OpenOmics

Una red global P2P para el almacenamiento y consulta en tiempo real de datos genómicos

Ignacio del Valle Alles

Introducción

- Un poco de biología
  - ADN > proteínas > fenotipo
  - 3000 millones ACGT, 99.9 % igual entre individuos. 0.1% diferencias a nivel genético > diferencias a nivel fenotípico

...ACGTACCAGTGTGTTAC...
...ACGTACCCGTGTGTTAC...
...ACGTACCAGTGTGTTAC...
...ACGTACCAGTGTGTTAC...

Introducción

- Panorama bioinformático actual
  - Abaratamiento secuenciación: (1000 dolares/genoma) + potencia demostrada genómica > número creciente de datos disponibles
  - Fase de genotipado e interpretación aislada (enfermedades monogénicas) > Fase *Big Data* e interpretación estadística (fenotipos complejos)
  - Proyectos de investigación trabajando aisladamente
  - Globalmente proceso no óptimo. Necesidad de maximizar datos y metadatos

OPENOMICS

DREAMgenics
Genomes made easy

Global Alliance
for Genomics & Health

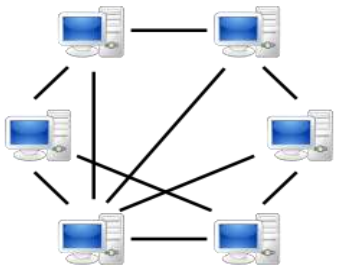
- Fundada en 2013
- Iniciativa internacional
- Compartición segura de datos genómicos y médicos
- White paper
  - Diversidad de opiniones entre individuos
  - Grandes diferencias legales entre países > Involucrar a organizaciones y gobiernos se prevé como un proceso lento y costoso

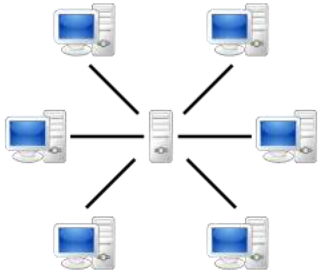
Openness

- Ley universal: Libertad para publicar voluntariamente información sobre uno mismo
- Base del camino a la colaboración global
- Necesidad de un sistema donde todos los voluntarios puedan ceder su información
- Sistema no propietario, abierto y gratuito
- Neutral
- No susceptible de ser controlado en beneficio de unos pocos, o por intereses ocultos, y basado en la colaboración y mérito objetivo de sus participantes
- Infraestructura ideal para este sistema: P2P

P2P

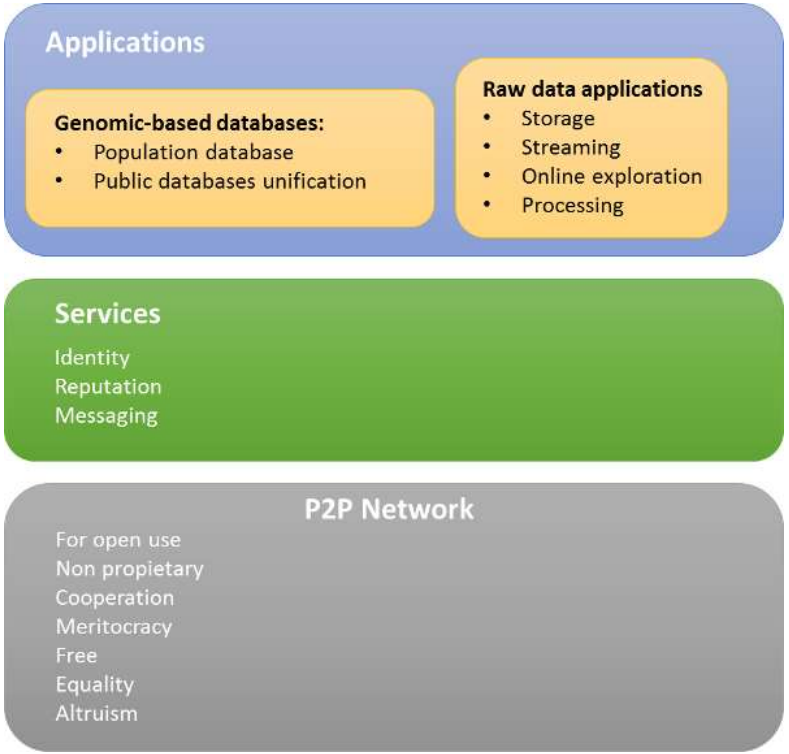
- P2P != Ilegalidad
- ¿Qué es P2P?





- Única infraestructura que garantiza el cumplimiento de los anteriores principios
- Su neutralidad y transparencia maximizaría el número de usuarios a nivel global

OpenOmics



- Proyecto abierto, colaborativo y meritocrático
- Infraestructura escalable y extensible orientada a la colaboración en bioinformática
- Especificación de protocolos + implementación de referencia
- Objetivo último: BBDD abierta de datos personales cedidos voluntariamente, directamente consultable.
- Otras aplicaciones muy útiles para la comunidad
- Adaptable a nuevas necesidades futuras.

- Reputación: Filtrado de datos y usuarios
- Mensajería: Colaboración y curación de datos

OpenOmics

Ideas clave (bio)

Representación genómica:

- Representación determinista del genotipo. Sistema directamente consultable
- Nueva nomenclatura para representar no sólo regiones variantes, sino regiones no variantes y desconocidas > frecuencias alélicas

Representación clínica o fenotípica:

- Utilización de ontologías

OpenOmics

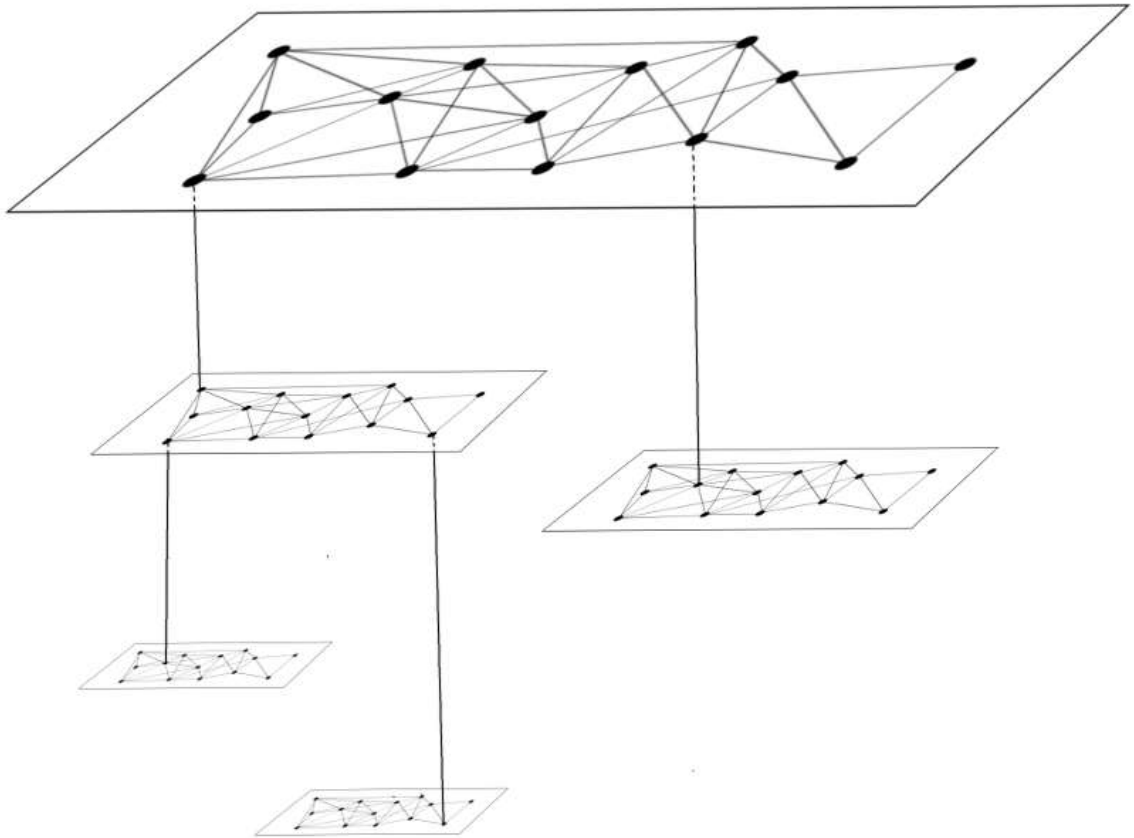
Ideas clave (técnicas)

- Nueva tecnología inspirada en *Apache Cassandra* y *Amazon's Dynamo*
- Consistencia eventual de los datos
- Modelo de consultas: Búsqueda facetada, ordenación y range queries > Algoritmo propio de particionado dinámico
- Interfaz web y API REST a través del peer local
- Private networks

OPENOMICS

DREAMgenics
Genomes made easy

OpenOmics



Conclusiones

- Beneficios (tipo y objeto)
  - Sociedad
  - Comunidad bioinformática
  - Participantes en el proyecto
  - Empresas

- Estado del proyecto
  - H2020. Liderado por Treelogic. 10 empresas y organizaciones EU
  - GA4GH
  - Participación abierta

https://code.google.com/p/openomics

¿Por ventura es la sociedad otra cosa que una gran compañía, en que cada uno pone sus fuerzas y sus luces, y las consagra al bien de los demás?

Gaspar Melchor de Jovellanos

Bootstrapping (unirse a una red)

1. Listado de nodos conocidos previamente en la red
2. Conexión a esos nodos y a otros nuevos conocidos por ellos
3. En el proceso de conexión se informa al otro nodo de la región del espacio de claves que el propio nodo está cubriendo (tiene indexada)
4. Tabla de enrutamiento: creación y mantenimiento (pooling) de una mapa en memoria de intervalo en espacio de claves de búsqueda a nodos activos (con redundancia) hasta tener cubierto todo el espacio de claves
5. Preguntar a nodos con solapamiento por actualizaciones de datos y actualización del índice local
6. Peer operativo

Introducción de datos a la red

1. El usuario introduce los datos (genómicos, fenotípicos, ambiente, tecnologías empleadas) a través de su peer local
2. Validación de los datos en forma (respecto a los esquemas definidos)
3. Por cada registro: validación, normalización, anotación transcritos, anotación ontológica
4. Validación global
5. Por cada registro:
    a. Firmado digital del registro
    b. Envío gossip del registro a los nodos cubriendo el espacio del registro
    c. Indexación local

Consulta de datos

1. La consulta se divide en subconsultas a enviar a algunos de los nodos conocidos.
2. Se agregan los resultados de cada subconsulta, redundancia opcional.
3. Clasificación facetada de los resultados

Representación genotipo

SNV, indels:


```
{"refVersion": "GRCh37",
 "seqId": "Chr1",
 "start": 156514,
 "end": 156514,
 "ref": "T",
 "alt": "CA"}
```


Gross deletion:


```
{ "refVersion": "GRCh37",
 "seqId": "Chr1",
 "start": 156514,
 "end": 556514,
 "alt": "-"}
```


Rearrangement:


```
{ "refVersion": "GRCh37",
 "seqId": "Chr1",
 "start": 156514,
 "end": 556514,
 "alt": "Chr2:505651-605651"}
```


CNVs:


```
{ "refVersion": "GRCh37",
 "seqId": "Chr1",
 "start": 156514,
 "end": 156516,
 "alt": "4>50"}
```


Reference match:


```
{ "refVersion": "GRCh37",
 "seqId": "Chr1",
 "start": 156514,
 "end": 156516,
 "alt": "."}
```


Unknown region


```
{ "refVersion": "GRCh37",
 "seqId": "Chr1",
 "start": 156514,
 "end": 156516,
 "alt": "?"}
```