Seminario III: R/Bioconductor Agosto-Diciembre 2009

Leonardo Collado Torres

Licenciado en Ciencias Genómicas,
UNAM, Cuernavaca, México

lcollado@lcg.unam.mx

9 de agosto de 2009

Ayudantes: Alejandro Reyes areyes@lcg.unam.mx, José Reyes jreyes@lcg.unam.mx y Víctor Moreno jmoreno@lcg.unam.mx

Resumen

El curso Seminario III: R/Bioconductor será impartido en la Licenciatura de Ciencias Genómicas (LCG) de la UNAM los viernes de 12:00 a 14:00. Dicho curso profundizará en Bioconductor que es el conjunto de herramientas públicas, montadas en R y desarrolladas para el estudio de la genómica. Para tomar esta clase es necesario tener un manejo básico de estadística por un lado y de R por el otro. En el segundo caso se impartió un curso de R introductorio a la sexta generación de la LCG disponible en http://www.lcg.unam.mx/~lcollado/E/. El orden en el que se cubrirá el material y el proyecto asociado a esta materia estará integrado y directamente relacionado al curso de Bioinformática y Estadística I de la LCG.

La página oficial del curso es http://www.lcg.unam.mx/~lcollado/B/. Allí pueden encontrar las presentaciones, los códigos asociados a las presentaciones, los ejercicios, las respuestas esperadas, el material de apoyo, y los datos que vayamos a usar.

Para cualquier duda, pregunta, sugerencia o para pedir asesorías, favor de hacerlo a través del foro del Nodo Nacional de Bioinformática (NNB) en esta sección.

Índice

1. Objetivos

- Introducir a los estudiantes al mundo de Bioconductor con el fin de que usen las herramientas más actualizadas de genómica montadas en `R`.
- Expander el conocimiento y habilidad de hacer gráficas con tres variables y de orden genómico.
- Aprender a importar datos públicos a `R` usando Bioconductor: `biomaRt`.
- Conocer y tener las habilidades básicas para el estudio de microarreglos usando Bioconductor.
- Manejar secuencias y el software desarrollado para el análisis de datos derivados de la secuenciación masiva en Bioconductor.
- Cimentar las bases de la investigación reproducible y la práctica de compartir los datos vía un paquete de Bioconductor.
- Entrenar y preparar a los futuros herederos de `R` y Bioconductor dentro de la LCG. En un año los ayudantes actuales serán los profesores y algunos alumnos tomarán el puesto de ayudantes.

2. Proyecto

Durante el semestre actual, los alumnos desarrollarán en equipo un proyecto involucrando `Perl`, `MySQL` y `PHP` para la materia de Bioinformática y Estadística I. En la página final de dicho proyecto deberán presentar un análisis estadístico hecho con `R` y posiblemente usando Bioconductor. El proyecto de Seminario III: R/Bioconductor consiste en armar un paquete de datos experimentales para Bioconductor basado en el proyecto de la otra materia. Dicho paquete debe cubrir todos los requisitos de Bioconductor los cuales incluyen una documentación tipo *vignette* hecha con `Sweave` y LaTeX. En dicho archivo, escrito en inglés[1], deben explicar la idea original que motivó su proyecto, cómo obtuvieron los datos, los análisis que hicieron incluyendo código de R, gráficas y conclusiones. Los datos los podrán importar de su base de datos de `MySQL` simplemente usando el paquete `RMySQL`. La pregunta que haya motivado al proyecto puede ser simple aunque la extracción de los datos no debe ser trivial.

En otras palabras, su proyecto será un ejercicio real que contribuirán a la comunidad internacional vía Bioconductor.

[1] En realidad todo debe estar en inglés, incluyendo los nombres de sus variables

3. Una clase ejemplo

Una clase *normal* se desarrollará de la siguiente forma. En los primeros minutos los ayudantes o el profesor le preguntarán a uno o varios alumnos sobre temas que les parecieron interesantes que surgieron en la *mailing list* de Bioconductor durante la semana o en algún artículo relacionado. Mientras, uno o dos alumnos se prepararán[2] y a continuación expondrán un paquete de Bioconductor[3]

- describiendo brevemente para que sirve
- que imágenes se pueden derivar de su uso
- porque les pareció interesante
- para que tipo de análisis se usa
- con que otros paquetes de Bioconductor se complementa o si hay algún paquete que sea parcialmente redundante

El o los alumnos que hayan expuesto deberán entregar un archivo tipo *vignette* en inglés con la anterior información el cual compartiremos con la clase vía la página oficial del curso. Posteriormente se procederá al tema de la clase que en general incluirá una descripción del paquete, ejemplos y prácticas. Finalmente se les pedirá a los alumnos que hagan una práctica avanzada/completa que muy probablemente terminarán en su casa como tarea.

4. Evaluación

Su calificación dependerá de cuatro factores:

Participación 20 %
: Su participación en las clases, en leer la *mailing list* y/o encontrar artículos relacionados de interés, en preguntar dentro de nuestro foro.

Tareas 30 %
: Toda tarea tendrá como fecha límite de entrega las 9 am[4] de los viernes. Deberán ser portables, osea que no dependa de su estructura de carpetas y que los datos estén disponibles vía en línea[5]. Para cada tarea entregarán

[2] Prender y conectar la lap para proyectar su presentación

[3] Sin repetirlos

[4] Tiempo servidor!!

[5] Ya sea en un sitio web, vía algún paquete como `biomaRt` o simplemente en su carpeta de `public_html` en el servidor de la LCG

dos archivos: el `pdf` generado con `Sweave` y LaTeX; el script .R generado con `Stangle`. Estos deberán estar nombrados con *username_XX_descrip* donde *XX* es el número de la tarea y *descrip* es lo que quieran poner. Por ejemplo: `lcollado_01_repaso.pdf` y `lcollado_01_repaso.R`.

La presentación de un paquete de Bioconductor 10 %
Dicha presentación debe cumplir los puntos mencionados en *Una clase ejemplo*. Consiste en una plática breve de aproximadamente 5 minutos y el archivo tipo *vignette* con la información mencionada en la presentación.

El proyecto 40 %
Elaborar un paquete de datos experimentales para Bioconductor. Debe cumplir sus requisitos[6] y el documento tipo *vignette* debe ser como un reporte. Es decir, debe tener:

- Un resumen o *abstract*
- Una introducción explicando de donde salió la idea/pregunta y cual es
- Describir como obtuvieron *minaron* los datos. Es decir, como los obtuvieron y porque escogieron esos.
- Un análisis con sus datos que contenga código de `R`, gráficas y resultados[7].
- Conclusiones

5. Programa *tentativo* de las clases

14 Ago Clase I

Repaso Iniciando el curso, repaso y funciones apply

1. Descripción del curso incluyendo el proyecto y la evaluación.
2. Buscar ayuda en `R`.
3. Ejercicio con `for` y un par de gráficas.
4. Familia de funciones apply.

21 Ago Clase II

Bioconductor y documentación Fomentando la investigación reproducible

1. Intro a Bioconductor

[6] Todo (variables, funciones, texto, etc) en inglés
[7] No se les olvide interpretarlos!!

2. Ayuda dentro de Bioconductor: *mailing lists*
3. Instalación básica de paquetes
4. Bases de la investigación reproducible y ejemplos tipo *vignette*
5. `Sweave` como interface de `R` con LaTeX
6. Corta introducción a LaTeX y Beamer

28 Ago Clase III

Gráficas Gráficas avanzadas de uso general

1. Panorama de las gráficas que se pueden hacer con `lattice`
2. Ejemplos de gráficas con `Plotrix`

4 Sept Clase IV

`biomaRt` Acceso desde `R` a biomart

1. Explorando biomart
2. Construcción básica de un mart
3. Una serie de ejemplos

11 Sept Clase V

Interacción con `MySQL` Usando `RMySQL` y aprendiendo `annotationdbi`

1. Instalación de `RMySQL`
2. Conexión a una base de datos con `RMySQL`
3. Uso de `R` para construir *queries* de `MySQL`
4. Descripción de `annotationdbi`

18 Sept Clase VI

Gráficas *genómicas* Visualización de muchos datos a la vez

1. Descripción de `GenomeGraphs`
2. Ligando `GenomeGraphs` con `biomaRt`
3. Interacción con *Genome Browsers* como el de `UCSC` vía `rtracklayer`

25 Sept Clase VII

Microarreglos El primer bastión de Bioconductor

1. Bases de las regresiones lineales
2. Correlaciones básicas

3. Uso de `limma` para encontrar los genes diferencialmente expresados
4. Paquete `affy`

2 Oct Clase VII

Análisis de secuencias Las herramientas básicas

1. Uso de `IRanges`
2. Generación de *vistas*
3. Manipulación de secuencias con `Biostrings`
4. Alineando secuencias con `Biostrings`

9 Oct Clase IX

`R` **y *HTS*** Control de calidad y algunos análisis

1. Control de calidad de datos de `Solexa` usando `ShortRead`
2. Un tipo de análisis usando `chipseq`

16 Oct Clase X

Paquetes de Bioc Construyendo un paquete de `Bioc`

1. Estructura de un paquete básico de `R`
2. Requisitos para un paquete de datos experimentales para Bioconductor

23 Oct Clase XI

GOs Análisis de GO

1. Uso de `BLAST`
2. Diversos análisis de GOs con `R`

30 Oct Clase XII

Estadística misc

1. Lowess y loess
2. Corrección al hacer múltiples pruebas

6 Nov Clase XIII

Microarreglos II Una sesión más detallada

1. Paquete `multtest`
2. Metiéndonos más en detalle

13 Nov Clase XIV

HTS: un caso Transcriptoma de *E. coli*

1. Detallando un análisis

20 Nov Clase XV

Clase no definida Abierta a sugerencias

1. Invitado

27 Nov Clase XVI

Clase no definida Abierta a sugerencias

1. por ver

30 Nov - 4 Dic Primera Semana de Exámenes

Asesorías Detallando su análisis

1. Asesorías para detallar el análisis estadístico de su proyecto de Bioinformática y Estadística I
2. Asesorías para armar su paquete de Bioconductor

7-11 Dic Segunda Semana de Exámenes

Proyectos Entrega y evaluación

1. Entrega del proyecto[8]
2. Evaluación del proyecto
3. Revisiones y correciones del proyecto
4. Mandarlo a Bioconductor[9]

[8]Posiblemente el lunes
[9]Posiblemente el viernes