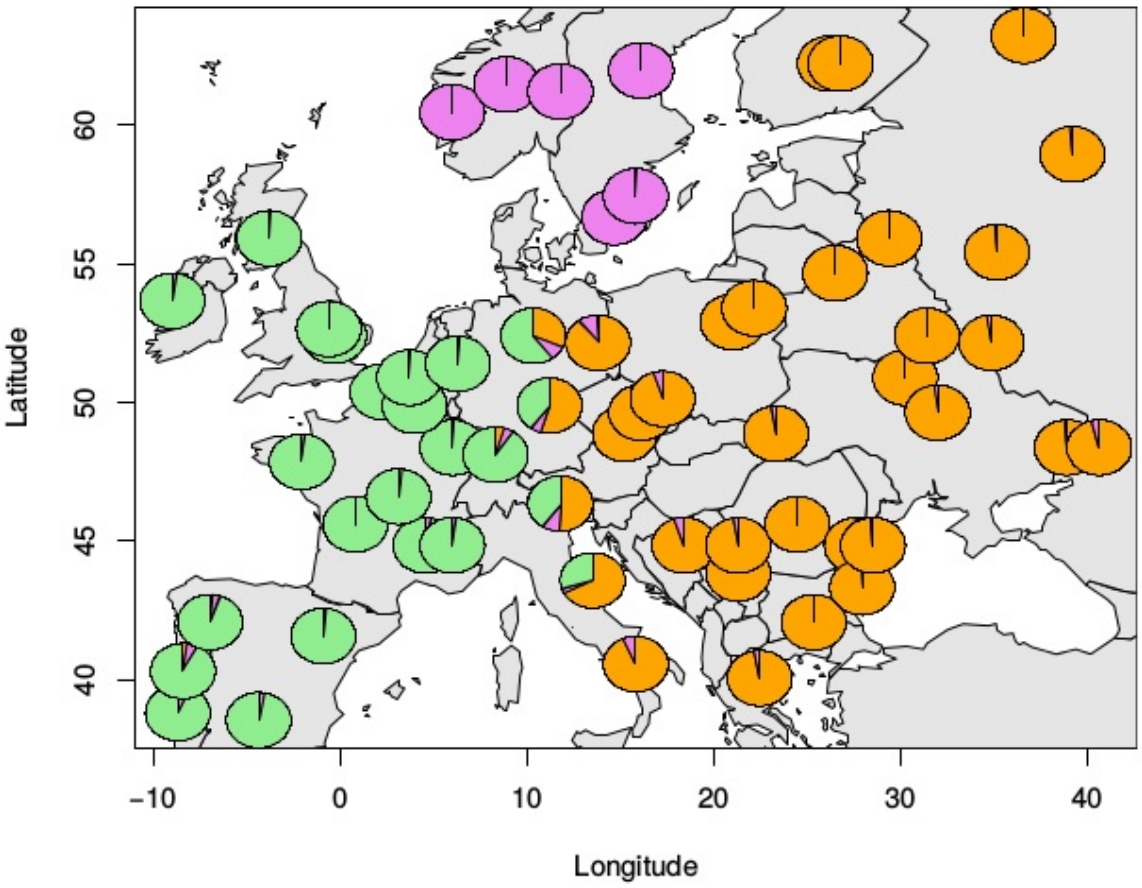
```
coord = read.table("coordinates.coord") pop = rep(1:60, each = 10)
```

Las estimaciones de los coeficientes de mezcla se pueden obtener mediante la adopción de los valores medios para cada una de las muestras de población 60.

```
 K = 3 Npop = length(unique(pop)) qpop = matrix(NA, ncol = K, nrow = Npop) coord.pop =
matrix(NA, ncol = 2, nrow = Npop) for (i in unique(pop)){ qpop[i,] = apply(qmatrix[pop ==
i,], 2, mean) coord.pop[i,] = apply(coord[pop == i,], 2, mean)}
```

En este código, el objeto qpop contiene los coeficientes de mezcla para cada muestra de la población. La dimensión de la matriz es ahora 60 × 3. Un mapeo geográfica de adm población.

```
plot(coord, xlab = "Longitude", ylab = "Latitude", type = "n")
map(add = T, col = "grey90", fill = TRUE)
for (i in 1:Npop){
  add.pie(z = qpop[i,], x = coord.pop[i,1], y = coord.pop[i,2],
labels = "",
  col = c("orange","violet","lightgreen"))}
```



Mapa de las estimaciones de la población mezcla utilizando K = 3 grupos. Las poblaciones de ascendencia mixta se identifican en Europa Central (zona de contacto).

Se debe tener en cuenta que la reproducción de estos comandos con sus propios datos, será necesario que se carga un vector de adición de etiquetas de población (un número entero para cada individuo).

```
pop = scan("mypop.txt")
```

Eligiendo el número de clusters

En la LEA, la eleccion del número de grupos que se basa en el criterio de entropía cruzada. Este criterio también es utilizado por el programa de mezclas. El criterio de entropía cruzada se basa en la predicción de una fracción de los genotipos enmascarados (terminación de la matriz), y en el enfoque de validación cruzada. Los valores más bajos del criterio de entropía cruzada por lo general significan mejores carreras. Llevamos a cabo carreras de 8 valores de K, y elegir el valor de K para el que la curva de entropía cruzada exhibe una meseta (K = 3, Figura 3).

```
 obj.snmf = snmf("secondary_contact.geno", K = 1:8, alpha = 100, project = "new")
plot(obj.snmf, col = "blue4", cex = 1.4, pch = 19)
```