MAESTRÍA EN BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

TALLER FINAL

Fecha de entrega Octubre 20

Profesor : Luis Fernando Rivera Serna

Tutor : Andres Felipe Londono

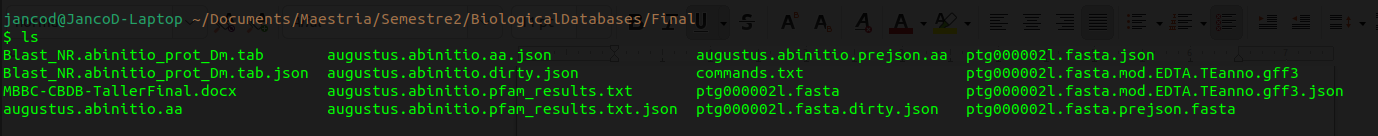
Trabajar los Mismos grupos asignados anteriormente.

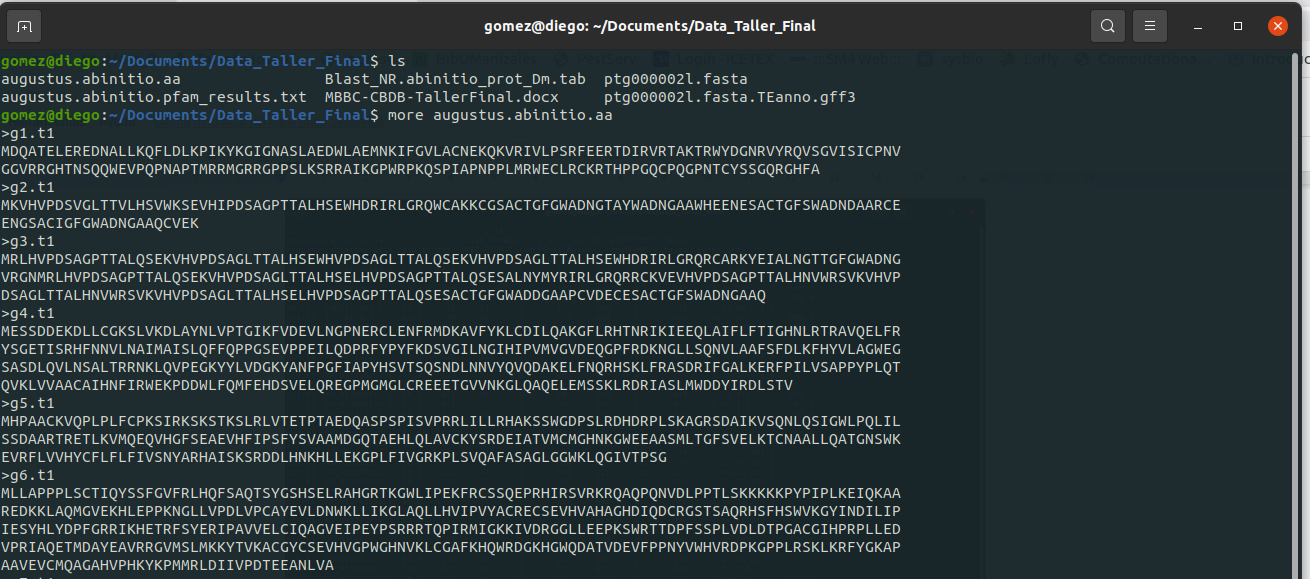
Se cuenta con la secuencia (DNA) de un fragmento de Cromosoma (archivo: ptg000002l.fasta.gz), perteneciente al Reino: *plantae*, División: *Magnoliophyta, Clase: Magnoliopsida con orden, familia, género y especie desconocidos.* Así mismo se cuenta con una gran variedad de información asociada a este fragmento cuyos archivos se listan a continuación : Herramientas

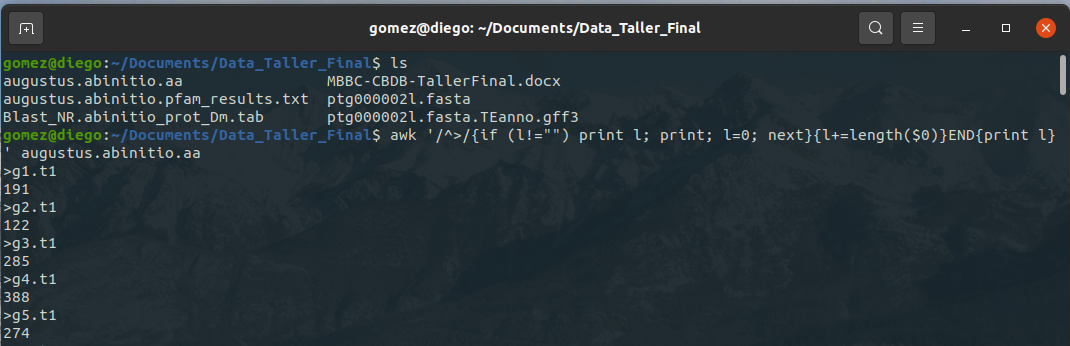
* ptg000002l.fasta.TEanno.gff3.g ## Anotación de TEs. (EDTA)
* augustus.abinitio.aa.gz ## Predicción de proteínas. (Augustus)
* Blast\_NR.abinitio\_prot\_Dm.tab.gz ## NCBI Blast Prot. predichas. (Diamond)
* augustus.abinitio.pfam\_results.txt.gz## Dominios de Proteínas . (pfam\_scan)

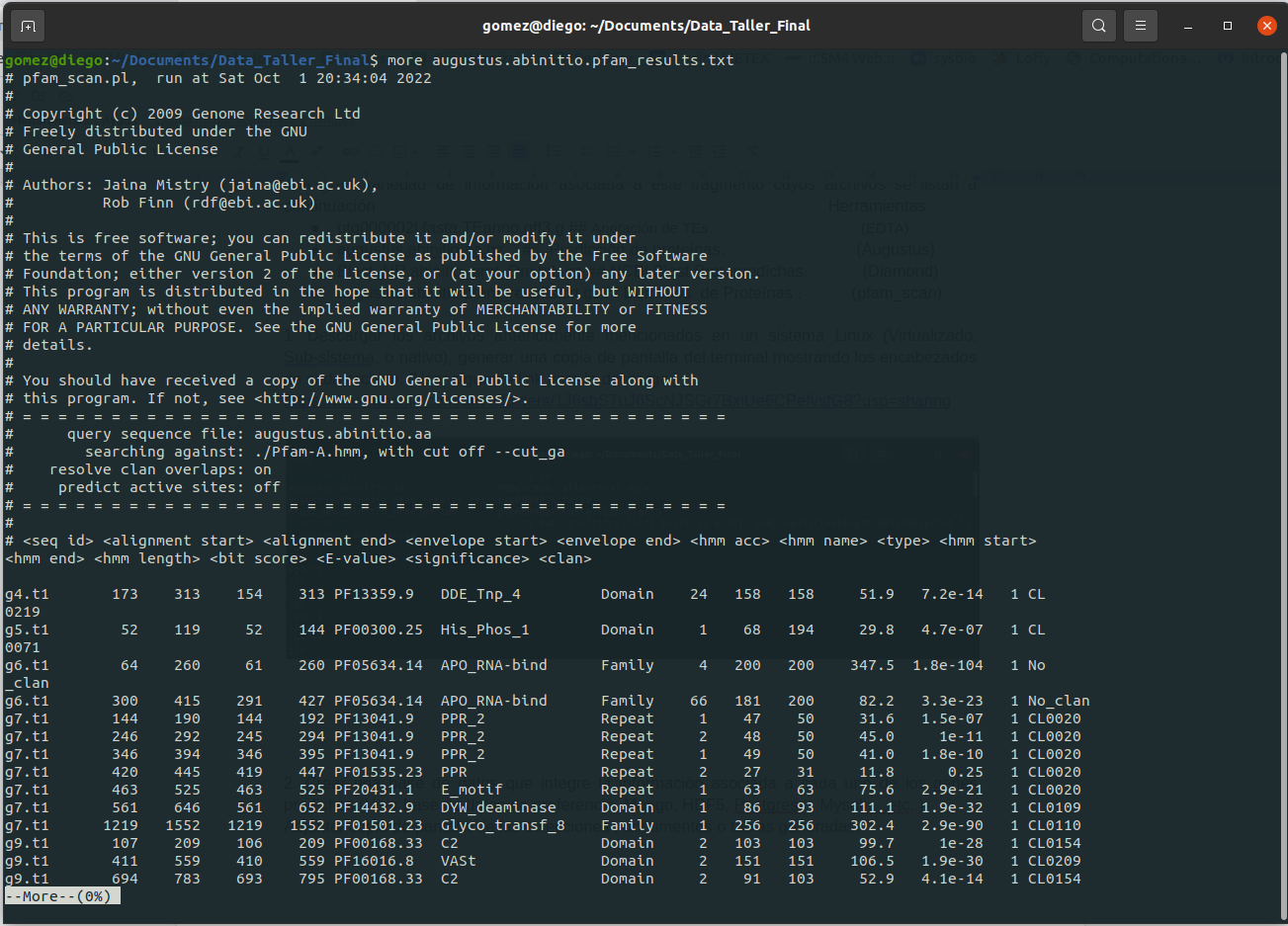
1. Descargar los archivos anteriormente mencionados en un sistema Linux (Virtualizado, Sub-sistema, o nativo), generar una copia de pantalla del terminal mostrando los encabezados de cada archivo. (Usar siguente link para la descarga )

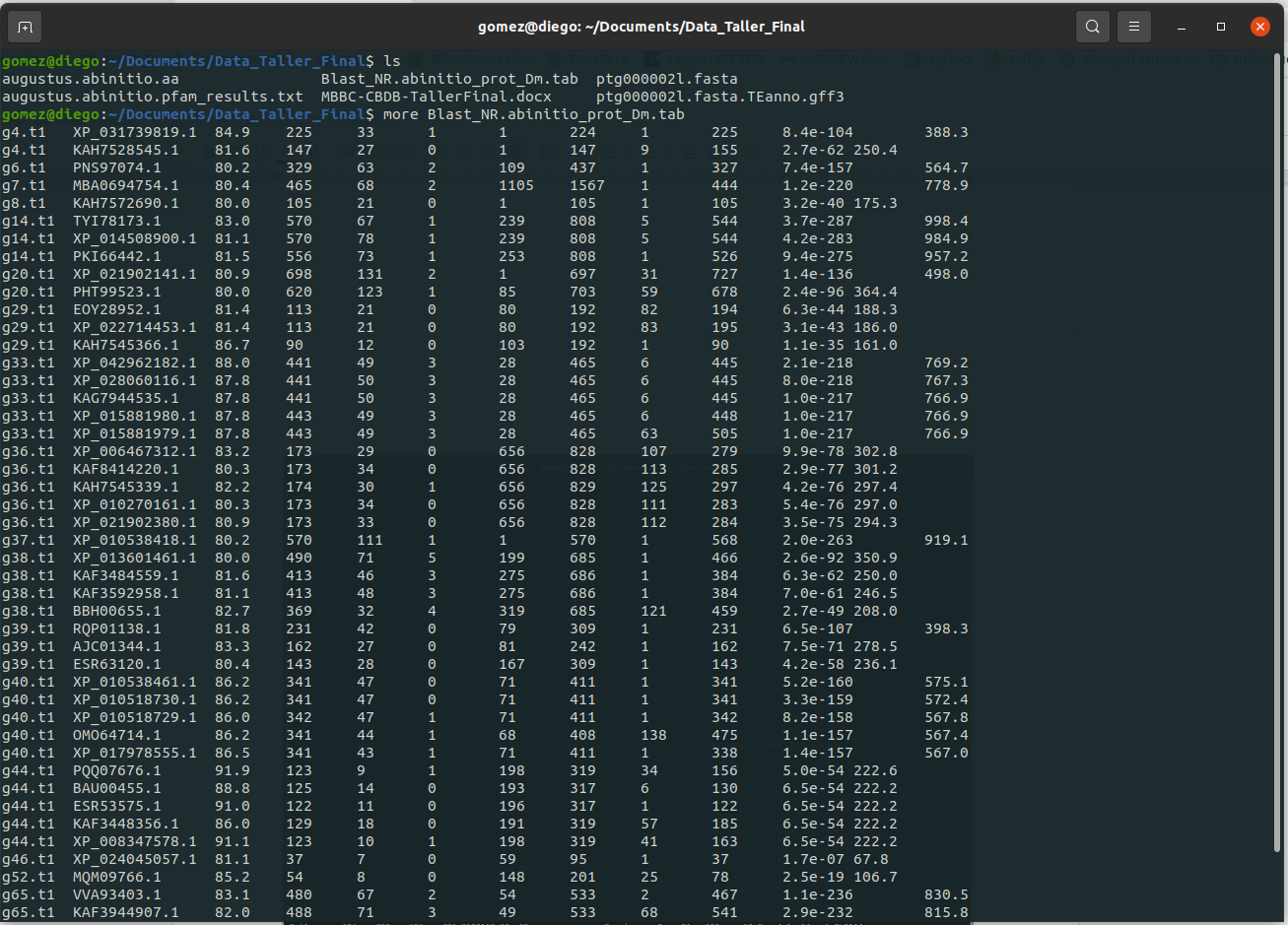
https://drive.google.com/drive/folders/1J6sbSTuJ6ScNJSGr7BxiUe6CPefvsfG8?usp=sharing

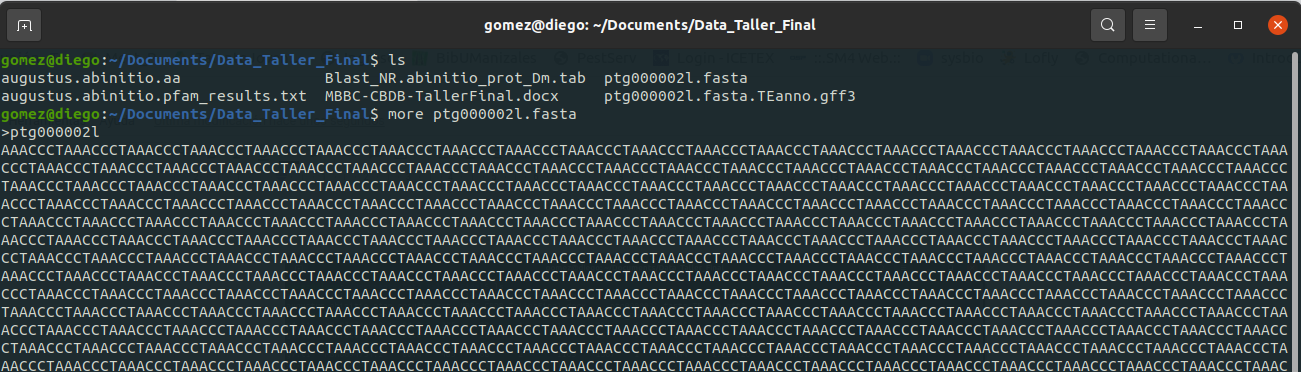


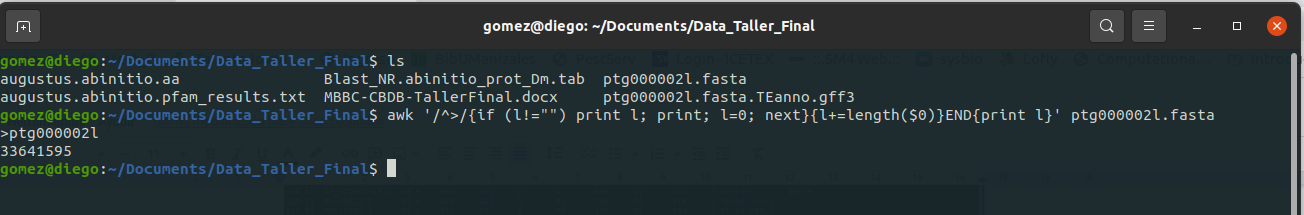


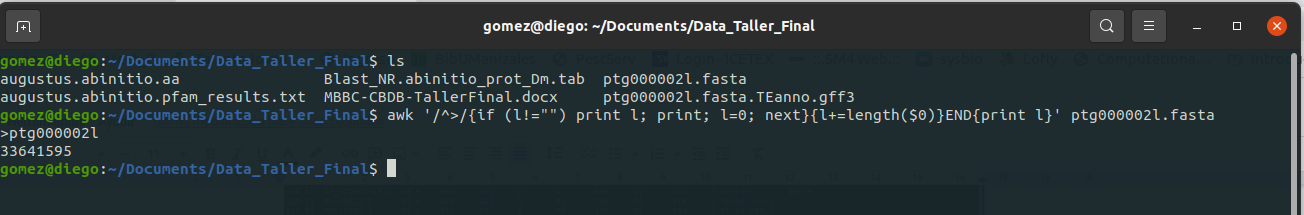






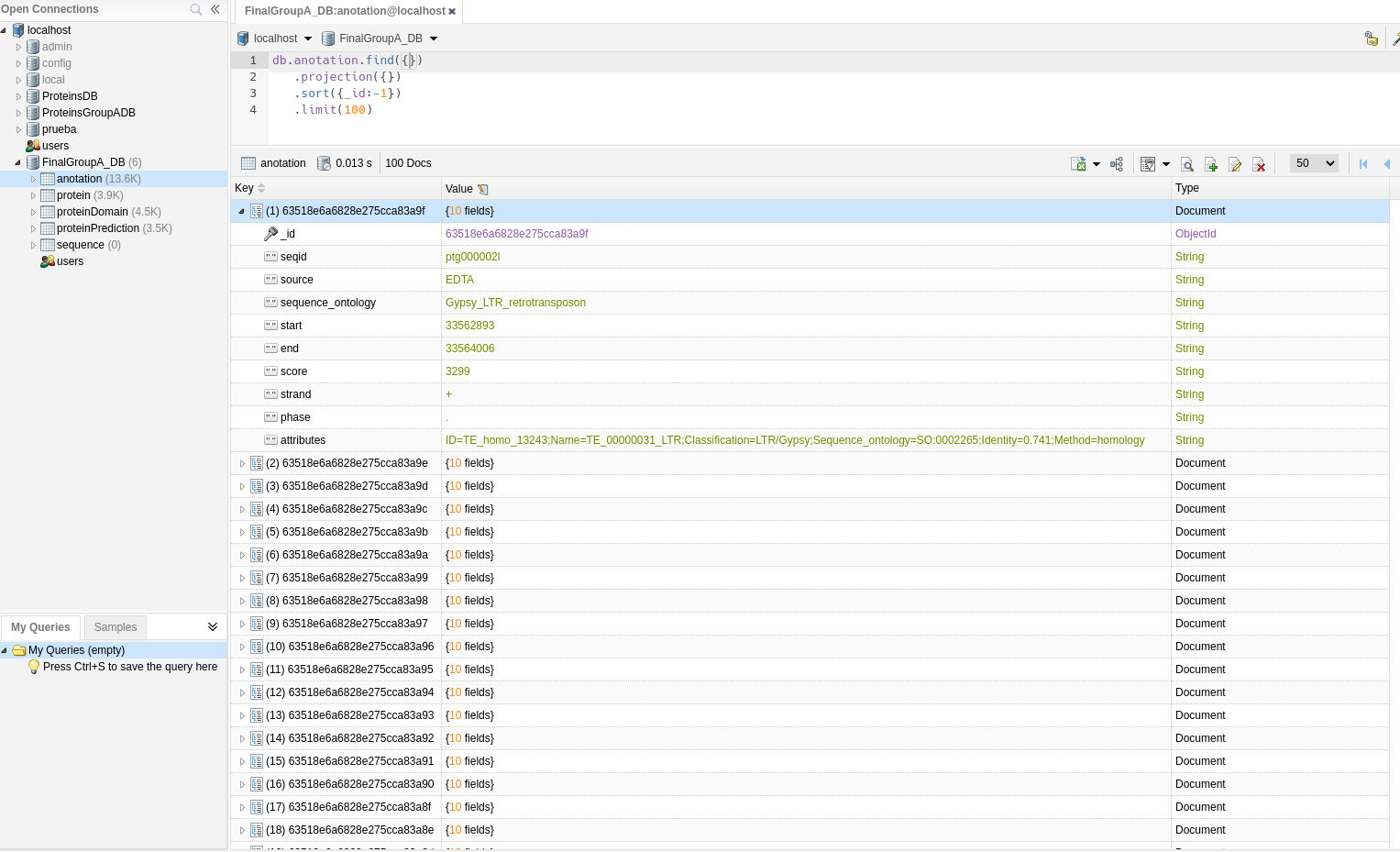


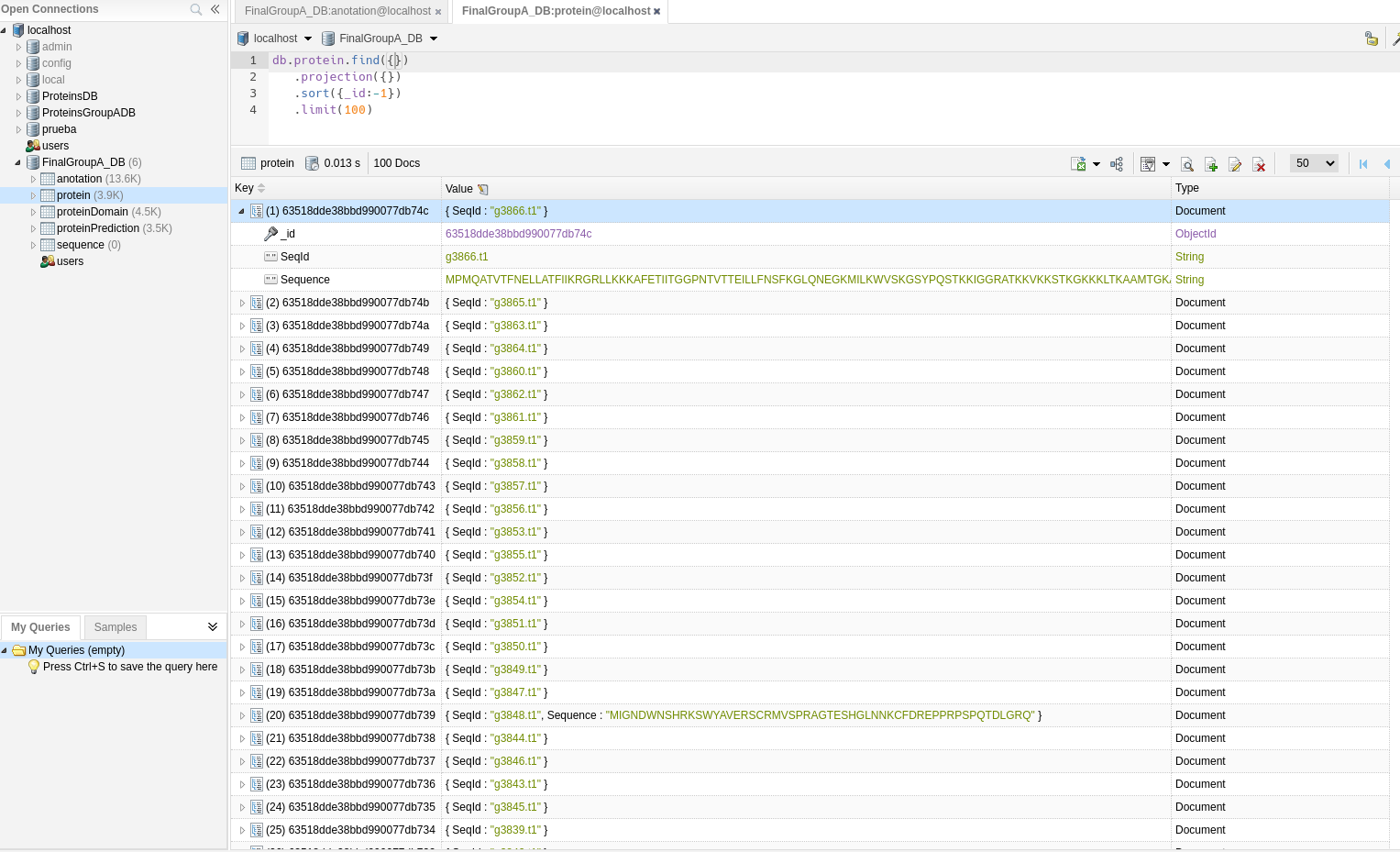


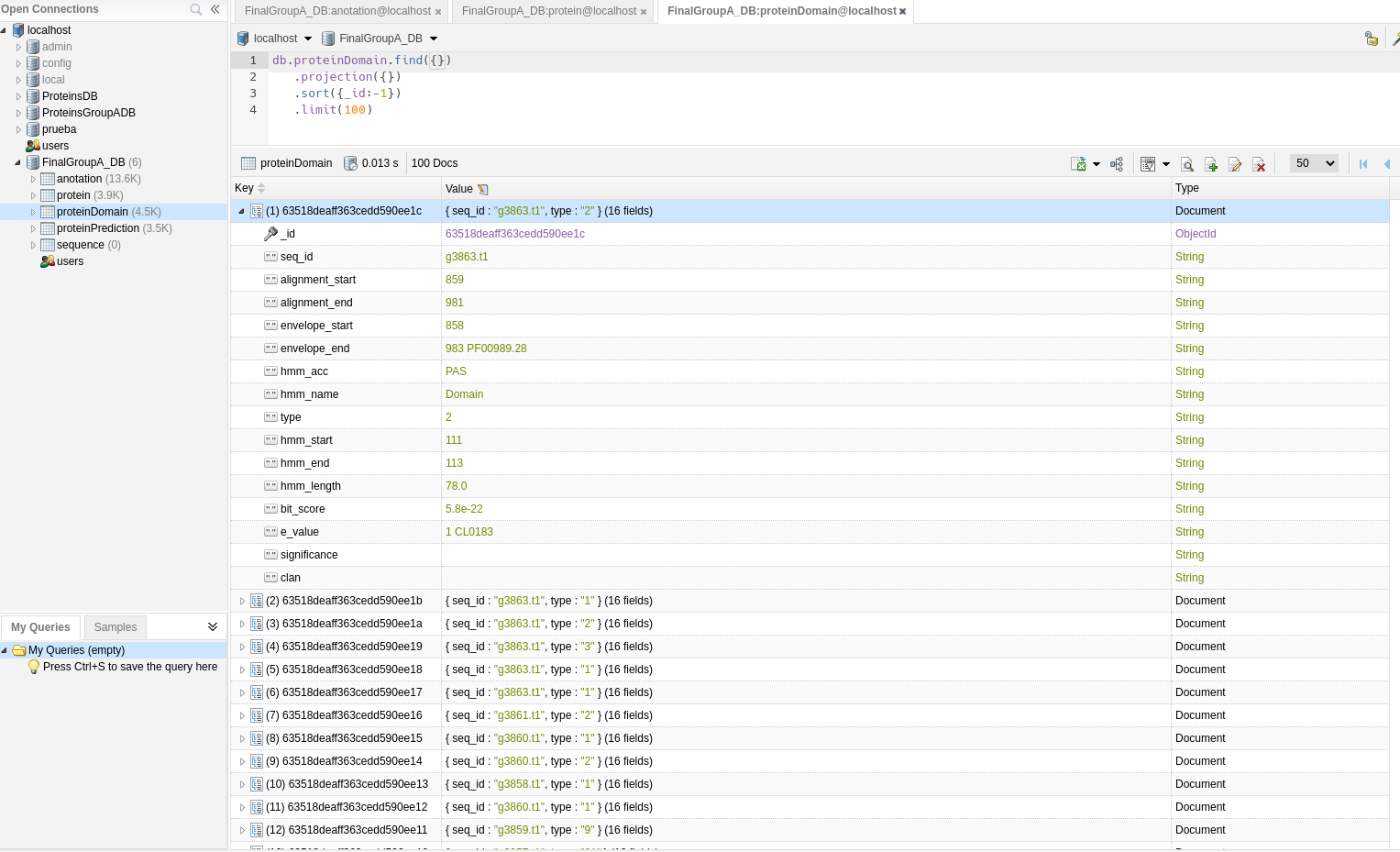


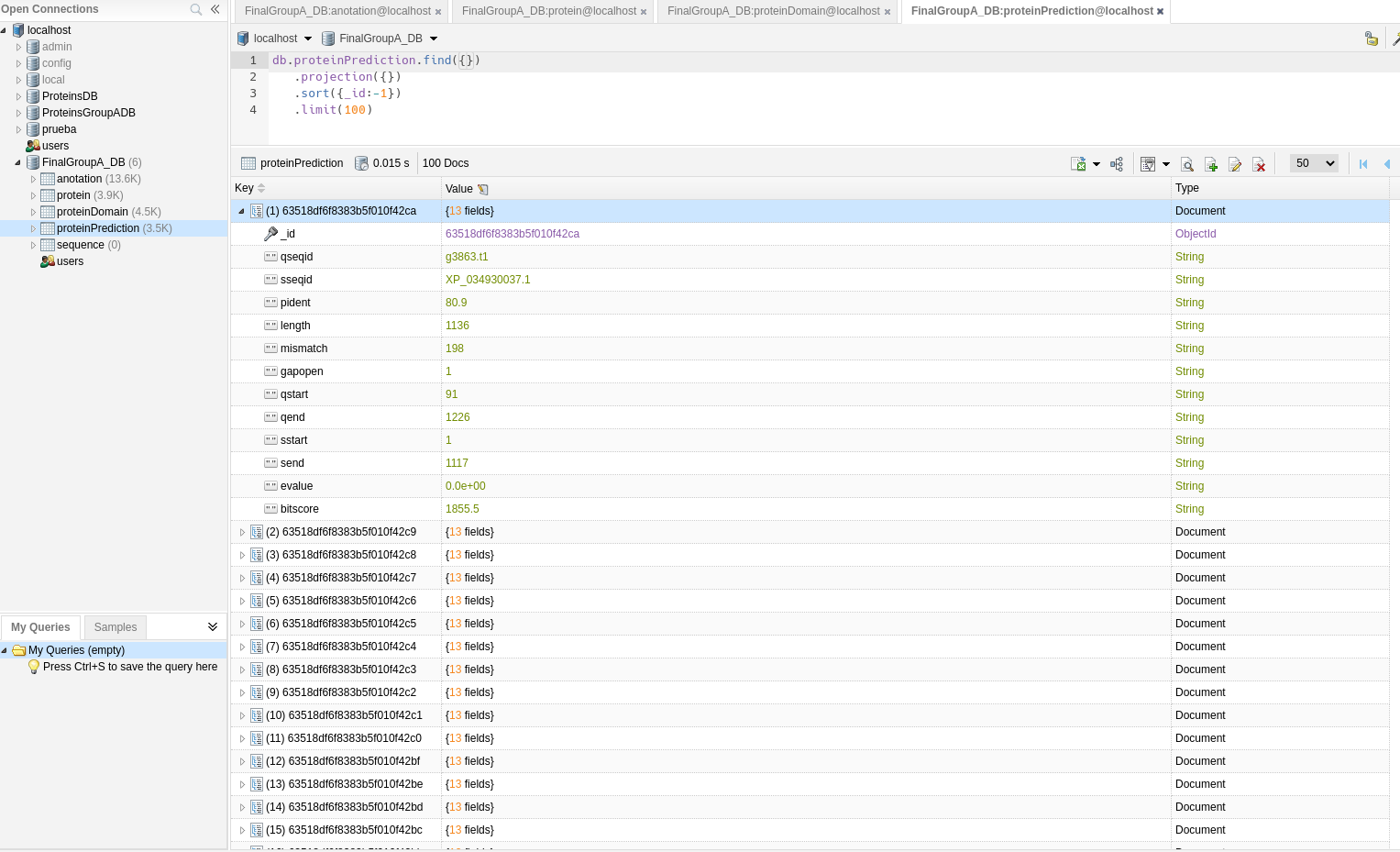
2. Crear una base de datos que integre la información asociada a cada uno de los genes predichos (Usar base de datos de preferencia, Mongo, HDF5, Postgress , Mysql ….etc..)

Adjuntar copia de pantalla de las colecciones , documentos o tablas generadas.









3. Realizar Consulta y copia de pantalla que provea la informacion asociada a la siguente lista de identificadores de gen:

g1026.t1, g1134.t1, g216.t1, g2485.t1, g2565.t1, g2724.t1, g2847.t1, g3154.t1, g3393.t1, g3633.t1

db.proteinPrediction.find({

$or: [{

qseqid: "g1026.t1"

}, {

qseqid: "g1134.t1"

}, {

qseqid: "g216.t1"

}, {

qseqid: "g2485.t1"

}, {

qseqid: "g2565.t1"

}, {

qseqid: "g2724.t1"

}, {

qseqid: "g2847.t1"

}, {

qseqid: "g3154.t1"

}, {

qseqid: "g3393.t1"

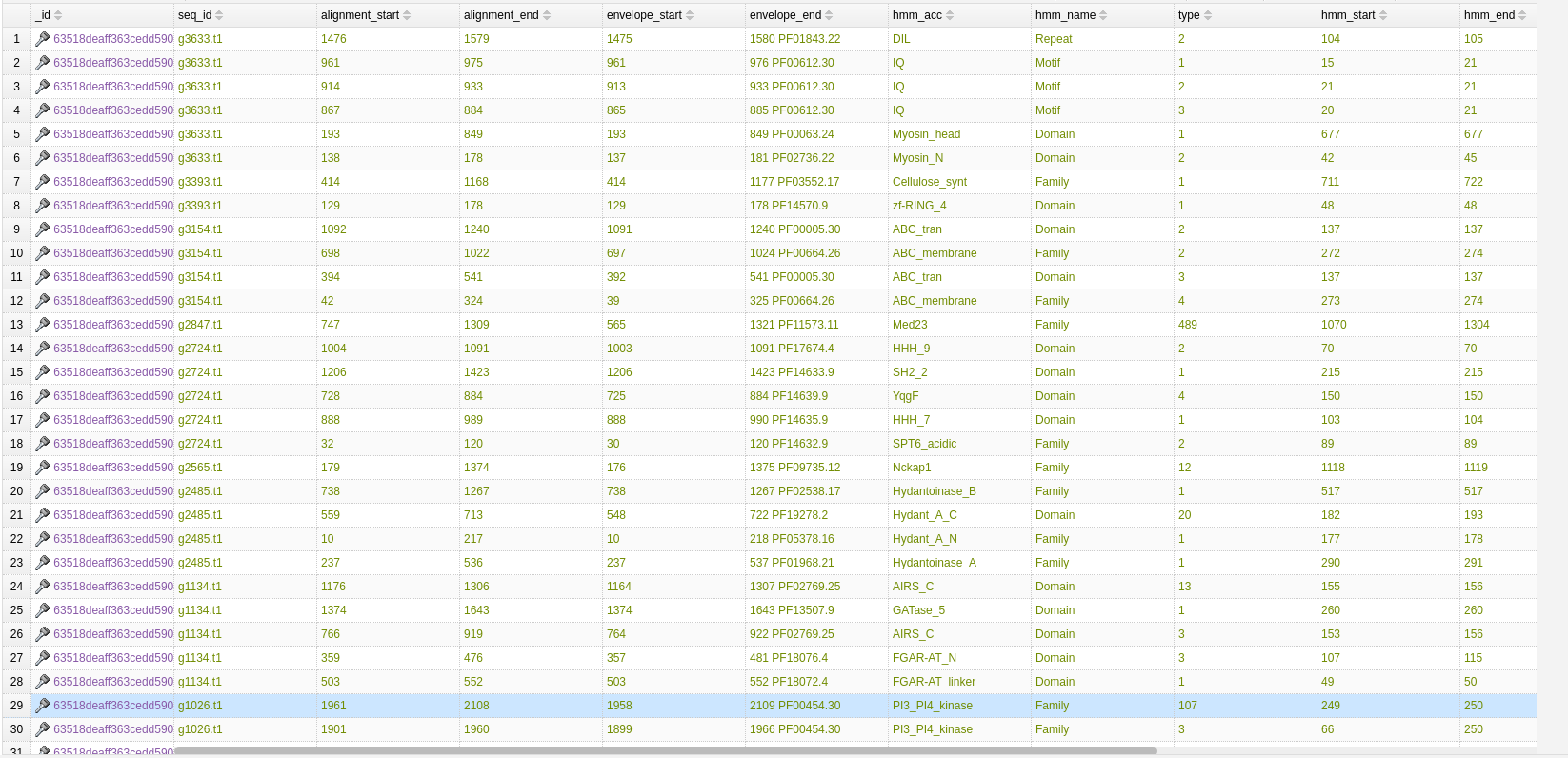
}, {

qseqid: "g3633.t1"

}

]

})



db.proteinDomain.find({

$or: [{

seq\_id: "g1026.t1"

}, {

seq\_id: "g1134.t1"

}, {

seq\_id: "g216.t1"

}, {

seq\_id: "g2485.t1"

}, {

seq\_id: "g2565.t1"

}, {

seq\_id: "g2724.t1"

}, {

seq\_id: "g2847.t1"

}, {

seq\_id: "g3154.t1"

}, {

seq\_id: "g3393.t1"

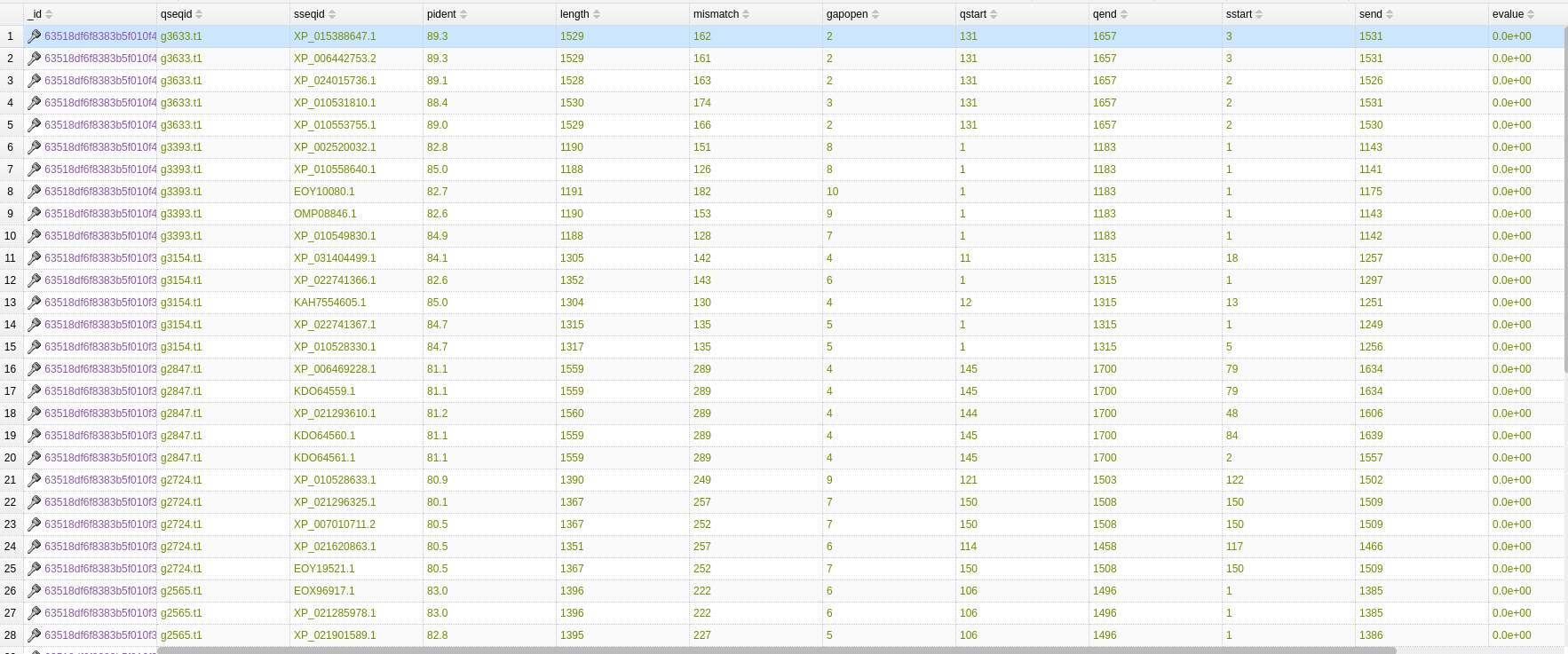
}, {

seq\_id: "g3633.t1"

}

]

})



db.protein.find({

$or: [{

SeqId: "g1026.t1"

}, {

SeqId: "g1134.t1"

}, {

SeqId: "g216.t1"

}, {

SeqId: "g2485.t1"

}, {

SeqId: "g2565.t1"

}, {

SeqId: "g2724.t1"

}, {

SeqId: "g2847.t1"

}, {

SeqId: "g3154.t1"

}, {

SeqId: "g3393.t1"

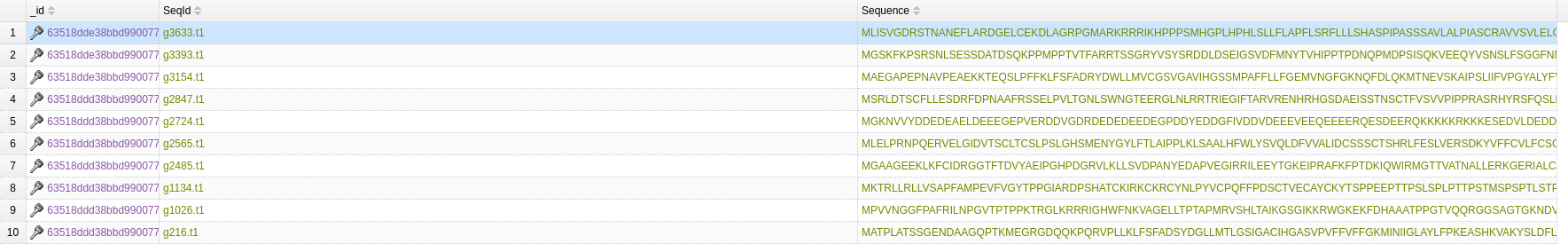
}, {

SeqId: "g3633.t1"

}

]

})

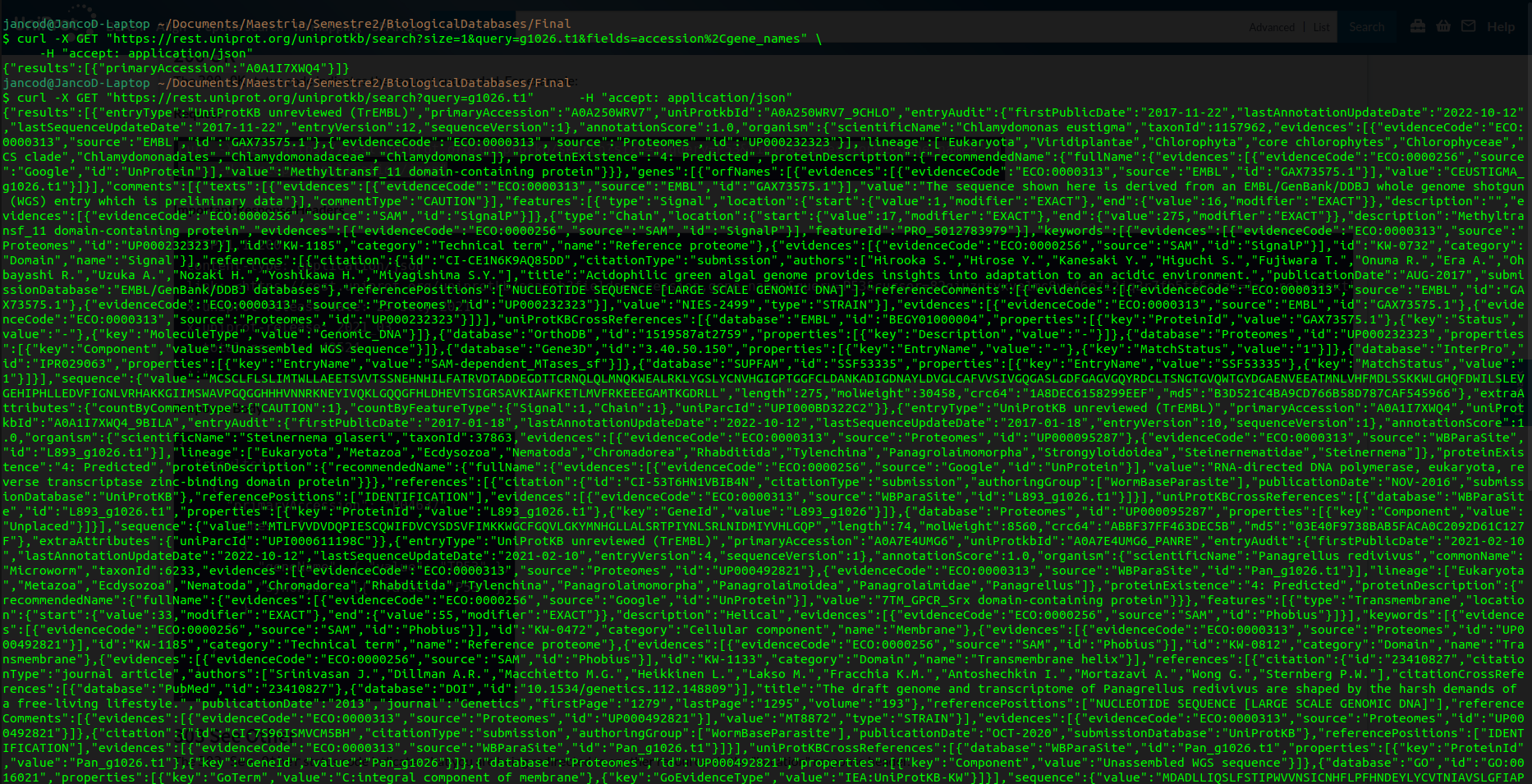


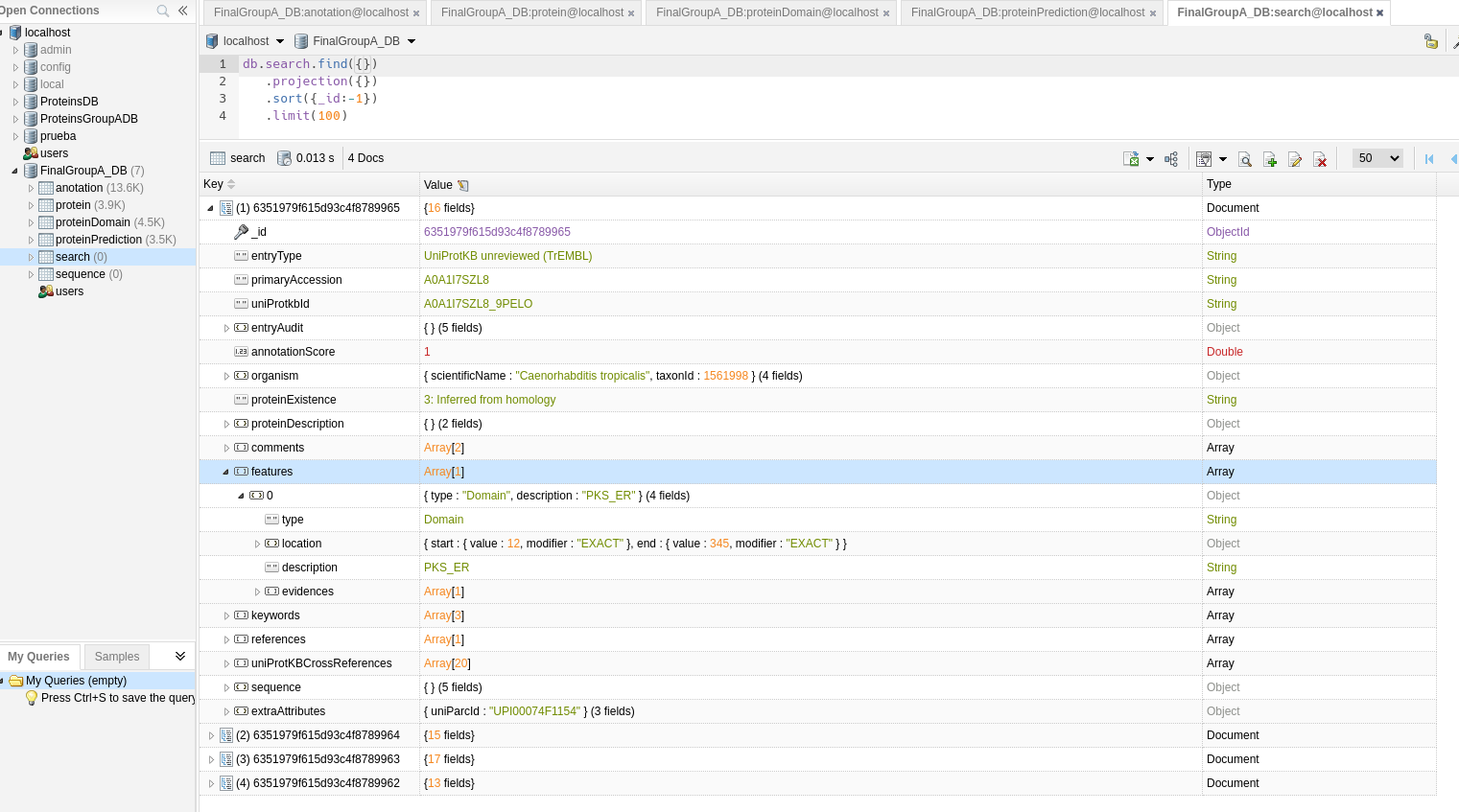
4. Hacer uso de bases de datos Primarias y secundarias para expandir la información asociada a los genes del punto anterios . Explicar estrategia y suministrar resultados.

Programaticamente consultando los endpoints disponibles en los servicios de bases de datos biológicas se puede consultar información relacionada para complementar y relacionar el gen con otros organismos, a continuacion se consulta la información de un gen y se almacena los resultados de búsqueda en Mongo

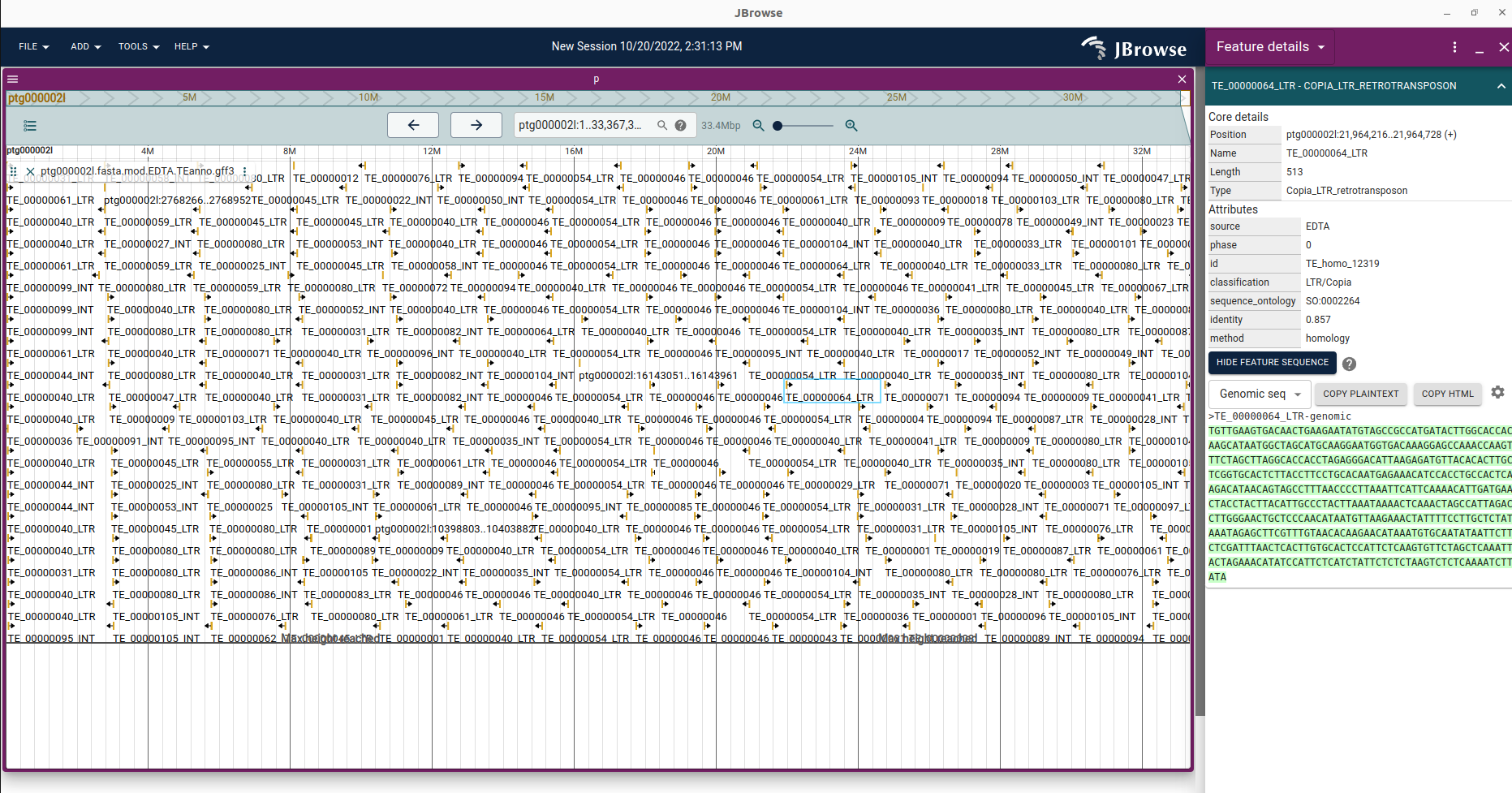
curl -X GET "https://rest.uniprot.org/uniprotkb/search?size=1&query=g1026.t1&fields=accession%2Cgene\_names" \

-H "accept: application/json"





5. (Opcional puntos adicionales), Realizar la instalación de la herramienta para visualizar genomas de preferencia (JBrowse, Gbrowse , etc..) , realizar la carga de la secuencia de referencia y las anotaciones de elementos repetitivos. Adjuntar copia de pantalla mostrando una región con TEs asociados.



6. Mediante consultas en bases de datos especializadas,tratar de asignar la taxonomía faltante (*orden, familia, género y especie*) a el fragmento de secuencia en cuestión (ptg000002l.fasta).

**Nota :** Trabajar por favor los mismos grupos asignados anteriormente.

*Theobroma cacao* L.

Reino Plantae

Subreino Viridiplantae

Infrareino Streptophyta

Superdivisión Embryophyta

División Tracheophyta

Subdivisión Spermatophytina

Clase Magnoliopsida

Superorden Rosanae

Orden Malvales

Familia Malvaceae

Género Theobroma

Especie *Theobroma cacao* L.

Se infiere que la secuencia pertenece a esta especie puesto que la predicción de genes se realiza con AUGUSTUS y en su conjunto de entrenamiento de solo *T. cacao* y *Arabidopsis thaliana* coinciden con la información previamente suministrada “Reino: *plantae*, División: *Magnoliophyta, Clase: Magnoliopsida*” (teniendo en cuenta que la organización jerárquica-taxonómica depende del sistema utilizado, aunque en general se mantienen algunos clados). Por otra parte, se realizaron búsqueda mediante Blastp con varias secuencias predichas mediante AUGUSTUS y aunque las coberturas rondaban el 85%, los demás parámetros son bastante aceptables y el E-value se mantenía en cero (0), *T. cacao* se mantenía como el único organismo frecuente entre las búsquedas mientras que *A. thaliana* no era frecuente o simplemente no aparecía en esta.