Índice

1. Introducción

1.1. Planteamiento del Problema

El cáncer de mama es una enfermedad compleja y heterogénea con más de 1,300,000 casos y 450,000 muertes cada año en todo el mundo. Esta enfermedad se caracteriza por diferentes ápectos biológicos como desregulazación de la expresión génica, alteraciones genómicas del ADN, etc. Todo esto da lugar al inicio y desarrollo del carcinoma de mama. En éstos últimos años el uso de datos ómicos, como los basados en microarreglos (microarrays) y secuenciación, esta en su pleno () en el campo de la biomedicina. Todos estos datos permiten estudiar enfermedades desde un punto de vista biomolecular. Con esto, se ofrecen grandes oportunidades para mejorar tanto la comprención de la enfermedad, como el desarrollo de nuevos métodos para el diagnóstico y tratamiento del paciente, sin embargo, el análisis de estos datos producidos por estas tecnologías, es bastante complejo por lo que es necesario la aplicación de avanzadas técnicas de análisis y cálculos computacionales que permiten obtener la información biológica disponibles. Hasta el día de hoy todos estos datos óptenidos de diversos experimentos, datos de muy alta calidad, datos clinicamente bien anotados, y una gran cantidad de datos de canceres analizados con el fin de encontrar anomalias recurrentes que sean importantes de la enfermedad y estos datos se guardan en diversas plataformas que permiten tener uns gran cantidad de información , pero estos datos no son tan faciles de analizar, por ello, la bioinformática ayuda a manejar estructurar y organizarla para que sea mas fácil de comprender. Circos plot es una de las herramientas que existen para la visualización datos, ideal para

explorar las relaciones entre objetos y posiciones. Esta herramienta es flexible, aunque originalmente fue diseñado para visualizar datos genómicos , se puede crear figúras a partir de datos en cualquier campo, desde la genómica hasta la visualización de la migración al arte matemático. Esta herramienta puede ser automatizada.

1.2. Justificación

Con el fin de mejorar el rendimiento y la cantidad de tiempo de los investigadores del Instituto Nacional de Medicina Genómica, es fundamental sistematizar este software para reducir el tiempo de programación y la investigación. La sistematización de circos plot, brindará la posibilidad de que el investigador ahorre en tiempo de programación o en dado caso que no se conozca nada del uso nativo de circos plot, leer todo el manual de uso de dicho software, para que ocupe su mayor de tiempo en la investigación y solo tome varios minutos para diseñar su grafica circular llamada circos plot.

1.3. Delimitación

1.4. Hipótesis

El desarrollo de una herramienta de Visualización basada en integral permitirá el análisis de la información de los datos derivados de múltiples plataformas.

1.5. Objetivos

1.5.1. Objetivo General

Implemetar una herramienta de visualización que integre datos genómicos derivados de múltiples plataformas.

1.5.2. Objetivos Específicos

- Realizar el pretatamiento de los datos de trascriptoma (Microarreglos).
- Realizar el pretatamiento de los datos de Metilación.
- Desarrollar algoritmo computacional para hacer la búsqueda en la referencia del genoma humano con los datos de trascriptoma y metilación

1.6. Aportaciones de la tesis

2. Estado del Arte

2.1. Tecnologías Genómicas

Las tecnologías genómicas es el conjunto de herramientas orientadas al estudio integral del funcionamiento, contenido, evolución del genoma. Es una de las áreas más vanguardistas de la biología. La genómica usa conocimientos derivados de distintas ciencias como la biología molecular, la bioquímica, la informática, la estadística, las matemáticas y la física. Para entender un poco más de estas tecnologias y de los datos que se obtiene de las antes mencionadas, hablaremos de las tecnologias genómicas que son: Microarreglos y Metilación.

2.2. Microarreglos

Un chip de ADN (del inglés DNA microarray) es una superficie sólida a la cual se une una colección de fragmentos de ADN. Las superficies empleadas para fijar el ADN son muy variables y pueden ser de vidrio, plástico e incluso de silicona. Los chips de ADN se usan para analizar la expresión diferencial de genes. Su funcionamiento consiste, básicamente, en medir el nivel de hibridación entre la sonda específica (probe, en inglés), y la molécula diana (target), y se indican generalmente mediante fluorescencia y a través de un análisis de imagen, lo cual indica el nivel de expresión del gen.

El tamño de los arreglos es de 1.28 cm x 1.28 cm, hay 500,000 ubicaciones en cada matriz y por lo general tiene millones de cadenas de ADN construidas en cada ubicación, cada cadena contiene 25 pares bases (Figura 1)

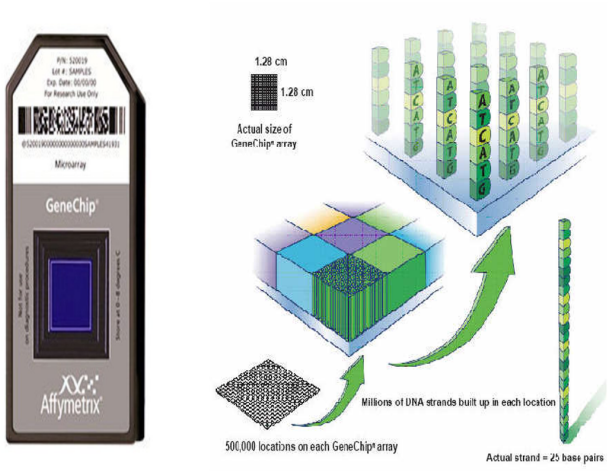

Figura 1: Chip de un Microarreglo.

2.2.1. Formato

Los archivos están disponibles en un formato de valores separados por comas (CSV). Estos son archivos de texto sin formato con cada fila terminada por un carácter de nueva línea. Los datos en campos separados están entre comillas y separados por comas. Ninguno de los campos de datos contiene ninguno de estos caracteres: comillas, nueva línea, retorno de carro o tabulación.

Estos archivos se usen principalmente en aplicaciones de hojas de cálculo y programas de bases de datos (como bases de datos SQL). Los datos estan formateados de tal manera que estos dos usos sean relativamente fáciles. Se tiene en cuenta que algunos de los archivos y los campos de datos en ellos son grandes.

La primera fila de cada archivo contiene los títulos de los campos que figuran en las filas siguientes.

Cada fila después de la primera fila contiene anotaciones para un solo conjunto de sondas. Todas las anotaciones para ese conjunto de sonda están contenidas en esa única fila. En algunos campos, como las anotaciones de dominio de proteínas, puede haber más de una anotación para un único conjunto de sondas. En este caso, los valores múltiples están separados por la cadena '///'.

En muchos tipos de anotaciones, los subcampos están separados por ’//’. Por ejemplo, una anotación para un "GO Biological Process"puede aparecer como "7155 // cell adhesion // predicted / computed". En este caso, las secciones corresponden a ÏD // Descripción // Evidencia", pero el significado de los subcampos varía entre los diferentes tipos de anotación, como se describe a continuación.

Los campos vacíos se indican con ’- - -’ . El hecho de utilizar una cadena de este tipo en lugar de dejar el campo vacío es que hace que la naturaleza columnar de los datos sea más visible en ciertos programas de hoja de cálculo. Algunas columnas en algunos archivos no contienen datos. Para ayudar a los usuarios a combinar datos de varios archivos, dichas columnas vacías no se eliminan. Por lo tanto, cada archivo tiene las mismas columnas en el mismo orden.

Algunos campos, como Çhip", contienen el mismo valor para cada conjunto de sonda en un archivo. Aunque estos datos son redundantes en cualquier archivo individual, son útiles para los usuarios que combinan datos de varios archivos.

2.3. Metilación

La metilacion del ADN es un proceso epigenético que participa en la regularización de la expresión génica de dos maneras, directamente al impedir la unio de factores de transcripción, e indirectamente proporcionando la estructura çerrada"de la cromatina.

3. Propuesta de Solución

El enfoque de solución ante la problemática planteada, consiste fundamentalmente en lo siguiente:

- Incorporar adecuadamente circos plot a un interface de usuario para su fácil uso.
- Los controles de manejo para el usario sencillos y factibles con todas las funciones posibles de circos plot.
- Incoporparar Circos plot a la plataforma de Galaxy.
- Instalar circos plot galaxy en Instiuto Nacional de Medicina Genómica.

4. Metodología

Circos es complejo de utilizar en su estado nativo, por lo que al automatizarlo se podrá usar de una forma mas sencilla.

Los pasos a seguir en esta metodología son los siguientes:

- Instalar Circos plot.
- Instalar Galaxy.
- Instalar la herramienta de circos galaxy.
- Errores comunes en instalación.

- Herramienta de galaxy para obtener formatos de circos.
- producto final.

4.1. Circos plot

.

Instalación de circos

Primero, descargamos Circos `http://circos.ca/software/download`. El contenido de la distribución se describe a continuación.

No necesitamos mover o editar ningún archivo en la distribución principal.

Listing 1: Suponiendo que desea instalar en ROOT= /software/circos

```
$ cd ~
$ mkdir software
$ mkdir software/circos
$ cd software/circos
# Descargar Circos y colocalo en el directorio software/circos
$ wget http://circos.ca/distribution/circos-0.69-6.tgz
# Descargar la versión mas actual (Recomendación).
#Descomprime
$ tar xvfz circos-0.69-6.tgz
...
circos-0.69-6/data/karyotype/karyotype.arabidopsis.txt
circos-0.69-6/data/karyotype/karyotype.zeamays.txt
circos-0.69-6/data/karyotype/karyotype.oryzasativa.txt
# Crea un enlace simbolico a current
$ ln -s circos-0.69-6 current
# Comprobamos si se creo el enlace simbolico.
$ ls -lh
drwxrwxr-x 9 juanjo juanjo 4.0K nov 29 10:36 circos-0.69-6/
-rw-rw-r-- 1 juanjo juanjo  22M nov 29 10:36 circos-0.69-6.tgz
lrwxrwxrwx 1 juanjo juanjo   13 nov 29 10:30 current -> circos-0.69-6/
# Borramos el archivo tgz, si ustede quiere
```

Para instalar los módulos GD y Perl en Ubuntu, usamos apt-get.

```
$ sudo apt-get -y install libgd2-xpm-dev
```

CORRIENDO CIRCOS

Circos utiliza banderas de línea de comandos, que son obligatorias. Por lo menos, debe especificar el archivo de configuración de imagen usando -conf.

Es una buena idea agregar el **bin/** directorio en la distribución para PATH que pueda ejecutar **bin/circos** desde cualquier lugar.

Añadimos al **root= /software/circos/current** como se describió anteriormente, añadimos esto a nuestro **/.bashrco /.bash-profile**.

```
$ export PATH=~/software/circos/current/bin:$PATH
```

Ejecutamos explícitamente cualquiera **/.bashrc /.bash-profile** para que esto surta efecto

```
$ .~/ .bashrc
# o
$ .~/ .bash_profile
```

Finalmente, probamos que nuestro PATH ha sido modificado,

```
$ cd ~
$ echo $PATH
~/software/circos/current/bin: ...
$ which circos
~/software/circos/current/bin/circos
```

Revisando si faltan módulos Perl

Verificamos si tenemos algún módulo faltante

```
$ circos -modules
ok       1.36 Carp
ok       0.38 Clone
ok       2.63 Config::General
ok       3.56 Cwd
ok      2.158 Data::Dumper
ok       2.54 Digest::MD5
ok       2.85 File::Basename
ok       3.56 File::Spec::Functions
ok     0.2304 File::Temp
ok       1.51 FindBin
ok       0.39 Font::TTF::Font
ok       2.53 GD
ok        0.2 GD::Polyline
ok       2.45 Getopt::Long
ok       1.16 IO::File
ok      0.413 List::MoreUtils
ok       1.41 List::Util
ok       0.01 Math::Bezier
ok     1.9997 Math::BigFloat
ok       0.07 Math::Round
ok       0.08 Math::VecStat
ok       1.03 Memoize
ok    1.53_01 POSIX
ok       1.26 Params::Validate
ok       1.64 Pod::Usage
ok       2.05 Readonly
ok 2016060801 Regexp::Common
ok       2.64 SVG
ok       1.19 Set::IntSpan
ok     1.6611 Statistics::Basic
ok    2.53_01 Storable
ok       1.20 Sys::Hostname
ok       2.03 Text::Balanced
ok       0.60 Text::Format
ok     1.9726 Time::HiRes
```

Cuando tenemos estas cosas ya tenemos circos plot instalado en nuestro SO

4.2. Galaxy

Galaxy es una plataforma abierta basada en la web para la investigación biomédica computacional accesible, reproducible y transparente.

Accesible: los usuarios sin experiencia en programación pueden especificar fácilmente parámetros y ejecutar herramientas y flujos de trabajo.
Reproducible: Galaxy captura información para que cualquier usuario pueda repetir y comprender un análisis computacional completo.
Transparente: los usuarios comparten y publican análisis a través de la web y crean páginas, documentos interactivos y basados en la web que describen un análisis completo.

Para obtener instalado galaxy necesitamos seguir los pasos siguientes:

Requisitos

- UNIX / Linux o Mac OSX
- Python 2.7

Empezar

Para producción o usuario único

Clonar Galaxy desde GitHub

```
$ git clone -b release_17.09 https://github.com/galaxyproject/galaxy.git
```

Comenzarlo
Galaxy requiere algunas cosas para ejecutar: un virtualenv, archivos de configuración y módulos dependientes de Python. Sin embargo, iniciar el servidor por primera vez creará / adquirirá estas cosas según sea necesario. Para iniciar Galaxy, simplemente ejecute el siguiente comando en una ventana de terminal:

```
$ cd ~
$ cd /galaxy
# En contadas ocaciones se requiere dara permisos de lectura escritura y ejecución para ello ponmos en
    terminal:
$ chmod -R 777 run.sh
$ sh run.sh
```

Esto iniciará el servidor Galaxy en el host local y el puerto 8080. Luego se puede acceder a Galaxy desde nuestro navegador web en **http:// localhost:8080.** Después de comenzar, el servidor de Galaxy imprimirá la salida a la ventana del terminal. Para detener el servidor Galaxy, se puede presionar las Ctrl+C en la ventana de la terminal desde la que se está ejecutando Galaxy.
Próximos pasos Convertirse en administrador

Para controlar Galaxy a través de la interfaz de usuario (instalación de herramientas, administración de usuarios, creación de grupos, etc.), los usuarios deben convertirse en administradores . Solo los usuarios registrados pueden convertirse en administradores. Para otorgar privilegios de administrador a un usuario, se completaron los siguientes pasos:

- En el directorio config/ viene el archivo de configuración pero lo encontramos como **galaxy.ini.sample** pero galaxy no lo reconoce así pero podemos tenerlo simplemente copiarlo como se muestra aqui:

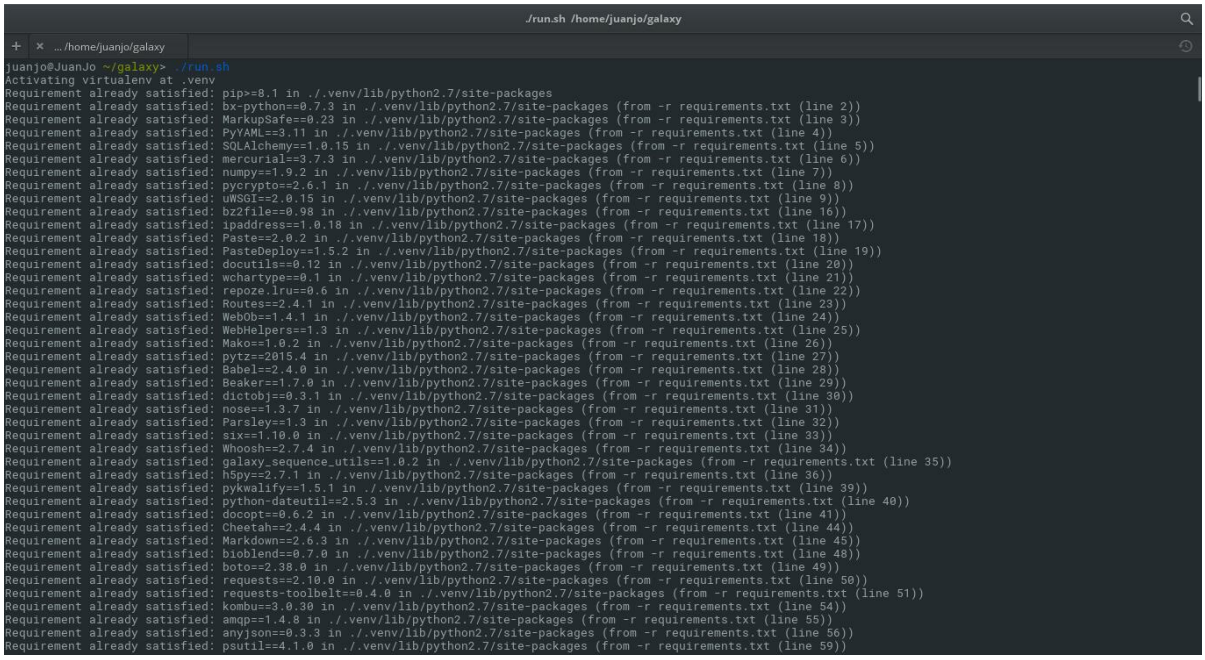

Figura 2: Consola corriendo galaxy.

```
$ cd /galaxy/config
$ cp galaxy.ini.sample galaxy.ini
$ vi galaxy.ini
```

- Agregamos el correo electrónico de inicio de sesión Galaxy del usuario al archivo de configuración config/galaxy.ini. Como se muestra aquí:

```
# Esta linea viene comentada por lo que hay que descomentarlar y ponemos:
admin_users = jjmartinez@inmegen.edu.mx
```

- Reinicié Galaxy después de modificar el archivo de configuración para que los cambios surtan efecto.

5. Cronograma

5.1. Referencias

TUSHER V. G., TIBSHIRANI R., CHU G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America, 98(9): 5116-21, 2001 Apr 24.

Metilación del ADN en cáncer de mama Diana Casandra Rodríguez-Ballesteros,* Sigfrid Leonardo García-Moreno-Mutio,* Joel Jaimes-Santoyo,* Rosa Elda Barbosa-Cobos,** Alberto de Montesinos Sampedro,*** Olga Beltrán-Ramírez*

Genómica, expresión génica y matrices de ADN David J. Lockhart 1 y Elizabeth A. Winzeler 1

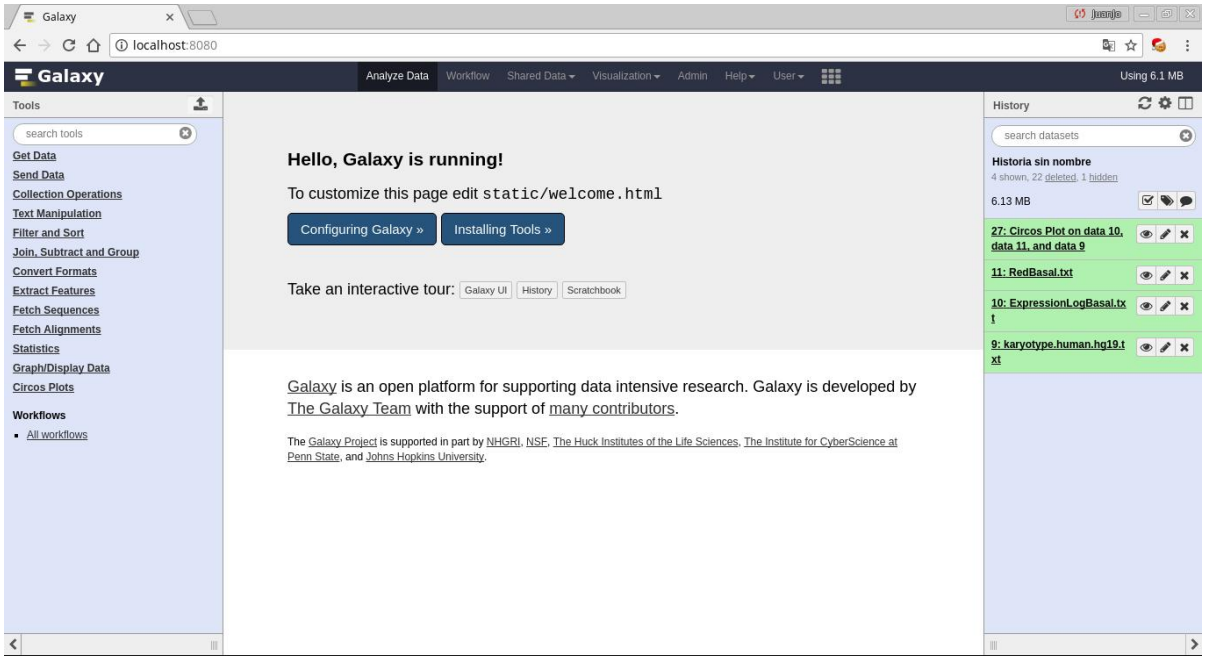

Figura 3: Pantalla principal de Galaxy.