Estirpex

Fundación COMPUTAEX
info@{computaex.es, cenits.es}
CénitS – Centro Extremeño de iNvestigación, Innovación Tecnológica y Supercomputación
Cáceres, Extremadura, España

Resumen—El proyecto Estirpex persigue la creación de una plataforma *online* que permita el acceso a un repositorio piloto de datos históricos y clínicos de ciudadanos de Extremadura. El objetivo principal es ayudar a que éstos consulten sus datos genealógicos y que los profesionales del ámbito sanitario puedan estudiar enfermedades hereditarias a través de la obtención y análisis de variaciones en la secuencia genética del exoma de las personas representadas, obtenida mediante tecnologías NGS (Next-Generation Sequencing). De esta forma, se logra conservar datos históricos, gracias a su digitalización e inclusión en un sistema de elevadas prestaciones y contribuir al estudio de enfermedades genéticas hereditarias.

Índice de Términos—NGS, DNA, exome sequencing, pedigree.

I. Introducción

Los avances en la investigación en el ámbito sanitario han sido siempre uno de los principales retos de la humanidad. Uno de los hallazgos más importantes ha sido determinar que el estudio de la secuencia de ADN (Ácido Desoxirribonucleico) permite conocer los procesos biológicos fundamentales del organismo y su funcionamiento.

Hasta hace escasos años secuenciar el ADN tenía un elevado coste económico y temporal. Ante estos problemas surgen las tecnologías NGS y, con ellas, un nuevo paradigma de secuenciación genética que permite la secuenciación del genoma o de secciones de éste a gran escala, con una importante disminución del tiempo y coste del procesamiento.

La Fundación COMPUTAEX ha emprendido varios proyectos con los Hospitales Infanta Cristina de Badajoz y San Pedro de Alcántara de Cáceres, centrados en el estudio de la secuencia genética de determinados pacientes, procesadas en el centro CénitS con el supercomputador LUSITANIA.

Con Estirpex se pretende utilizar la infraestructura de CénitS para albergar un repositorio digital piloto con información histórica y clínica, que permita a los ciudadanos generar automáticamente sus árboles genealógicos y consultar el historial vital de sus antecesores, mediante una plataforma *online*. Además, permitirá a los especialistas en genética profundizar en el estudio de determinadas enfermedades hereditarias, gracias a las posibilidades de filtrado, extracción y visualización de datos genéticos de alto nivel procedentes del árbol genealógico de sus pacientes.

II. Análisis pormenorizado

En esta fase se ha realizado el estudio de las implicaciones técnicas y legales asociadas a cada actividad del proyecto.

A. Repositorio de datos históricos

Para la creación del repositorio de datos históricos se requiere la consulta del Registro Civil [1] y los archivos parroquiales fundamentalmente. Aunque la iniciativa del Registro Civil en Línea, impulsada por el Ministerio de Industria, Energía y Turismo, contiene información digitalizada sobre los ciudadanos, se requieren datos adicionales que deben obtenerse para la construcción de los árboles genealógicos.

El equipo de CénitS ha estimado, a modo de piloto, el coste económico asociado a dicha digitalización para cuatro localidades extremeñas, utilizando para dicha estimación los datos generales de coste del programa Registro Civil en Línea y los datos de población del Instituto Nacional de Estadística[1]. Según este programa, se han digitalizado 6.000 hojas al día, lo cual supone 750 hojas cada hora (teniendo en cuenta una jornada laboral de 8 horas), con un coste de digitalización de 490,34 € por hora. También se han determinado los requisitos del repositorio en sí, estableciendo en tres el número de documentos que tiene asociada cada persona viva. La Tabla 1 muestra los resultados de la estimación del tiempo y coste asociados a la digitalización de documentos en cuatro localidades extremeñas.

	Cáceres	**Coria**	**Guadalupe**	**Campillo de Deleitosa**
Nº Hojas[2]	287.004	39.303	5.997	204
Coste temporal	383h=48d	53h=7d	8h=1 día	17'
Coste económico	187.801 €	25.988 €	3.922 €	138,92 €

Tabla 1: Coste de digitalización de documentos.

B. Implicaciones legales

Para un correcto cumplimiento de la legislación vigente ha

[1] Obtenidos de la revisión del Padrón Municipal a 1 de enero de 2012.

[2] Población multiplicada por el número máximo de hojas asociadas a una persona (tres para persona viva: partida de nacimiento, bautizo y matrimonio).

sido necesario analizar las implicaciones legales asociadas al almacenamiento y tratamiento de los datos clínicos vinculados a las familias representadas en los árboles genealógicos.

Al tratarse de información de carácter personal y referente a la salud, los datos debían tratarse como información de *nivel alto*, según lo estipulado por la LOPD (Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de carácter personal) [2]. Si bien, para aumentar la seguridad de los datos clínicos se ha llegado a la conclusión de que una medida muy eficaz a tomar es la de disociarlos (es decir, no almacenar ningún dato que permitiera identificar la persona de la que proceden), según la Agencia Española de Protección de Datos [3] la Ley no se aplicaría, aunque se tomarán las mismas medidas que en el caso de que los datos no se hubieran anonimizado.

C. Proceso de ultra-secuenciación genética

La secuenciación de ADN es un conjunto de métodos y técnicas bioquímicas cuya finalidad es la determinación de los nucleótidos de un oligonucleótido de ADN. En el caso de la ultra-secuenciación, cuyo proceso se muestra en la Figura 1, se trata de nuevas técnicas que permiten la obtención de las secuencias genéticas en un tiempo muy reducido y con un coste económico menor.

En el análisis primario se genera una biblioteca de clones de secciones de interés del ADN, para obtener millones de secuencias de longitud finita denominadas *lecturas*.

En el análisis secundario se realiza un alineamiento de las *lecturas* obtenidas en el análisis primario respecto a una secuencia de referencia de consenso, obteniendo un fichero de alineamiento BAM (Binary Alignment/Map). En cambio, en el terciario se realiza la detección de variaciones en las secuencias alineadas (tanto SNPs, Single Nucleotide Polymorphism, como small indels, inserciones o borrados de cadenas de nucleótidos dentro de las mencionadas secuencias) [4], así como su anotación.

El modelo de prestación de servicios desplegado en CénitS permite acometer estas dos últimas fases del flujo de trabajo de una manera eficaz y eficiente. Éste permite a los usuarios acceder a recursos físicos bajo demanda (IaaS, Infrastructure as a Service) y la ejecución de herramientas HPC (High Performance Computing) que logran el procesamiento de datos genómicos a través de los recursos solicitados en el menor tiempo posible, lo que se conoce como HPC2 (High Performance Cloud Computing). Concretamente, la máquina virtual destinada a la realización de estas dos últimas fases consta de 24GB de memoria RAM, 16 cores y 3TB de HDD y permite llevar a cabo el flujo de trabajo completo (descrito en la Figura 1) de las secuencias genéticas de un exoma de 247 millones de *lecturas* de 50x35 *base pair* en 4 días.

D. Consejo genético

El consejo genético es un proceso comunicativo para informar y apoyar a pacientes que tienen una enfermedad genética hereditaria o presentan un riesgo de padecerla.

Para proporcionar consejo genético se utilizan los pedigríes, es decir, árboles genealógicos en los que se obtiene información sobre una enfermedad hereditaria en la familia. Desde CénitS se ha trabajado en la búsqueda de una herramienta que facilite la construcción y consulta de pedigríes, ya que habitualmente es necesaria la utilización de herramientas poco intuitivas o incluso la construcción del mismo manualmente.

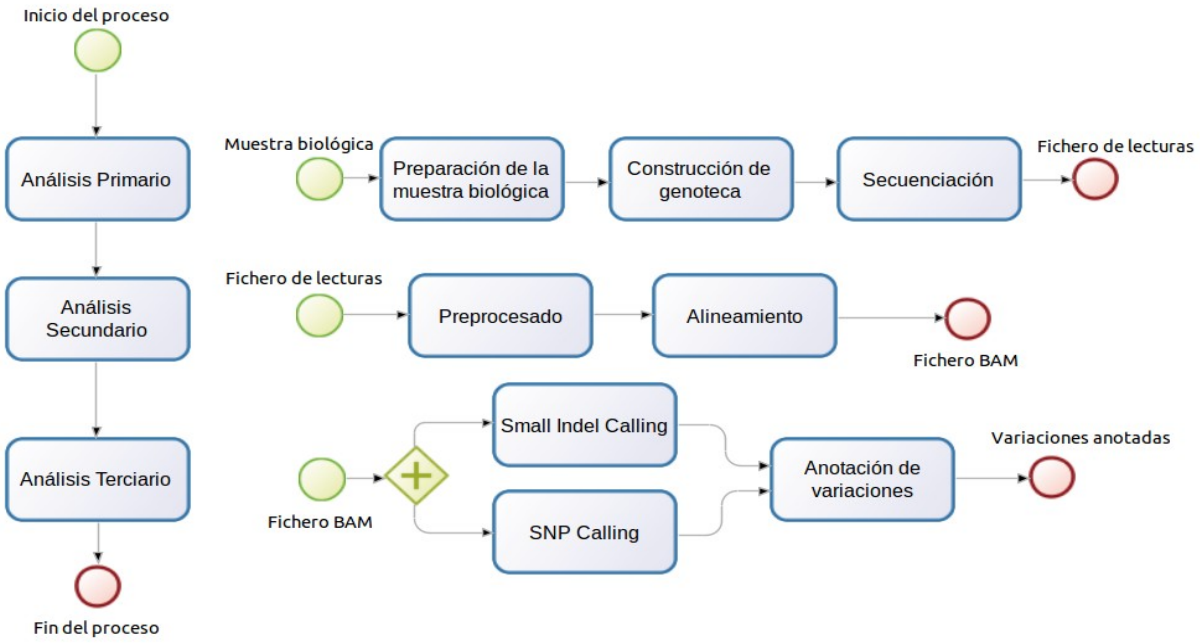

Figura 1: Fases del proceso de ultra-secuenciación genética.

De todas las posibles aplicaciones cuyas características se adaptaran a los requisitos establecidos, se ha optado por utilizar Madeline 2 [5] porque permite la generación automática de pedigríes a cuyos nodos (familiares) se puede asignar información específica de enfermedades que padezcan, así como la integración de esa funcionalidad en la implementación de software adicional.

III. Desarrollo e implementación

Durante la ejecución del proyecto fue necesario el desarrollo de varias soluciones software que apoyaran el despliegue de la plataforma de acceso a datos históricos y clínicos.

A. Base de datos de variaciones

El principal objetivo de la base de datos de variaciones es almacenar toda la información generada en el procesamiento de las *lecturas* de un exoma, tanto las variaciones como sus anotaciones, obtenidas estas últimas a partir de la ejecución de dos *scripts* (implementados con el lenguaje Perl) que hacen uso de la API v75 del Ensembl Project [6].

B. Software de inserción de variaciones anotadas

Esta aplicación, desarrollada con Java, se encarga de trasladar el contenido de los ficheros resultantes del procesamiento de las secuencias de exomas a la base de datos de variaciones creada para tal efecto.

C. Visor de variaciones

Tras analizar las necesidades de los especialistas en genética, se concluyó que era necesario desarrollar una herramienta que les facilitara la consulta de las variaciones detectadas en un exoma concreto cuyas *lecturas* hayan sido procesadas previamente en la infraestructura de LUSITANIA.

Habiendo sido implementado también con Java, el visor permite la consulta de toda la información asociada a las variaciones que se encuentran en la base de datos, así como la aplicación de filtros para acotar la búsqueda de los resultados. Además de la comparación de las variantes detectadas en los exomas procesados, se pueden obtener las variaciones comunes y desechar las detectadas en alguno de ellos, lo cual es muy útil a la hora de investigar la incidencia de una enfermedad genética hereditaria en el ámbito de una familia.

IV. conclusiones y líneas futuras

El desarrollo del proyecto Estirpex ha supuesto el despliegue de software para procesar la secuencia genética de determinados exomas, almacenar la información generada por dicho procesamiento y filtrar y visualizar los resultados obtenidos.

Para el trabajo realizado ha sido necesario procesar las secuencias genéticas generadas por varias plataformas propietarias de ultra-secuenciación genética, concretamente la SOLiD 4 System de LifeTechnologies [7] y la HiSeq 2500 de Illumina [8], lo cual enriquece las herramientas a las que los usuarios de la infraestructura del centro pueden acceder para procesar sus datos genéticos.

De acuerdo con la filosofía del centro CénitS, se ha hecho un esfuerzo importante para aportar aplicaciones acordes a los principios del software libre.

Con este proyecto también se sientan las bases para unas prometedoreas líneas de trabajo futuro:

- Digitalización y conservación de archivos históricos para permitir su consulta, por parte de las familias representadas, a través de una plataforma *online*.
- Estudio e implantación de medidas de seguridad que garanticen el cumplimiento de la legislación vigente en materia de protección de datos.
- Estudio de mecanismos de automatización del procesamiento de secuencias genéticas de exomas procedentes de secuenciadores alternativos, para ampliar la procedencia de los exomas cuyas secuencias genéticas pueden ser procesadas por los usuarios de la infraestructura de CénitS.
- Automatización del procesamiento de las secuencias genéticas de los familiares representados en los pedigríes, para reducir los costes económico y temporal del flujo de trabajo que culmina con el consejo genético a las familias por parte de los genetistas.
- Diseño e implementación de una aplicación para la construcción y consulta de pedigríes.

V. Referencias

[1] “Registro Civil” http://www.registrocivil.es/
[2] “Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de carácter personal” http://www.agpd.es/portalwebAGPD/canaldocumentacion/legislacion/estatal/common/pdfs/LOPD_consolidada.pdf
[3] “Agencia Española de Protección de Datos” https://www.agpd.es/portalwebAGPD/index-ides-idphp.php
[4] “Human Genome Project Information Archive Glossary” http://web.ornl.gov/sci/techresources/Human_Genome/glossary.shtml
[5] “Madeline 2.0” http://eyegene.ophthy.med.umich.edu/madeline/madeline2.0.php
[6] “Ensembl Project” http://www.ensembl.org/index.html
[7] “SOLiD 4 System” http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-4-system.html
[8] “Illumina HiSeq 2500” http://www.illumina.com/systems/hiseq_2500_1500.ilmn