Progettazione della Base di Dati

Università La Sapienza, Informatica

A.A 2022/2023

Relazione Associata al Tirocinio: "Identificazione di Biomarcatori dai Dati di Espressione Genica"

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica
Dipartimento di Informatica
Corso di laurea in Informatica

Valerio Mesiti - 1936543

Responsabile – Maurizio Mancini

Sommario

PROGETTAZIONE CONCETTUALE

Richiesta dei Requisiti

Si vuole realizzare una base di dati per analizzare la mole di dati derivanti dai maggiori siti di biomedicina riguardanti la “gene expression” e la “protein expression”.

Tutto il database è diviso in progetti (project) che hanno un codice univoco ed un nome, i progetti contengono tutti le analisi (Analysis) sotto forma di file e i casi (case) che ne fanno parte

Di un caso sappiamo l’id, il tipo di malattia, il primary_site. Da questo paziente (che può avere un’anagrafica) scaturiscono più Campioni Biologici (Biospecimen). Questi prelievi primari, chiamati sample, possono essere trattati e resi o slide o porzioni (Portion), quest’ultime possono diventare analiti (Analyte) che infine con determinati trattamenti diventano aliquot (Aliquot)

Per i campioni (sample) identificati da un id e da un uuid, vogliamo sapere il tipo del campione e di che tumore si tratta. Per i derivati di ciascuno avremo i loro identificatori e per gli analiti e le aliquot anche la concentrazione nel campione.

Rappresentiamo quindi l’analisi sotto forma di file, dove è contenuto un determinato insieme di expression presi da uno o più biospecie e per ognuno di essi da chi proviene. Le analisi hanno un codice, un nome, un submitter_id data_category, file_size, created_datetime, updated_datetime, data_type ed una experimental_strategy

Per le due espressioni (geni e proteine), contenute in più file e identificate da un codice univoco, si vuole memorizzare il nome e le misurazioni che ne derivano.

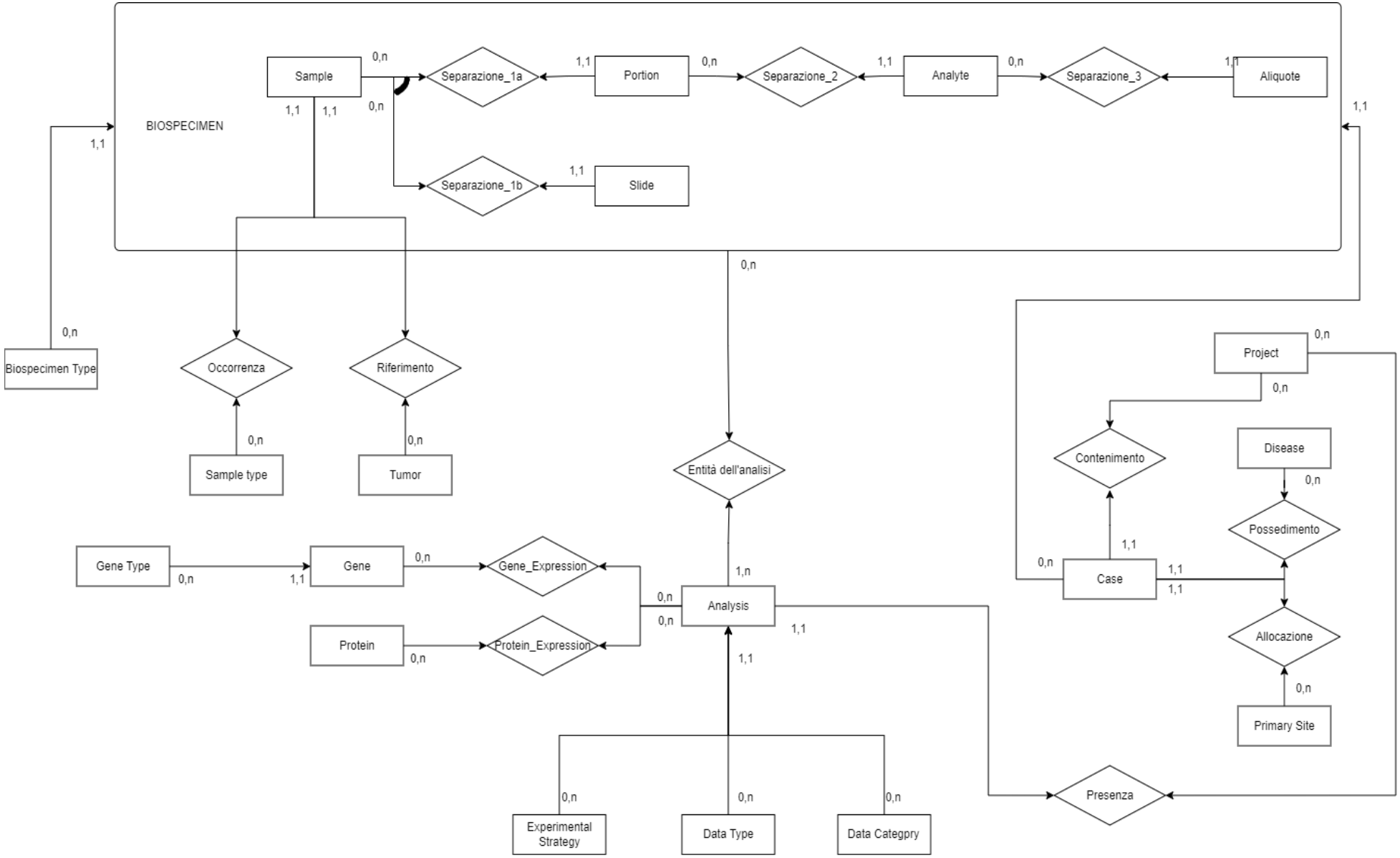
Per scelta progettuale inoltre:

- Decidiamo di creare un’entità Tipo del campione/tumore/malattia così da differenziarle con un codice univoco
- Decidiamo di creare tre entità: Experimental Strategy, Data Type e Data Category per futuri ampliamenti di database

Glossario dei Termini

Termine	Descrizione	Sinonimi	Collegamenti
Project	Questa entità rappresenta un progetto biomedico con un codice univoco e un nome. Può contenere una serie di analisi, casi, e altre informazioni specifiche del progetto.	Progetto	Case, Analysis
Case	Questa entità rappresenta un caso o un paziente coinvolto nel progetto. Contiene informazioni come il tipo di malattia, il sito primario e la sua anagrafica. Può essere collegato a uno o più campioni biologici.	Paziente, Caso	Project, Biospecimen, Disease, Primary Site
Primary Site	Contiene i dati sul sito primario di analisi del caso preso in questione	Sito, Locazione	Case
Disease	Contiene i dati sulla malattia del caso preso in questione	Malattia	Case
Biospecimen	Questa entità rappresenta campioni biologici prelevati da un paziente/caso. Contiene dettagli sui campioni biologici, ma può essere suddiviso in parti sempre più piccole.	Campione, Porzione, Analita, Aliquota, Entità, Slide	Case, File, Biospecimen Type
Analysis	Rappresenta i file che contengono dati di espressione genica o proteica. Contiene informazione come il nome del file, il suo submitter, la dimensione del file e altro. È collegato a campioni biologici o altri elementi	Insieme di espressioni, File	Project, Biospecimen, Gene, Protein
Gene	Questa entità rappresenta i dati di un gene specifico. Contiene informazioni sul gene, come il suo ID e il nome.	Espressione	File
Protein	Questa entità rappresenta i dati di una proteina specifica. Contiene informazioni sul gene, come il suo ID e il nome.	Espressione	File
Biospecimen Type	Tipologia di un Campione Biologico: Sample, Slide, Portion, Analyte o Aliquot	Tipo di Campione	Biospecimen

Modello Concettuale



Dizionario dei Dati (Entità)

Entità	Descrizione	Attributi	Identificatore
Project	Questa entità rappresenta un progetto biomedico con un codice univoco e un nome. Può contenere una serie di file, casi, e altre informazioni specifiche del progetto.	project_id, project_name	project_id
Case	Questa entità rappresenta un caso o un paziente coinvolto nel progetto. Contiene informazioni come il tipo di malattia, il sito primario e le immagini associate al caso. Può essere collegato a uno o più campioni biologici.	case_id, Ethnicity, Gender, Race, Vital Status	case_id
Biospecimen	Questa entità rappresenta campioni biologici prelevati da un paziente o un caso. Contiene dettagli sui campioni biologici, ma può essere collegato a campioni (sample) più piccoli.	Id, uuid	Id
Disease	Contiene i dati sulla malattia del caso preso in questione	Id, Type	Id
Primary Site	Contiene i dati sul sito preso in analisi del caso in questione	Id, Site	Id
Sample IS-A Biospecimen	Rappresenta un campione biologico specifico, identificato da un ID univoco. Contiene informazioni sul tipo di campione e il tipo di tumore, se applicabile. È collegato a un caso o paziente.	(Ereditati da Biospecimen)	Biospecimen.Id
Sample Type	Tipologia di un campione	Id, Type	Id
Tumor	Contiene i dati sul tumore del campione preso in questione	Id, Name, Descriptor	Id

Portion IS-A Biospecimen	Rappresenta una porzione o una parte di un campione biologico. Contiene informazioni sulla porzione del campione, ma può essere collegato ad analiti più piccoli.	(Ereditati da Biospecimen)	Biospecimen.Id
Analyte IS-A Biospecimen	Rappresenta un analita, una parte più piccola di una porzione di campione. Contiene informazioni sull'analita.	(Ereditati da Biospecimen), Concentration	Biospecimen.Id
Aliquot IS-A Biospecimen	Questa entità rappresenta una piccola porzione di un analita con una concentrazione specifica. Contiene informazioni sulla concentrazione dell'aliquota.	(Ereditati da Biospecimen), Concentration	Biospecimen.Id
Slide IS-A Biospecimen	Rappresenta una porzione o una parte di un campione biologico.	(Ereditati da Biospecimen), Image	Biospecimen.Id
Biospecimen Type	Tipologia di un Campione Biologico: Sample, Slide, Portion, Analyte o Aliquot	Id, Type	Id
Analysis	Rappresenta i file che contengono dati di espressione genica o proteica. Contiene informazione come il nome del file, il suo submitter, la dimensione del file e altro. È collegato a campioni biologici o altri elementi	Id, Filename, submitter_id, file_size, created_datetime, updated_datetime	Id
Experimental Strategy	Contiene i dati sulla strategia sperimentale usata nell’analisi	Id, Strategy	Id
Data Type	Contiene i dati sul tipo di dati nel file dell’analisi	Id, Type	Id
Data Category	Contiene i dati sulla categoria di dati nel file dell’analisi	Id, Category	Id
Gene	Questa entità rappresenta i dati di un gene specifico.	gene_id, gene_name	gene_id

	Contiene informazioni sul gene, come il suo ID e il nome.		
Gene Type	Tipologia di un Gene	Id, Type	Id
Protein	Questa entità rappresenta i dati di una proteina specifica. Contiene informazioni sul gene, come il suo ID e il nome.	AGID	AGID

Dizionario dei Dati (Relazioni)

Relazioni	**Descrizione**	**Componenti**	**Attributi**
Contenimento	Associa un progetto ai casi che contiene	Project, Case	
Presenza	Associa un progetto ai file che presenta	Project, File	
Possedimento	Associa un caso alla malattia che possiede	Case, Disease	
Allocazione	Associa un caso al sito dove si alloca la malattia	Case, Primary Site	
Scaturire	Associa dei casi alle biospecie che hanno scaturito	Case, Biospecimen	
Occorrenza	Associa un campione biologico al tipo di occorrenza che è	Sample/Analyte/Aliquot, Type	
Riferimento	Associa un tumore al campione che fa riferimento	Sample, Tumor	
Separazione_1	Associa un campione alle porzioni separate da esso	Sample, Portion	
Separazione_2	Associa una porzione agli analiti separati da esso	Portion, Analyte	
Separazione_3	Associa un analita alle aliquot separate da esso	Analyte, Aliquot	
Entità dell’analisi	Associa più campioni biologici ai file dove è contenuto	Biospecimen, File	
Gene Expression	Associa più geni ai file dove sono contenute le misurazioni	File, Gene	unstranded, stranded_first,

			stranded_second tpm, fpkm, fpkm_uq
Protein Expression	Associa più proteine ai file dove sono contenute le misurazioni	File, Protein	peptide_target, expression

Tavola dei Volumi (Indicativa solo alla piattaforma GDC)

Concetto	Tipo	Volume (indicativo)	Spiegazione dei volumi
Project	E	100	Scelta basata sulle analisi fatte nella piattaforma
Case	E	20000	Scelta basata sulle analisi fatte nella piattaforma
Disease	E	200	Scelta basata sulle analisi fatte nella piattaforma
Primary Site	E	68	Scelta basata sulle analisi fatte nella piattaforma
Biospecimen	E	30000	Ogni caso produce circa 1,5 Biospecie
Sample IS-A Biospecimen	E	15000	50% delle Biospecie
Sample Type	E	20	Numero dei campionamenti più comuni
Tumor	E	33	Scelta basata sulle analisi fatte nella piattaforma
Portion IS-A Biospecimen	E	6000	20% delle Biospecie
Analyte IS-A Biospecimen	E	3000	10% delle Biospecie
Aliquot IS-A Biospecimen	E	3000	10% delle Biospecie
Slide IS-A Biospecimen	E	3000	10% delle Biospecie
Biospecimen Type	E	5	Sample, Portion, Analyte, Aliquot and Slide
Analysis	E	35000	Scelta basata sulle analisi fatte nella piattaforma (circa 100GB)
Experimental Strategy	E	12	Scelta basata sulle analisi fatte nella piattaforma
Data Type	E	30	Scelta basata sulle analisi fatte nella piattaforma
Data Category	E	11	Scelta basata sulle analisi fatte nella piattaforma

Gene	E	22588	Geni presenti sulla piattaforma
Gene Type	E	20	Numero dei tipi di geni più comuni
Protein	E	200000	Il numero esatto di proteine presente è difficile da determinare con precisione, ma si stima che ci siano centinaia di migliaia di proteine
Contenimento	R	20000	Ogni caso appartiene a un progetto quindi per ogni caso ci sarà un record
Presenza	R	N.A	N.A
Possedimento	R	N.A	N.A
Allocazione	R	N.A	N.A
Scaturire	R	30000	Ogni campione biologico scaturisce da un caso quindi per ogni caso ci sarà un record
Occorrenza	R	N.A	N.A
Riferimento	R	N.A	N.A
Separazione_1	R	7500	Ogni porzione si separa da un campione quindi per ogni caso ci sarà un record
Separazione_2	R	4500	Ogni porzione si separa da un campione quindi per ogni caso ci sarà un record
Separazione_3	R	3000	Ogni porzione si separa da un campione quindi per ogni caso ci sarà un record
Entità dell'analisi	R	N.A	N.A
Gene Expression	R	N.A	N.A
Protein Expression	R	N.A	N.A

Dizionario dei Vincoli Esterni

1. (Biospecimen)
   Per aggiungere un record alle tabelle sottostanti campione deve essere presente un record a cui si riferiscono a cascata

RISTRUTTURAZIONE DELLO SCHEMA CONCETTUALE

Modifiche schema ER

Per la ristrutturazione del diagramma ER ho effettuato le seguenti modifiche:

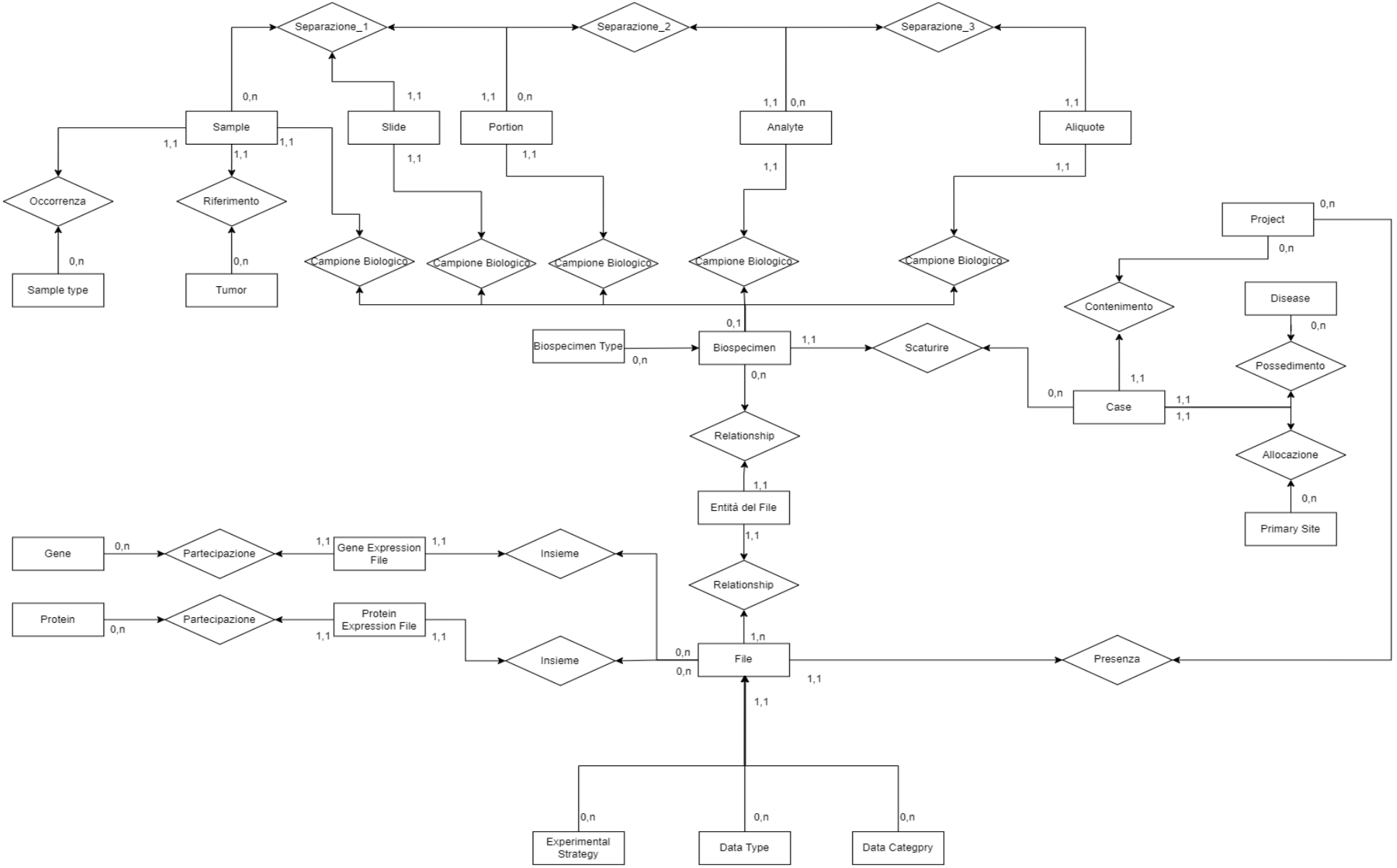
Entità

- Creazione “Entità del File”: dalla risoluzione della relazione n-n (Entità del File)
- Creazione “Gene Expression File” dalla risoluzione della relazione n-n (Gene Expression)
- Creazione “Protein Expression File” dalla risoluzione della relazione n-n (Protein Expression)

Relazioni

- Aggiunta “Campione Biologico”: applicato il metodo 3 (Sostituzione) dalla risoluzione della generalizzazione di Biospecimen
- Aggiunta “Insieme”: tra File e Gene/Protein Expression File

Diagramma ER Ristrutturato



Dizionario dei Dati (Entità)

Entità	Descrizione	Attributi	Identificatore	Vincoli Esterni
Project	Questa entità rappresenta un progetto biomedico con un codice univoco e un nome. Può contenere una serie di file, casi, e altre informazioni specifiche del progetto.	project_id, project_name	project_id	
Case	Questa entità rappresenta un caso o un paziente coinvolto nel progetto. Contiene informazioni come il tipo di malattia, il sito primario e le immagini associate al caso. Può essere collegato a uno o più campioni biologici.	case_id, Ethnicity, Gender, Race, Vital Status	case_id	
Biospecimen	Questa entità rappresenta campioni biologici prelevati da un paziente o un caso. Contiene dettagli sui campioni biologici, ma può essere collegato a campioni (sample) più piccoli.	Id, uuid	Id	Se un id è presente già in una tabella sottostante a Biospecimen non può essere in un’altra
Disease	Contiene i dati sulla malattia del caso preso in questione	Id, Type	Id	
Primary Site	Contiene i dati sul sito preso in analisi del caso in questione	Id, Site	Id	
Sample IS-A Biospecimen	Rappresenta un campione biologico specifico, identificato da un ID univoco. Contiene informazioni sul tipo di campione e il tipo di tumore, se	(Ereditati da Biospecimen)	Biospecimen.Id	Per aggiungere un record alle tabelle sottostanti campione deve essere presente un record a cui

	applicabile. È collegato a un caso o paziente.			si riferiscono a cascata
Sample Type	Tipologia di un campione	Id, Type	Id	
Tumor	Contiene i dati sul tumore del campione preso in questione	Id, Name, Descriptor	Id	
Portion IS-A Biospecimen	Rappresenta una porzione o una parte di un campione biologico. Contiene informazioni sulla porzione del campione, ma può essere collegato ad analiti più piccoli.	(Ereditati da Biospecimen)	Biospecimen.Id	Per aggiungere un record alle tabelle sottostanti porzione deve essere presente un record a cui si riferiscono a cascata
Analyte IS-A Biospecimen	Rappresenta un analita, una parte più piccola di una porzione di campione. Contiene informazioni sull'analita.	(Ereditati da Biospecimen), Concentration	Biospecimen.Id	Per aggiungere un record alle tabelle sottostanti analita deve essere presente un record a cui si riferiscono a cascata
Aliquot IS-A Biospecimen	Questa entità rappresenta una piccola porzione di un analita con una concentrazione specifica. Contiene informazioni sulla concentrazione dell'aliquota.	(Ereditati da Biospecimen), Concentration	Biospecimen.Id	
Slide IS-A Biospecimen	Rappresenta una porzione o una parte di un campione biologico.	(Ereditati da Biospecimen), Image	Biospecimen.Id	
Biospecimen Type	Tipologia di un Campione Biologico: Sample, Slide, Portion, Analyte o Aliquot	Id, Type	Id	
Analysis	Rappresenta i file che contengono dati di espressione genica o proteica. Contiene	Id, Filename, submitter_id, file_size,	Id	

	informazione come il nome del file, il suo submitter, la dimensione del file e altro. È collegato a campioni biologici o altri elementi	created_datetime, updated_datetime		
Experimental Strategy	Contiene i dati sulla strategia sperimentale usata nell'analisi	Id, Strategy	Id	
Data Type	Contiene i dati sul tipo di dati nel file dell'analisi	Id, Type	Id	
Data Category	Contiene i dati sulla categoria di dati nel file dell'analisi	Id, Category	Id	
Gene	Questa entità rappresenta i dati di un gene specifico. Contiene informazioni sul gene, come il suo ID e il nome.	gene_id, gene_name	gene_id	
Gene Type	Tipologia di un Gene	Id, Type	Id	
Protein	Questa entità rappresenta i dati di una proteina specifica. Contiene informazioni sul gene, come il suo ID e il nome.	AGID	AGID	
Gene Expression File	Associa più geni ai file dove sono contenute le misurazioni	unstranded, stranded_first, stranded_second, tpm, fpkm, fpkm_uq	FK	
Protein Expression File	Associa più proteine ai file dove sono contenute le misurazioni	Expression	FK	

Dizionario dei Dati (Relazioni)

Relazioni	**Descrizione**	**Componenti**

Contenimento	Associa un progetto ai casi che contiene	Project, Case
Presenza	Associa un progetto ai file che presenta	Project, File
Possedimento	Associa un caso alla malattia che possiede	Case, Disease
Allocazione	Associa un caso al sito dove si alloca la malattia	Case, Primary Site
Scaturire	Associa dei casi alle biospecie che hanno scaturito	Case, Biospecimen
Occorrenza	Associa un campione biologico al tipo di occorrenza che è	Sample/Analyte/Aliquot, Type
Riferimento	Associa un tumore al campione che fa riferimento	Sample, Tumor
Separazione_1	Associa un campione alle porzioni separate da esso	Sample, Portion
Separazione_2	Associa una porzione agli analiti separati da esso	Portion, Analyte
Separazione_3	Associa un analita alle aliquot separate da esso	Analyte, Aliquot
Entità dell'analisi	Associa più campioni biologici ai file dove è contenuto	Biospecimen, File
Gene Expression	Associa più geni ai file dove sono contenute le misurazioni	File, Gene
Protein Expression	Associa più proteine ai file dove sono contenute le misurazioni	File, Protein
Insieme	Associa un File alle concentrazioni in esso	File, Gene/Protein Expression File
Partecipazione	Associa un Gene/Proteina ai file dov'è contenuto	Gene/Protein, Gene/Protein Expression File

Tavola dei Volumi (2 Versione)

Concetto	Tipo	Volume (indicativo)
Project	E	100
Case	E	20000
Biospecimen	E	30000
Disease	E	200

Primary Site	E	68
Sample IS-A Biospecimen	E	15000
Sample Type	E	20
Tumor	E	33
Portion IS-A Biospecimen	E	6000
Analyte IS-A Biospecimen	E	3000
Aliquot IS-A Biospecimen	E	3000
Slide IS-A Biospecimen	E	3000
Biospecimen Type	E	5
Analysis	E	35000
Experimental Strategy	E	12
Data Type	E	30
Data Category	E	11
Gene	E	22588
Gene Type	E	20
Protein	E	200000
Gene Expression File	E	25000
Protein Expression File	E	10000
Contenimento	R	20000
Presenza	R	N.A.
Possedimento	R	N.A.
Allocazione	R	N.A.
Occorrenza	R	N.A.
Scaturire	R	30000
Riferimento	R	N.A.
Entità dell’analisi	R	N.A.
Separazione_1	R	7500
Separazione_2	R	4500

Separazione_3	R	3000
Gene Expression	R	N.A.
Protein Expression	R	N.A.
Insieme	R	N.A.
Partecipazione	R	N.A.
Contenimento	R	N.A.

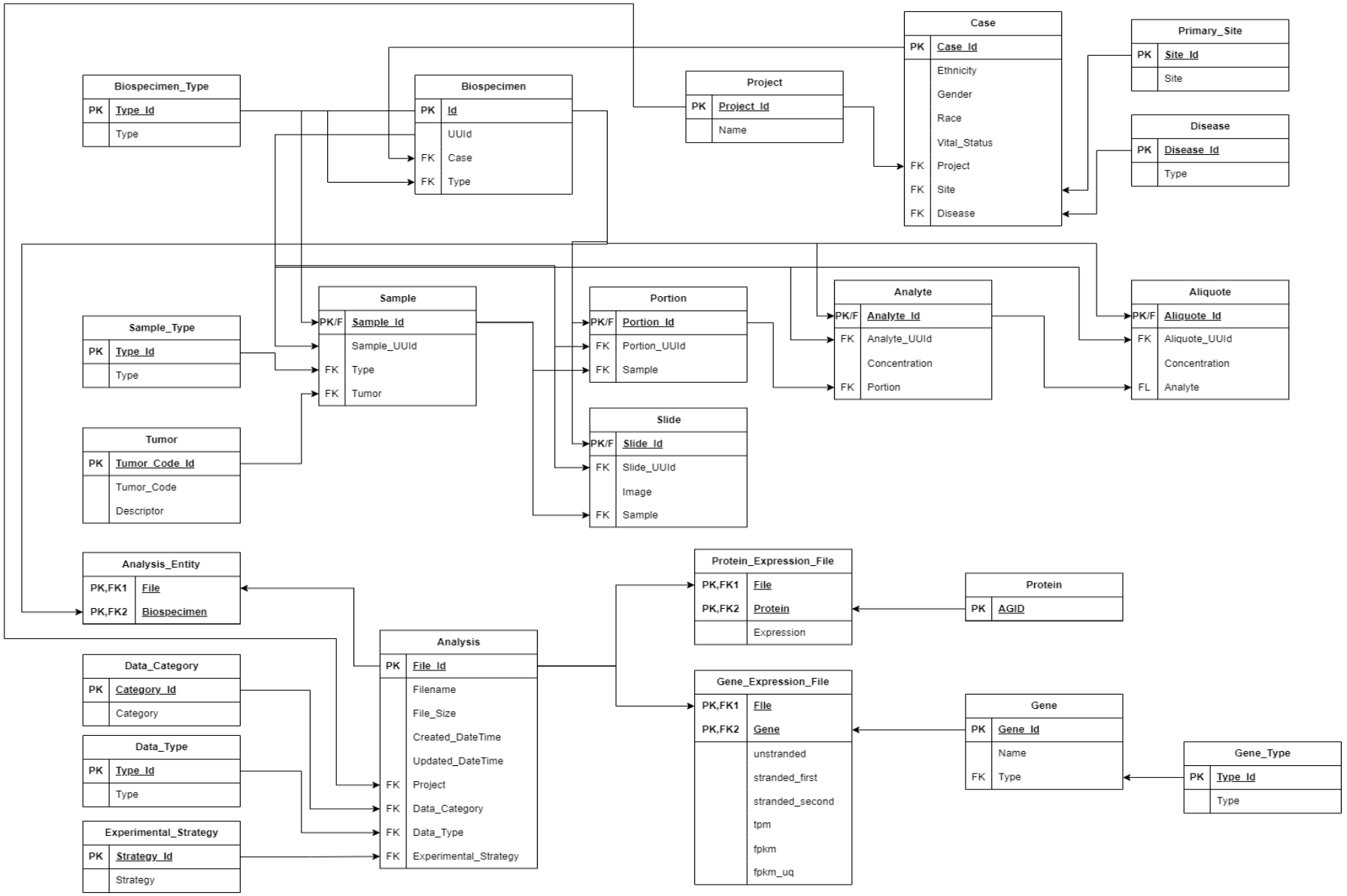
MODELLO RELAZIONALE

Schema Logico

Project(Project_Id, Name)
Case(Case_Id, Ethnicity, Gender, Race, Vital_Status, Project, Site, Disease) **Foreign Key** Project **references** Project(Project_Id) **Foreign Key** Site **references** Primary_Site(Site_Id) **Foreign Key** Disease **references** Disease(Disease _Id)
Disease(Disease_Id, Type)
Primary_Site(Site_Id, Site)
Biospecimen(Id, UUId, Case, Type) **Foreign Key** Case **references** Case(Case_UUID) **Foreign Key** Case **references** Biospecimen_Type(Id)
Biospecimen_Type(Id, Type)
Slide(Slide_Id, Slide_UUId, Image, Sample) **Foreign Key** Slide_Id, Slide_UUId **references** Biospecimen(Id, UUId) **Foreign Key** Sample **references** Sample(Sample_Id)
Sample(Sample_Id, Sample_UUId, Type, Tumor) **Foreign Key** Sample_Id, Sample_UUId **references** Biospecimen(Id, UUId) **Foreign Key** Type **references** Sample_Type(Type_Id) **Foreign Key** Tumor **references** Tumor(Tumor_Code)
Portion(Portion_Id, Portion_UUId, Sample) **Foreign Key** Portion_Id, Portion_UUId **references** Biospecimen(Id, UUId) **Foreign Key** Sample **references** Sample(Sample_Id)
Analyte(Analyte_Id, Analyte_UUId, Concentration, Portion) **Foreign Key** Analyte_Id, Analyte_UUId **references** Biospecimen(Id, UUId) **Foreign Key** Portion **references** Portion(Portion_Id)
Aliquot(Aliquot_Id, Aliquot_UUId, Concentration, Analyte) **Foreign Key** Alquote_Id, Alquote_UUId **references** Biospecimen(Id, UUId) **Foreign Key** Analyte **references** Analyte(Analyte_Id)
Sample_Type(Type_Id, Type)
Tumor(Tumor_Code, Tumor_Code_Id, Descriptor)
Analysis_Entity(File, Biospecimen) **Foreign Key** Biospecimen **references** Biospecimen(Id) **Foreign Key** File **references** File(File_Id)
Analysis(File_Id, Filename, Data_Category, File_Size, Created_DateTime, Updated_DateTime, Project, Data_Type, Data_Category, Experimental_Strategy)

Foreign Key Project **references** Project(Project_Id) **Foreign Key** Data_Type **references** Data_Type(Type_Id) **Foreign Key** Data_Category **references** Data_Category(Cayegory_Id) **Foreign Key** Experimental_Strategy **references** Experimental_Strategy(Strategy_Id)
Data_Type(Type_Id, Type)
Data_Category(Cayegory_Id, Category)
Experimental_Strategy(Strategy_Id, Strategy)
Gene_Expression_File(Gene, File, unstranded, stranded_first, stranded_second tpm, fpkm, fpkm_uq) **Foreign Key** Gene **references** Gene(Gene_Id) **Foreign Key** File **references** Analysis(File_Id)
Protein_Expression_File(Protein, File, Expression) **Foreign Key** File **references** Analysis(File_Id) **Foreign Key** Protein **references** Protein(Protein_Id)
Gene(Gene_Id, Name, Type) **Foreign Key** Type **references** Gene_Type(Type_Id)
Gene_Type(Type_Id, Type)
Protein(AGID, lab_id, catalog_number, set_id, peptide_target)

Schema Relazionale



Ristrutturazione dello Schema Relazionale

Dato che il numero degli accessi per le operazioni è ridotto al minimo e il carico generale dell'applicazione non è eccessivo, non è necessaria una ristrutturazione dello schema relazionale.

SPECIFICA DEL DATABASE IN SQL

Creazione Database

```
CREATE DATABASE IF NOT EXIST "GDC";

USE "GDC";


CREATE TABLE public."Aliquote" (
    "Aliquote_Id" text PRIMARY KEY,
    "Aliquote_UUId" text PRIMARY KEY,
    "Analyte_Id" text NOT NULL,
    "Analyte_UUId" text NOT NULL,
    "Concentration" numeric
    FOREIGN KEY ("Analyte_Id", "Analyte_UUId") REFERENCES public."Analyte"("Analyte_Id", "Analyte_UUId");
    FOREIGN KEY ("Aliquote_Id", "Aliquote_UUId") REFERENCES public."Biospecimen"("Id", "UUId");
);


CREATE TABLE public."Analysis" (
    "File_Id" text PRIMARY KEY,
    "Filename" text,
    "File_Size" numeric,
    "Created_DateTime" date,
    "Updated_DateTime" date,
    "Project" text NOT NULL,
    "Data_Category" integer,
    "Data_Type" integer,
    "Experimental_Strategy" integer,
    CONSTRAINT "CK_Data" CHECK (("Updated_DateTime" >= "Created_DateTime"))
```

```
    FOREIGN KEY ("Project") REFERENCES public."Project"("Project_Id");

    FOREIGN KEY ("Data_Category") REFERENCES public."Data_Category"("Category_Id")
NOT VALID;

    FOREIGN KEY ("Data_Type") REFERENCES public."Data_Type"("Type_Id") NOT VALID;

    FOREIGN KEY ("Experimental_Strategy") REFERENCES
public."Experimental_Strategy"("Strategy_Id") NOT VALID;

);


CREATE TABLE public."Analysis_Entity" (

    "Analysis" text PRIMARY KEY,

    "Biospecimen_Id" text PRIMARY KEY,

    "Biospecimen_UUId" text PRIMARY KEY

    FOREIGN KEY ("Biospecimen_Id", "Biospecimen_UUId") REFERENCES
public."Biospecimen"("Id", "UUId");

    FOREIGN KEY ("File") REFERENCES public."File"("File_Id");

);


CREATE TABLE public."Analyte" (

    "Analyte_Id" text PRIMARY KEY,

    "Analyte_UUId" text PRIMARY KEY,

    "Portion_Id" text NOT NULL,

    "Portion_UUId" text NOT NULL,

    "Concentration" numeric

    FOREIGN KEY ("Analyte_Id", "Analyte_UUId") REFERENCES public."Biospecimen"("Id",
"UUId");

    FOREIGN KEY ("Portion_Id", "Portion_UUId") REFERENCES public."Portion"("Portion_Id",
"Portion_UUId");

);


CREATE TABLE public."Biospecimen" (

    "Id" text PRIMARY KEY,

    "UUId" text PRIMARY KEY,

    "Case" text NOT NULL
```

```
    "Type" integer NOT NULL
    FOREIGN KEY ("Case") REFERENCES public."Case"("Case_UUId") NOT VALID;
    FOREIGN KEY ("Type") REFERENCES public."Biospecimen_Type"("Type_Id") NOT
VALID;
);

CREATE TABLE public."Biospecimen_Type" (
    "Type_Id" integer AUTO_INCREMENT PRIMARY KEY,
    "Type" text
);

CREATE TABLE public."Case" (
    "Case_UUId" text PRIMARY KEY,
    "Case_Id" text NOT NULL,
    "Ethnicity" text,
    "Race" text,
    "Gender" text,
    "Vital_Status" text,
    "Project" text NOT NULL,
    "Site" integer,
    "Disease" integer,
    CONSTRAINT "CK_Gender" CHECK (("Gender" = ANY (ARRAY['Male'::text, 'Female'::text,
'Unknown'::text, 'Not reported'::text]))),
    CONSTRAINT "CK_Vital_Status" CHECK (("Vital_Status" = ANY (ARRAY['Alive'::text,
'Dead'::text])))
    FOREIGN KEY ("Disease") REFERENCES public."Disease"("Disease_Id");
    FOREIGN KEY ("Project") REFERENCES public."Project"("Project_Id");
    FOREIGN KEY ("Site") REFERENCES public."Primary_Site"("Site_Id");
);

CREATE TABLE public."Data_Category" (
    "Category_Id" integer AUTO_INCREMENT PRIMARY KEY,
```

```
    "Category" text
);

CREATE TABLE public."Data_Type" (
    "Type_Id" integer AUTO_INCREMENT PRIMARY KEY,
    "Type" text
);

CREATE TABLE public."Disease" (
    "Disease_Id" integer AUTO_INCREMENT PRIMARY KEY,
    "Type" text
);

CREATE TABLE public."Experimental_Strategy" (
    "Strategy_Id" integer AUTO_INCREMENT PRIMARY KEY,
    "Strategy" text
);

CREATE TABLE public."Gene" (
    "Gene_Id" text PRIMARY KEY,
    "Name" text,
    "Type" integer
    FOREIGN KEY ("Type") REFERENCES public."Gene_Type"("Type_Id") NOT VALID;
);

CREATE TABLE public."Gene_Expression_File" (
    "File" text PRIMARY KEY,
    "Gene" text PRIMARY KEY,
    "TPM" numeric,
    "FPKM" numeric,
    "FPKM_UQ" numeric
```

```
    "unstranded" integer,
    "stranded_first" integer,
    "stranded_second" integer
    FOREIGN KEY ("Analysis") REFERENCES public."Analysis"("File_Id");
    FOREIGN KEY ("Gene") REFERENCES public."Gene"("Gene_Id");
);

CREATE TABLE public."Gene_Type" (
    "Type_Id" integer AUTO_INCREMENT PRIMARY KEY,
    "Type" text
);

CREATE TABLE public."Portion" (
    "Portion_Id" text PRIMARY KEY,
    "Portion_UUId" text PRIMARY KEY,
    "Sample_Id" text NOT NULL,
    "Sample_UUId" text NOT NULL
    FOREIGN KEY ("Portion_Id", "Portion_UUId") REFERENCES public."Biospecimen"("Id",
"UUId");
    FOREIGN KEY ("Sample_Id", "Sample_UUId") REFERENCES public."Sample"("Sample_Id",
"Sample_UUId");
);

CREATE TABLE public."Primary_Site" (
    "Site_Id" integer AUTO_INCREMENT PRIMARY KEY,
    "Site" text
);

CREATE TABLE public."Project" (
    "Project_Id" text PRIMARY KEY,
    "Name" text
);
```

```
CREATE TABLE public."Protein" (
    "AGID" text PRIMARY KEY
    "lab_id" integer,
    "catalog_number" text,
    "set_id" text,
    "peptide_target" text
);

CREATE TABLE public."Protein_Expression_File" (
    "File" text PRIMARY KEY,
    "Protein" text PRIMARY KEY,
    "Expression" numeric
    FOREIGN KEY ("Analysis") REFERENCES public."Analysis"("File_Id");
    FOREIGN KEY ("Protein") REFERENCES public."Protein"("AGID");
);

CREATE TABLE public."Sample" (
    "Sample_Id" text PRIMARY KEY,
    "Sample_UUId" text PRIMARY KEY,
    "Type" integer,
    "Tumor" integer
    FOREIGN KEY ("Sample_Id", "Sample_UUId") REFERENCES public."Biospecimen"("Id",
"UUId");
    FOREIGN KEY ("Tumor") REFERENCES public."Tumor"("Tumor_Code_Id");
    FOREIGN KEY ("Type") REFERENCES public."Sample_Type"("Type_Id");
);

CREATE TABLE public."Sample_Type" (
    "Type_Id" integer PRIMARY KEY,
    "Type" text
```

```
);

CREATE TABLE public."Slide" (
    "Slide_Id" text PRIMARY KEY,
    "Slide_UUId" text PRIMARY KEY,
    "Sample_Id" text NOT NULL,
    "Sample_UUId" text NOT NULL,
    "Image" json
);

CREATE TABLE public."Tumor" (
    "Tumor_Code_Id" integer PRIMARY KEY,
    "Code" text,
    "Descriptor" text
);
```