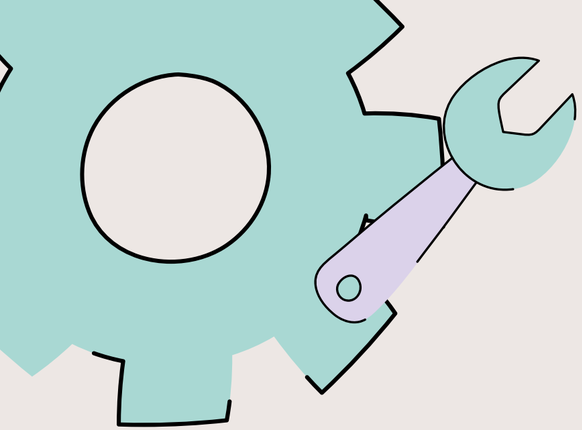
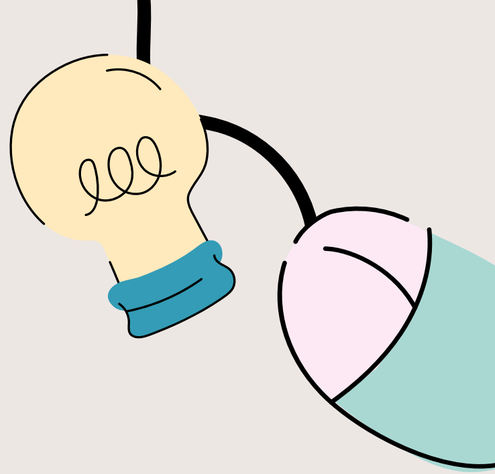
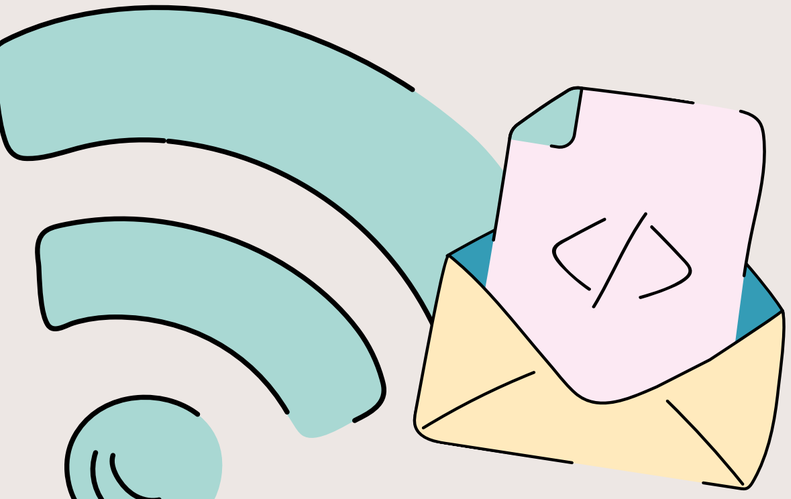


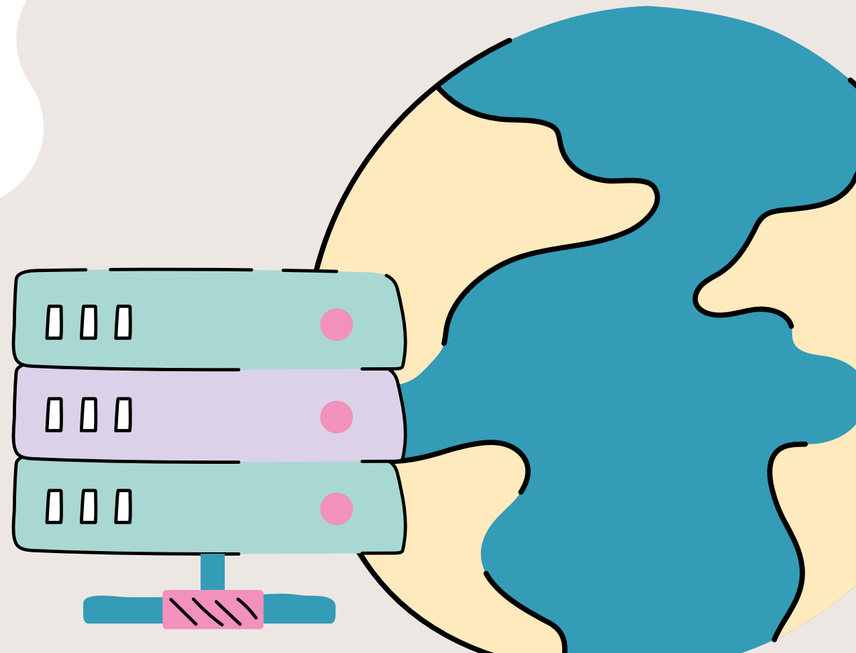


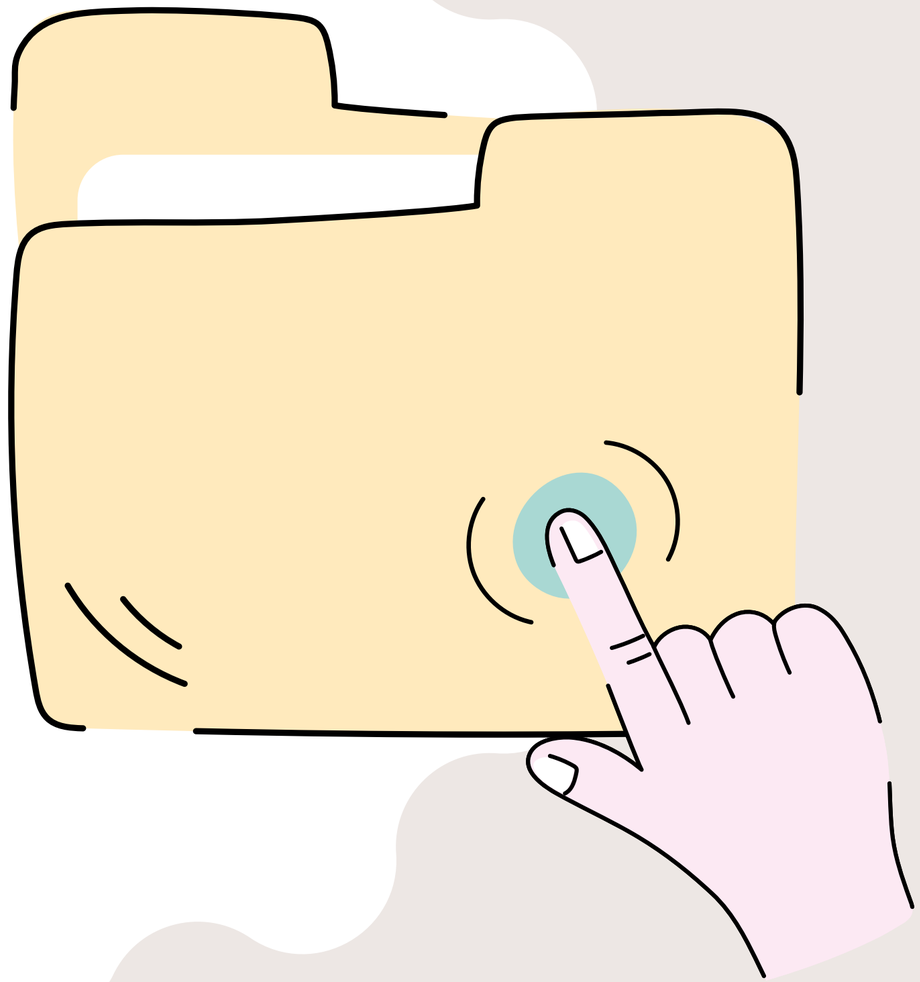
Mutaciones humanas

Database

Nerea Martin Serrano
David Cubilos Del Toro







Contenidos

1. Introducción
2. Modelo relacional MySQL
3. Inserción de datos
4. Query design
5. Optimización de la base de datos
6. Diseño en XML
7. NoSQL
8. Conclusiones

1. Introduccion

Base de datos que contiene información sobre mutaciones en humanos. Contiene información sobre:

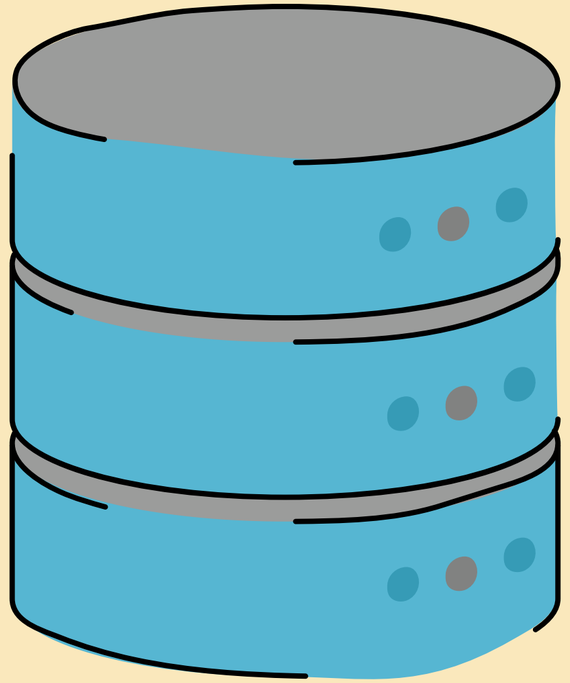
El gen mutado

El gen original

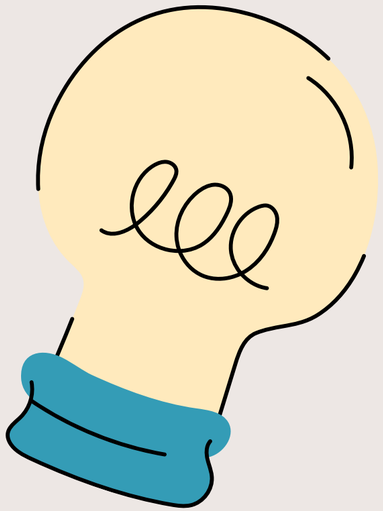
Proteína afectada

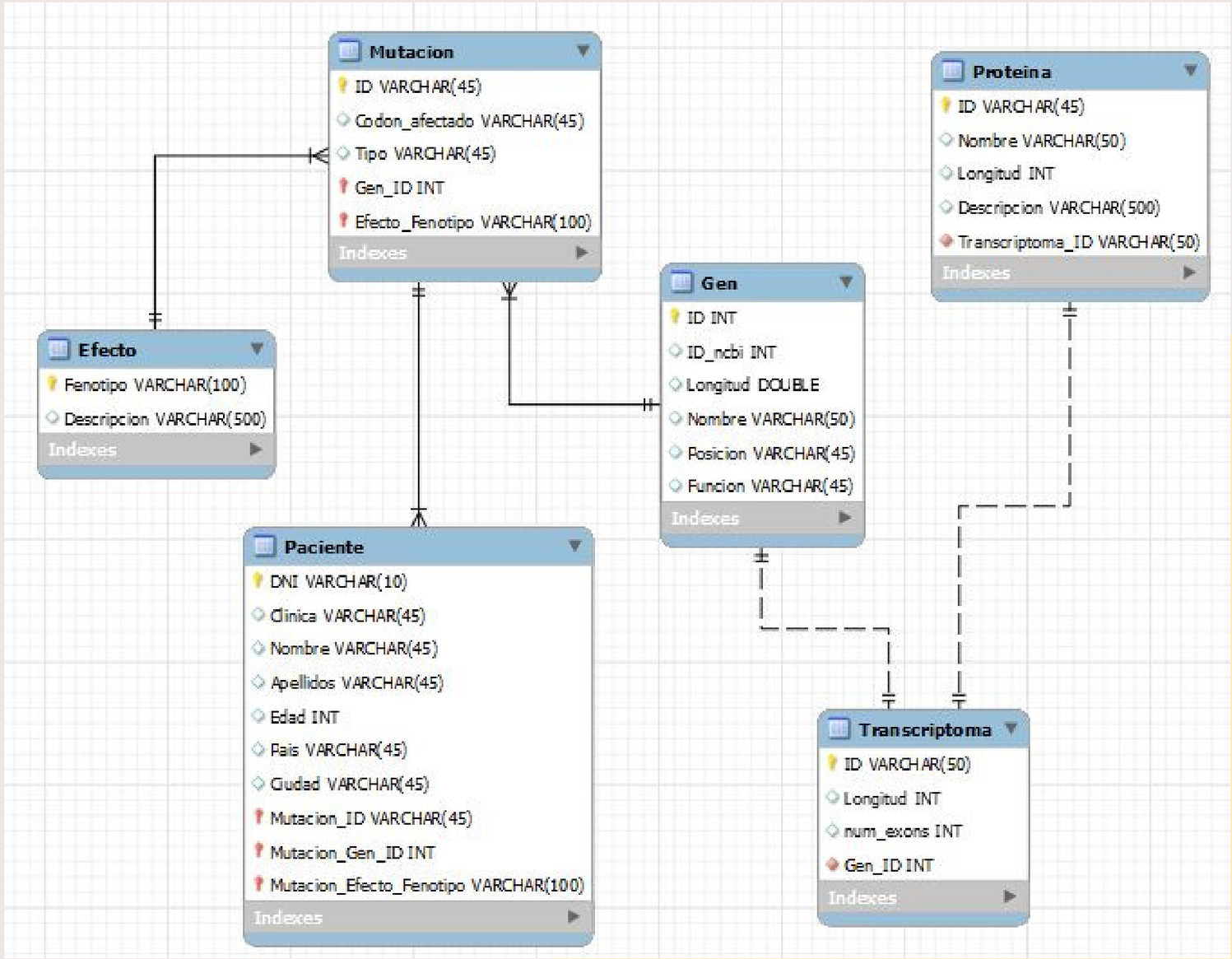
Enfermedad que puede causar la mutación

Pacientes



2. Modelo relacional





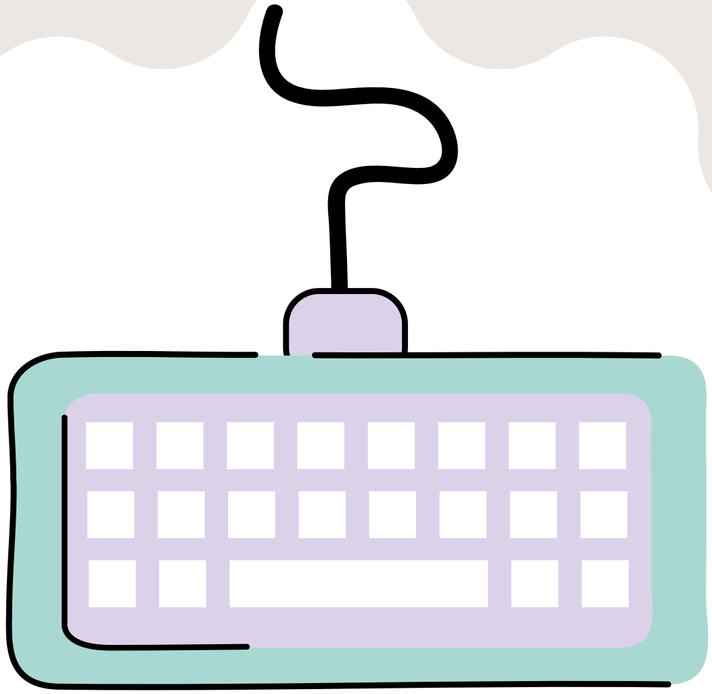
3. Inserción de Datos

Una vez tenemos el modelo relacional, habrá que almacenar los datos en la base.

- Debemos insertar las tablas en orden por sus foreign key, la primera es gen y las siguientes aquellas que almacenen Gen_Id

- En un principio, la insercion fue manual, sacada de bases de datosd de genes online como OMIM, NCBI o HGMD

- Para ampliar los datos se ha recurrido a mockaroo.com, una web que genera datos aleatorios con el tamaño que deseemos de forma automatizada



4. Query Desing

Con cláusulas Where

```
1 •  SELECT Gen.Nombre AS Gene_Symbol, Gen.Longitud AS Longitud_Gen,
2        Proteina.Nombre AS Nombre_Proteina
3    FROM Gen, Mutacion, Proteina, Transcriptoma
4    WHERE  Mutacion.Tipo = "Missense/nonsense"
5        AND Gen.ID = Transcriptoma.Gen_ID
6        AND Gen.ID = Mutacion.Gen_ID
7        AND Transcriptoma.ID = Proteina.Transcriptoma_ID
8        AND Gen.Longitud <= 200
9    ORDER BY Gen.Longitud ASC;
```

esult Grid | Filter Rows: | Export: | Wrap Cell Content:

Gene_Symbol	Longitud_Gen	Nombre_Proteina
ARHGDIA	3.62	Rho GDP-dissociation inhibitor 1
ACTG2	26.69	Actin, gamma-enteric smooth muscle
ACTG2	26.69	Actin, gamma-enteric smooth muscle
ATP7A	139.7	Copper-transporting ATPase 1

Query 1

Gene Symbol y la proteína que codifica de los genes con longitud menor a 200 kb que sean afectados por mutaciones tipo missense

```
1 •  SELECT AVG(Pacientes.Edad) AS Edad_Media, Mutacion_Efecto_Fenotipo
2    FROM Pacientes
3    WHERE Mutacion_Efecto_Fenotipo = "Motor neuropathy, distal";
```

esult Grid | Filter Rows: | Export: | Wrap Cell Content:

Edad_Media	Mutacion_Efecto_Fenotipo
36.6667	Motor neuropathy, distal

Query 2

Edad media de los pacientes que sufren de Motor neuropathy

Con cláusulas Join

```sql
SELECT Gen.Nombre AS Gene_Symbol, Gen.ID_ncbi AS NCBI,
    Proteina.Nombre AS Proteina_Nombre, Proteina.Descripcion
FROM Gen, Proteina
JOIN Transcriptoma
ON Transcriptoma.ID = Proteina.Transcriptoma_ID
WHERE Gen.ID = Transcriptoma.Gen_ID;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Gene_Symbol	NCBI	Proteina_Nombre	Descripcion
ACTG2	72	Actin, gamma-enteric smooth muscle	Actins are highly...
ABAT	18	4-aminobutyrate aminotransferase, mitochondrial	Catalyzes the co...
OSBP	5007	Oxysterol-binding protein 1	Lipid transporter...
ATP7A	538	Copper-transporting ATPase 1	ATP-driven copp...
ORC1	4998	Origin recognition complex subunit 1	Component of th...
ARHGDIA	396	Rho GDP-dissociation inhibitor 1	Controls Rho pro...

Query 3

Lista completa de los genes junto con las proteínas que codifican

```sql
SELECT Paciente.Nombre, Paciente.Ciudad, Efecto.Fenotipo,
        Efecto.Descripcion AS 'Descripcion enfermedad'
From Paciente, Efecto
LEFT JOIN Mutacion
ON Mutacion.Efecto_Fenotipo = Efecto.Fenotipo
Where Paciente.Ciudad= "Madrid"
AND Paciente.Mutacion_Efecto_Fenotipo=Efecto.Fenotipo ;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Nombre	Ciudad	Fenotipo	Descripcion enfermedad
MARIA DEL CARMEN	Madrid	Motor neuropathy, distal	The distal hereditary motor neuropathies (dHM...
JUAN PABLO	Madrid	Motor neuropathy, distal	The distal hereditary motor neuropathies (dHM...

Query 4

Busca aqeuellos pacientes que esten siendo tratados en Madrid

Subqueries and Where conditions

```sql
SELECT DISTINCT Proteina.Nombre, Efecto_Fenotipo
FROM Proteina, Mutacion
WHERE Longitud >= 700 AND Transcriptoma_ID  IN (SELECT ID FROM Transcriptoma
    WHERE num_exons > 5)
AND Tipo = "Splicing" AND Gen_ID IN (SELECT Gen_ID FROM Transcriptoma WHERE num_exons > 5)
ORDER BY Proteina.Nombre
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Nombre	Efecto_Fenotipo
a	Menkes syndrome
a	Meier-Gorlin syndrome
a feugiat	Menkes syndrome
a feugiat	Meier-Gorlin syndrome
a libero	Menkes syndrome
a libero	Meier-Gorlin syndrome
a nibh	Menkes syndrome
a nibh	Meier-Gorlin syndrome
a pede	Menkes syndrome
a pede	Meier-Gorlin syndrome
ac	Menkes syndrome
ac	Meier-Gorlin syndrome

Query 5

Efecto que tienen las mutaciones producidas por splicingen proteínas grandes (con más de 500 aminoácidos) y con un número de exones mayor a 5,

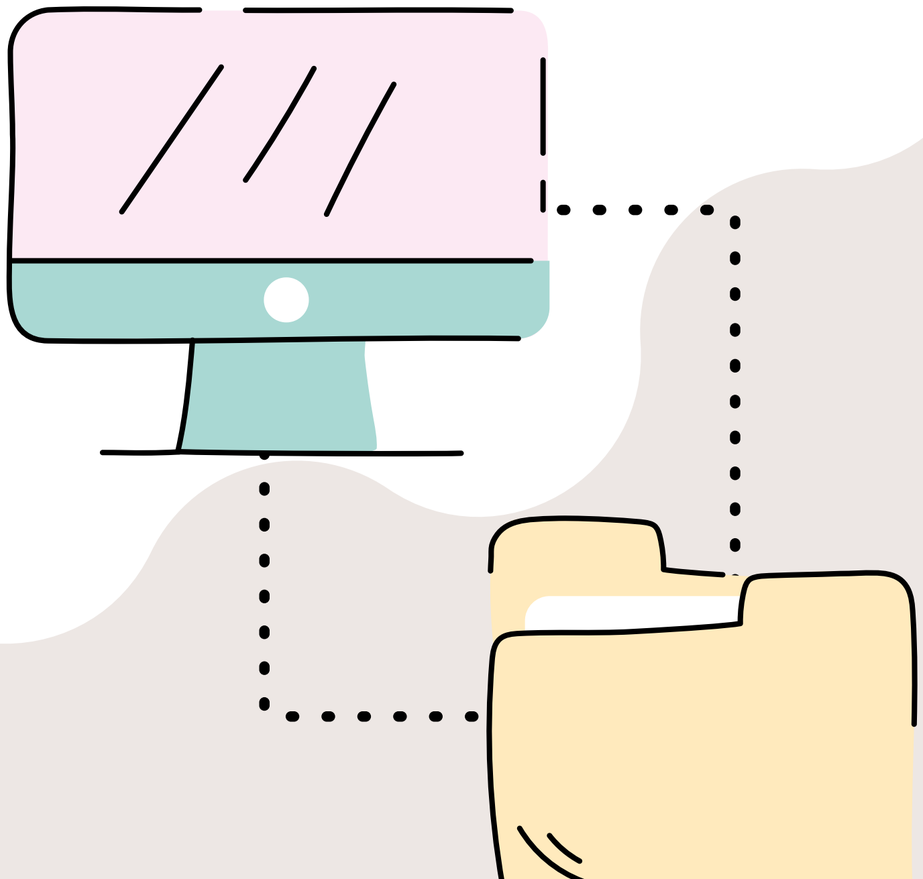
```sql
SELECT DISTINCT
Paciente.Apellidos AS 'Paciente',Paciente.Edad,
Paciente.Mutacion_Efecto_Fenotipo AS 'Enfermedad',
Paciente.Mutacion_Gen_ID AS ' ID Gen afectado'

FROM  Paciente
WHERE Paciente.Mutacion_ID in (
  SELECT Mutacion_ID
  FROM Mutacion
  WHERE  Gen_ID IN (SELECT Gen_ID
                    FROM Gen
                    WHERE Gen.Longitud<900
                    AND Gen.Longitud>=500)
  AND Mutacion.Tipo= 'Missense/nonsense')
AND Paciente.DNI in ( SELECT Paciente.DNI
                      FROM  Paciente
                      WHERE Paciente.Edad >17
                      AND Paciente.Edad <60)
```

Query 6

Se busca estudiar las mutacions del tipo **missense/nosense** que puedan aparecer en pacientes con edades comprendidas **entre 17 y 60 años** , que ademas tengan un gen asociado con una longitud **ente 500 y 900 pares de bases**

5 . Optimizacion de la base de datos



Métodos de optimización

- Correcto diseño de las tablas
- Uso optimizado de indices
- Motores de busqueda adecuados

Tipos de almacenamiento

- Innodb
- Memory
- MyISAM

Índices

Query 1

.	Inicial	Index X
Tiempo (ms)	0.0263246	0.0011939
Tuplas devueltas/examinadas	11/996	11/37

Query 3

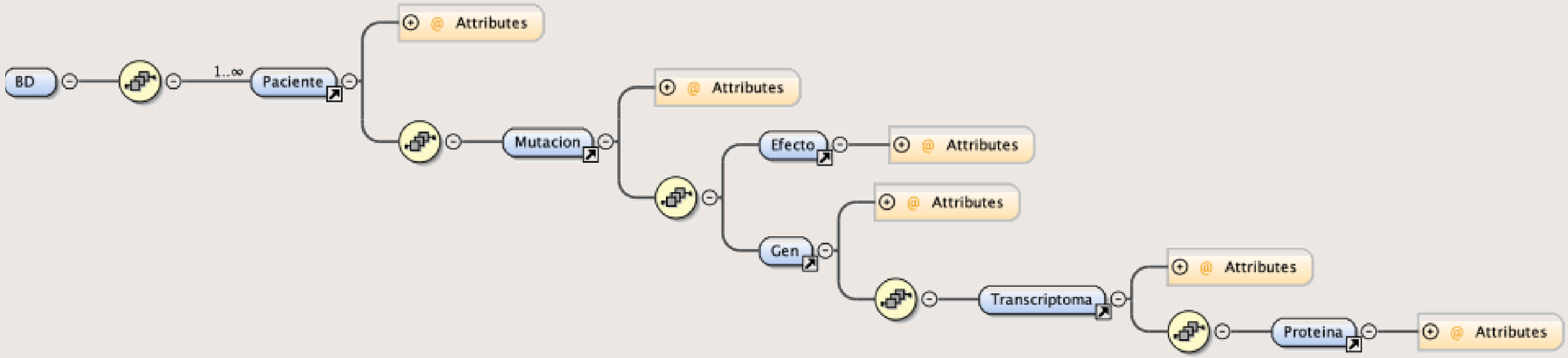
.	Inicial	Index X
Tiempo (ms)	0.0147838	0.0033916
Tuplas devueltas/examinadas	2/1935	2/967

Query 5

*	Inicial	Index X	Index Y
Index Z	**Index X+Z**		
Tiempo (ms)	0.0236228	0.0186792	0.4293
0.04736	0.03		
Tuplas devueltas/examinadas	930/5231	930/4963	930/593986
930/4155	930/3887		

6. Diseño en XML

XML significa eXtensible Markup Language. XML fue diseñado para almacenar y transportar datos de manera que pudiera ser legible tanto por humanos como por máquinas



Conversión a xml

```
SELECT CONCAT ('<BD>',group_concat('<Paciente DNI= "' , Paciente.DNI, '" Clinica="',Paciente.Clinica,
        '" Nombre="',Paciente.Nombre,'" Apellidos="',Paciente.Apellidos,'" Edad="',Paciente.Edad,
                '" Pais="',Paciente.Pais,'" Ciudad="',Paciente.Ciudad, '">',
                (       SELECT CONCAT (
                  group_concat('<Mutacion ID= "' , Mutacion.ID, '" Codon_afectado="',Mutacion.Codon_afectado,
                                '" Tipo="',Mutacion.Tipo,'"> ',(
                                 SELECT CONCAT (
                                   group_concat('<Efecto Fenotipo= "' , Efecto.Fenotipo,'">' '','' separator''),
                                   '</Efecto>')  FROM Efecto
                                 GROUP BY Efecto.Fenotipo
                                 HAVING Efecto.Fenotipo= Mutacion.Efecto_Fenotipo
```

Se ha generado una query que explore todas las tablas del proyecto y devuelva un string con formato XML.

Concat

la funcion conca, como su nombre indica, concatena uno o mas strings

TablaPacientes.xml — Modificado

```
<BD>
        <Paciente DNI= "57743585I " Clinica="Hospital Pellegrin Bordeaux" Nombre="OSCAR  " Apellidos="GARCIA PALACIOS
" Edad="12" Pais="France  " Ciudad="Bourdeaux"><Mutacion ID= "CM112288" Codon_afectado="TTT-TCT" Tipo="Missense/
nonsense "> <Efecto Fenotipo= "Microcephalic primordial dwarfism"></Efecto><Gen ID= "601902" Longitud= "867" Nombre=
"ORC1" Posicion= "1p32" Funcion= "coding gene"><Transcriptoma ID= "NM_004153" Longitud="3192" num_exons="17"><Proteina
ID= "Q13415" Nombre="Origin recognition complex subunit 1" Longitud="861"></Proteina></Transcriptoma></Gen></
Mutacion></Paciente>

        <Paciente DNI= "85595548X" Clinica="Npkeo lzbnqh" Nombre="Cmgfqe" Apellidos="Rkyujo" Edad="5" Pais="Qcixug"
Ciudad="Msngpn"><Mutacion ID= "CM1618630" Codon_afectado="GAC-ACC" Tipo="Missense/nonsense   "> <Efecto Fenotipo=
"Neurodegeneration"></Efecto><Gen ID= "137150" Longitud= "500" Nombre= "ABAT" Posicion= "16p13.2" Funcion= "coding
gene"><Transcriptoma ID= "ABAT-V1" Longitud="4831" num_exons="16"><Proteina ID= "P80404" Nombre="4-aminobutyrate
aminotransferase, mitochondrial" Longitud="500"></Proteina></Transcriptoma></Gen></Mutacion></Paciente>

<Paciente DNI= "17284929F " Clinica="Fundacion Jimenez Diaz" Nombre="MARIA DEL CARMEN  " Apellidos="PEREZ TORDESILLAS
" Edad="28" Pais="Spain  " Ciudad="Madrid"><Mutacion ID= "CM172480" Codon_afectado="TAC-TGC" Tipo="Missense/nonsense">
<Efecto Fenotipo= "Motor neuropathy, distal"></Efecto><Gen ID= "300011" Longitud= "1500" Nombre= "ATP7A" Posicion=
"Xq13.2-q13.3" Funcion= "coding gene"><Transcriptoma ID= "NM_000052" Longitud="8488" num_exons="23"><Proteina ID=
"Q04656" Nombre="Copper-transporting ATPase 1" Longitud="1500"></Proteina></Transcriptoma></Gen></Mutacion></Paciente>

        <Paciente DNI= "98929839H " Clinica="Hospital Edouard Herriot" Nombre="ANA MARIA" Apellidos="CALZADA RIVERA"
Edad="22" Pais="France  " Ciudad="Lyon"><Mutacion ID= "CM1615168" Codon_afectado="CCCTCC" Tipo="Missense/
nonsense "> <Efecto Fenotipo= "GABA-transaminase deficiency"></Efecto><Gen ID= "137150" Longitud= "500" Nombre= "ABAT"
Posicion= "16p13.2" Funcion= "coding gene"><Transcriptoma ID= "ABAT-V1" Longitud="4831" num_exons="16"><Proteina ID=
"P80404" Nombre="4-aminobutyrate aminotransferase, mitochondrial" Longitud="500"></Proteina></Transcriptoma></Gen></
Mutacion></Paciente>

        <Paciente DNI= "18296427J " Clinica="Hospital San Carlos" Nombre=" JUAN PABLO" Apellidos="MORENCO SALAS"
Edad="32" Pais="Spain  " Ciudad="Madrid"><Mutacion ID= "CM172480" Codon_afectado="TAC-TGC" Tipo="Missense/nonsense">
<Efecto Fenotipo= "Motor neuropathy, distal"></Efecto><Gen ID= "300011" Longitud= "1500" Nombre= "ATP7A" Posicion=
"Xq13.2-q13.3" Funcion= "coding gene"><Transcriptoma ID= "NM_000052" Longitud="8488" num_exons="23"><Proteina ID=
"Q04656" Nombre="Copper-transporting ATPase 1" Longitud="1500"></Proteina></Transcriptoma></Gen></Mutacion></Paciente>
        <Paciente DNI= "89796179G " Clinica="Hospital San Paolo" Nombre="LUCIA   " Apellidos="PEREZ GALDOS   "
Edad="46" Pais="italy  " Ciudad="Milano"><Mutacion ID= "CM1615168" Codon_afectado="CCCTCC" Tipo="Missense/
nonsense "> <Efecto Fenotipo= "GABA-transaminase deficiency"></Efecto><Gen ID= "137150" Longitud= "500" Nombre= "ABAT"
Posicion= "16p13.2" Funcion= "coding gene"><Transcriptoma ID= "ABAT-V1" Longitud="4831" num_exons="16"><Proteina ID=
"P80404" Nombre="4-aminobutyrate aminotransferase, mitochondrial" Longitud="500"></Proteina></Transcriptoma></Gen></
Mutacion></Paciente>

        <Paciente DNI= "92793892P " Clinica="Aurelia Hospital" Nombre="PEDRO   " Apellidos="SANCHEZ CASTEJON   "
Edad="50" Pais="Italy  " Ciudad="Roma"><Mutacion ID= "CM172480" Codon_afectado="TAC-TGC" Tipo="Missense/nonsense">
<Efecto Fenotipo= "Motor neuropathy, distal"></Efecto><Gen ID= "300011" Longitud= "1500" Nombre= "ATP7A" Posicion=
"Xq13.2-q13.3" Funcion= "coding gene"><Transcriptoma ID= "NM_000052" Longitud="8488" num_exons="23"><Proteina ID=
"Q04656" Nombre="Copper-transporting ATPase 1" Longitud="1500"></Proteina></Transcriptoma></Gen></Mutacion></Paciente>
```

XQuery

```
1  xquery version "3.1";
2
3  for $paciente in doc("/db/BD_mutaciones/TablaPacientes.xml")/BD/Paciente
4  where $paciente/@Ciudad ="Madrid"
5  return <Resultado><Paciente>{data($paciente/@Nombre)}</Paciente>
6      <Fenotipo>{data($paciente/Mutacion/Efecto/@Fenotipo)}</Fenotipo>
7  </Resultado>
8
```

/db/BD_mutaciones/consultaMadrid

Adaptive Output | Indent | Live Preview | Highlight Index Matches

```
1 <Resultado>
      <Paciente>MARIA DEL CARMEN  </Paciente>
      <Fenotipo>Motor neuropathy, distal</Fenotipo>
  </Resultado>
2 <Resultado>
      <Paciente> JUAN PABLO</Paciente>
      <Fenotipo>Motor neuropathy, distal</Fenotipo>
  </Resultado>
```

XQuery 1: Mutaciones en Madrid

Pacientes que viven en Madrid y la enfermedad que padecen por culpa de una mutación

```
1  xquery version "3.1";
2
3  for $paciente in doc("/db/BD_mutaciones/TablaPacientes.xml")/BD/Paciente
4
5  where xs:integer($paciente/Mutacion/Gen/Transcriptoma/Proteina/@Longitud) > 700
6  and xs:integer($paciente/Mutacion/Gen/Transcriptoma/@num_exons) > 5
7  and $paciente/Mutacion/@Tipo = "Missense/nonsense "
8  return <Resultado>
9      <Proteina>{data($paciente/Mutacion/Gen/Transcriptoma/Proteina/@Nombre)}</Proteina>
10     <Fenotipo>{data($paciente/Mutacion/Efecto/@Fenotipo)}</Fenotipo>
11 </Resultado>
```

/db/BD_mutaciones/consultaProteina

Adaptive Output | Indent | Live Preview | Highlight Index Matches

```
1 <Resultado>
      <Proteina>Origin recognition complex subunit 1</Proteina>
      <Fenotipo>Microcephalic primordial dwarfism</Fenotipo>
  </Resultado>
2 <Resultado>
      <Proteina>Copper-transporting ATPase 1</Proteina>
      <Fenotipo>Occipital horn syndrome</Fenotipo>
  </Resultado>
3 <Resultado>
      <Proteina>Origin recognition complex subunit 1</Proteina>
      <Fenotipo>Microcephalic primordial dwarfism</Fenotipo>
  </Resultado>
```

XQuery 2: Query sobre Proteína

Efecto que tienen las mutaciones producidas por splicingen proteínas grandes (con más de 500 aminoácidos) y con un número de exones mayor a 5,

```
1  xquery version "3.1";
2  avg(
3  for $paciente in doc("/db/BD_mutaciones/TablaPacientes.xml")/BD/Paciente
4  where contains($paciente/Mutacion/Efecto/@Fenotipo, "Motor neuropathy")
5  let $avg := avg(xs:integer($paciente/@Edad))
6  return $avg)
7
```

__new__1

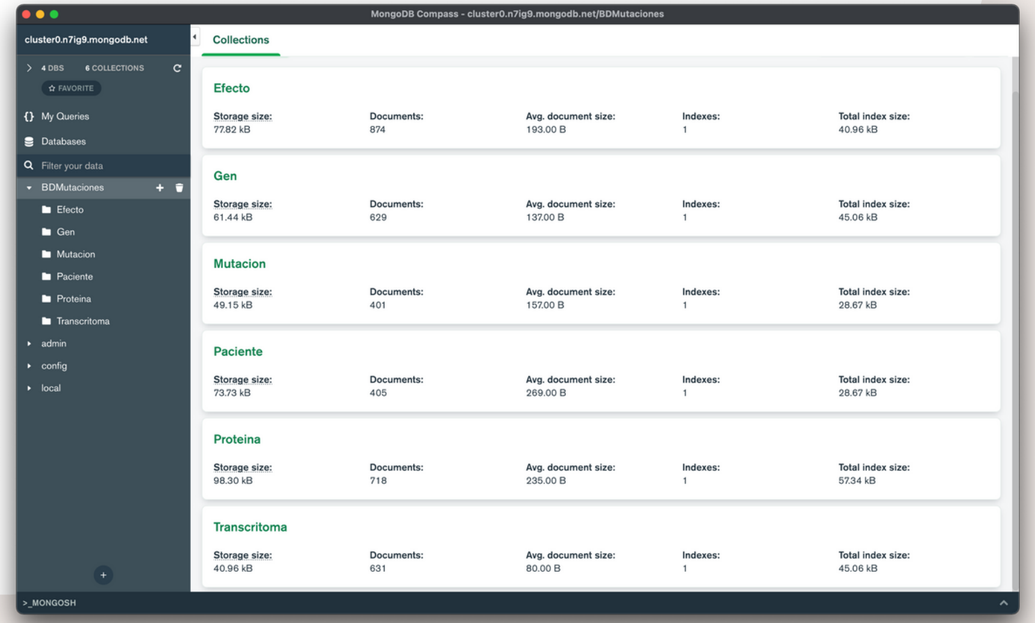
Adaptive Output ☑ Indent ☐ Live Preview ☑ Highlight Index Matches

```
1  36.666666666666666667
```

XQuery 3: Query Media Edad

Edad media de los pacientes que sufren de Motor neuropathy

7. NoSQL



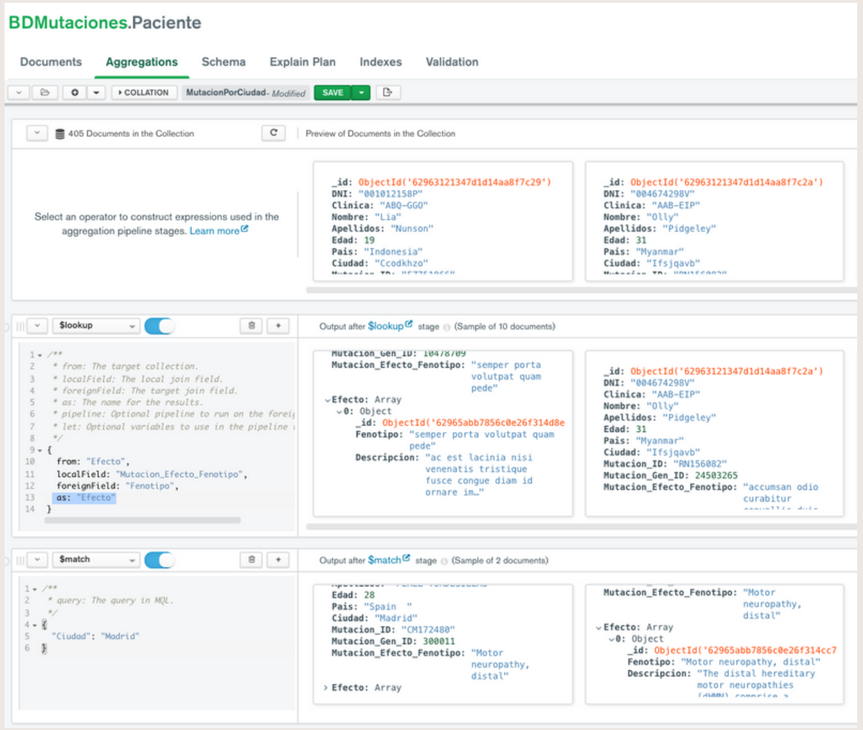
Migracion de datos

- En primer lugar, crearemos las colecciones para almacenar las tablas correspondientes.

- Desde MySQL, con el export wizzard, se han pasado las tablas a formato JSON, format que mongo admite en sus colecciones.
- Al final, las colecciones con los datos se verían como la imagen

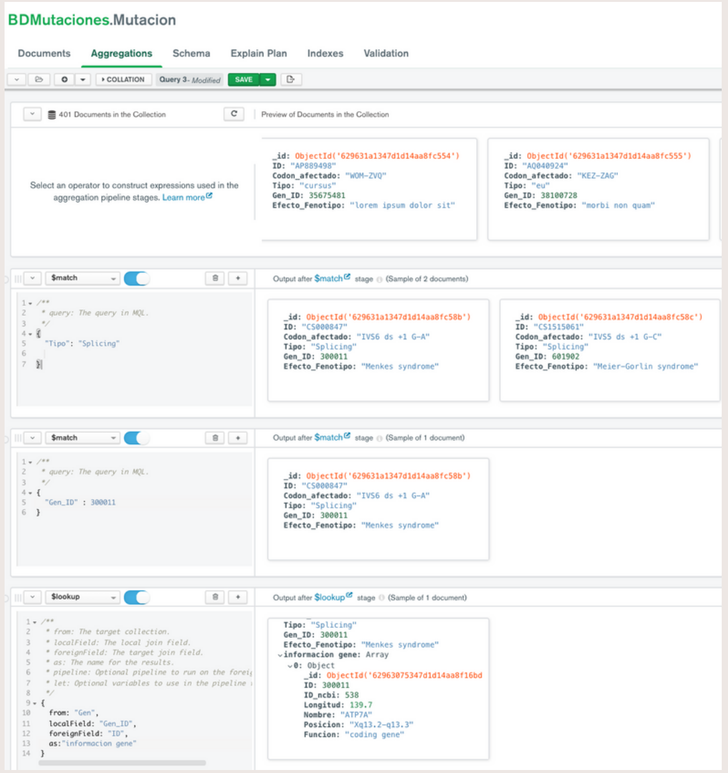
Ahora construiremos un par de queries implmentadas anteriormente .
A diferencia de las queries implemnetadas en otros lenguajes, Mongo db nos obliga a hacerlas en forma de pipeline

Consultas



Query 4: MutacionesPorCiudad

Uso de Lookup para añadir efecto y de match para filtrar por ciudad



Query3

Lookup para insertar información sobre el gen asociado y dos match para filtrar el id del gen y el tipo de mutación

Conclusión

Nerea Martín Serrano
David Cubillos del Toro