

**Relazione del progetto per il corso**  
**Basi di dati**  
**Rete di centri di ricerca**

Roberto Lepore  
roberto.lepore2@studio.unibo.it  
0000970366

Giugno 2022



# Indice

<b>1</b>	<b>Analisi dei requisiti</b>	<b>5</b>
1.1	Intervista . . . . .	5
1.2	Glossario dei termini . . . . .	6
1.3	Rielaborazione . . . . .	6
<b>2</b>	<b>Progettazione concettuale</b>	<b>9</b>
2.1	Partecipazione a un team di ricerca . . . . .	9
2.2	Presenza di un gene in un organismo . . . . .	10
2.3	Sviluppo di una ricerca . . . . .	11
2.4	Schema concettuale finale . . . . .	12
<b>3</b>	<b>Progettazione logica</b>	<b>13</b>
3.1	Stima dei volumi . . . . .	13
3.2	Frequenza delle operazioni . . . . .	13
3.3	Tabelle degli accessi . . . . .	14
3.4	Studio della ridondanza . . . . .	18
3.5	Ristrutturazione dello schema . . . . .	19
3.6	Traduzione in relazioni . . . . .	20
3.7	Schema relazionale finale . . . . .	21
3.8	Query SQL . . . . .	22
<b>4</b>	<b>Sviluppo dell'applicativo</b>	<b>25</b>
4.1	Architettura e funzionalità . . . . .	25



# 1 Analisi dei requisiti

## 1.1 Intervista

Si vuole creare una base di dati per la raccolta e condivisione delle ricerche sviluppate in una rete di centri di ricerca. In ogni centro di ricerca lavorano dei ricercatori, i quali si organizzano in team di ricerca con altri scienziati appartenenti alla rete. Un team, di cui un membro è titolare, sviluppa una ricerca riguardante un insieme di geni o specifiche varianti di geni. Ogni gene è identificato da un nome e un codice ufficiali (univoci per un organismo), e se ne memorizzano il tipo di gene, l'organismo da cui proviene, il suo locus genico, una descrizione e un insieme di condizioni associate al gene. Una variante di un gene è caratterizzata da un nome, una lunghezza espressa in coppie di basi (bp), la sequenza di basi azotate, il tipo di molecola in cui si esprime, e un insieme di regioni codificanti (che possono essere sconosciute). Una ricerca, una volta completata, viene pubblicata su una rivista scientifica e gli scienziati che hanno preso parte allo studio possono partecipare a nuovi team e sviluppare nuove ricerche.

Deve essere possibile visualizzare varie informazioni sulle ricerche, come il numero di pubblicazioni seguite da un dato ricercatore o dai ricercatori di un determinato centro di ricerca, oppure il numero di studi che hanno analizzato un gene o un organismo. Inoltre si vuole gestire la pubblicazione di una ricerca e la creazione di un nuovo team di ricerca.

Nella figura 1.1 si mostra un esempio di variante e relativo gene, con gli attributi che ci interessa memorizzare

Nome	Gene	Organismo	Lunghezza	Sequenza	Tipo di molecola	CDS
Variante di trascrizione Dp140c	Distrofina	Homo Sapiens	7.086	acactagcaa...	mRNA	1042..4389

Nome	Codice	Organismo	Tipo	locus	Descrizione	Condizioni associate
Distrofina	DMD	Homo Sapiens	Codificante Proteina	Xp21.2	This gene spans a genomic range of greater than 2 Mb...	Distrofia muscolare di Becker

Figura 1.1: Scheda informativa di una variante e relativo gene

## 1.2 Glossario dei termini

Di seguito si elencano i concetti fondamentali del dominio e una loro breve descrizione

Termine	Descrizione	Sinonimi
Centro di ricerca	Centro di laboratori che condivide i propri dati di ricerca con gli altri centri della rete	Laboratorio, università
Ricerca	Attività di studio nell'ambito genetico condotta da ricercatori	Tesi, attività di ricerca
Ricercatore	Studioso che segue progetti di ricerca nell'ambito genetico	Scienziato, biologo
Team di ricerca	Gruppo di ricercatori che collaborano per una specifica ricerca	Team, gruppo di ricerca
Gene	Sequenza di DNA presente nei cromosomi di un organismo che codificano un RNA o una proteina	
Variante	Specifica sequenza di nucleotidi in cui un gene si può presentare dentro l'organismo	
Organismo	Essere vivente contenente un genoma e soggetto a evoluzione	

Figura 1.2: Glossario dei termini

## 1.3 Rielaborazione

Di ogni centro di ricerca si memorizzano il nome e il paese in cui si trova la sede principale. Un ricercatore lavora per un unico centro di ricerca e può partecipare a più team di ricerca alla volta. Di ogni ricercatore si memorizzano nome, cognome e codice identificativo, univoco nell'intera rete. Un team, di cui un membro è anche titolare, si concentra su una ricerca alla volta, la quale ha un nome univoco e si concentra su un insieme di varianti di geni. Una ricerca, una volta completata, viene pubblicata su un numero di una ricerca scientifica di cui si vuole memorizzare nome e data di pubblicazione. Di ogni gene sono memorizzati nome e codice identificativi ufficiali, il tipo, l'organismo da cui proviene, il locus genico, una descrizione e un insieme di condizioni associate sull'organismo. Due geni diversi presenti in due organismi differenti possono avere lo stesso nome e codice identificativi. Una variante di un gene è caratterizzata da un nome, una lunghezza espressa in coppie di basi (bp), la sequenza di basi azotate, il tipo di molecola in cui si esprime, e un insieme di regioni codificanti (che possono essere sconosciute). Di ogni organismo si memorizza nome scientifico, nome comune, tassonomia, rango (genere o specie), nome dello scienziato che ha assegnato il nome e in quale anno.

Dev'essere possibile eseguire le seguenti operazioni:

1. Visualizzazione del numero di ricerche a cui hanno contribuito ricercatori di un dato centro di ricerca, divise per rivista di pubblicazione.
2. Visualizzazione del numero di ricerche sviluppate da un dato ricercatore
3. Chiusura di una ricerca.
4. Visualizzazione delle ricerche riguardanti un dato gene
5. Visualizzazione delle ricerche riguardanti un dato organismo
6. Visualizzazione del titolare e del numero di pubblicazioni per ogni team
7. Visualizzazione delle varianti di geni associate ad una data condizione.
8. Visualizzazione delle ricerche aperte di un ricercatore.
9. Creazione di un nuovo team.
10. Visualizzazione del ricercatore che complessivamente ha lavorato su più geni in tutta la rete





## 2 Progettazione concettuale

### 2.1 Partecipazione a un team di ricerca

Un team è formato da più ricercatori, dei quali uno solo è anche titolare del team. Ogni ricercatore è dipendente di un centro di ricerca e può partecipare a più team alla volta. Non è espresso il vincolo dell'appartenenza del titolare al team stesso, che dovrà essere verificato a livello applicazione. L'attributo *numero pubblicazioni* è ridondante e si dovrà decidere se mantenerlo in fase di progettazione logica.

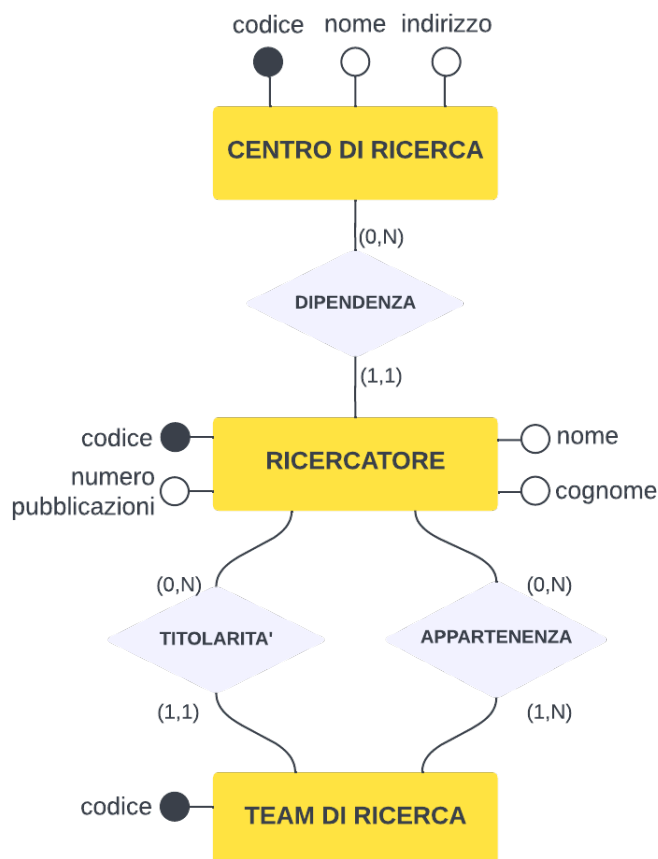


Figura 2.1: Partecipazione a un team

## 2.2 Presenza di un gene in un organismo

In un organismo sono presenti numerosi geni, i quali a loro volta possono essere espressi in diverse varianti. Un gene può essere identificato univocamente rispetto a un organismo sia dal nome che dal codice, per cui sono entrambi chiavi. Una variante è identificata univocamente dalla combinazione del suo nome e del gene che esprime.

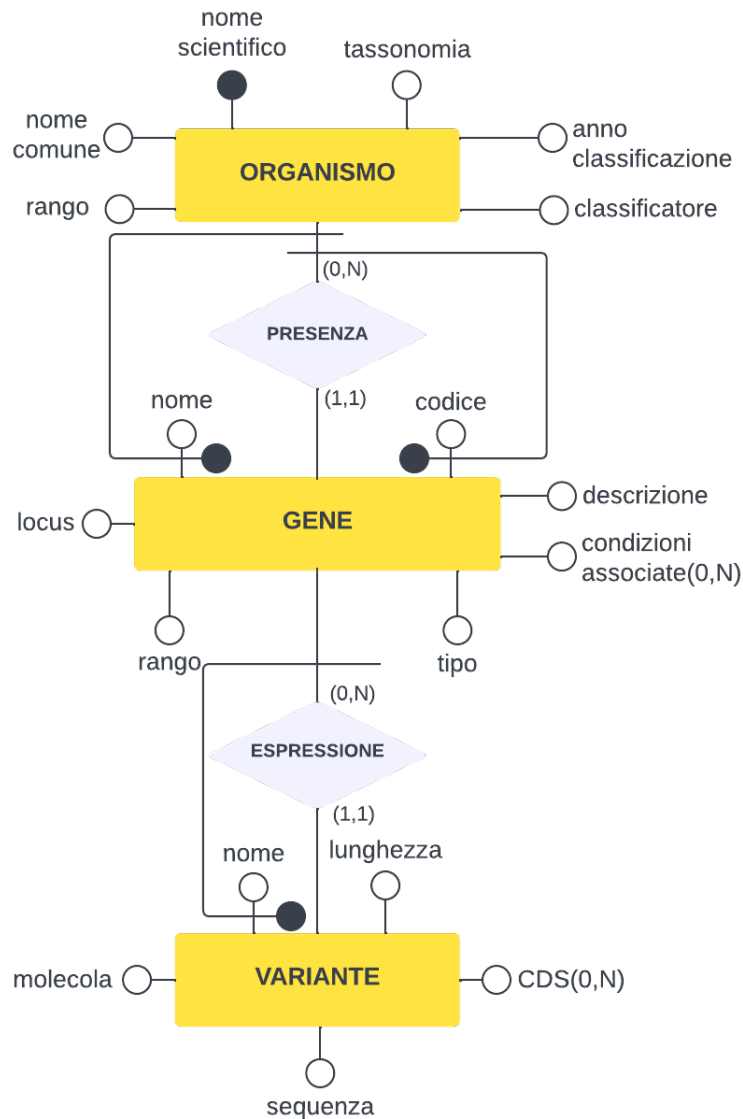


Figura 2.2: Geni e organismi

## 2.3 Sviluppo di una ricerca

Un team di ricerca sviluppa una sola ricerca alla volta. Quando una ricerca viene chiusa, questa viene mantenuta nello storico, legata al team che l'ha seguita attraverso la relazione **SVILUPPO COMPLETATO**. Questa dinamicità non è ideale ma, dato che una ricerca una volta chiusa rimane in questo stato, è facilmente gestibile. Una ricerca studia sempre almeno una variante.

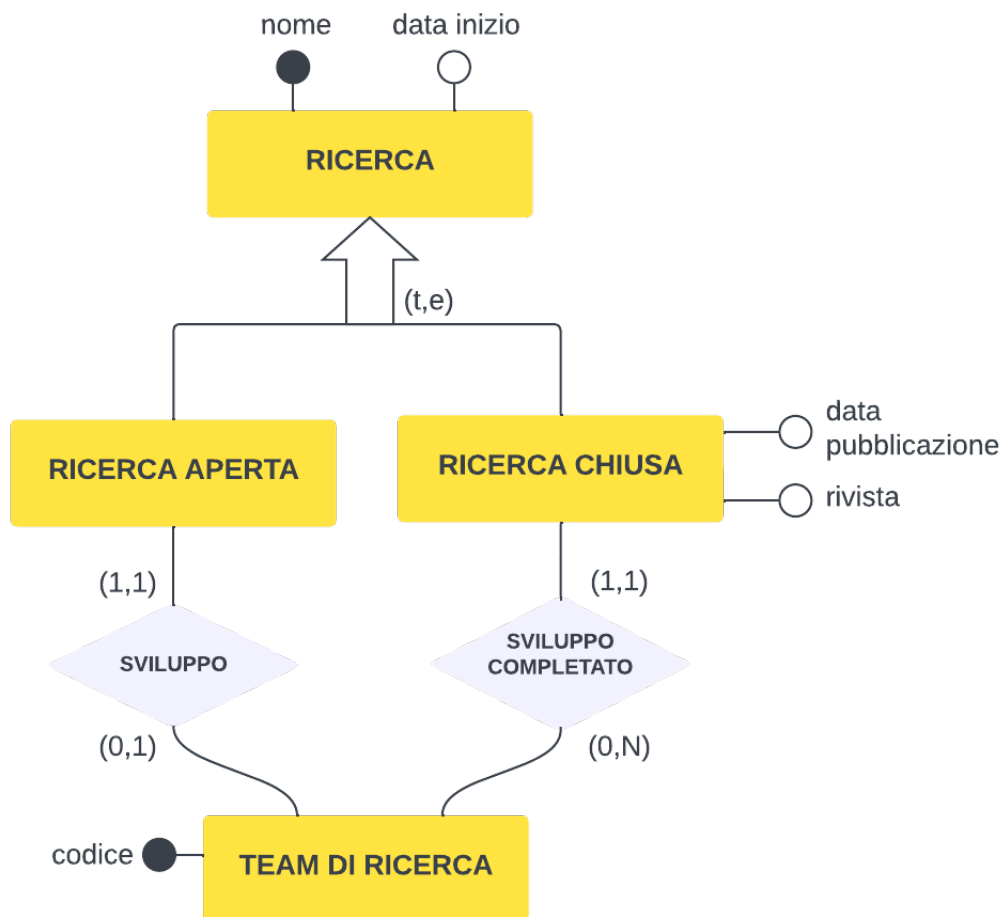


Figura 2.3: Sviluppo ricerche

## 2.4 Schema concettuale finale

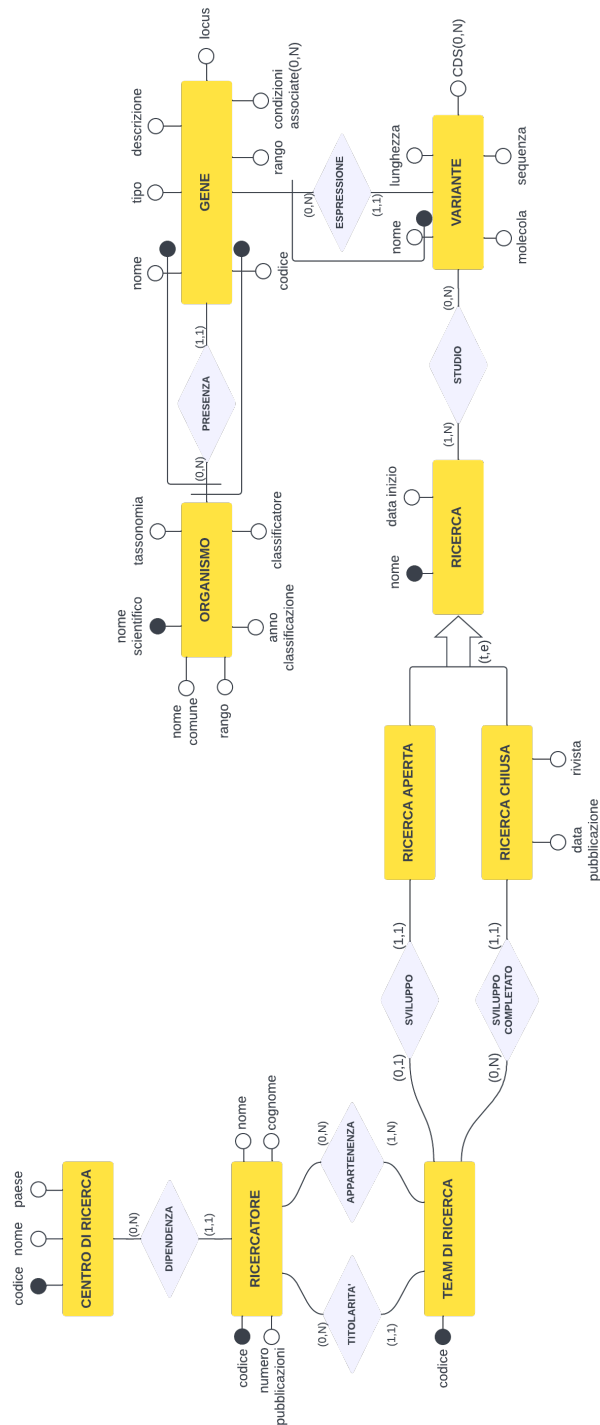


Figura 2.4: Schema concettuale finale

## 3 Progettazione logica

### 3.1 Stima dei volumi

Di seguito si indica il volume stimato per ogni entità e associazione del dominio.

Concetto	Costrutto	Volume
Centro di ricerca	E	100
Dipendenza	R	2.000
Ricercatore	E	2.000
Titolarità	R	4.000
Appartenenza	R	10.000
Team di ricerca	E	4.000
Sviluppo	R	4.000
Sviluppo completato	R	36.000
Ricerca	E	40.000
Ricerca chiusa	E	36.000
Ricerca aperta	E	4.000
Studio	R	120.000
Variante	E	50.000
Espressione	R	50.000
Gene	E	5.000
Presenza	R	5.000
Organismo	E	500

Figura 3.1: Tabella dei volumi

### 3.2 Frequenza delle operazioni

Di seguito si indica la frequenza stimata per ogni operazione che dev'essere implementata.

Codice	Operazione	Frequenza
1	Numero di pubblicazioni dato un centro di ricerca divise per rivista	100 al giorno
2	Numero di pubblicazioni di un dato ricercatore	10 al giorno
3	Chiusura di una ricerca	5 al giorno
4	Ricerche riguardanti un dato gene	300 al giorno
5	Ricerche riguardanti un dato organismo	200 al giorno
6	Titolare e numero di ricerche per team	5 al giorno
7	Varianti di geni associati a una data condizione	2 al giorno
8	Ricerche aperte di un ricercatore	100 al giorno
9	Creazione di un nuovo team di ricerca	5 a settimana
10	Ricercatore che ha studiato più geni	1 a settimana

Figura 3.2: Tabella delle frequenze

### 3.3 Tabelle degli accessi

In questa sezione si stima il numero di accessi ai dati necessari per ogni operazione. Le operazioni 2 e 3 considerano un elemento ridondante, per cui verranno esaminate a parte.

**Operazione 1** Visualizzazione del numero di ricerche a cui hanno contribuito ricercatori di un dato centro di ricerca, divise per rivista di pubblicazione.

Concetto	Costrutto	Accessi	Tipo
Centro	E	1	L
Dipendenza	R	20	L
Ricercatore	E	20	L
Appartenenza	R	100	L
Team	E	100	L
Sviluppo compl.	R	900	L
Totale:		1.141	L
114.100 accessi al giorno			

Figura 3.3: Tabella degli accessi per l'operazione 1

**Operazione 4** Visualizzazione delle ricerche riguardanti un dato gene.

Concetto	Costrutto	Accessi	Tipo
Gene	E	1	L
Espressione	R	10	L
Variante	E	10	L
Studio	R	24	L
Ricerca	E	24	L
Totale:		69	L
20.700 accessi al giorno			

Figura 3.4: Tabella degli accessi per l' operazione 4

**Operazione 5** Visualizzazione delle ricerche riguardanti un dato organismo.

Concetto	Costrutto	Accessi	Tipo
Organismo	E	1	L
Presenza	R	10	L
Gene	E	10	L
Espressione	R	100	L
Variante	E	100	L
Studio	R	240	L
Ricerca	E	240	L
Totale:		701	L
140.000 accessi al giorno			

Figura 3.5: Tabella degli accessi per l' operazione 5

**Operazione 6** Visualizzazione del titolare e del numero di pubblicazioni per ogni team.

Concetto	Costrutto	Accessi	Tipo
Team	E	1	L
Titolarità	R	1	L
Ricercatore	E	1	L
Sviluppo compl.	R	9	L
Ricerca chiusa	E	9	L
Totale:		21 L	
		105 accessi al giorno	

Figura 3.6: Tabella degli accessi per l' operazione 6

**Operazione 7** Visualizzazione delle varianti di geni associate ad una data condizione.

Concetto	Costrutto	Accessi	Tipo
Gene	E	5000	L
Totale:		5.000 L	
		10.000 accessi al giorno	

Figura 3.7: Tabella degli accessi per l' operazione 7

**Operazione 8** Visualizzazione delle ricerche aperte di un ricercatore.

Concetto	Costrutto	Accessi	Tipo
Ricercatore	E	1	L
Appartenenza	R	5	L
Team	E	5	L
Sviluppo	R	5	L
Ricerca aperta	E	5	L
Totale:		21 L	
		2.100 accessi al giorno	

Figura 3.8: Tabella degli accessi per l' operazione 8



**Operazione 9** Creazione di un nuovo team di ricerca.

Concetto	Costrutto	Accessi	Tipo
Team	E	1	S
Titolarità	R	1	S
Appartenenza	R	2,5	S
Totale:		4,5	S
45 accessi a settimana			

Figura 3.9: Tabella degli accessi per l' operazione 9

**Operazione 10** Ricercatore che ha studiato complessivamente più geni di tutti i ricercatori nella rete.

Concetto	Costrutto	Accessi	Tipo
Ricercatore	E	1	L
Appartenenza	R	5	L
Team	E	5	L
Sviluppo compl.	R	9	L
Ricerca chiusa	E	9	L
Studio	R	3,3	L
Variante	E	3,3	L
Espressione	R	3,3	L
Totale:		38,9	L
38,9 accessi a settimana			

Figura 3.10: Tabella degli accessi per l' operazione 10

### 3.4 Studio della ridondanza

Le operazioni 2 e 3 considerano un elemento ridondante. Per valutare se mantenerlo o rimuoverlo si esaminano entrambi i casi.

#### Senza ridondanza

**Operazione 2** Visualizzazione del numero di ricerche sviluppate da un dato ricercatore.

Concetto	Costrutto	Accessi	Tipo
Ricercatore	E	1	L
Appartenenza	R	5	L
Team di ricerca	E	5	L
Sviluppo completato	E	45	L
Totale:		56 L	
560 accessi al giorno			

Figura 3.11: Tabella degli accessi per l'operazione 2 senza ridondanza

**Operazione 3** Chiusura di una ricerca.

Concetto	Costrutto	Accessi	Tipo
Sviluppo	R	1	S
Ricerca aperta	E	1	S
Sviluppo compl.	R	1	S
Ricerca chiusa	E	1	S
Totale:		4 L	
40 accessi al giorno			

Figura 3.12: Tabella degli accessi per l'operazione 3 senza ridondanza

#### Con ridondanza

**Operazione 2** Visualizzazione del numero di ricerche sviluppate da un dato ricercatore.

Concetto	Costrutto	Accessi	Tipo
Ricercatore	E	1	L
Totale:		1 L	
10 accessi al giorno			

Figura 3.13: Tabella degli accessi per l'operazione 2 con ridondanza

**Operazione 3** Chiusura di una ricerca.

Concetto	Costrutto	Accessi	Tipo
Sviluppo	R	1	S
Ricerca aperta	E	1	S
Sviluppo completato	R	1	S
Ricerca chiusa	E	1	S
Ricercatore	E	1	L
Ricercatore	E	1	S
Totale:		1 L	
		5 S	
		55 accessi al giorno	

Figura 3.14: Tabella degli accessi per l'operazione 3 con ridondanza

Quindi risulta vantaggioso mantenere la ridondanza dell'attributo *numero pubblicazioni*.

## 3.5 Ristrutturazione dello schema

Per ottenere lo schema logico finale ho eseguito le seguenti operazioni di ristrutturazione sullo schema concettuale:

### Scelta di una chiave primaria in GENE

L'entità *GENE* può essere identificata rispetto ad un organismo sia dal nome che dal codice. Ho scelto quindi l'attributo *codice* come chiave primaria.

### Eliminazione della gerarchia

Ho effettuato un collasso verso il basso per eliminare la gerarchia *RICERCA*, in quanto le operazioni prendono sempre in considerazione le ricerche pubblicate e quelle in via di sviluppo separatamente. Si noti che viene perso il vincolo di unicità del nome tra tutte le ricerche.

### Trasformazione in entità degli attributi multipli

Nello schema concettuale sono presenti due attributi multipli: *condizioni associate* in *GENE* e *cds* in *VARIANTE*. Una *CONDIZIONE ASSOCIATA* è identificata dal suo nome ed è in relazione molti a molti con l'entità *GENE*. Mentre *CDS* è un intervallo che non può essere ripetuto per una stessa variante, per cui è identificato dal suo valore e dalla variante a cui si riferisce.

## 3.6 Traduzione in relazioni

- CENTRI DI RICERCA( nome, paese, codice )
- RICERCATORI( nome, cognome, codice, numero pubblicazioni, id centro: CENTRI DI RICERCA)
- TEAM DI RICERCA( codice, titolare: RICERCATORI )
- APPARTENENZE( id ricercatore: RICERCATORI, id team: TEAM DI RICERCA)
- RICERCHE APERTE ( nome, data inizio, id team: TEAM DI RICERCA )  
**Unique**(id team)
- RICERCHE CHIUSE ( nome, data inizio, data pubblicazione, rivista, id team: TEAM DI RICERCA)
- STUDI ( nome ricerca: RICERCHE APERTE, (nome, codice gene, id organismo): VARIANTI)
- STUDI COMPLETATI ( nome ricerca: RICERCHE CHIUSE, (nome, codice gene, id organismo): VARIANTI)
- ORGANISMI ( nome scientifico, nome comune, tassonomia, rango, classificatore, anno classificazione)
- GENI ( nome, codice, id organismo: ORGANISMO , tipo, locus, descrizione )  
**Unique**(nome)
- VARIANTI ( nome, (codice gene, id organismo): GENE, lunghezza, sequenza, tipo di molecola)
- CONDIZIONI ASSOCIATE ( nome )
- MANIFESTAZIONI ( nome condizione: CONDIZIONI ASSOCIATE, (codice gene, id organismo): GENE )
- CDS( intervallo, (nome, codice gene, id organismo):VARIANTI)

### 3.7 Schema relazionale finale

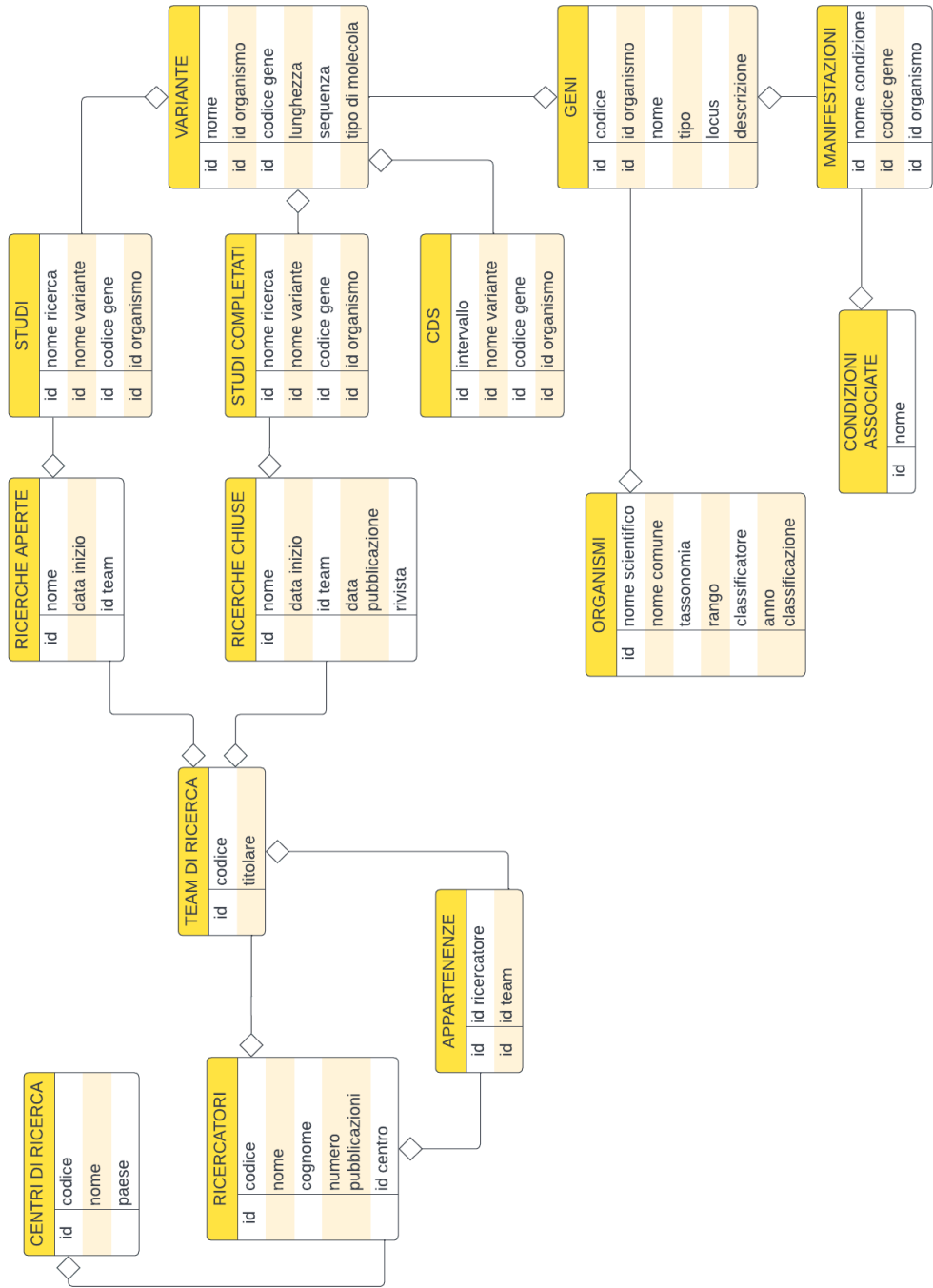


Figura 3.15: Schema relazionale

## 3.8 Query SQL

**Operazione 1:** Numero di ricerche a cui hanno contribuito ricercatori di un dato centro di ricerca, divise per rivista di pubblicazione.

```

1  SELECT rc.rivista, COUNT(distinct rc.nome)
2  FROM centriricerca c, ricercatori r, appartenenze a, team t,
3     ricerchechiuse rc
4  WHERE c.nome = <nome centro>
5         AND r.idcentro = c.codice
6         AND a.idricercatore = r.codice
7         AND a.idteam = t.codice
8         AND rc.idteam = t.codice
9  GROUP BY rc.rivista

```

**Operazione 2:** Numero di ricerche sviluppate da un dato ricercatore.

```

1  SELECT r.nome, r.cognome, c.nome, numeropubblicazioni
2  FROM ricercatori r, centriricerca c
3  WHERE r.nome = <nome> AND r.cognome = <cognome>
4  AND c.codice = r.idcentro

```

**Operazione 3:** Chiusura di una ricerca.

Per semplificare la query ho creato dei trigger all'interno del database, uno che si attiva al momento dell'eliminazione di una ricerca aperta e uno al momento dell'eliminazione di uno studio.

```

1  DELIMITER $$
2  CREATE TRIGGER chiudi_ricerca
3  BEFORE DELETE on ricercheaperte
4  FOR EACH ROW
5  BEGIN
6      INSERT INTO ricerchechiuse(nome, datainizio,
7         datapubblicazione, rivista, idteam)
8         VALUES(OLD.nome,OLD.datainizio,now(),"Nature",OLD.idteam);
9      DELETE FROM studi where numericerca = OLD.nome;
10     UPDATE ricercatori
11         SET numeropubblicazioni = numeropubblicazioni + 1
12         WHERE codice IN (
13             SELECT idricercatore
14             FROM appartenenze
15             WHERE idteam = OLD.idteam
16         );
17     END$$
18     DELIMITER ;

```

```

1  DELIMITER $$
2  CREATE TRIGGER chiudi_studio
3  BEFORE DELETE ON studi
4  FOR EACH ROW
5  BEGIN
6      INSERT INTO studicompletati(numericerca, nomevariante,
7      idgene, idorganismo)
8      VALUES(OLD.numericerca, OLD.nomevariante, OLD.idgene,
9      OLD.idorganismo);
10 END$$
11 DELIMITER ;

```

L'applicativo esegue la seguente query:

```

1  DELETE FROM ricercheaperte WHERE nome = <nome>;
2  UPDATE ricerchechiuse SET rivista = <rivista>
3  WHERE nome = <nome>;

```

**Operazione 4:** Visualizzazione delle ricerche riguardanti un dato gene.

Viene indicato solo il codice del gene, quindi possono risultare ricerche riguardanti a organismi diversi.

```

1  SELECT DISTINCT s.idorganismo, s.numericerca, r.idteam
2  FROM studicompletati s, ricerchechiuse r
3  WHERE s.numericerca = r.nome
4  AND s.idgene = <gene>

```

**Operazione 5:** Visualizzazione delle ricerche riguardanti un dato organismo.

```

1  SELECT DISTINCT o.nomescientifico, s.numericerca, r.idteam
2  FROM studicompletati s, ricerchechiuse r, organismi o
3  WHERE s.numericerca = r.nome
4  AND o.nomescientifico = s.idorganismo
5  AND o.nomecomune = <organismo>

```

**Operazione 6:** Titolare e numero ricerche per team.

```

1  SELECT t.codice, r.cognome, count(rc.nome)
2  FROM team t LEFT OUTER JOIN ricerchechiuse rc
3  ON (t.codice = rc.idteam), ricercatori r
4  WHERE r.codice = t.titolare
5  GROUP BY t.codice, r.cognome
6  ORDER BY t.codice

```

**Operazione 7:** Visualizzazione delle varianti di geni associate ad una data condizione.

```

1  SELECT m.codicegene, m.idorganismo, v.nome
2  FROM manifestazioni m, varianti v
3  WHERE v.codicegene = m.codicegene
4  AND v.idorganismo = m.idorganismo
5  AND m.condizione = <nome condizione>
6  ORDER BY m.codicegene, m.idorganismo

```

**Operazione 8:** Elenco ricerche aperte di un ricercatore.

```
1  SELECT t.codice, ra.nome, ra.datainizio
2  FROM ricercatori r, appartenenze a, team t, ricercheaperte ra
3  WHERE a.idricercatore = r.codice
4  AND a.idteam = t.codice
5  AND ra.idteam = t.codice
6  AND r.nome = <nome>
7  AND r.cognome = <cognome>
8  ORDER BY ra.datainizio
```

**Operazione 9:** Creazione di un nuovo team.

```
1  INSERT INTO team VALUES (<codice team>, <titolare>)
```

Per ogni ricercatore da inserire nel team si esegue:

```
1  INSERT INTO appartenenze values (<ricercatore>,<team>)
```

**Operazione 10:** Ricercatore che complessivamente ha lavorato su più geni in tutta la rete.

```
1  SELECT TOP 1 WITH TIES r.nome, r.cognome, r.codice, COUNT(
2  distinct s.idgene, s.idorganismo) AS "Numero geni"
3  FROM ricercatori r, appartenenze a, team t, ricerchechiuse rc,
4  studicompletati s
5  WHERE r.codice = a.idricercatore
6  AND t.codice = a.idteam
7  AND rc.idteam = t.codice
8  AND s.nomericerca = rc.nome
9  GROUP BY r.nome, r.cognome, r.codice
10 ORDER BY "Numero geni"
```



## 4 Sviluppo dell'applicativo

### 4.1 Architettura e funzionalità

L'applicativo è una web app basata su un server Node.js. Il server comunica con un database MySQL locale attraverso la libreria di ORM Sequelize. Questo tool permette sia di rappresentare le tabelle come modelli, sia di eseguire direttamente query MySQL.

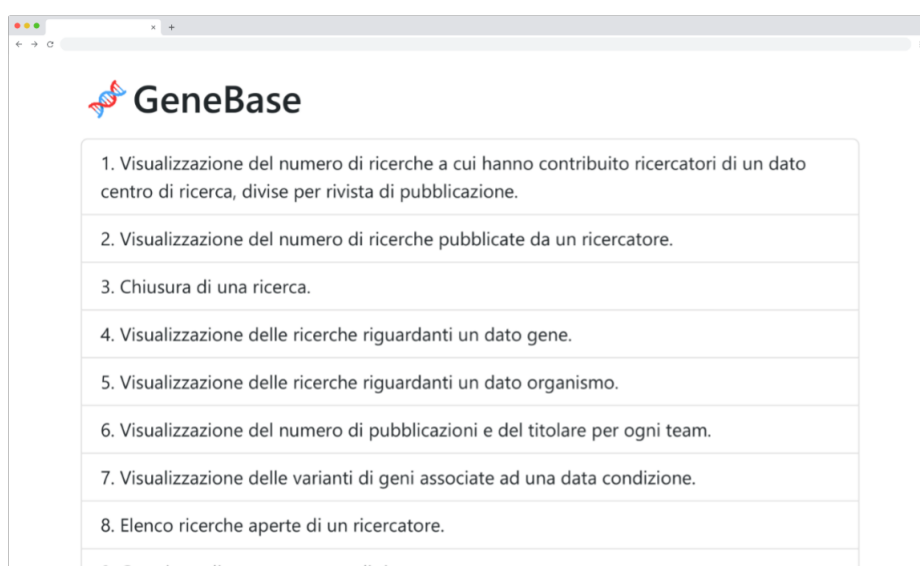
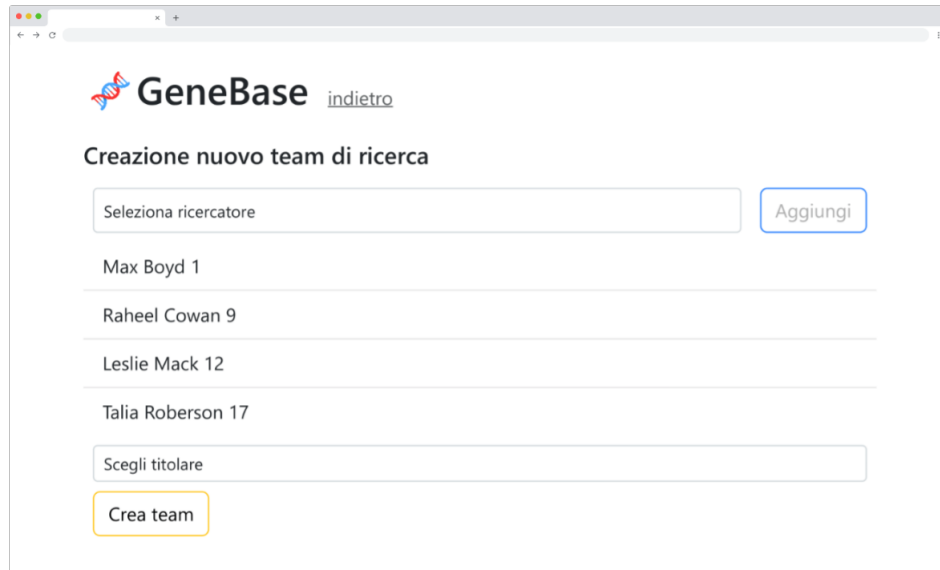


Figura 4.1: Schermata principale

Nel caso sia necessario, le pagine del sito richiedono di compilare un form, il quale manda i dati inseriti all'API del server che a sua volta esegue la query, restituendo alla pagina il risultato.



The screenshot shows a web browser window with the GeneBase logo and a navigation link 'indietro'. The main heading is 'Creazione nuovo team di ricerca'. Below it is a form with a dropdown menu labeled 'Seleziona ricercatore' and an 'Aggiungi' button. The dropdown menu is open, showing a list of researchers: 'Max Boyd 1', 'Raheel Cowan 9', 'Leslie Mack 12', and 'Talia Roberson 17'. Below the list is a 'Scegli titolare' dropdown menu and a 'Crea team' button.

Seleziona ricercatore
Max Boyd 1
Raheel Cowan 9
Leslie Mack 12
Talia Roberson 17

Scegli titolare

Crea team

Figura 4.2: Pagina con form per la creazione di un nuovo team

I controlli necessari sono eseguiti dall'applicativo prima di eseguire le query. Le funzionalità messe a disposizione dall'applicativo sono quelle elencate nella fase di analisi, alle quali sono state aggiunte operazioni banali per la visualizzazione dei dati all'interno del database.