FACULTAD DE MEDICINA
UNIVERSIDAD DE CHILE

GENOMED-Lab

ICBM
INSTITUTO DE CIENCIAS BIOMÉDICAS

TumorSec: Paquete tecnológico para el análisis y búsqueda de variantes genéticas en pacientes con cáncer de mama

Alejandro E. Blanco

Ricardo A. Verdugo

15 de septiembre de 2015

GENOMED-Lab— Independencia 1027, Independencia, Santiago

T: +56 (2) 2978-9528 **E:** alejandroblanco@med.uchile.cl

1. TumorSec

TumorSec es un paquete tecnológico para datos de secuenciación masiva en cáncer de mama. Permite el análisis y descubrimiento de mutaciones somáticas con gran exactitud y sensibilidad. Para obtener variantes somáticas confiables, se utilizan datos de secuenciación de una muestra de sangre y otra del tejido tumoral de un mismo individuo, permitiendo contrastar las variantes germinales, presentes en las células mononucleadas de sangre periférica del individuo, de las somáticas, presentes en las celulas tumorales. Está diseñado para trabajar con datos de secuenciación masiva dirigida ultra profunda de amplicones de PCR (Ultra Deep Targeted Sequencing, UDTseq (Harismendy et al., 2011)). TumorSec está compuesto por varias herramientas libres diseñadas para el análisis de datos genómicos, además de bases de datos públicas para realizar las anotaciones de las mutaciones detectadas.

2. Infraestructura Computacional

Con las nuevas tecnologías de secuenciación, los proyectos genómicos generan cada vez más volúmenes de datos, requiriendo la implementación de algoritmos sofisticados para su análisis. Por lo tanto es necesario contar con la infraestructura computacional adecuada que permita la obtención de resultados en un tiempo compatible con práctica clínica. Cuatro aspectos a considerar son: 1) Transferencia de datos desde el secuenciador, 2) Almacenamiento de datos, 3) Sistema de Respaldo y 4) Procesamiento y análisis de datos.

2.1. Transferencia de archivos desde el secuenciador.

Una vez que la secuenciación ha terminado, es necesario cambiar los datos al computador en donde se realizarán los análisis. Dicha transferencia puede ser a través de discos duros externos, pendrives o redes locales. En el Laboratorio de Patología Molecular del Cáncer de la Facultad de Medicina de la Universidad de Chile, el secuenciador MiSeq se encuentra conectado directamente al servidor de cómputo, ubicado en el Laboratorio de Bioinformática Genomed-Lab, a través de fibra óptica. La velocidad máxima de transferencia teórica en dicha red es de 10 Gbps si ambos equipos tienen el hardware necesario. En nuestro caso, el secuenciador MiSeq tiene una tarjeta de red que limita la transferencia de los datos a 1 Gbps. La transferencia de un archivo fastq de 49 Mb demora 1 segundo en ser transferido a una tasa de descarga promedio de 46968.1 KiloBytes/sec.

Un procedimiento importante de realizar luego de la transferencia de los datos, es revisar que los archivos fastq no estén truncados debido a alguna falla durante la transferencia desde el secuenciador. Se recomienda hacer un checksum de los datos antes de comenzar los análisis de búsqueda de variantes.

2.2. Almacenamiento de datos.

En promedio, por corrida de MiSeq, se generan 2.5 Gb de datos. El almacenamiento es un factor importante a considerar si se va a secuenciar una gran cantidad de muestras. A esto se le debe sumar el tamaño de archivos utilizados durante el proceso de análisis de los datos, como el genoma de referencia (~3Gb para la versión hg19), bases de datos como dbSNP (~11Gb para dbSNP versión 138 descomprimido) y el tamaño de los archivos resultantes del análisis con Mutascope, los cuales van a variar dependiendo del tamaño y número de reads de cada muestra. Como referencia, para una corrida de MiSeq con un output de ~2.6 Gb de datos, Mutascope genera ~2.2 Gb de datos.

2.3. Sistema de respaldo.

Los datos pueden ser almacenados en los servidores por el tiempo que estén en uso o mientras se estime conveniente. En el Laboratorio de Bioinformática de la Facultad de Medicina de la Universidad de Chile, contamos con un servidor de respaldo con 6 HDD de 3TB montados en RAID5, con lo que se tiene una partición de 14.5 TB. Este servidor es utilizado para realizar respaldos diarios de la carpeta home de seis computadores de escritorio y dos servidores con los que cuenta el laboratorio. Además, poseemos un grabador de cintas para realizar respaldos a largo plazo. Una cinta de respaldo tiene 1.6 TB de capacidad y asegura una integridad de los datos por décadas.

2.4. Procesamiento y análisis de datos.

En el Laboratorio de Bioinformática de la Facultad de Medicina de la Universidad de Chile, disponemos para proyectos UDTseq de un servidor SGI H2106-G7, que cuenta con 4 procesadores AMD Opteron (tm) Processor 6376 @2.3GHz 16 núcleos, 264 GB de RAM y 6 discos de 3 TB. Este equipamiento se ubica en una sala de servidores con aire acondicionado permanente y supervisión profesional constante. Dado los recursos disponibles, nos es posible analizar varias muestras simultáneamente. Por ejemplo, el análisis para un set de 212 amplicones (~35 kb) y una muestra pareada normal-tumor con ~510000 lecturas para los reads1 y reads2 de las muestras normal y tumor (~2 millones de lecturas en total), utilizando 1 núcleo, demora en el servidor SGI H2106-G7 174 min y 10 seg., con un tiempo de CPU de 179 min y 52 seg. Al utilizar el mismo servidor, pero esta vez con 8 núcleos para la etapa de alineamiento de las lecturas al genoma de referencia y 1 núcleo para el resto de los módulos de Mutascope, el mismo análisis demora 123 min y 17 seg. Adicionalmente, la misma muestra fue analizada en un computador de escritorio, con un procesador Intel Core i7-3770 CPU @ 3.40GHz x 8 núcleos, 16 GB de RAM, 1 SSD de 250GB y 1 HDD de 3 TB, utilizando 8 núcleos para la etapa de alineamiento y 1 núcleo para el resto de los módulos de Mutascope, demora 56 min y 55 seg. Evidentemente el computador de escritorio fue más rápido que el servidor al analizar una muestra, esto se debe a la velocidad del procesador y a la disponibilidad de un SSD que permite operaciones

de lectura escritura mucho más rápido. Sin embargo, la disponibilidad de un servidor permite el análisis de varias muestras al mismo tiempo, de este modo al ejecutar Mutascope utilizando el servidor SGI H2106-G7 para una corrida de secuenciación con 17 muestras pareadas normal/tumor, demora 283 min y 4.3 seg.

3. Software

El paquete tecnológico de análisis de datos TumorSec, tiene 3 niveles de software/herramientas que permiten realizar: 1) el análisis de detección de variantes, 2) control de calidad, 3) generación de reporte por corrida de secuenciación, y 4) reportes por paciente. Estos 4 niveles vienen empaquetados e integrados en una imagen Docker con el sistema operativo Ubuntu 14.04. Para información sobre la imagen de Docker, dirigirse a la sección 4. sobre empaquetamiento del flujo de trabajo TumorSec en una imagen de Docker.

3.1. Análisis de detección de variantes

Este nivel se basa en Mutascope (Yost et al., 2013), una suite diseñada para el análisis de datos de secuenciación masiva de amplicones de PCR, con un énfasis en la comparación de tejido normal y tumoral, permitiendo la identificación de mutaciones de baja prevalencia. Mutascope es un flujo de trabajo que consiste de 8 módulos programados en Perl (Figura 1) y utiliza herramientas desarrolladas por terceros como SAMTools (Li et al., 2009), BWA (Li and Durbin, 2010), Picard Tools, UCSCTools (Kent et al., 2002), R y GATK (McKenna et al., 2010). Para su correcto funcionamiento, este software requiere que el genoma de referencia a utilizar este previamente indexado (bwa index) y con su formato ".2bit" (utilizando el script *faToTwoBit* de UCSCTools). El indexado del genoma de referencia, solo se debe realizar una vez. Los 8 módulos de Mutascope pueden ser ejecutados en forma automática utilizando el módulo principal *runPipeline*. De manera alternativa, cada módulo puede ser ejecutado de manera independiente entregándoles los datos de entrada correspondientes y utilizando la estructura de datos requerida por Mutascope (Figura 1). Para mayor información acerca de Mutascope referirse al Manual oficial.

Con el objetivo de disminuir los tiempos de análisis de los datos de secuenciación de una corrida en el MiSeq, TumorSec viene con scripts escritos en bash que permiten la ejecución en paralelo de Mutascope para varias muestras al mismo tiempo, haciendo uso de toda la capacidad de cómputo que se encuentre disponible. Dentro de la imagen de Docker, hay una serie de pequeños scripts que ayudan a realizar dichas tareas de automatización. La ubicación de los scripts dentro de la imagen es:

```
/opt/Mutascope_tools
```

A continuación se muestran los pasos si se quisiera correr Mutascope con el módulo "*runPipeline*" en varias muestras al mismo tiempo. Si se tiene una carpeta con los archivos fastq a procesar, se puede utilizar el script (*create-jobs.sh*) para crear los trabajos en la carpeta actual.

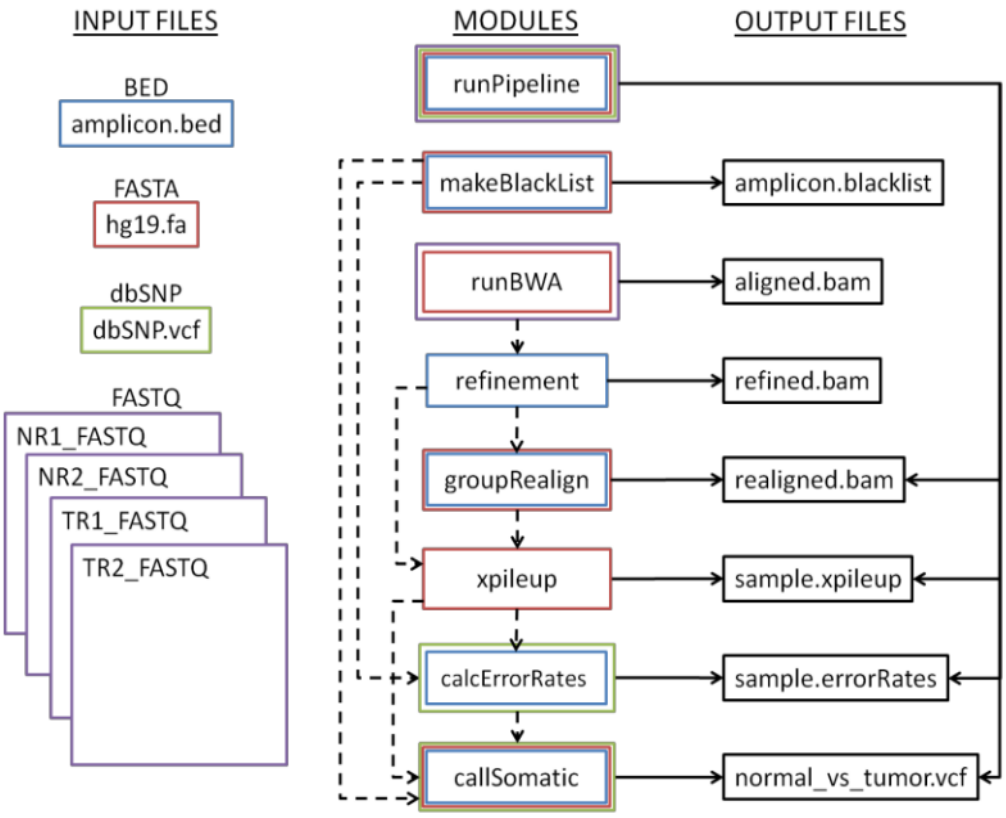

Figura 1. Flujo de trabajo general de Mutascope. Las 3 columnas corresponden a los archivos de entrada y salida, como también los módulos de Mutascope. Los módulos están marcados con diferentes colores, indicando los archivos de entrada necesarios. Las flechas continuas indican archivos de resultado o intermediarios generados por el módulo. Las flechas discontinuas indican las distintas secuencias de módulos posible. Por ejemplo, callSomatic requiere que se haya ejecutado y generado los archivos de salida de los módulos makeBlacklist, xpileup y calcErrorRates.

"*create-jobs.sh":* Este script recibe como parámetro la ruta de la carpeta donde están los archivos fastq, la ruta de la carpeta donde está el VCF de dbSNP, la ruta al genoma de referencia indexado con su formato .2bit, y por último la ruta al archivo .bed con los amplicones diseñados. Ejemplo de ejecución:

```
./create-jobs.sh  /path/to/fastq_folder/  /path/to/dbSNP.vcf
/path/to/reference.fa /path/to/amplicons.bed
```

Este script asume que los nombres de los archivos fastq tienen el formato presentado más abajo. Para un individuo secuenciado con una librería *paired-end* para sus muestras normal y tumoral, deberían existir los siguientes archivos:

```
AA0926G_S2_L001_R1_001.fastq
AA0926G_S2_L001_R2_001.fastq
AA0926T_S3_L001_R1_001.fastq
AA0926T_S3_L001_R2_001.fastq
```

Donde “AA0926” corresponde al ID del individuo, seguido de una “G” o “T”, que indica si es la muestra normal (Germinal) o Tumoral, respectivamente; “R1” indica que son las lecturas de un extremo, y “R2” corresponden a las lecturas del otro extremo, debido a que es una secuenciación *paired-end*. Al ejecutar este script, se generarán N carpetas, donde N es el número de muestras. Dentro de cada carpeta se crea un archivo que contiene el comando para correr el pipeline de Mutascope con el módulo “*runPipeline*”. El comando de Mutascope viene con la opción “–t 8” por defecto, es decir que utilizará 8 núcleos para la etapa de alineamiento con Bwa. Para cambiar el número de núcleos se debe cambiar manualmente el script.

A continuación, con la ayuda de otro script se pueden ir ejecutando los procesos para cada individuo. Esto también puede ser automatizado y paralelizado para correr múltiples individuos al mismo tiempo. El script “*masive-submits-parallel.sh*” hace este trabajo, permitiendo además controlar el número de individuos que se quieren analizar al mismo tiempo. Este script debe ser ejecutado en la raíz donde se crearon las carpetas para cada individuo con el script “*create-jobs.sh”*. Por defecto, el script “*masive-submits-parallel.sh*” permite la ejecución de Mutascope para 6 individuos al mismo tiempo. El número de procesos al mismo tiempo puede cambiarse modificando la línea:

```
…
MAX_NPROC=6    # Number of process to execute at the same time.
…
```

Para determinar el número máximo de individuos a ejecutar al mismo tiempo, se debe considerar lo siguiente: si un sistema Linux consta de 32 núcleos, el número de núcleos para correr Bwa dentro de Mutascope es definido como 8 (“-t 8”) y el número de individuos a analizar al mismo tiempo se define como 6, quiere decir que estaremos ocupando 8*6=48 núcleos como máximo en las etapas donde se utilice Bwa. Sin embargo, el sistema tiene 32 núcleos, por lo que se debe estimar y definir a priori el número de núcleos e individuos a paralelizar con tal de no exceder el máximo de recursos disponibles. Cabe destacar que el modo de ejecución de Mutascope mediante el script “*masive-submits-parallel.sh*”, funcionará en un servidor o sistema Linux sin gestor de colas. Si va a utilizar Mutascope en un clúster, debe ejecutarlo de acuerdo a los requerimientos del sistema.

Mutascope genera los resultados ordenados en tres carpetas. Una carpeta llamada “intermediate”, la cual contendrá archivos intermediarios que fueron utilizados durante el análisis (archivos de alineamiento BAM, xpileup y tasas de error). La segunda carpeta que genera se llama “quality”, en esta carpeta Mutascope guarda una serie de archivos de texto plano y gráficos con información relativa a la calidad de la secuenciación, mapeo de lecturas y llamado de variantes. Por último, la carpeta “results” contiene un archivo VCF con el llamado de variantes final, más otro archivo con el conteo de lecturas por amplicon tanto en la muestra normal como en la tumoral. Una vez finalizada la ejecución de Mutascope, recomendamos juntar todos los output de cada muestra en una misma carpeta (manteniendo la estructura de las carpetas “intermediate”, “quality” y “results”) para facilitar los análisis posteriores.

Una vez finalizado el análisis de Mutascope, el siguiente paso es la anotación del efecto para cada variante/mutación detectada. Esto puede ser realizado con el programa SnpEff (Cingolani et al., 2012). Para anotar los archivos VCF con SnpEff, se puede utilizar el script *"run-snpEff.sh"*. Este script recibe como parámetro la ruta donde se encuentran los archivos VCF obtenidos con Mutascope. Ejemplo ejecución:

```
./run-snpEff.sh /path/to/the/results/
```

El script anterior generara archivos VCF anotados con SnpEff, para cada individuo analizado, en una carpeta llamada "results_SnpEff" en la ubicación desde donde se ejecutó el script.

Finalmente, es posible generar una tabla final con las variantes detectadas en todos los individuos secuenciados en una corrida. Esta tabla puede ser generada con el script "*variants-table.sh*", el cual ejecutará VCFTools (Danecek et al., 2011) para extraer la información relevante de los VCF anotados con SnpEff, generando una tabla en formato CSV (Comma-Separated Values) para facilitar su manipulación con Excel, LibreOffice o algún programa parecido. Ejemplo para ejecutar el script "*variants-table.sh":*

```
./variants-table.sh /path/to/the/vcf-snpEff_files/
```

Este script generará una tabla en formato CSV con todas las variantes detectadas en todos los individuos. Dicha tabla tiene las siguientes columnas:

Nombre Columna	**Definición**	**Nombre Columna**	**Definición**
Ind	Id del individuo	fep	p-value somático calculado en el módulo callSomatic
chr	Cromosoma de la variante	nad	Cobertura del alelo referencia, alternativo en la muestra normal
pos	Posición (bp) de la variante	tad	Cobertura del alelo referencia, alternativo en la muestra tumoral
id	rsID in dbSNP	ngt	Genotipo en la muestra normal
ref	Alelo de referencia	tgt	Genotipo en la muestra tumoral
alt	Alelo alternativo	nfreq	Frecuencia alélica en la muestra normal
vt	Tipo de variante (SNP, INS o DEL)	tfreq	Frecuencia alélica en la muestra tumoral
filter	Filtros aplicados por Mutascope	nss	Estado de validación final en la muestra normal
qual	Puntaje de calidad en escala phred para el alelo alternativo	tss	Estado de validación final en la muestra normal
dp	Profundidad de cobertura total	snpEff	Anotación del efecto con SnpEff

3.2. Control de Calidad

En general, la calidad de las lecturas entregadas por el secuenciador es bastante buena, por lo que en la mayoría de los casos no es necesario filtrar lecturas antes de empezar los análisis con Mutascope. Sin embargo, pueden darse algunos casos que requieren de un filtro previo de lecturas, como cuando una corrida de secuenciación falla por algún motivo técnico generando un número inusual de lecturas de mala calidad, motivo que causa un alto porcentaje de lecturas que no pueden ser alineadas al genoma de referencia o que son alineadas en múltiples sitios. También puede darse el caso de que queden algunos adaptadores de la secuenciación en las lecturas, lo que va a influir en las etapas posteriores del análisis. En estos casos, es preferible filtrar por calidad y realizar un *trimming* de los adaptadores, previamente antes de comenzar los análisis de búsqueda de variantes, lo que mejora considerablemente el número de lecturas que se alinean a la referencia, en la mayoría de los set de datos (Schirmer et al., 2015).

Se ha demostrado que la mayoría de los errores de secuenciación se deben a problemas durante la etapa de preparación de la librería y en la amplificación por PCR (Schirmer et al., 2015). Un ejemplo al cual nos vimos enfrentados, fueron 3 muestras pareadas normal/tumor, que fallaron en una corrida de secuenciación (Figura 2). El problema fue debido a que para esas muestras, la concentración de ADN utilizada en la preparación de la librería fue demasiado baja. Además, en la misma corrida se utilizaron dos *filter plate* durante la preparación de la librería, uno de ellos era nuevo y otro reutilizado. Coincidentemente, las muestras que fallaron, fueron cargadas en el *filter plate* reutilizado junto con otras 3 muestras pareadas, las cuales presentaron un número de lecturas por debajo del promedio observado en las muestras cargadas en el *filter plate* nuevo (Figura 2). En este caso la solución fue volver a secuenciar las muestras que fallaron en una nueva corrida.

Otro problema que puede suceder es que se confunda el orden de algunas muestras en el momento previo a su carga en el secuenciador, generando inconsistencia de los datos. Por otro lado, el algoritmo encargado de separar las lecturas de acuerdo a los *barcodes* puede asignar erróneamente lecturas a otra muestra. Si se llegara a detectar o sospechar alguno de estos problemas se recomienda solucionarlos y limpiar los datos antes de seguir con el proceso de análisis.

Si se tiene una buena secuenciación, las lecturas de mala calidad que queden son filtradas en el mismo software Mutascope, al momento del alineamiento. Además, Mutascope genera varios archivos con diversas métricas que permiten evaluar la calidad de la secuenciación. Estos archivos tienen información variada sobre el alineamiento (lecturas totales secuenciadas, lecturas que fueron alineadas al genoma de referencia, lecturas que fueron alineadas en una sola región), número de lecturas por amplicon, uniformidad, sensibilidad y desviación estándar del número de lecturas por amplicon. Además es posible determinar coberturas promedios por gen (todos los amplicones que pertenecen a un gen determinado), para buscar si algún gen falló por completo o solo algún(os) de los amplicones diseñados.

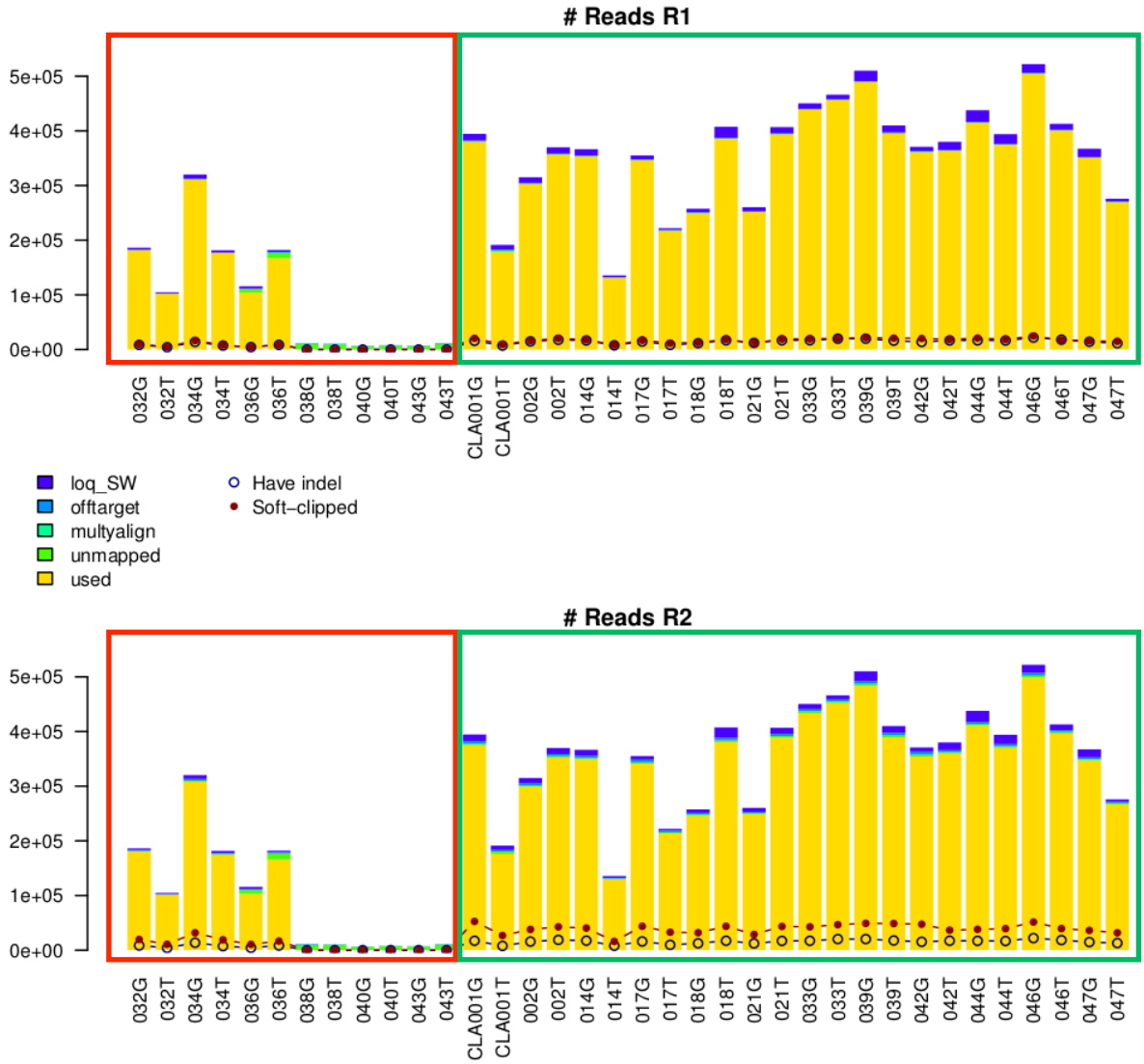

Figura 2. Número de lecturas en una corrida de secuenciación. Arriba se muestras las lecturas del read1, y abajo las del read2. El recuadro rojo enmarca las muestras que se cargaron en el *filter plate* reutilizado, mientras que el verde son las del *filter plate* nuevo.

En la imagen de Docker viene el script "*TruSeq_qc.R*", que permite generar figuras y contenido sobre el control de calidad de una corrida de secuenciación, y que será útil en la confección de los reportes por corrida de secuenciación. El script "*TruSeq qc.R*", recibe como parámetros la ruta donde están los archivos de calidad generados por Mutascope (carpeta "*quality*"), la tabla con todas las variantes generada con el script "*variants-table.sh*" y el orden de las muestras en el *filter plate*. Luego, el script es ejecutado de la siguiente manera:

```
Rscript TruSeq_qc.R /path/to/the/quality_files/
/path/to/the/variants.csv order_samples.txt
```

El orden de las muestras en el plate debe ser un archivo de texto plano con una columna que contenga los ID de los individuos en el orden establecido. El script genera como resultado una carpeta llamada "*TruSeq_qc*" con el siguiente contenido:

```
./TruSeq_qc
├── amplicons_coverage_boxplots.pdf
├── amplicons_coverage_cor_Blood.pdf
├── amplicons_coverage_cor_Tumor.pdf
├── amplicons_coverage.pdf
├── coverage.txt
├── reads_barplot.pdf
├── reads_statsAll.csv
├── reads_stats_by_side.csv
├── sensitivity.csv
├── subjects_per_variants.txt
├── uniformity.csv
├── variants_called.bed
├── variants_called.txt
├── variants_counts.txt
├── variants_freqs_hist.pdf
├── variants_freqs.pdf
├── variants_origin.txt
└── varplots
    ├── 010_variants_freqs.pdf
    ├── 038_variants_freqs.pdf
    ├── 040_variants_freqs.pdf
    ├── 043_variants_freqs.pdf
    ├── 045_variants_freqs.pdf
    ├── 048_variants_freqs.pdf
    ├── 049_variants_freqs.pdf
    ├── 050_variants_freqs.pdf
    ├── 051_variants_freqs.pdf
    ├── 052_variants_freqs.pdf
    ├── 053_variants_freqs.pdf
    ├── 054_variants_freqs.pdf
    └── 055_variants_freqs.pdf
```

Este es un ejemplo de los archivos generados por el script para una corrida en particular. La carpeta "*varplots*" tendrá N archivos dependiendo del número de individuos secuenciados en una corrida. Los archivos contenidos en la carpeta "*varplots*" corresponden a gráficos de las frecuencias alélicas de las variantes que pasaron todos los filtros en cada individuo de una corrida.

3.3. Reporte por corrida de secuenciación

Una vez obtenidos los resultados con Mutascope, se genera un reporte con los siguientes puntos básicos a considerar:

1) **Introducción:** Una breve descripción con información sobre el origen de las muestras, equipo utilizado y características generales del ensayo de Secuenciación Dirigida de Alta Profundidad.
2) **Materiales y Métodos:** Detalle de las muestras secuenciadas, kit de secuenciación utilizado, software y algoritmos utilizados para analizar los archivos FASTQ

3) **Resultados:**
    a. Control de Calidad: Se detalla aspectos específicos de la secuenciación. Concentraciones iniciales del ADN utilizado, junto con el orden de las muestras en el *filter plate.* Esto es útil para establecer correlaciones con los resultados de la secuenciación. También se analiza el número de lecturas pareadas generadas por cada muestra. Estadísticas sobre mapeo de las lecturas generadas en el genoma de referencia.
    b. Cobertura: Se detalla la cobertura promedio obtenida a nivel de genes y para cada amplicon entre todas las muestras secuenciadas en una misma corrida de MiSeq. Además se entregan estadísticas sobre la uniformidad, representada como la fracción de amplicones cubiertos por un número de lecturas dentro de 2 veces el promedio de lecturas por amplicon en una muestra. La sensibilidad también es reportada y corresponde al número de amplicones con cobertura menor a 50X, 100X, 500X, 1000X y mayor a 1000X por amplicon. Por último se crean *scatterplots* de correlación entre la cobertura por amplicón entre todas las muestras secuenciadas en una misma corrida.
    c. Detección de variantes: Se muestra el número total de variantes detectadas por individuo, clasificadas de acuerdo a si son somáticas o germinales. Además se detalla el número de variantes que pasaron, y las que no, todos los filtros aplicados por el llamador de variantes implementado en Mutascope. Se generan gráficos de las frecuencias alélicas de las variantes que pasaron todos los filtros en la muestra normal v/s tumoral detectadas en cada paciente.
    d. Frecuencia de filtros aplicados: Se muestran histogramas con la frecuencia de los filtros de Mutascope en las variantes descartadas. Esto con el objetivo de determinar que filtros son más determinantes en el llamado de variantes final.

Todo el contenido nombrado, es obtenido con el script "*TruSeq_qc.R*".

3.4. Reportes por paciente

En nuestra experiencia, no existe una base de datos especializada para la revisión de variantes de utilidad clínica en cáncer de mama y las que hay deben utilizarse en conjunto y ser revisadas cuidadosamente. Todas las variantes encontradas deben ser investigadas, puesto que algunas mutaciones que se encuentran localizadas en regiones intrónicas pueden ser de relevancia clínica. Por lo tanto es difícil generar un reporte automatizado por paciente que tenga todos los antecedentes relevantes a ser considerados por el clínico. Se requiere una revisión bibliográfica de las variantes encontradas y los resultados ser analizados por un equipo multidisciplinario que asesoren al oncólogo tratante en sus decisiones terapéuticas. Estos equipos constituyen los denominados "Tumor Boards", y suelen incluir a todos los profesionales involucrados en la generación y análisis de los datos genómicos, así como los médicos tratantes y expertos en el área.

No obstante de lo recién descrito, es posible realizar búsquedas automatizadas en las bases de datos disponibles para las variantes encontradas en cada paciente. Para dicho objetivo,

aconsejamos considerar, al menos, los recursos indicados en el Anexo 1. En la imagen de Docker, viene instalado el programa Annovar (Wang et al., 2010) con dos de las bases de datos descritas en el Anexo1 (COSMIC y ClinVar), facilitando la anotación de manera rápida para un archivo VCF. Además Annovar tiene una gran cantidad de bases de datos descargables para anotar variantes, y permite la creación de bases de datos personalizadas que estén en formato GFF3 (Generic Feature Format versión 3), para integrar información como PharmGkB. Como alternativa a PheGenI, Annovar posee una base de datos para encontrar variantes que han sido previamente asociadas a enfermedades o fenotipos en estudios de GWAS. Cabe destacar que es necesario mantener actualizadas las bases de datos de Annovar.

4. Empaquetamiento del flujo de trabajo TumorSec en una imagen de Docker.

Utilizando Docker ("docker," n.d.), se creó una imagen basada en Ubuntu 14.04, la cual contiene todas las librerías y programas necesarios para ejecutar Mutascope.
Para utilizar esta imagen es necesario tener instalado Docker en el sistema donde se quieren realizar los cálculos. Una vez instalado, la imagen debe ser descargada, esto se logra utilizando el siguiente comando:

```
docker pull tumorsec/mutascope
```

Esto descargará la imagen que pesa ~848 Mb. Una vez descargada, puede ser utilizada de inmediato. Para ejecutar la imagen se usa el siguiente comando:

```
docker run -it tumorsec/mutascope:latest bash
```

Este comando ejecutará la imagen, que es idéntica a un sistema Ubuntu Linux. Esta imagen viene con el software indicado en la tabla 1. Adicionalmente, se adjuntó un conjunto de scripts en bash y R para automatizar ciertos procesos del flujo de trabajo y/o generar resultados respectivos a controles de calidad (referirse a secciones 3.2 y 3.3 para detalles de scripts adicionales). La imagen también puede ser ejecutada montando uno o varios directorios del sistema Host en la imagen de Docker. Esto es útil para cargar los datos de secuenciación, genoma de referencia, dbSNP y todos los requerimientos especificados en el manual de Mutascope. Además, al montar directorios, los cambios realizados dentro de la imagen, se verán reflejados en el sistema Host, de esta manera si se monta un directorio vacío para guardar los resultados de los análisis, estos quedaran guardados en el sistema host y no se perderán al cerrar Docker. Para ejecutar la imagen montando un directorio del sistema host se ejecuta el siguiente comando:

```
docker run -it -v /ruta/directorio/en/host:/ruta/en/imagen
tumorsec/mutascope:latest bash
```

Ahora se podrán utilizar todos los datos del directorio que montamos dentro de la imagen de Docker. Si el directorio de la imagen especificado en el comando anterior no existe, es creado automáticamente.
Luego de cerciorarse de tener todo lo necesario para correr Mutascope, ejecutamos la imagen, la cual estará lista para ejecutar el flujo de trabajo completo. La imagen de docker

utiliza el mismo hardware del sistema host, por lo que están disponibles la misma memoria RAM y número de núcleos, facilitando paralelizar ciertos procesos del análisis como el alineamiento de las lecturas al genoma de referencia con Bwa.

Programa	Versión	Adicionales
gcc	4.8.4	-
Java	7	-
Perl	5.18.2	-
Bwa	0.7.12-r1039	-
Samtools	1.2	-
R	3.0.2	Paquete mclust
SnpEff	4.1	-
VCFtools	0.1.12b	-
Mutascope	1.0.2	-
Annovar	2015-06-17	Bases de datos Clinvar y Cosmic

Tabla 1. Software preinstalado en la imagen de Docker "tumorsec"

5. Bibliografía

Altman, R.B., 2007. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. Nat. Genet. 39, 426. doi:10.1038/ng0407-426

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92. doi:10.4161/fly.19695

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Group, 1000 Genomes Project Analysis, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi:10.1093/bioinformatics/btr330

docker [WWW Document], n.d. . docker. URL https://www.docker.com/ (accessed 8.7.15).

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., Stratton, M.R., 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al CHAPTER, Unit–10.11. doi:10.1002/0471142905.hg1011s57

Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., Teague, J.W., Stratton, M.R., Futreal, P.A., 2010. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. 38, D652–657. doi:10.1093/nar/gkp995

Forbes, S.A., Tang, G., Kok, C., Jia, M., Bamford, S., Cole, J., Dawson, E., Menzies, A., Teague, J.W., Stratton, M.R., Futreal, P.A., 2008. An Introduction to COSMIC, the Catalogue of Somatic Mutations in Cancer. NCI Nat. Pathw. Interact. Database. doi:10.1038/pid.2008.3

Harismendy, O., Schwab, R.B., Bao, L., Olson, J., Rozenzhak, S., Kotsopoulos, S.K., Pond, S., Crain, B., Chee, M.S., Messer, K., Link, D.R., Frazer, K.A., 2011. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol. 12, R124. doi:10.1186/gb-2011-12-12-r124

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, and D., 2002. The Human Genome Browser at UCSC. Genome Res. 12, 996–1006. doi:10.1101/gr.229102

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., Maglott, D.R., 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, D980–985. doi:10.1093/nar/gkt1113

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 26, 589–595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079. doi:10.1093/bioinformatics/btp352

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. doi:10.1101/gr.107524.110

Ramos, E.M., Hoffman, D., Junkins, H.A., Maglott, D., Phan, L., Sherry, S.T., Feolo, M., Hindorff, L.A., 2014. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur. J. Hum. Genet. EJHG 22, 144–147. doi:10.1038/ejhg.2013.96

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. gku1341. doi:10.1093/nar/gku1341

Thorn, C.F., Klein, T.E., Altman, R.B., 2010. Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics 11, 501–505. doi:10.2217/pgs.10.15

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164–e164. doi:10.1093/nar/gkq603

Yost, S.E., Alakus, H., Matsui, H., Schwab, R.B., Jepsen, K., Frazer, K.A., Harismendy, O., 2013. Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing. Bioinformatics. doi:10.1093/bioinformatics/btt305

Nombre	Descripción	URL	Acceso*	Ref.
dbSNP	Base de datos de polimorfismos SNPs y pequeños indels. Es mantenida por el “National Center for Biotechnology Information” (NCBI).	http://www.ncbi.nlm.nih.gov/SNP/index.html	P	-
COSMIC	Catálogo de Mutaciones Somáticas en Cáncer (COSMIC) diseñado para almacenar información sobre mutaciones somáticas relacionadas con el cáncer humano.	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/	P	(Forbes et al., 2010; S.A. Forbes et al., 2008; Simon A Forbes et al., 2008)
PharmGkB	Base de datos que se encarga de recolectar, curar y diseminar conocimiento sobre el impacto de las variaciones genéticas humanas sobre las respuestas a drogas. La información no está 100% curada y las relaciones no representan necesariamente una asociación directa entre las variables en cuestión.	https://www.pharmgkb.org	A	(Altman, 2007; Thorn et al., 2010)
ClinVar	ClinVar fue diseñado para proveer un acceso libre y público a archivos con reportes de relaciones entre variaciones humanas y fenotipos (estados de salud).	https://www.ncbi.nlm.nih.gov/clinvar/	P	(Landrum et al., 2014)
PheGenI	Integrador de fenotipo y genotipo. Une la información de estudios GWAS con una variada cantidad de bases de dato alojadas en NCBI (Gene, dbGaP, OMIM, GTEx y dbSNP).	http://www.ncbi.nlm.nih.gov/gap/phegeni#pgGAP	P	(Ramos et al., 2014)

*P: público; S: suscripción; A: bajo autorización del equipo.