OPTIMIZACIÓN DEL PREPROCESAMIENTO DE LECTURAS DE SECUENCIACIÓN DE NUEVA GENERACIÓN

OPTIMIZING THE PRE-PROCESSING OF NEXT GENERATION SEQUENCING (NGS) READS

RESUMEN

En este documento se presenta una herramienta bioinformática que integra y facilita el uso de utilidades de preprocesamiento de lecturas de secuenciación de nueva generación por medio de una interfaz web, y ayuda a los investigadores a enfocarse en el procesamiento y análisis de secuencias sin necesidad del manejo de las herramientas por medio de la consola. Se evalúa la funcionalidad de la herramienta y si compara con otras similares; se concluye que el aporte de este desarrollo es la integración de todas las etapas necesarias en el preprocesamiento y la facilidad de manejo por parte del usuario.

Palabras clave: control de calidad, *interleave*, *clipping*, lecturas, normalización, preprocesamiento, secuenciación de nueva generación, *trimming*.

ABSTRACT

In this paper is presented a bioinformatics tool that integrates and facilitates the use of pre-processing utilities for next generation sequencing reads through a web interface, researchers to focus on processing and sequence without the management tools through the console. The functionality of the tool is evaluated and compared to other similar tools; it concludes that the contribution generated by this development is the integration of all necessary pre-processing steps in and ease of handling by the user.

Key words: quality Control, interleave, clipping, reads, normalization, pre-processing, next generation sequencing, trimming.

José Nelson Pérez Castillo
Doctor en Informática
Docente Universidad Distrital Francisco José de Caldas
nelsonp@udistrital.edu.co
Bogotá, Colombia

Nelson Enrique Vera Parra
Estudiante Doctorado en Ingeniería
Docente Universidad Distrital Francisco José de Caldas
neverap@udistrital.edu.co
Bogotá, Colombia

Luis Miguel Gutiérrez Ramírez
Estudiante de Ingeniería de Sistemas
Universidad Distrital Francisco José de Caldas
lmgutierrezr@correo.udistrital.edu.co
Bogotá, Colombia

Tipo: Artículo corto

Fecha de Recepción: Junio 6 de 2014
Fecha de Aceptación: Octubre 25 de 2014

1. INTRODUCCIÓN

El preprocesamiento es un paso previo al ensamblaje que permite limpiar (substraer datos no pertinentes adicionados en el proceso de secuenciación) y acondicionar los archivos de lecturas para mejorar los tiempos de ejecución y reducir los errores en el proceso de ensamblaje. El preprocesamiento involucra varias etapas: *trimming*, *clipping*, normalización, control de calidad, *interleave* y conversión de formatos, lo cual obliga al investigador a tener destreza en varias herramientas bioinformáticas.

En este documento se presenta un software que integra y facilita el uso de herramientas de preprocesamiento por medio de una interfaz Web y ayuda a los investigadores a enfocarse en el procesamiento y análisis de secuencias sin necesidad del manejo de las herramientas por medio de la consola.

Este documento se organiza en 4 secciones: inicialmente se presenta una descripción general de NGSpre y se exponen sus funcionalidades, luego se describe su arquitectura, donde se explica cada uno de sus módulos, y posteriormente se realiza una evaluación de las funcionalidades del software y sus ventajas con respecto a otros similares.

2. DESCRIPCIÓN GENERAL

NGSpre es un software bioinformático orientado a la web (figura 1) que integra y facilita el uso de las herramientas necesarias para hacer limpieza y optimización de archivos de secuencias para su posterior ensamblado. NGSpre puede ser usado de forma local o en la nube.

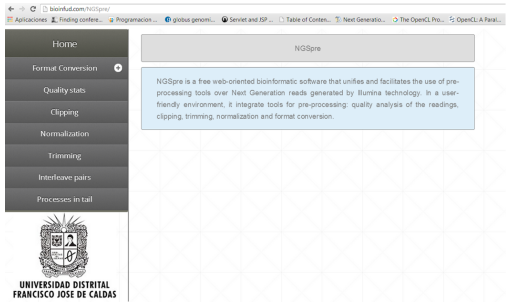

Figura 1. NGSpre

3. ARQUITECTURA

NGSpre está compuesto por 4 módulos que forman un flujo de trabajo (figura 2): el primero es una interfaz Web el segundo es el módulo de herramientas, el tercero es el módulo de gestión de base de datos y, por último, está el módulo para el procesamiento de las peticiones de los usuarios.

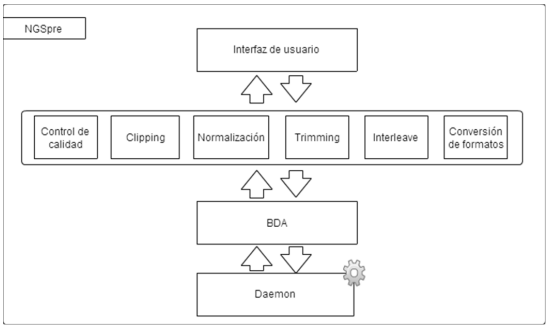

Figura 2. Arquitectura de NGSpre

3.1. Módulo de interfaz Web

Este módulo ofrece una interfaz gráfica que integra e intermedia la interacción entre el usuario y las herramientas de preprocesamiento. Para su funcionamiento utiliza 3 componentes: el primero es un servidor apache (http://httpd.apache.org/), el segundo es php (http://www.php.net/) para el procesamiento de las peticiones, y el último es html (http://www.w3.org/html/) para la interfaz gráfica con la que interactúa el usuario.

3.2. Módulo de herramientas

En este módulo se encuentran las herramientas de pre-procesamiento:

Control de calidad: esta herramienta realiza una asociación y organización entre las secuencias y los puntajes de calidad aportados por el secuenciador. Está basada en el script FastQC de Babrahan Bioinformatics [1]. La entrada de esta herramienta es un archivo con secuencias crudas y la salida es un archivo en formato FASTQ.

Clipping: en el proceso de secuenciación se uti-

lizan unos adaptadores que finalmente quedan agregados a las lecturas; estos adaptadores son innecesarios en las siguientes etapas de procesamiento. Esta herramienta se encarga de eliminar estos adaptadores utilizando el script Fastx_Clipper de FASTX_toolkit [2]. Tanto la entrada como la salida de esta herramienta son archivos de formato FAST o FASTQ.

Normalización: cuando se está trabajando con grandes sets de datos, como aquellos que exceden 300M pares bases, utilizar normalización in silico es lo más adecuado antes del ensamblado. La normalización es el proceso que se encarga de reducir las secuencias a una cobertura específica para hacer el proceso de ensamblaje menos intensivo computacionalmente, reduciendo los requerimientos de memoria y mejorando significativamente los tiempos de ejecución. Esta herramienta está basada en el script insilico_read_normalization de Trinity [3]. La entrada de esta herramienta es un archivo FASTQ o FASTA con las secuencias, una cobertura y un k-mer, y la salida es un archivo FASTA o FASTQ normalizado.

Trimming: esta herramienta le permite al investigador establecer un umbral de calidad para filtrar las secuencias que no estén dentro de él. Está basado en el script fastx_trimming de FASTX_toolkit y permite el procesamiento de archivos FASTA y FASTQ.

Interleave pairs: esta herramienta divide los archivos en secuencias pareadas y no pareadas (huérfanas) específicamente para las lecturas de Illumina. Está basado en el script interleave_pairs.py de lexnederbragt (https://github.com/lexnederbragt). Las entradas a esta herramienta son dos archivos FASTQ o FASTA, uno con las lecturas derechas y el otro con las lecturas izquierdas; y la salida son dos archivos FASTA o FASTQ, uno con las secuencias pareadas y otro con las secuencias huérfanas.

Conversión de formatos: esta herramienta es un conjunto de utilidades para la conversión de los siguientes formatos: .bam a .fastq / .bam a fasta / fastq a fasta

3.3. Módulo BDA

Módulo encargado de conectarse a la base de datos MySQL (www.mysql.com) y traer los datos necesarios para la ejecución de las herramientas por petición del usuario. Primero se encarga de recibir datos por parte de interfaz de las herramientas para hacer una cola de ejecución y luego le proporciona la información al Daemon para ejecutar las peticiones que se encuentren en la cola.

3.4. Módulo Daemon

El módulo Daemon es un script hecho en Python que se encarga de "escuchar" constantemente las peticiones de los usuarios por medio del módulo BDA y de esta forma procesar la cola de ejecución. Al finalizar cada proceso se notifica al usuario y continúa con el siguiente proceso, si lo hay.

4. METODOLOGÍA

Para evaluar NGSpre se compararon sus características y funcionalidades con 3 programas: ngsShoRT [4], NGSQC toolkit [5] y SeqTrimNext [6]. Las características comparadas fueron: portabilidad, interfaz amigable con el usuario, fácil instalación, fácil configuración y ejecución, *pipeline* a elección del usuario y disponibilidad Web. Las funcionalidades fueron: *trimming*, *clipping*, normalización, control de calidad, *interleave* y conversión de formatos, que son las más usadas para preprocesamiento. Para evaluar NGSpre se compararon sus características y funcionalidades con 3 programas: ngsShoRT [4], NGSQC toolkit [5] y SeqTrimNext [6]. Las características comparadas fueron: portabilidad, interfaz amigable con el usuario, fácil instalación, fácil configuración y ejecución, pipeline a elección del usuario y disponibilidad web. Las funcionalidades fueron: trimming, clipping, normalización, control de calidad, interleave y conversión de formatos, que son las más usadas para preprocesamiento

5. RESULTADOS Y ANÁLISIS

5.1. Análisis de características

En la tabla 1 se puede observar el resultado del comparativo entre los software . El resultado muestra, en cuanto a la portabilidad, que NGSpre está diseñado para funcionar operativamente sobre Linux debido a las herramientas que utiliza, además cuenta con una interfaz orientada a la web que la hace intuitiva para el usuario y con un fácil acceso a cada herramienta desde cualquier navegador en distintos sistemas operativos, lo cual le da una ventaja significativa, ya que la mayoría de herramientas para preprocesamiento sólo se puede ejecutar por medio de línea de comandos. Aunque la instalación de NGSpre necesita conocimientos básicos en Linux y el manejo de la consola, el usuario final nunca requerirá estos conocimientos.

Tabla 1. Arquitectura para NGSpre

Característica	ngsShoRT	NGSQC	SeqTrim Next	NGSpre
Portabilidad	NO	SI	NO	NO
Interfaz amigable con el usuario	NO	NO	SI	SI
Fácil instalación	SI	SI	NO	NO
Pipeline a elección del usuario	N/A	SI	SI	SI
Disponibilidad Web	NO	NO	SI	SI
Fácil configuración y ejecución	NO	NO	NO	SI

Por otro lado, gracias a la arquitectura de NGSpre, el usuario puede ejecutar las herramientas individualmente, permitiendo la ejecución de *pipelines* al gusto del investigador. Finalmente, en cuanto su configuración y ejecución, NGSpre resalta entre las otras herramientas, ya que para ejecutar un proceso tan sólo se necesita llenar un pequeño formulario con los parámetros necesarios.

5.2. Análisis de funcionalidades

Como se puede evidenciar en la tabla 2, NGSpre cuenta con las herramientas más usadas por los investigadores para preprocesamiento. A diferencia de los demás software que están enfocados en cierta parte del preprocesamiento, NGSpre intenta abarcar todos los aspectos posibles para dar al investigador una herramienta integral.

Tabla 2. Arquitectura para NGSpre

Característica	ngsShoRT	NGSQC	SeqTrim Next	NGSpre
Control de calidad	NO	SI	NO	SI
Clipping	NO	NO	SI	SI
Normalización	NO	NO	NO	SI
Trimming	SI	SI	SI	SI
Interleave	NO	NO	NO	SI
Conversión de formatos	NO	SI	NO	SI

6. CONCLUSIONES

Se ha logrado desarrollar un software que reúne las herramientas necesarias para el paso de preprocesamiento con una interfaz amigable con el usuario. Además, al analizar las características funcionales de este tipo de software se halló que NGSpre cuenta con la mayoría de éstas, lo que ayuda a a los investigadores en el proceso de análisis de datos genómicos y transcriptómicos.

Referencias

[1] S. Andrews. (2010). FastQC: A quality control tool for high throughput sequence data. Recuperado de http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] A. Gordon and G. J. Hannon. (2010).

"Fastx-toolkit. FASTQ/A short-reads pre-processing tools". Recuperado de:http://hannonlab. cshl. edu/fastx_toolkit.

[3] B. J. Haas et al., "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis". *Nature Protocols,* vol 8, no. 8, pp. 1494-1512, 2013.

[4] C. Chen, S. S. Khaleel, H. Huang, and C. H. Wu, "ngsShoRT: A Software for Pre-processing Illumina Short Read Sequences for De Novo Genome Assembly." *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, p. 706. ACM . Sept. 2013.

[5] R. K. Patel and M. Jain, "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data". *PLoS One*, vol. 7, no. 2, e30619 , 2012.

[6] J. Falgueras, A. J. Lara, N. Fernández-Pozo, F. R, Cantón, G. Pérez-Trabado y M. G. Claros, "SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read". BMC *bioinformatics*, vol. 11, no. 1, 38 , 2010.