**Apelidos, nome /Apellidos, nombre:** Galego Torreiro, Roi

**Título:**
Desarrollo de una herramienta para el control de calidad de secuencias genéticas en entornos Big Data

**Mención cursada:**
Ingeniería del Software

**Resumo / Resumen:**
La obtención de secuencias de ADN de seres vivos es, en muchas ocasiones, el primer paso en los estudios desarrollados por biólogos y bioinformáticos. En la actualidad, las tecnologías *Next Generation Sequencing* (NGS) permiten generar cientos de millones de secuencias en una única ejecución, disminuyendo drásticamente su coste. Sin embargo, la calidad de la secuenciación no es excesivamente alta en todos los casos. El resultado de los análisis bioinformáticos puede verse dañado a causa de una baja calidad de secuenciación en algunos fragmentos de ADN. Por ello, los análisis genéticos actuales suelen comenzar con una primera etapa donde se realiza un control sobre el conjunto de datos genómicos de entrada, eliminado o modificando aquellas secuencias que no son útiles.
Actualmente existen varias herramientas que permiten realizar este paso previo, cada una de ellas permitiendo aplicar diferentes tipos de controles. Sin embargo, con la explosión de datos genómicos disponibles debido al desarrollo de las tecnologías NGS, dichas herramientas pueden necesitar mucho tiempo de computación para procesar grandes conjuntos de datos. Otro defecto de algunas de estas herramientas es el de carecer de una interfaz gráfica que facilite su uso a los usuarios más inexpertos en cuanto al uso de herramientas por líneas de comando.
El resultado obtenido en este Trabajo de Fin de Grado se trata de **SeQual**, una herramienta paralela, implementada usando el framework *open-source* **Apache Spark,** que permite realizar diversos controles de calidad sobre conjuntos de datos de secuencias genómicas de una forma eficiente. La herramienta está orientada a trabajar con cantidades de datos masivos o *Big Data* en entornos distribuidos, de cara a ofrecer el mejor rendimiento posible.
SeQual ofrece principalmente cuatro grupos de operaciones a realizar sobre las secuencias genéticas: **Filtros**, **Recortadores**, **Formateadores** y **Estadísticas**. A parte de estas operaciones, también se ofrecen otras funcionalidades más transversales y de configuración, de cara a ofrecer un mayor control al usuario sobre la herramienta y los procesos. Las 42 funcionalidades disponibles se explican en profundidad en el capítulo 5 de la memoria. Además, y con el objetivo de facilitar el uso de la herramienta, SeQual dispone de una interfaz gráfica de usuario, la cual permite el procesamiento de los ficheros genómicos de una forma sencilla e intuitiva para usuarios no familiarizados con entornos de consola.
De cara a medir la eficiencia de SeQual, se realizaron diversas comparaciones con una herramienta estado del arte llamada **PRINSEQ**, una de las más utilizadas y citadas (con más de 2500 citas según Google Scholar), y la cual dispone de varias funcionalidades similares a las de SeQual. Tras ejecutar diversas operaciones sobre conjuntos de datos, y en una única máquina, SeQual demostró necesitar incluso 50 veces menos tiempo que la herramienta de referencia para llevar a cabo las operaciones. Sin embargo, SeQual fue diseñada para trabajar en entornos distribuidos. Ejecutando las operaciones más costosas en el clúster de alto rendimiento de la Facultad de Informática, SeQual las realizó hasta 190 veces más rápido que la herramienta de referencia. En el capítulo 7 de la memoria se ofrecen los resultados de dicha comparación, así como una valoración más exhaustiva.
Los usuarios objetivo de esta aplicación son, principalmente, biólogos, bioinformáticos y científicos que necesiten procesar grandes cantidades de secuencias genómicas para obtener conjuntos de datos con los que poder trabajar y que cumplan con los estándares de calidad oportunos o que se ajusten a sus necesidades. Gracias a las mejoras de rendimiento que proporciona SeQual ante otras herramientas similares, el tiempo dedicado a esta clase de procesamiento inicial disminuye drásticamente, reduciendo procesamientos que podrían necesitar varias horas con otras herramientas a tan solo minutos.
La herramienta desarrollada en este TFG se encuentra disponible bajo la licencia GNU GPL en el siguiente repositorio Git: https://github.com/roigalegot/SeQual.

**Posibles aplicacións / Posibles aplicaciones:**
La herramienta se puede aplicar en diversos ámbitos relacionados con los controles de calidad sobre secuencias genéticas:

- **Filtrado**, eliminado aquellas secuencias que no cumplan con ciertos parámetros establecidos por el usuario.
- **Recortado**, reduciendo el tamaño de las secuencias hasta alcanzar el indicado por el usuario, o hasta cumplir alguna condición de calidad concreta.
- **Formateado**, transformando las secuencias entre varios formatos disponibles.
- **Estadísticas**, permitiendo calcular y obtener propiedades del conjunto de datos analizados.

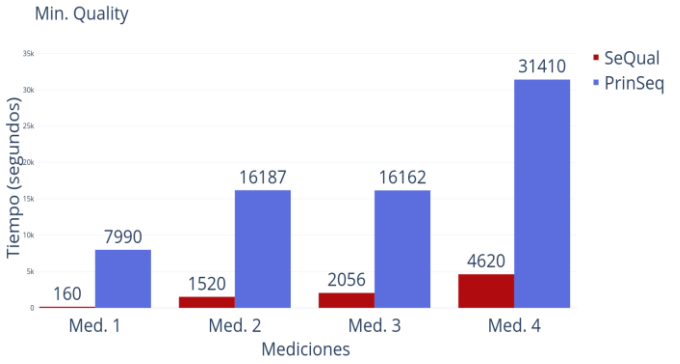
Además, SeQual está enfocada a trabajar con Big Data, o ficheros de gran tamaño, ofreciendo una gran eficiencia gracias a su procesamiento en paralelo, lo que le permite exprimir al máximo los recursos de las máquinas en las que se ejecuta. También soporta la ejecución en clústers de alto rendimiento, entorno para el que fue diseñada, y donde la ventaja que ofrece sobre otras herramientas del mismo ámbito es aún más notable. Todo esto contribuye a reducir al máximo la gran cantidad de tiempo que científicos y biólogos invierten actualmente en este paso previo a poder trabajar con los conjuntos de datos.
Por otro lado, el proyecto se ha liberado bajo una licencia *open-source* con el objetivo de que cualquier usuario pueda, además de acceder a él libremente, añadir funcionalidades o modificarlo para cubrir sus necesidades. Para facilitar dicha modificación, el proyecto se ha implementado siguiendo una estructura modular, la cual permite la adición de nuevas funcionalidades a los grupos ya existentes, así como la creación y uso de nuevos grupos de funcionalidades, de forma transparente.

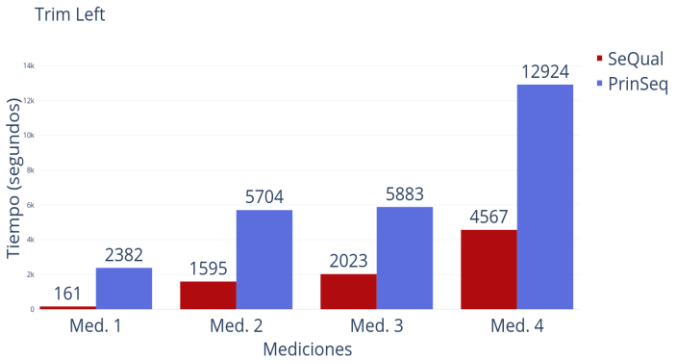
**Etapas para o seu desenvolvemento futuro / Etapas para su desarrollo futuro:**
Para SeQual se han planteado varias líneas de trabajo futuro, como, por ejemplo, las siguientes:

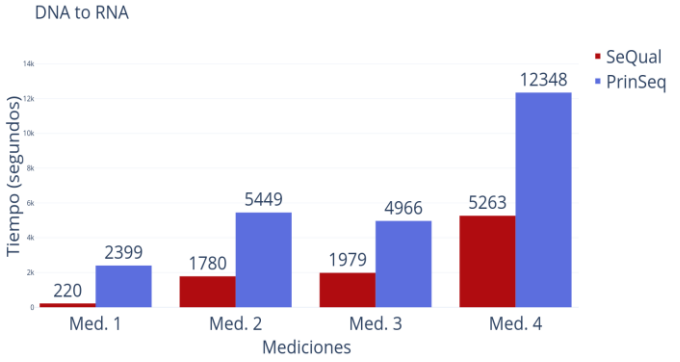
- Realizar una implementación en modo *streaming* de las funcionalidades actuales, de cara a poder realizar el procesamiento de los ficheros a la par que se descargan desde una URL de internet.
- Permitir la lectura y el análisis de ficheros comprimidos.
- Ampliar los formatos de codificación de ficheros genéticos aceptados.

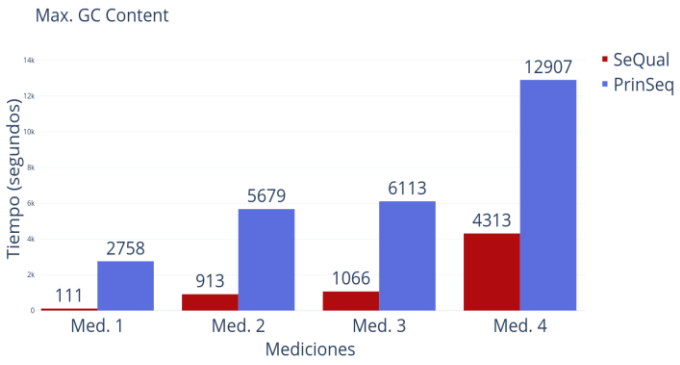
**Imaxes representativas / Imágenes representativas:**

Ejemplos de comparativas con PRINSEQ:

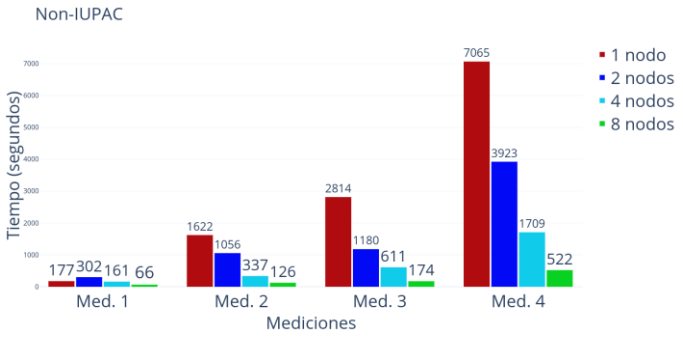


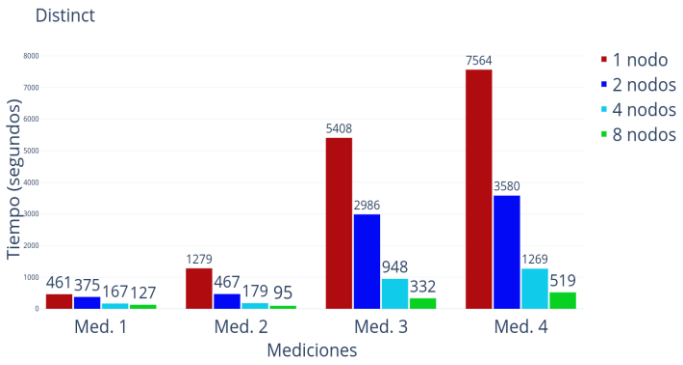



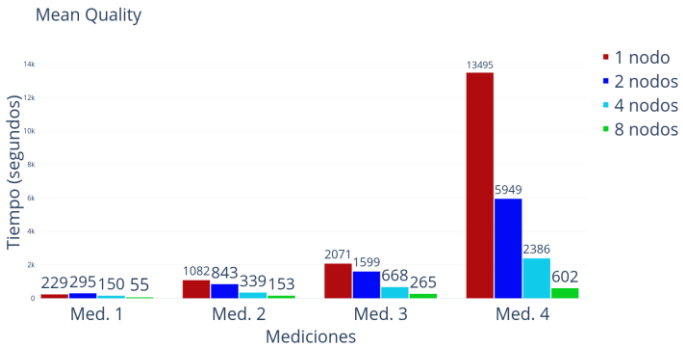



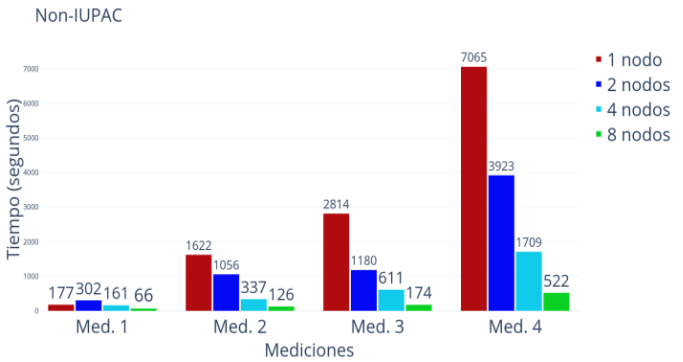




Ejemplos de escalabilidad ejecutándose en un clúster de alto rendimiento:

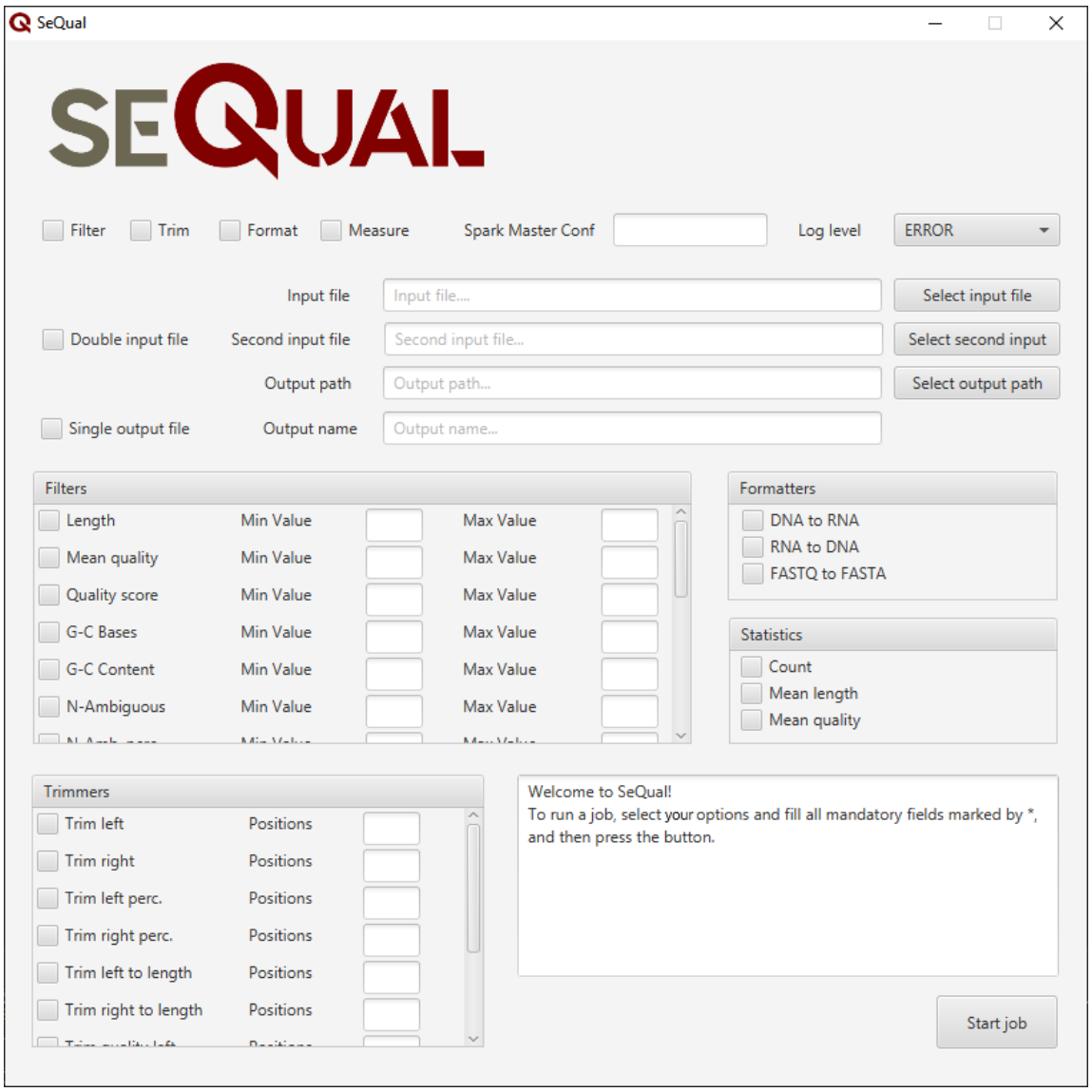


Interfaz gráfica de SeQual:




| | |
|---|---|
| X | **Autorizo a consulta por parte dos membros da comisión evaluadora da memoria do meu proxecto / Autorizo la consulta por parte de los miembros del tribunal de la memoria de mi proyecto.** |

**Instruccións para o depósito da memoria / Instrucciones para el depósito de la memoria:**

Débese depositar no OneDrive da UDC, dentro da carpeta co seu nome de usuario incluida en:

5 edición Premio TFG aplicado

Se debe depositar en el OneDrive de la UDC, dentro de la carpeta con su nombre de usuario incluida en:

5 edición Premio TFG aplicado