Estirpex-2

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

Resumen—El proyecto Estirpex-2 persigue, bajo la Estrategia RIS3 (Estrategia de Investigación e Innovación para la Especialización Inteligente de Extremadura [1], Áreas de Excelencia de la salud y de Excelencia de las TIC), la continuidad del trabajo desarrollado en el proyecto Estirpex [2]. Concretamente, los objetivos de Estirpex-2 se corresponden con el estudio, desarrollo y despliegue de servicios para sectores económicos relevantes en la región que puedan beneficiarse de la tecnología de secuenciación genética masiva (NGS, Next-Generation Sequencing), apoyada en el uso de la supercomputación.

Índice de Términos—Precision Medicine, NGS services, genetic counseling, autoctonal species.

I. INTRODUCCIÓN

El proyecto Estirpex supuso el despliegue de software libre, en la infraestructura de LUSITANIA, para procesar la secuencia genética de determinados exomas, almacenar la información generada y filtrar y visualizar los resultados obtenidos; pero también la apertura de una serie de líneas de trabajo destinadas a la optimización de los procesos de obtención de información de alto nivel relacionada con los estudios de secuenciación masiva realizados.

La experiencia que la Fundación COMPUTAEX ha demostrado en el desarrollo de otros proyectos del ámbito de la sanidad se ve reforzada por el auge que el concepto *Precision Medicine* está experimentando, de lo que se deduce que el momento en el que se enmarca el desarrollo de Estirpex-2 es inmejorable. Merece la pena destacar iniciativas nacionales, como la lanzada por el presidente Barack Obama, que con una inversión de 215 millones de dólares (para el presupuesto de 2016) y la secuenciación del genoma de 1 millón de personas, pretende liderar la investigación de tratamientos adaptados lo máximo posible a los pacientes [3].

El diseño de un catálogo de servicios que permita que los sectores económicos más relevantes de la región sean capaces de acceder a las ventajas de la secuenciación masiva ha sido la clave del trabajo llevado a cabo durante Estirpex-2. En ese sentido, se ha prestado especial atención al establecimiento de medidas para maximizar la utilización de los recursos computacionales por parte de los usuarios que puedan acceder a esos servicios. Además, se ha realizado un análisis de viabilidad del despliegue y uso de los servicios del catálogo.

II. ANÁLISIS PORMENORIZADO

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

A. Plataformas de ultra-secuenciación genética.

La base de los servicios de secuenciación masiva está representada por las plataformas con las que llevar a cabo estudios de secuenciación genética masiva. Por ello, la primera fase del análisis pormenorizado ha consistido en el estudio integral de las mismas. Concretamente, en los siguientes aspectos:

- Técnicas y métodos bioquímicos para llevar a cabo un proceso de ultra-secuenciación genética, así como el equipamiento necesario para llevar a cabo el proceso.
- Prestaciones de los secuenciadores disponibles en el mercado (coste, precisión, tiempo empleado, cantidad de pares de bases generadas, etc.).
- Estudios de secuenciación masiva más apropiados según la plataforma utilizada (resecuenciación, ensamblado de genomas, etc.).

Del estudio de los secuenciadores del mercado actual se ha determinado que existen dos gamas de plataformas, en función del coste y la versatilidad que presentan en cuanto a los tipos de estudios que permiten realizar.

Es destacable mencionar que la utilización de la secuenciación masiva en clínica necesita la confirmación de los resultados por métodos de secuenciación genética con un porcentaje de error casi nulo. Por ello, en esta fase del análisis se ha incluido el estudio de una plataforma de secuenciación con el método tradicional de Sanger, que se usaría en un fragmento de ADN lo suficientemente pequeño como para confirmar los resultados de secuenciación masiva, sin encarecer el coste final del estudio de forma considerable.

B. Procesamiento de secuencias genéticas.

El procesamiento de secuencias genéticas (*reads*) representa las tareas necesarias para que, dado un conjunto de lecturas generadas por un secuenciador, el especialista pueda obtener información de alto nivel, de utilidad en su investigación.

Durante el transcurso de Estirpex-2, se ha querido profundizar en el estudio del software necesario para procesar las secuencias genéticas de los estudios más típicos que se pueden llevar a cabo con los secuenciadores que hay

actualmente en el mercado, a saber: resecuenciación, ensamblado de genomas (*de Novo sequencing*) y *RNA-Seq*.

C. NGS y cloud computing.

Los dos apartados anteriores del Análisis Pormenorizado ponen de manifiesto que se necesita disponer de una infraestructura que permita no sólo la obtención de resultados genéticos de alto nivel en un tiempo lo más bajo posible, sino que el uso de los recursos computacionales se optimice.

El modelo de prestación de servicios de *cloud computing* es ideal para ofrecer recursos computacionales a usuarios de los servicios, teniendo en cuenta que se necesita que esos recursos se puedan compartir y que su uso tiene que ser flexible.

D. Valor estratégico de servicios.

A la hora de diseñar el catálogo de servicios de ultra-secuenciación genética, se ha querido tener en cuenta el valor añadido que tendría el acceso de determinados sectores económicos de la región a esos servicios.

Si bien la Fundación COMPUTAEX, en el campo del NGS, tiene más experiencia en proyectos del ámbito de la sanidad, se ha hecho hincapié en el valor añadido que la secuenciación del genoma de especies autóctonas, animales y vegetales, podría tener en ciertos sectores económicos de la región.

En ese sentido, la tecnología NGS ha supuesto un importante auge de la secuenciación, por primera vez, de esas especies, como se muestra en la Figura 1.

Como conclusiones del estudio, se puede determinar que, dado el peso que el sector primario posee en la economía de la región, la colaboración entre miembros del SECTI (Sistema Extremeño de Ciencia, Tecnología e Innovación) u otras entidades especializadas en el mundo animal y vegetal con COMPUTAEX podría ser muy enriquecedora para la región.

III. Desarrollo e Implementación

Durante el desarrollo de Estirpex-2 se han implantado soluciones software para apoyar el despliegue de un servicio NGS del catálogo que se propondrá en sucesivas secciones.

A. FI4VDI – Prototipo para servicio de resecuenciación

El proyecto europeo FI4VDI[1] (Federation Infrastructure - Virtual Desktop Infrastructure) [4], en el que participa la Fundación COMPUTAEX, ha supuesto el despliegue de una federación (Figura 2) de infraestructuras computacionales con OpenNebula [5], para lanzar máquinas virtuales bajo demanda, a partir de un conjunto de imágenes virtuales.

Una de esas imágenes es el prototipo de un entorno para automatizar el lanzamiento de las tareas de las que constan las fases computacionales de un estudio de resecuenciación genética. Además, se proporcionan herramientas de apoyo a la interpretación, manualmente, de los resultados obtenidos.

[1] La federación utiliza recursos de distintos centros de cálculo intensivo situados en el territorio del Espacio SUDOE, para garantizar la protección de datos de los usuarios, la conformidad y la seguridad de la información.

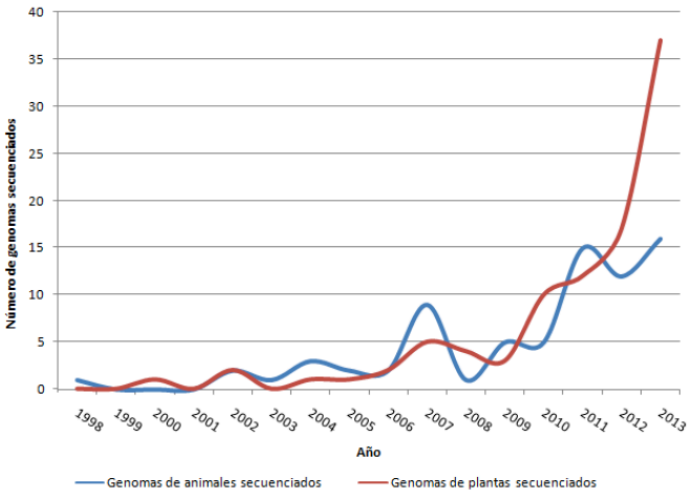

Figura 1: Número de genomas secuenciados de animales y plantas por primera vez en el ámbito internacional.

Las ventajas del *cloud computing*, y la posibilidad de crear imágenes personalizadas para el despliegue de máquinas virtuales, permiten que se pueda desarrollar un conjunto de imágenes para lanzar trabajos de procesamiento de la información generada por cada servicio del catálogo.

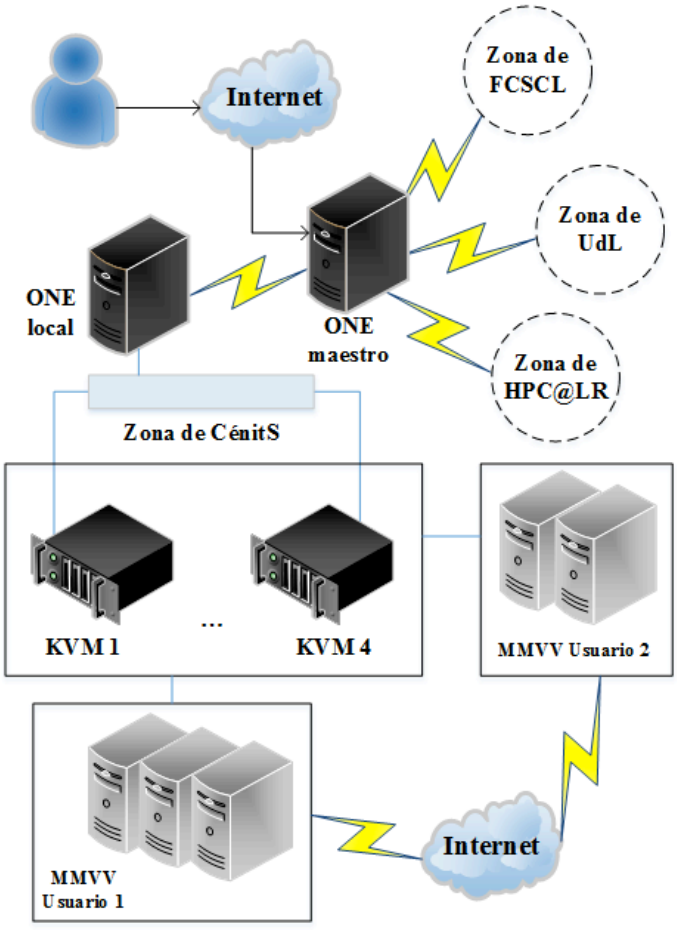

Figura 2: Esquema de la federación desplegada en el FI4VDI [4].

B. Estudio de seguridad

El avance y desarrollo constante de los medios de almacenamiento, tratamiento y transmisión facilita la gestión de la información, pero aumenta la inseguridad y el riesgo asociado. Por ello, después del despliegue del prototipo para el servicio de resecuenciación, se ha realizado un estudio del riesgo de los sistemas de información involucrados.

La metodología adoptada ha sido MAGERIT [6], y ha permitido establecer una serie de acciones a realizar para alcanzar un nivel de riesgo de los sistemas de información tolerable. Ello se ha conseguido, entre otros, gracias a la elaboración de un inventario de activos, al establecimiento de amenazas sobre los mismos y a la identificación del nivel de madurez de las salvaguardas existentes en CénitS. La Figura 3 muestra el resultado de la aplicación de MAGERIT para analizar los riesgos a los que se ven expuestos los sistemas de información implicados en el prototipo de servicio desplegado.

C. PedigreeX

En consejo genético la construcción y análisis de pedigríes[2] es de vital importancia. No obstante, las herramientas software existentes o son poco intuitivas o disponen de una licencia cuyo coste es prohibitivo. Por ello, durante el transcurso de Estirpex-2 se ha desarrollado un software que trate de paliar las carencias existentes al respecto.

La herramienta se ha desarrollado con Java y consta de una interfaz a través de la cual el usuario especifica qué acciones quiere realizar para editar su pedigree. Esas acciones pasan a ser los datos de entrada del software Madeline 2[3] [7], que es la aplicación que realmente genera el pedigree resultante. Posteriormente, el diagrama se muestra por pantalla al usuario.

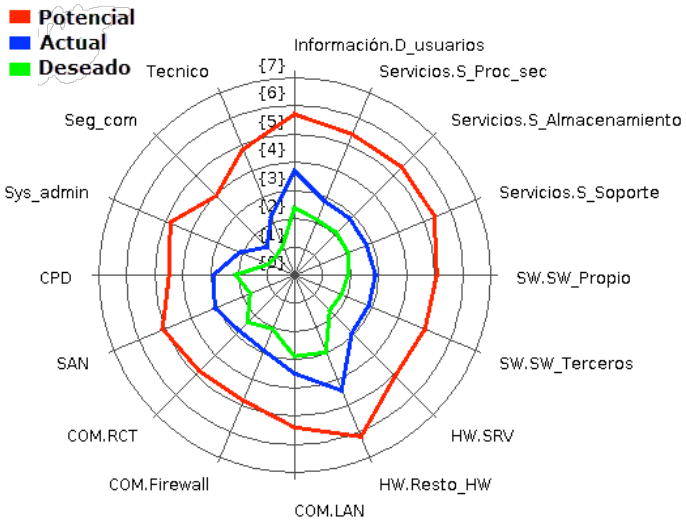

Figura 3: Riesgo potencial, actual y deseado de los sistemas de información asociados al prototipo de servicio de resecuenciación.

[2] Diagrama para representar y analizar las relaciones genealógicas de una familia, para determinar cómo se hereda y manifiesta una característica genética.

[3] La herramienta funciona por línea de comandos, por lo que su uso, por parte de especialistas sin conocimientos informáticos, puede ser muy tedioso.

IV. SERVICIOS DE SECUENCIACIÓN MASIVA

A. Catálogo de servicios

Se ha propuesto un catálogo de servicios de secuenciación masiva, apoyados en el uso de herramientas HPC y en el *cloud computing*, dentro de la infraestructura de LUSITANIA. Los tipos de servicio del catálogo propuesto son los siguientes:

- Servicios de resecuenciación.
- Servicios de genómica del cáncer.
- Servicios de genómica de la agricultura y ganadería.
- Servicios de genética forense.
- Servicios de genética microbiana.
- Servicios de bioinformática.

B. Análisis de viabilidad

El análisis de viabilidad ha consistido en la evaluación de los servicios desde varios puntos de vista distintos.

Desde el punto de vista económico, Estirpex-2 presenta un análisis del precio mínimo que deberían tener los servicios propuestos para que sean rentables, según el coste que supone llevar a cabo los estudios necesarios para ello con el secuenciador elegido.

En cuanto a la viabilidad tecnológica, es evidente que la tecnología NGS se encuentra en su mayor apogeo; por ello, es necesario continuar el desarrollo de recursos computacionales que permitan satisfacer las demandas de los usuarios.

V. HEREDITY

Gracias a los trabajos e investigaciones realizados durante el desarrollo de los proyectos Estirpex y Estirpex-2, se ha conseguido presentar una propuesta para la convocatoria de Proyectos Europeos H2020-PHC-2015 (Personalising health and care), *topic* "Digital representation of health data to improve disease diagnosis and treatment" (PHC-30-2015).

HEREDITY: "HPC to reduce VUS rates in genetic counseling", es una propuesta liderada por la Fundación COMPUTAEX, creada por un consorcio de 10 entidades tecnológicas y del ámbito de la salud de 5 países.

El objetivo de la propuesta es desplegar un servicio gratuito para la reducción del porcentaje de VUS (Variants of Uncertain Significance) obtenidas en consejo genético. El servicio estará soportado por un sistema de apoyo a la toma de decisiones en el diagnóstico de enfermedades de origen genético, así como en la integración de fuentes heterogéneas de datos genéticos (véase la Figura 4) y la aplicación de técnicas de *big data*, para lograr predecir, cuando sea posible, la patogenicidad de una VUS.

Para satisfacer las demandas de los proveedores de servicios NGS, se pretende desplegar una red federada de recursos HPC entre los centros tecnológicos del consorcio.

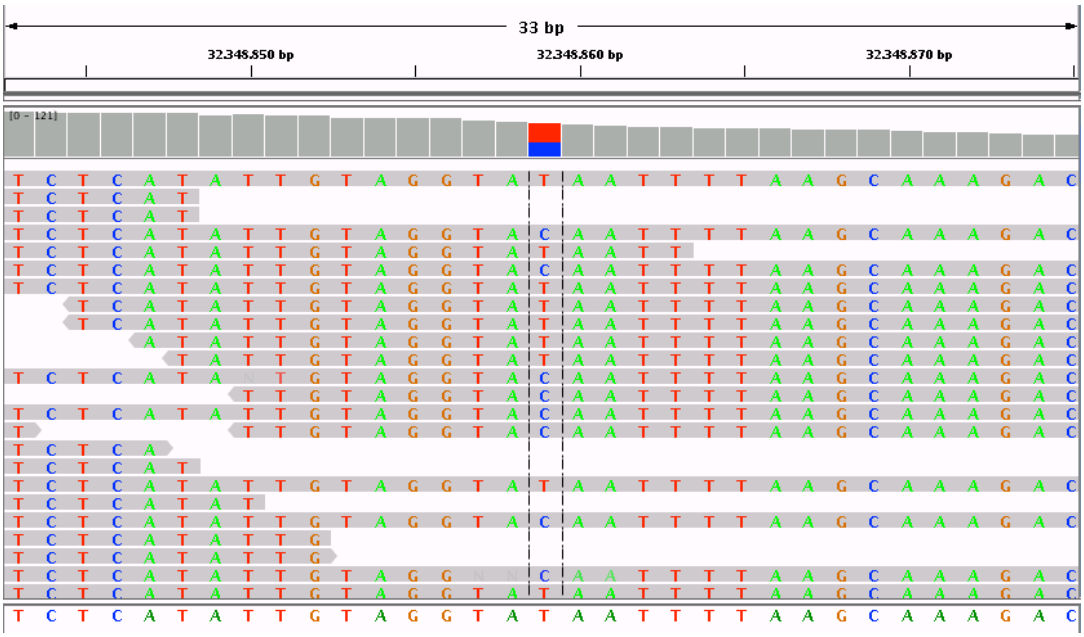

Figura 4: Ejemplo de variación en la secuencia del gen BRCA2.

VI. Conclusiones y líneas futuras

El desarrollo del proyecto Estirpex-2 ha supuesto realizar una propuesta y el análisis de viabilidad de un catálogo de servicios de secuenciación masiva, aprovechando, para el procesamiento de las secuencias genéticas generadas, la infraestructura del supercomputador LUSITANIA en CénitS. Se ha hecho especial hincapié en aportar valor añadido a sectores del ámbito de la salud y a aquellos en los que se trabaje con especies autóctonas que sean de especial interés.

Así mismo, se ha desarrollado un prototipo, en la plataforma OpenNebula, para el procesamiento automático de secuencias generadas por un servicio de resecuenciación, incluyendo herramientas manuales de interpretación. De esta forma, la interpretación final de las variaciones detectadas se deja en manos de los especialistas. Este prototipo incluye un estudio de seguridad de los activos y sistemas de información asociados con el servicio.

Además, se ha desarrollado un software para la construcción y edición de pedigríes por parte de los especialistas en consejo genético.

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

- Integrar la información representada en el pedigree de una familia con la obtenida al secuenciar el ADN de alguno de sus miembros a través del servicio de resecuenciación.
- Establecimiento de flujos de trabajo para procesar secuencias genéticas de otros tipos de estudios NGS.
- Creación de un *pool* de imágenes, en la plataforma desplegada con OpenNebula, para el despliegue de máquinas virtuales destinadas a procesar los datos generados por cada servicio del catálogo.
- Estudiar la seguridad de los activos implicados en cada servicio y aplicar medidas correctoras en cada caso.
- Trabajar en la secuenciación de especies autóctonas, especialmente con centros especializados en Extremadura.
- Aplicación de técnicas de *big data* para procesar datos heterogéneos que ayuden a la interpretación de la patogenicidad de las VUS obtenidas en estudios de resecuenciación, cuando sea posible.

VII. Referencias

[1] "Estrategia RIS3" one.gobex.es/docs/Estrategia_RIS3_Extremadura.pdf

[2] "Estirpex" www.cenits.es/sites/cenits.es/files/publicaciones/resumen-estirpex.pdf

[3] Francis S. Collins, M.D., Ph.D., and Harold Varmus, M.D., "A New Initiative on Precision Medicine," The NEW ENGLAND JOURNAL of MEDICINE, 3 pages, 2015.

[4] "FI4VDI" fi4vdi-sudoe.org/

[5] IaaS Cloud Architecture: From Virtualized Datacenters to Federated Cloud Infrastructures, R. Moreno-Vozmediano, R. S. Montero, I. M. Llorente. IEEE Computer, vol. 45, pp. 65-72, Dec. 2012.

[6] "MAGERIT" www.ccn-cert.cni.es/publico/herramientas/pilar5/magerit/

[7] "Madeline 2.0 PDE: A new program for local and web-based pedigree drawing," Edward H. Trager; Ritu Khanna; Adrian Marrs; Lawrence Siden; Kari E.H. Branham; Anand Swaroop; Julia E. Richards. Bioinformatics 2007; doi: 10.1093/bioinformatics/btm242