Requisitos Técnicos para Puesta en Marcha de Sistemas Beacon

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
Instituto de Salud Carlos III

IMPaCT
Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología

Programa	IMPacT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología		
Nombre Proyecto	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT		
Expediente	IMP/00019		
Duración	Enero 2021 – Diciembre 2023		
Paquete Trabajo	WP5 – Integración de datos		
Tarea	T5.2 - Implantación de mecanismos para la identificación automática de información genómica relevante, o contenido de las imágenes médicas que den respuestas anonimizadas al estilo beacon		
Entregable	E5.4. Requisitos Técnicos para Puesta en Marcha de Sistemas Beacon		
Versión	1.1.1		
Fecha Entrega	31/12/2021	***Fecha Aprobación***	17/05/2023
Responsable	BSC		
Nivel Diseminación	X	PU	Público
		CO-IMP	Confidencial, sólo participantes de los pilares de IMPaCT, incluyendo la comisión de evaluación de IMPaCT.
		CO-DATA	Confidencial, sólo participantes de IMPaCT-Data, incluyendo la comisión de evaluación de IMPaCT.

Autores		
Organización	**Nombre**	**Rol**
BSC	Lidia López	Coordinación
FPS	Javier Perez Florido	Autor
BSC	Salvador Capella-Gutierrez	Revisor
CRG	Jordi Rambla	Revisor

Historial de versiones			
Nro.	**Fecha**	**Descripción**	**Autor**
v 0.0	08/10/2021	Creado	L.López (BSC)
v 0.1	01/12/2021	Texto añadido para su revisión	Javier Pérez Florido
v 1.0	27/12/2021	Comentarios de los revisores añadidos en el texto	Javier Pérez Florido
v 1.1	17/05/2023	Cambio visibilidad a público y aprobado	Comité Directivo
v 1.1.1	14/06/2023	Cambio de formato para publicar en la Web de IMPaCT	David Velasco (ISCIII)

Contenido

Figuras

Resumen Ejecutivo

Un sistema *Beacon* es una herramienta para el descubrimiento de variantes genómicas que considera, bajo una misma entidad, numerosos conjuntos de datos genómicos. La API de *Beacon* habilita la búsqueda de variantes genómicas e información asociada sin comprometer la privacidad del conjunto de datos de forma que cualquier institución puede hacer que sus conjuntos de datos ómicos sean descubribles.

Los sistemas Beacon han evolucionado a lo largo de los últimos años, partiendo de un sistema básico (Beacon v1) donde se pregunta acerca de la presencia de una determinada variante genómica obteniendo una respuesta afirmativa o negativa hasta nuestros días donde la versión v2 permite consultas más complejas de forma que el protocolo Beacon cubre diferentes entidades y detalles asociados a las mismas.

Aquellas instituciones como centros de investigación, hospitales o grupos que deseen instalar un sistema Beacon, deben tener en cuenta una serie de requisitos técnicos, así como cuestiones relacionadas con su implementación. Todo ello será brevemente descrito en este entregable.

Introducción

Audiencia

Este entregable está dirigido a todas aquellas entidades (hospitales, centros de investigación, grupos de investigación) que deseen compartir datos genómicos y metadatos asociados (al nivel que decidan) sin comprometer la privacidad de los mismos mediante un sistema Beacon. Proporcionamos la información básica de los requerimientos técnicos necesarios para ello.

Ámbito

Este entregable supone un punto de partida para aquellas entidades que participarán, en el contexto del proyecto IMPaCT-Data, en la compartición de datos genómicos e información asociada. La compartición de este tipo de información es fundamental para el desarrollo del programa de medicina personalizada en los diferentes sistemas de atención sanitarios.

Relación con otros Entregables

Este entregable es el primero del WP5 y el trabajo indicado en este documento se complementará con el entregable E5.6 del mismo WP. Por otro lado, está relacionado con el entregable E3.1 del WP3 dado que corresponde a un trabajo para la implementación de la compartición de datos genómicos.

Estructura Entregable

El entregable tiene la siguiente estructura:

1. Conceptos básicos de sistemas Beacon. ¿Qué es y para qué sirve?
2. Información proporcionada por sistemas Beacon v1.
3. Información proporcionada por sistemas Beacon v2
4. Requisitos técnicos para la puesta en marcha de un sistema Beacon.

1 Sistemas Beacon

1.1 ¿Qué es un sistema Beacon y por qué es necesario?

Un *Beacon* es una herramienta para el descubrimiento de variantes genómicas que considera, bajo una misma entidad, numerosos conjuntos de datos genómicos. Uno de los mayores retos a los que se enfrenta hoy día la investigación en genética humana es la falta de datos, no porque no se generen los suficientes, sino porque los datos no son compartidos en la mayoría de los casos entre la comunidad científica. Los datos genómicos son identificables y deben estar protegidos, pero debido a la falta de infraestructura de seguridad sobre ellos y las buenas prácticas en su tratamiento y uso, hace que los investigadores y médicos no los compartan, de forma que los progresos realizados no son conocidos.

En tiempos de la medicina personalizada, el diagnóstico, el pronóstico y las estrategias terapeúticas no tiene sentido no compartir los datos. La API (Application Programming Interfaces) de *Beacon* trata de solucionar estas limitaciones de forma que habilita la búsqueda de variantes genómicas e información asociada sin comprometer la privacidad del conjunto de datos. De esta forma, cualquier institución (de investigación u hospital) puede hacer que sus conjuntos de datos ómicos sean “beaconizados” sin comprometer la privacidad del propietario de los datos. Así, la comunidad global se ve beneficiada al disponer de muchos más datos que si no estuvieran accesibles.

Beacon es una API (en algunas ocasiones extendida con una interfaz web) que permite, por tanto, el descubrimiento de datos genómicos y datos fenoclínicos. El proyecto Beacon se desarrollado bajo la iniciativa de GA4GH (Global Alliance for Genomics and Health) y de ELIXIR.

1.2 Beacon v1

En sus orígenes, el protocolo *Beacon* permitía a los usuarios obtener información de la presencia o ausencia de una mutación genómica en un conjunto de datos de pacientes de una determinada enfermedad o de la población en general (Figura 1).

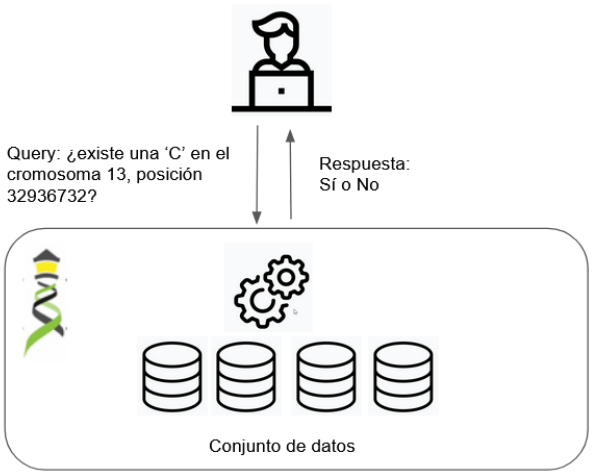

Figura 1. Ejemplo de instancia Beacon v1

Por otro lado, una red de *Beacons* nos permite realizar búsquedas en diferentes instancias Beacon, de forma que, con una única consulta sobre una variante genómica, obtenemos una respuesta agregada de todos los Beacons integrados en la red (Figura 2).

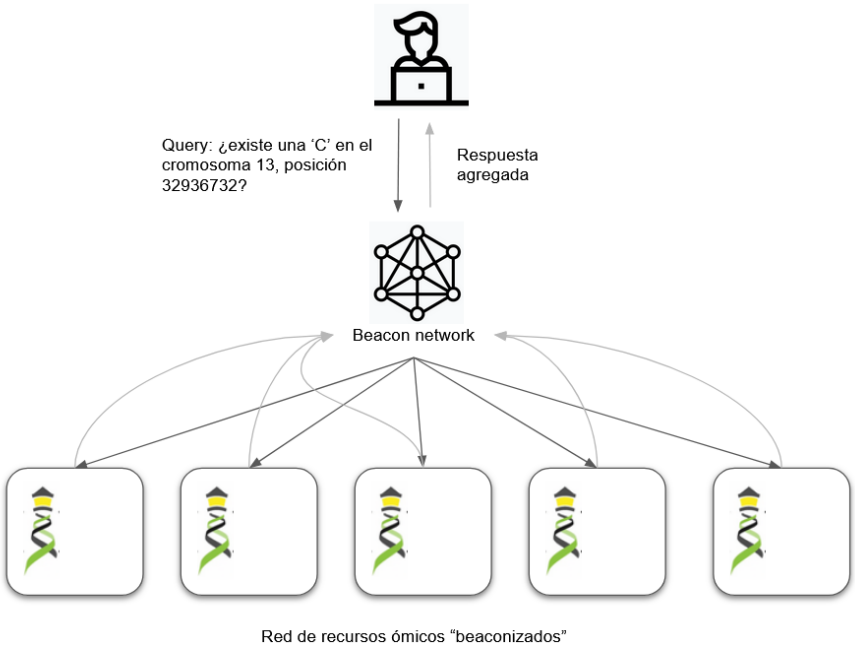

Figura 2. Representación esquemática de una red de Beacons, con una pregunta y una respuesta agregada

Un ejemplo de una red de Beacon, lo encontramos en la página de la "Beacon Network" (https://beacon-network.org) donde podemos hacer consultas de variantes genómicas sobre diferentes bases de datos distribuidas por el mundo (Figura 3).

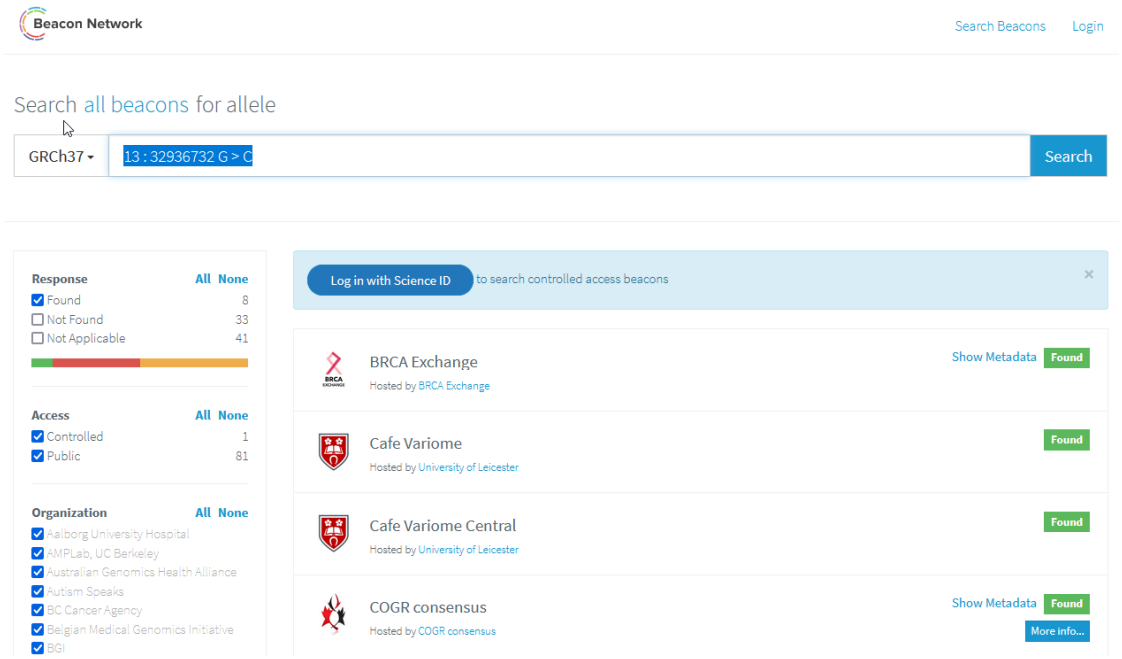

Figura 3. Ejemplo de consulta de la variante 13 : 32936732 G > C en la red de Beacons

1.3 Beacon v2

La versión 1 de Beacon solo permite preguntar por la presencia o ausencia de una determinada variante genómica en una instancia Beacon o en una red d-e Beacons, lo cual hace que su capacidad sea muy limitada. Sin embargo, la versión 2 del protocolo Beacon, actualmente en evaluación por parte de la GA4GH, supone un paso adelante en la consulta de variantes genómicas y datos fenoclínicos. Esta nueva versión, permitirá:

- Realizar preguntas o *queries* más complejas, como por ejemplo preguntar por sexo o edad o realizar filtrados por los términos que establece cada nodo de la red Beacon. Por ejemplo, un nodo puede permitir filtrar por términos HPO (Human Phenotype Ontology) y otro diferente puede permitir filtrar por términos ICD-10 (International Classification of Diseases).
- La posibilidad de conocer el propietario de los datos o las condiciones de uso de los mismos.
- La posibilidad de permitir el salto a otro tipo de consultas. Por ejemplo, si la instancia Beacon es para uso interno de un hospital, sería interesante proporcionar el ID del registro de salud electrónico de los pacientes que tienen las mutaciones de interés.
- Anotaciones de las variantes encontradas, por ejemplo, la patogeneicidad de la variante genómica consultada
- Información acerca de cohortes de pacientes.

Así, para la versión 2, el protocolo Beacon ha evolucionado para cubrir diferentes conceptos o entidades y los detalles asociados a ellas. En la actualidad, el modelo se encuentra en el borrador versión 4 e incluye las siguientes entidades:

- Conjunto de datos (Dataset): agrupa variantes o individuos (sujetos) que tienen algún aspecto en común. Podemos definir el aspecto en común o relación como queramos: puede ser tan débil como por ejemplo que no estén en el mismo repositorio. O también puede ser fuerte, como por ejemplo, pertenencia al mismo estudio.
- Cohorte (Cohort): un conjunto de características que describen una cohorte, la cual es definida como un conjunto de individuos que pueden pertenecer a uno o más conjuntos de datos
- Individuo (Individual): describe individuos que se almacenan en el repositorio, incluyendo información clínica, como enfermedad, tratamiento o características fenotípicas.
- Biomuestra (Biosample): describe muestras tomadas de individuos, incluyendo detalles sobre el procedimiento de extracción y fechas.
- Experimento (Experiment): incluye detalles del procedimiento utilizado para la secuenciación de una biomuestra.
- Análisis (Analyses): contiene detalles del procedimiento bioinformático para identificar variantes genómicas.
- Variantes genómicas (Genomic Variations): describe cómo una variante está presente en una determinada muestra y si se considera más o menos relevante en el diagnóstico de un caso. Adicionalmente, se incluyen anotaciones sobre el efecto de la variante en un fenotipo dado.

Este modelo descrito representa un modelo lógico y no es un ejemplo de una implementación de base de datos para la versión 2 de Beacon. Las relaciones en el modelo son aquellas que serán utilizadas en la respuesta a una consulta Beacon. Así pues, diferentes implementaciones físicas pueden ser compatibles con este modelo lógico de entidades de la versión 2 de Beacon.

La especificación Beacon se compone de dos partes: el entorno o framework Beacon[1] y el modelo Beacon[2]. El entorno es la parte que describe la estructura global de las peticiones, respuestas, parámetros, etc, de la API. El modelo, por otro lado, describe el conjunto de conceptos incluidos en una versión de Beacon, como por ejemplo individuo o biomuestra.

[1] https://github.com/ga4gh-beacon/beacon-framework-v2

[2] https://github.com/ga4gh-beacon/beacon-v2-Models

2 Aspectos técnicos de un sistema Beacon

¿Qué requisitos técnicos son necesario para disponer de una instancia Beacon? En primer lugar y de acuerdo a la especificación 2 de Beacon, necesitamos definir qué entidades se van a consultar. Una vez definidas, necesitamos:

- Un servidor de aplicaciones. Este servidor de aplicaciones (por ejemplo, Apache Tomcat) proporciona los servicios web que implementan los *endpoints* de las entidades. Los endpoints definen una forma de solicitar información a un servicio.
- Una base de datos de variantes, con toda la información necesaria de acuerdo a las entidades definidas. De esta forma, los *endpoints* consultan esta base de datos. Adicionalmente, el tipo de base de datos (por ejemplo, relacional o no relacional) se selecciona en función de las características del proyecto sobre el que se va a implementar un Beacon y de la complejidad de los datos que se vayan a almacenar.

Así, con un servidor de aplicaciones y una base de datos de variantes, sería suficiente para tener una instancia Beacon (Figura 4).

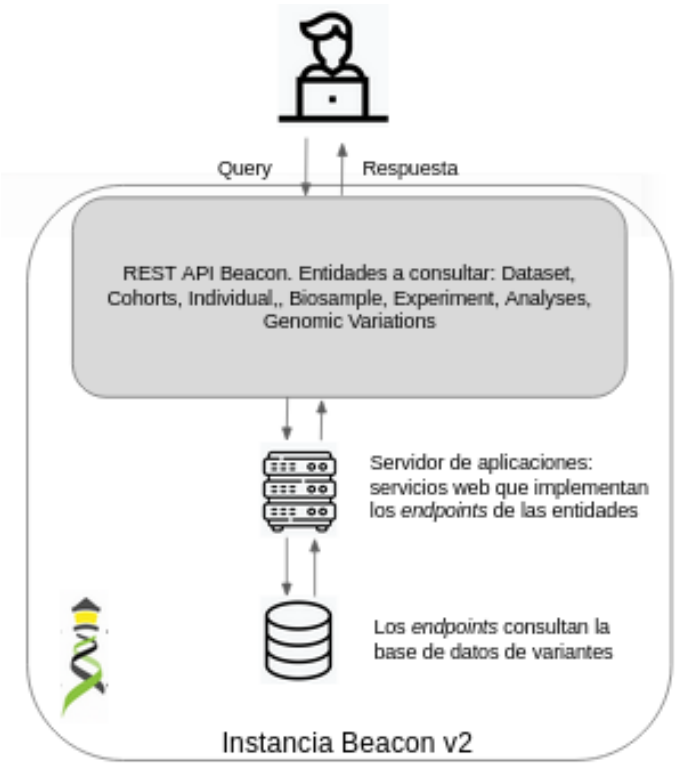

Figura 4. Componentes software de una instancia Beacon

Por otro lado, los recursos computacionales necesarios para la puesta en marcha de un sistema Beacon dependen de varios factores, como por ejemplo:

- La cantidad de datos (muestras)
- El tipo de datos: genomas, exomas o paneles de genes
- La complejidad de las consultas permitidas, que pueden ir desde la existencia de una variante concreta hasta la posibilidad de buscar a los pacientes con un cierto fenotipo y una variante con cierta significancia clínica, por ejemplo.

En función de todos estos factores, se definirán los recursos computacionales. En general, a mayor cantidad de muestras, muestras tipo genoma y más complejidad en las consultas, más recursos serán necesarios.

3 Casos de uso

En este apartado, vamos a hacer una breve descripción de dos casos de uso: uno donde se ha implementado un sistema Beacon v1 y otro donde se está desarrollando un sistema Beacon v2.

3.1 Instancia Beacon en CSVS

CSVS [1] es una iniciativa colaborativa que proporciona información acerca de la variabilidad genómica de población española a la comunidad científica. Almacena, en la actualidad, información de más de 2,000 secuencias (entre exomas y genomas) de individuos españoles no relacionados, ya sean sanos o con alguna enfermedad. La base de datos contiene frecuencias alélicas correspondientes a diferentes consorcios y proyectos como Medical Genome Project, ENoD del CIBERER y NAGEN 1000 genomas entre otras iniciativas, así como grupos de investigación en España. El repositorio es utilizado como una población pseudo-control para la búsqueda de nuevas variantes y genes responsables de enfermedad.

CSVS es descubrible a través de la Beacon Network (https://beacon-network.org/) mediante una instancia de Beacon v1 que se comunica con la base de datos de variantes de CSVS. De esta forma, para una variante genómica introducida en la página web de Beacon Network, si existe en CSVS, se reportará únicamente su presencia. Si no existe, la instancia Beacon dará la correspondiente repuesta negativa (Figura 5).

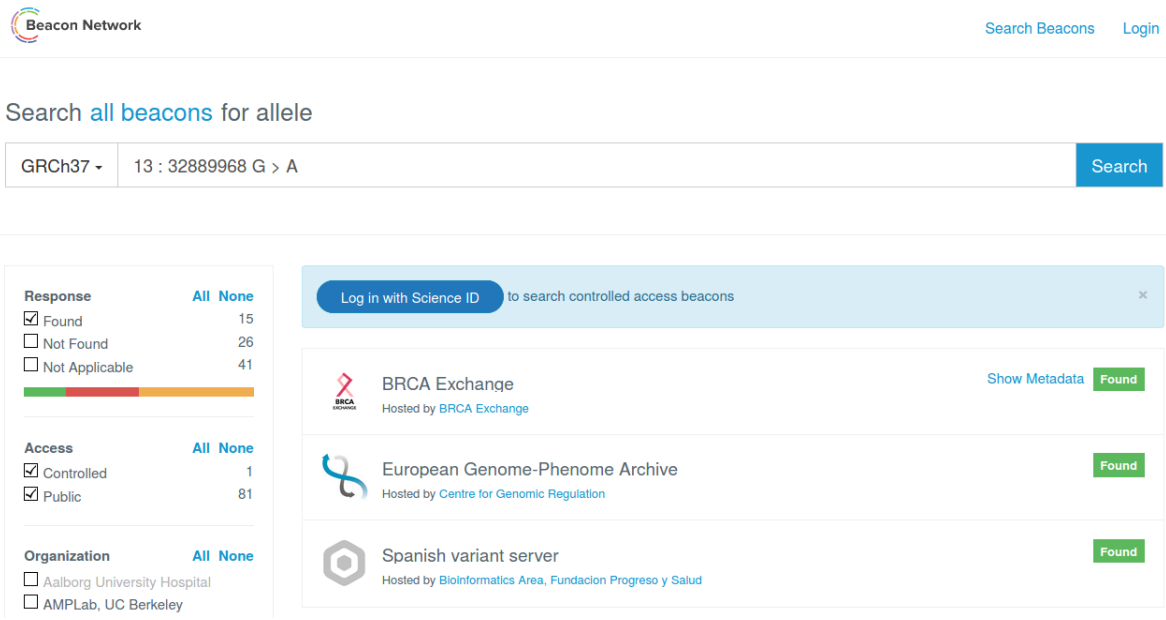

Figura 5. Ejemplo de consulta de variante 13:32889968 G > a través de la "Beacon Network", donde se ha reportado a través de la instancia Beacon v1 asociada a CSVS, que la variante existe en la base de datos de población española

3.2 Instancia Beacon en el programa EnOD

El programa ENoD (Enfermedades no Diagnosticadas) del CIBERER, nació como un modelo de gestión conjunta para casos sin diagnóstico de origen multicéntrico gracias a la estructura distribuida del Centro de Investigación Biomédica en Red en Enfermedades Raras (CIBERER). Basado en comités transversales, con una herramienta en línea de registro y fenotipado HPO de los casos clínicos y dotado de recursos propios, ofrece orientación al diagnóstico y consejo experto, reinterpretación de datos genómicos previos y nuevas pruebas diagnósticas.

Los datos crudos (ficheros FASTQs) de la secuenciación genómica de estas muestras (exomas clínicos, exomas completos y genomas) son analizados por el Área de Bionformática de la Fundación Progreso y Salud. El análisis de las variantes se realiza en dicha área por herramientas de priorización de variantes desarrolladas para ello y esas variantes genómicas se encuentran almacenadas en una base de datos NoSQL.

Sobre esa base de datos que incluye una cohorte de unas 400 muestras en la actualidad, se está gestionando una instancia Beacon v2 donde se implementarán las entidades biomuestra, individuo, variante en muestra y conjunto de datos. Esta instancia Beacon permitirá que las variantes recopiladas en el contexto del programa EnOD, sean accesibles por todas las entidades participantes en el mismo y a su vez, si así se considera por parte de los responsables del programa, sea visible a la comunidad científica en un esfuerzo por compartir información que pueda ser útil en el diagnóstico de las enfermedades raras.

4 Conclusiones

En conclusión, este documento describe de forma muy breve qué es un sistema Beacon, qué puede aportar en el contexto del proyecto IMPaCT-Data y cuáles son los requisitos técnicos necesarios para su implantación. Como ejemplo de caso de uso de Beacon v2, se propone el que se está desarrollando en el programa ENoD del CIBERER, pero se definirán sistemas Beacon en los diferentes grupos invoclucrados en la generación de datos genómicos en el contexto del proyecto IMPaCT y que se irán detallado en sucesivos entregables.

Referencias

1. Peña-Chilet M, Roldán G, Perez-Florido J, et al. CSVS, a crowdsourcing database of the Spanish population genetic variability. Nucleic Acids Res. 2021;49(D1):D1130-D1137

Glosario

API	Application Programming Interfaces
CSVS	Collaborative Spanish Variation Server
ENoD	Enfermedades no Diagnosticadas
GA4GH	Global Alliance for Genomics and Health
HPO	Human Phenotype Ontology
ICD	International Classification of Diseases
IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT