BINF6211/8211 Design and Implementation of Biological Databases Spring 2021 Instructor: Weller

Problem Sets

Relational Algebra/SQL Practice problems – Set 1 (36 points)

The problem sheet shows 3 very simple tables and a set of 10 results for which you are asked to provide the question that the result output would answer and then formulate the query in the relational algebra notation and in sql. For one question you are asked to suggest 2 approaches and for a second question you are asked to provide 3 approaches to the RA and SQL. Thus, there are 36 points available.

Given these 3 tables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene |  |  | Transcript | |  | Disease |  |  |
| G-ID | g\_symbol |  | G-ID | T-ID |  | T-ID | D-name | |
| 1234 | ALB |  | 1234 | Q4540 |  | Q4540 | Ollier Syndrome | |
| 4000 | LTL1a |  | 1234 | M9090 |  | M9090 | Alzheimers disease | |
| 4040 | APOE4 |  | 1234 | R3280 |  | R3280 | Ollier Syndrome | |
|  |  |  | 4000 | Q1256 |  | Q1256 | Cardiomyopathy | |
|  |  |  | 4000 | Q9090 |  | Q9090 | Hypercholesterolemia | |

For each result, provide these 3 items:

1. What is the question being asked?
2. What relational algebra operations were used to obtain the following results?
3. Write the equivalent SQL also.
4. G\_symbol

ALB

LTL1a

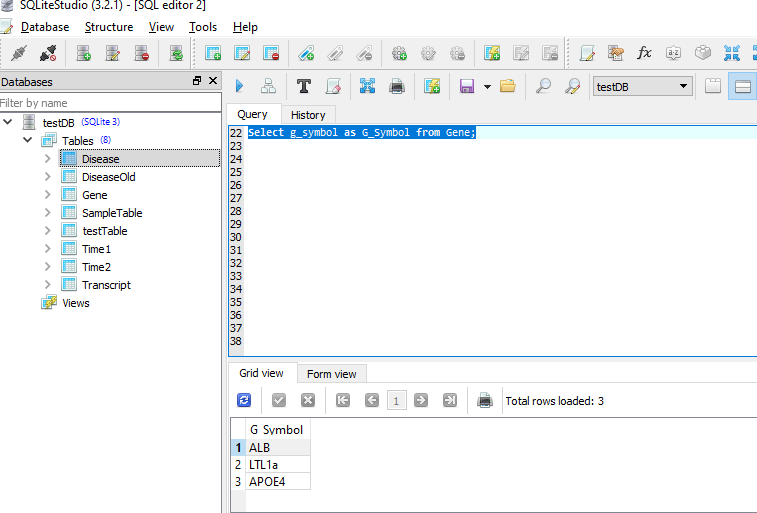
APOE4

**Question:** Get the Gene Symbols from GENE table

Using Π Projection Output columns from all rows of the input relation (remove duplicate tuples)

Πg\_symbol(Gene)

**Sql:** Select g\_symbol as G\_Symbol from Gene



Note: If column has duplicate value then select distinct is needed in the query

1. D-name

Ollier Syndrome

Alzheimers Disease

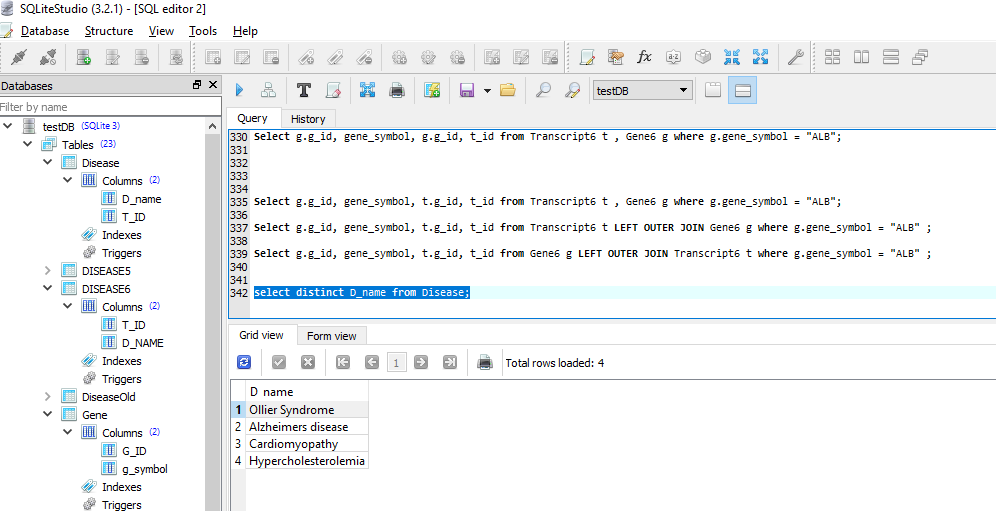
Cardiomyopathy

Hypercholesterolemia

**Question:** Get all disease names from DISEASE table

ΠD-name(Disease)

**Sql:** select distinct D\_name from Disease;



1. T-ID | D\_name .

R3280 | Ollier Syndrome

Q4540 | Ollier Syndrome

(There are 2 ways to do this – show both)

**Question:** Get the Transcript id and disease name for the disease Ollier Syndrome

**way 1:**

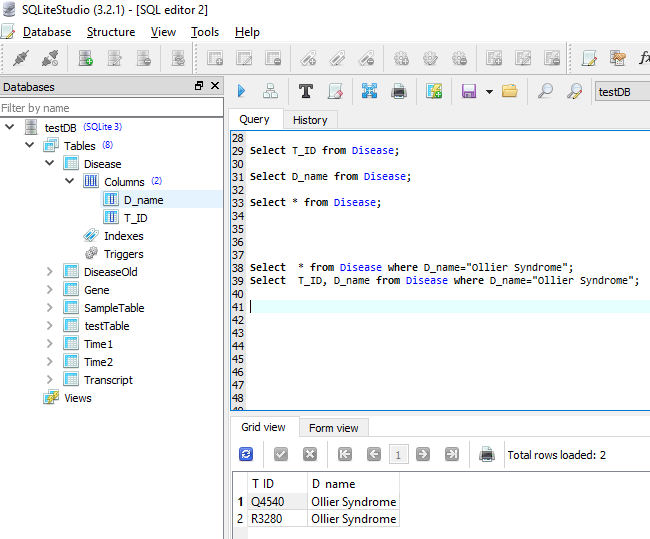
using σ Selection Return rows of the input relation that are TRUE (satisfy the predicate)

Retrieve from Disease table with matching disease name

σD-name=”Ollier Syndrome”(Disease)

**Sql:** Select \* from Disease where D\_name="Ollier Syndrome";

Select T\_ID, D\_name from Disease where D\_name="Ollier Syndrome";



**way 2:**

By Joining tables

ρSELTRANSC 🡨 (Gene  GENE.G-ID = TRANSCRIPT.G-IDTranscript)

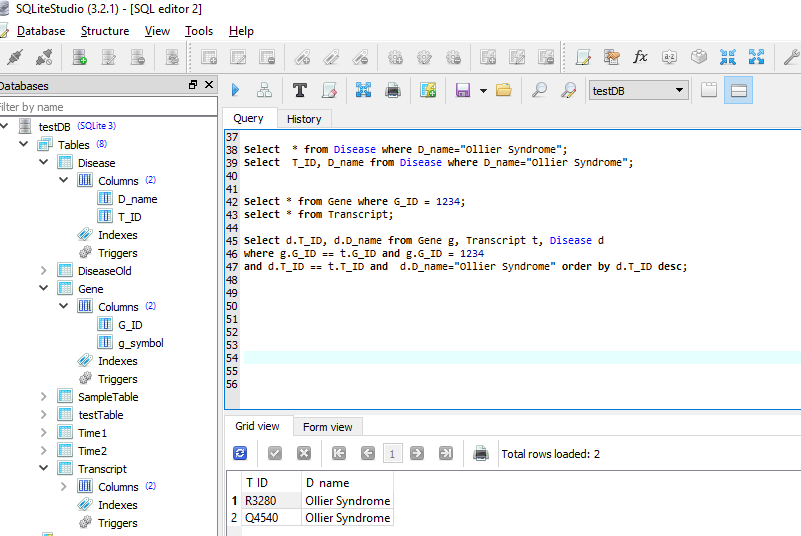
ρSELDISES 🡨 (SELTRANSC  SELTRANSC.T-ID = Disease.T-IDDisease)

Result 🡨 ΠT-ID,D-name (σ(G-ID =”1234”) and (D-name=”Ollier Syndrome”) ) (SELDISES)

**Sql:** Select d.T\_ID, d.D\_name from Gene g, Transcript t, Disease d

where g.G\_ID == t.G\_ID and g.G\_ID = 1234

and d.T\_ID == t.T\_ID and d.D\_name="Ollier Syndrome" order by d.T\_ID desc;



1. G-ID | g-symbol
   1. | ALB

4000 | LTL1a

(there are 3 ways to get this table, show all)

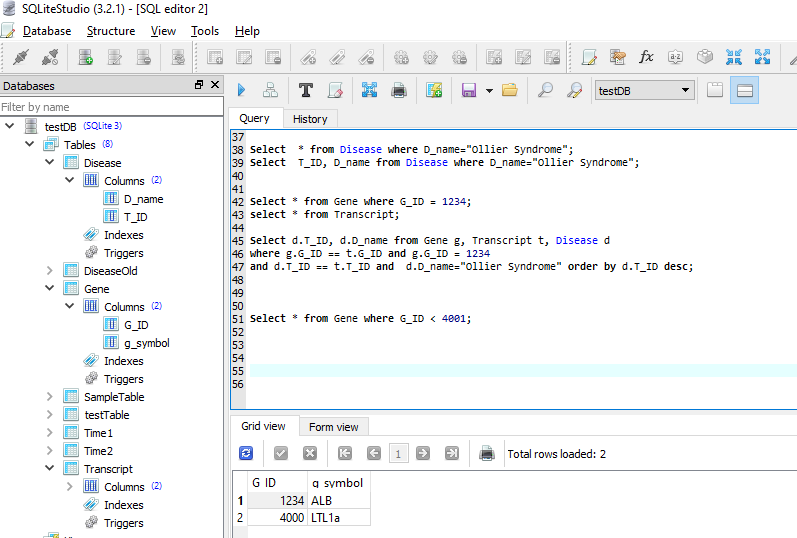
**Way 1:**

**Question:** Get the Gene id and gene symbol with Gen Id is < 4001

σG-ID<4001(Gene)

Getting directly from Gene table

**Sql:** Select \* from Gene where G\_ID < 4001;



**Way 2:**

**Question:** Get the distinct/unique Gene id and gene symbol of impacted Genes with any of the diseases.

By joining tables

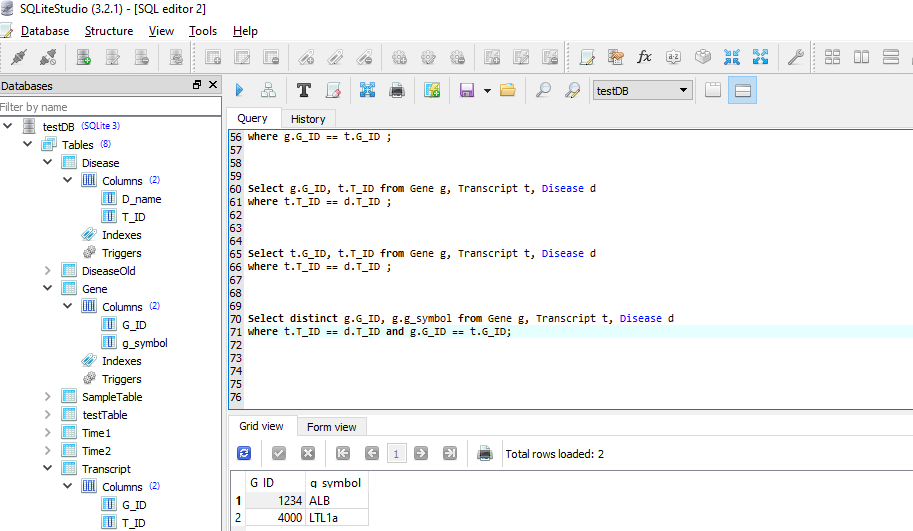
ρSELTRANSC 🡨 (Transcript  TRANSCRIPT.T-ID = DISEASE.T-IDDisease)

ρSELGENE 🡨 (SELTRANSC  SELTRANSC.G-ID = GENE.G-IDGENE)

Result 🡨 ΠG-ID,G-SYMBOL(SELGENE)

**Sql:** Select distinct g.G\_ID, g.g\_symbol from Gene g, Transcript t, Disease d

where t.T\_ID == d.T\_ID and g.G\_ID == t.G\_ID;



**Way 3:**

**Question:** Get the Gene id and gene symbol using Intersect functionality

EXCEPT operator can’t be applied for both gene id and symbol because of column constraint so I am using sub query

**Sql:** SELECT g.G\_ID, g.g\_symbol

FROM Gene g where g.G\_ID in(

SELECT g.G\_ID

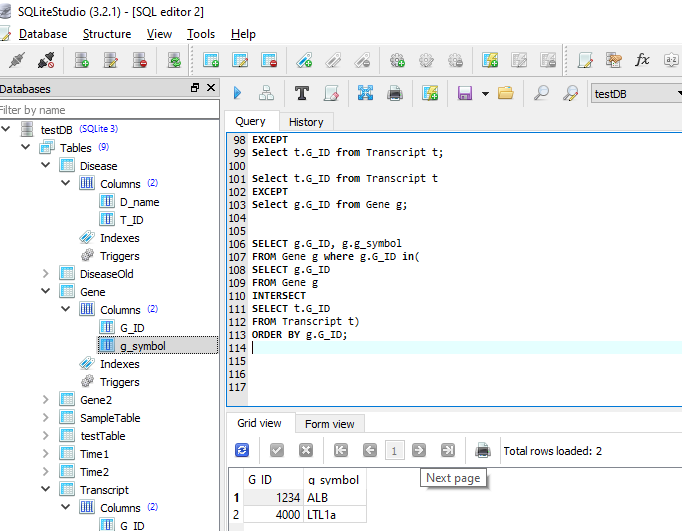
FROM Gene g

INTERSECT

SELECT t.G\_ID

FROM Transcript t)

ORDER BY g.G\_ID;



1. G-ID | g\_symbol | G-ID | T-ID .
   1. |ALB | 1234 |Q4540

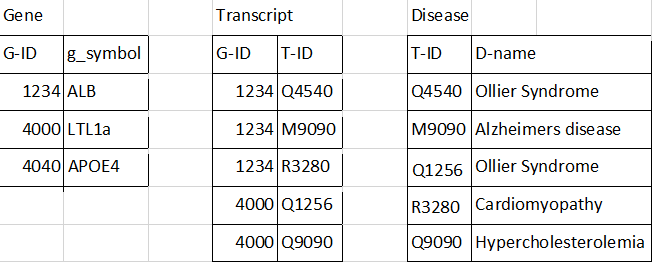
1234 |ALB |1234 |M9090

1234 |ALB |1234 |R3280

1234 |ALB |1234 |Q1256

1234 |ALB |1234 |Q9090

It’s not possible directly as 1234 is not associated with Q1256 and Q9090 in transcript table.



g-id can be duplicated while projecting the result.

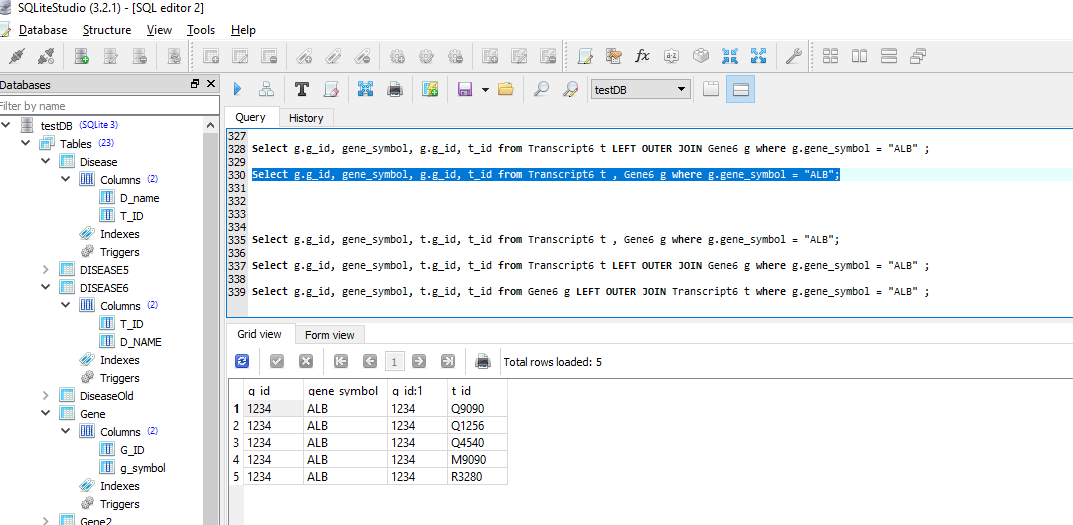
**Question**: By joining tables GENE and TRANSCRIPT table and Gen Id, gene symbol, transcript id for ALB Gene

ρSELGENE 🡨 (Transcript TRANSCRIPT.G-ID = GENE.G-IDGENE)

Result 🡨 Π SELGENE.G-ID, SELGENE.g-symbol, SELGENE.G-ID, SELGENE.G-ID, SELGENE.T-ID

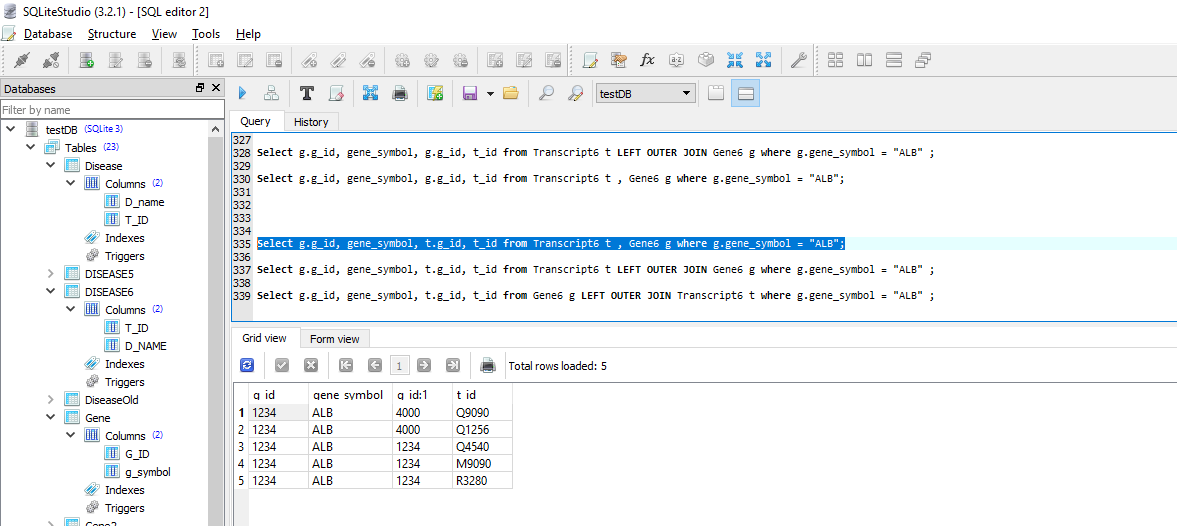
(σ SELGENE.G\_Symbol=”ALB”)(SELGENE)

**Sql :** Select g.g\_id, g\_symbol, g.g\_id, t\_id from Transcript t , Gene g where g.g\_symbol = "ALB";



If I try to display g-ID of both tables then 4000 also gets displayed

Select g.g\_id, g\_symbol, t.g\_id, t\_id from Transcript t , Gene g where g.g\_symbol = "ALB";



1. G-ID | g\_symbol | G-ID | T-ID .
   1. |ALB | 1234 |Q4540

1234 |ALB |1234 |M9090

1234 |ALB |1234 |R3280

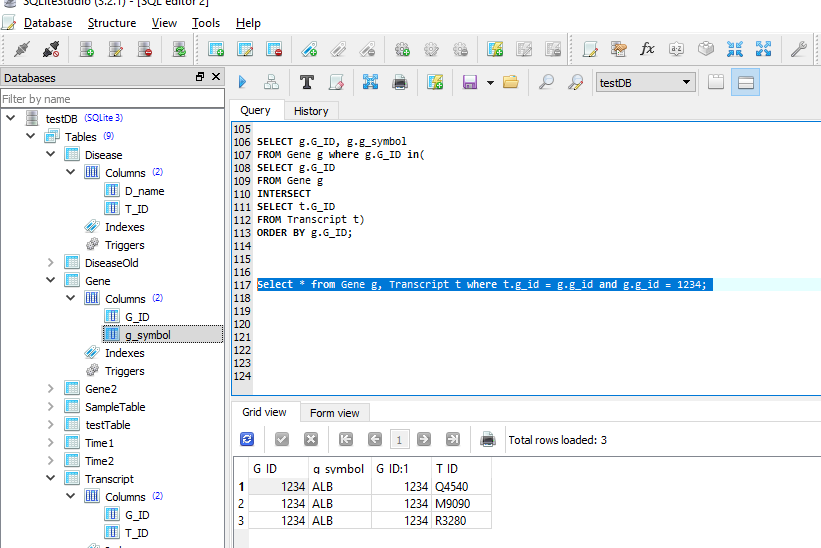
**Question**: By joining tables GENE and TRANSCRIPT table and get all records with Gen Id 1234

By joining tables then select

ρSELGENE 🡨 (Transcript  TRANSCRIPT.G-ID = GENE.G-IDGENE)

Result 🡨 σ SELGENE.G-ID=1234(SELGENE)

**Sql:** Select \* from Gene g, Transcript t where t.g\_id = g.g\_id and g.g\_id = 1234;



1. G-ID | g\_symbol | T-ID .
   1. | ALB |Q4540

1234 |ALB |M9090

1234 |ALB |R3280

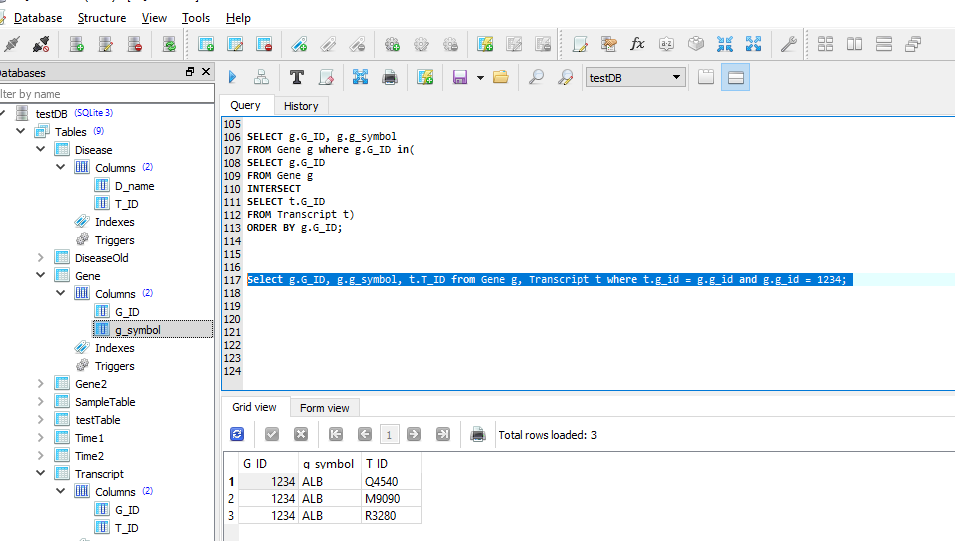
**Question:** Get the Gene id, gene symbol, transcript id for gene id 1234 by joining tables

Doing joining tables then select and project

ρSELGENE 🡨 (Transcript  TRANSCRIPT.G-ID = GENE.G-IDGENE)

Result 🡨 Π SELGENE.G-ID, SELGENE.g-symbol(σ SELGENE.G-ID=1234) (SELGENE)

**Sql:** Select g.G\_ID, g.g\_symbol, t.T\_ID from Gene g, Transcript t where t.g\_id = g.g\_id and g.g\_id = 1234;



1. G-ID | T-ID .

1234 |Q4540

* 1. |M9090

1234 |R3280

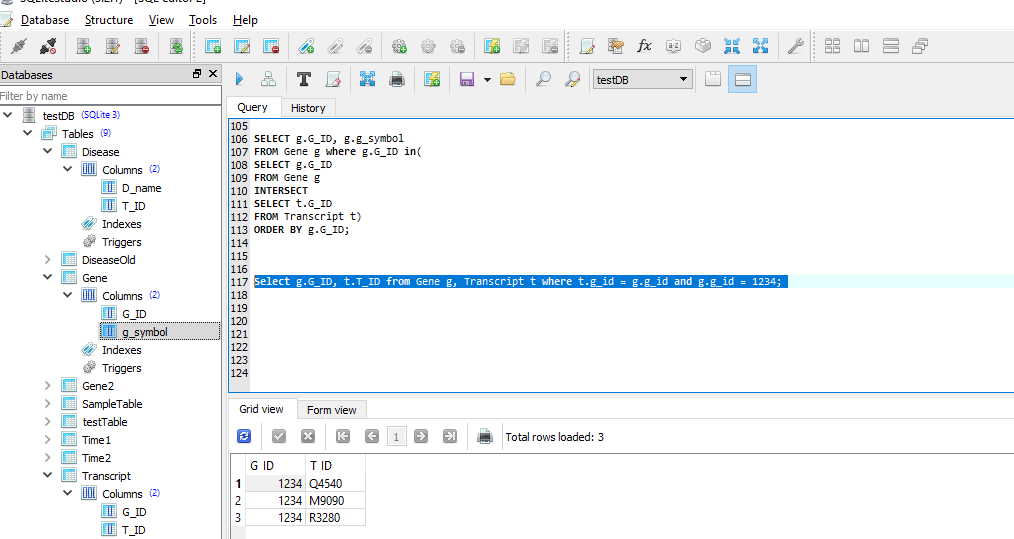
**Question:** Get the Gene id, transcript id for gene id 1234 by joining tables

Doing joining tables then select and project

ρSELGENE 🡨 (Transcript  TRANSCRIPT.G-ID = GENE.G-IDGENE)

Result 🡨 Π SELGENE.G-ID, SELGENE.T-ID(σ SELGENE.G-ID=1234) (SELGENE)

**Sql:** Select g.G\_ID, t.T\_ID from Gene g, Transcript t where t.g\_id = g.g\_id and g.g\_id = 1234;



1. G-ID | g\_symbol | T-ID | D-name .

4000 |LTL1a |Q1256 |Cardiomyopathy

4000 |LTL1a |Q9090 |Hypercholesterolemia

**Question:** Get the Gene id, gene symbol, transcript id, disease name for gene id 4000 by joining tables

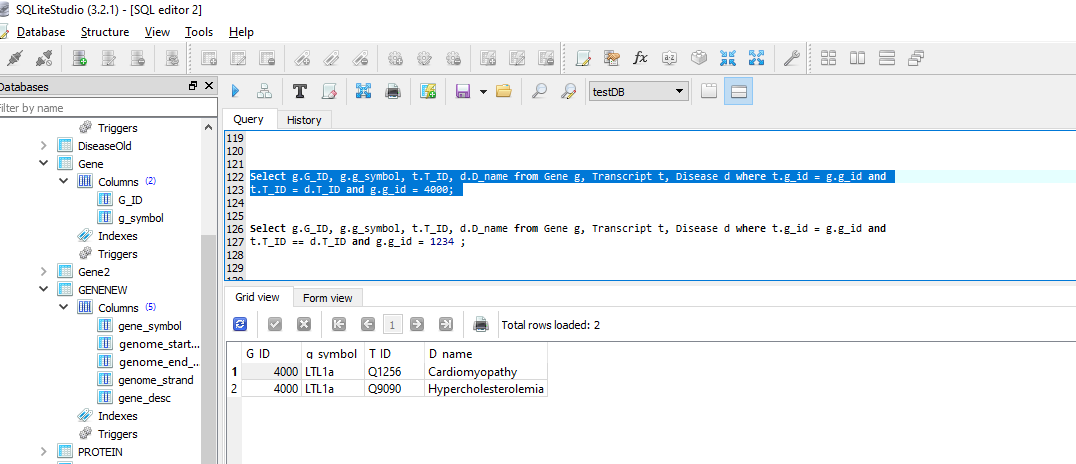
ρSELGENE 🡨 (Transcript  TRANSCRIPT.G-ID = GENE.G-IDGENE)

ρTRANS 🡨 (SELGENE  SELGENE.T-ID = DISEASE.T-IDDisease)

Result 🡨 ΠTRANS.G-ID, TRANS.g-symbol, TRANS.T-ID, TRANS.D-name(σTRANS.G-ID=4000) (TRANS)

**Sql:** Select g.G\_ID, g.g\_symbol, t.T\_ID, d.D\_name from Gene g, Transcript t, Disease d where t.g\_id = g.g\_id and

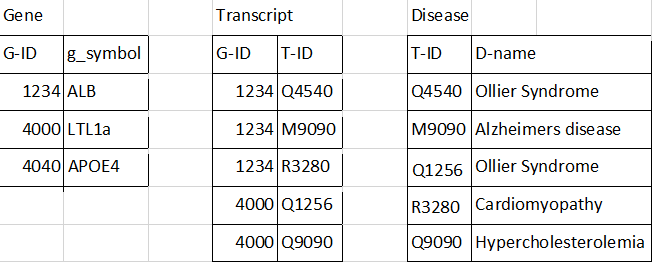
t.T\_ID = d.T\_ID and g.g\_id = 4000;



1. g-symbol | D-name .

ALB |Ollier syndrome

LTL1a | Ollier syndrome



**Question:** by joining tables , Get the gene symbol, disease name for Ollier Syndrome

ρSELGENE 🡨 (Transcript  TRANSCRIPT.G-ID = GENE.G-IDGENE)

ρTRANS 🡨 (SELGENE  SELGENE.T-ID = DISEASE.T-IDDisease)

Result 🡨 ΠTRANS.g-symbol, TRANS.D-name(σTRANS. D-name=”Ollier Syndrome” ) (TRANS)

**Sql:** Select g.g\_symbol, d.D\_name from Gene g, Transcript t, Disease d where t.g\_id = g.g\_id and

t.T\_ID = d.T\_ID and d.D\_name = 'Ollier Syndrome' order by g.g\_symbol

