Q&A RareSeq- Important

Pipeline incorporating the DBs
Connecting to DBs
Via SQL? Work on this
Drugs for the DEG
Find the pertubants

Pipeline incorporating the DBs

1. Data Preprocessing:
    - Obtain RNA-seq data from patient samples.
    - Perform quality control and filtering to remove low-quality reads or artifacts.
    - Normalize the data to account for sequencing depth and sample-to-sample variation.
2. Differential Gene Expression Analysis:
    - Compare gene expression levels between patient samples and control samples using appropriate statistical methods (e.g., edgeR, DESeq2, limma).
    - Identify differentially expressed genes (DEGs) that show significant changes in expression between the two groups.
3. Query CMAP and L1000 Databases:
    - Utilize the Connectivity Map (CMAP) database and its updated version, L1000, which provide gene expression profiles of various drugs and their effects on gene expression.
    - Query the databases using the identified DEGs to find drugs that induce gene expression changes opposite to the disease signature.
    - Obtain a ranked list of potential drugs based on their similarity to the disease signature reversal.
4. Query Creeds Database:

- Access the Creeds database, which contains gene sets related to biological processes, pathways, and diseases.
- Utilize the identified DEGs to query the Creeds database and find relevant gene sets associated with the disease.
- Extract drugs associated with the enriched gene sets and prioritize them as potential therapeutic options.

5. Query Recursion Database:
    - Access the Recursion database, which provides a comprehensive library of small molecules and their effects on cellular phenotypes.
    - Utilize the identified DEGs to query the Recursion database and identify small molecules that reverse the disease-associated phenotypes.
    - Prioritize small molecules based on their efficacy in reversing the cellular phenotypes associated with the disease.
6. Drug Prioritization:
    - Combine the results from querying the CMAP/L1000, Creeds, and Recursion databases to generate a comprehensive list of potential therapeutic drugs.
    - Rank the drugs based on their relevance to the identified DEGs, gene sets, and cellular phenotypes.
    - Consider additional factors such as drug properties, clinical relevance, and available therapeutic options to further refine the prioritization.

Here's a pseudo code representation of the pipeline:

```
# Step 1: Data Preprocessing
preprocess_data()

# Step 2: Differential Gene Expression Analysis
differential_expression_analysis()

# Step 3: Query CMAP and L1000 Databases
cmap_l1000_query_results = query_cmap_l1000_databases(degs)

# Step 4: Query Creeds Database
creeds_query_results = query_creeds_database(degs)

# Step 5: Query Recursion Database
```

```
recursion_query_results = query_recursion_database(degs)

# Step 6: Drug Prioritization
potential_drugs = prioritize_drugs(cmap_l1000_query_results, creeds_query_results, recursi
on_query_results)

# Display or output the list of potential therapeutic drugs
display_potential_drugs(potential_drugs)
w
```

Connecting to DBs

The L1000 database is a valuable resource for querying gene expression profiles and identifying potential therapeutic drugs. To connect to and query the L1000 database, you can follow these general steps:

1. Access the L1000CDS^2 website: The L1000CDS^2 (LINCS L1000 Characteristic Direction Signature Search Engine) is an online platform that provides access to the L1000 database. Visit the L1000CDS^2 website at https://clue.io/l1000cds2/ to access the database and query interface.
2. Perform a Gene Set Query:
    - On the L1000CDS^2 website, click on the "Explore Signatures" tab.
    - Enter your gene set of interest, which could be the list of differentially expressed genes (DEGs) identified from your RNA-seq analysis.
    - Select the appropriate gene ID type (e.g., Ensembl, Entrez, Gene Symbol) for your input gene set.
    - Choose the desired perturbation type (e.g., upregulated, downregulated) to specify the direction of gene expression changes you want to explore.
    - Submit the query to obtain results.
3. Analyze and Interpret Results:
    - The L1000CDS^2 platform will provide you with a ranked list of small molecules or drugs that induce gene expression changes similar to or opposite to your input gene set.
    - Examine the results, including the similarity score, p-value, and other relevant information, to evaluate the potential therapeutic relevance of the identified

small molecules.

- Consider additional information, such as drug properties, mechanisms of action, and known therapeutic applications, to further assess the suitability of the identified drugs.

The L1000CDS^2 website provides additional features and functionalities, such as exploring predefined gene sets, visualizing signature similarities, and accessing detailed information about individual small molecules and their gene expression profiles. Take advantage of these resources to gain further insights into the results and to aid in the interpretation of the findings.

It's important to note that the L1000 database and its associated tools may evolve over time. Therefore, it's recommended to refer to the official L1000CDS^2 website for the most up-to-date instructions, guidelines, and available features.

Via SQL? Work on this

Yes, it is possible to query the L1000 database using SQL. The L1000 database provides a downloadable file called "L1000FWD.db," which is a SQLite database containing gene expression data and associated metadata.

Here's a general outline of how you can query the L1000 database using SQL:

1. Download the L1000FWD.db file: Visit the L1000CDS^2 website (https://clue.io/l1000cds2/) and navigate to the "Download" section. Download the "L1000FWD.db" file, which contains the SQLite database.
2. Set up SQLite: Install SQLite on your local machine if you haven't already. SQLite provides a command-line interface for executing SQL queries.
3. Connect to the L1000FWD.db database: Open the command-line interface or terminal and navigate to the directory where you downloaded the "L1000FWD.db" file. Use the following command to connect to the database:

```
sqlite3 L1000FWD.db
```

4. Explore the available tables: Once connected to the L1000 database, you can list the available tables using the following command:

```
.tables
```

This will display the table names that you can query.

5. Write SQL queries: You can now write SQL queries to retrieve specific information from the L1000 database. Here's an example query to retrieve gene expression data for a specific gene:

```
SELECT
  gene_symbol,
  perturbation,
  perturbation_type,
  fold_change
FROM
  gene_info
WHERE
  gene_symbol = 'YOUR_GENE_SYMBOL';
```

Replace `'YOUR_GENE_SYMBOL'` with the gene symbol you are interested in.

6. Execute the SQL query: Copy and paste your SQL query into the SQLite command-line interface and press Enter to execute the query. The results will be displayed in the command-line interface.

By utilizing SQL queries, you can retrieve various information from the L1000 database, including gene expression data, perturbations, metadata, and more. It allows for more advanced and customized querying based on specific criteria or research needs.

Note that the specific table names, columns, and available data may vary depending on the version of the L1000 database you are using. It's recommended to refer to the documentation or accompanying resources provided with the database to understand the schema and structure of the database for your specific version.

Drugs for the DEG

Utilize the LINCS L1000 database to identify potential drugs that may have corrective effects on the dysregulated genes represented in your list of differentially expressed genes (DEGs). Here's an approach to accomplish this:

1. Retrieve Gene Expression Signatures:

- Query the LINCS L1000 database using your DEG list to retrieve gene expression signatures associated with those genes.
- Obtain the sig_id values or other relevant identifiers for the gene expression signatures linked to your DEGs.

2. Find Perturbagens with Inverse Expression Profiles:
    - Use the retrieved sig_id values to identify perturbagens (drugs or compounds) in the LINCS L1000 database that exhibit inverse gene expression profiles compared to your DEGs.
    - Look for perturbagens that show downregulation of genes that are upregulated in your DEG list, and vice versa.
    - These perturbagens may have the potential to correct or normalize the dysregulation observed in your gene expression data.

PERT_DESC

HAVE NOT COME ACROSS A WAY TO DO THIS SO FAR

```
# DEG -> Gene Sig_ID -> L1K for inverse gene action
```

Find the pertubants

Utilize the LINCS Query interface provided by the LINCS Data Portal. Here's how you can access the relevant information:

1. Visit the LINCS Data Portal website: Go to the LINCS Data Portal at https://lincsportal.ccs.miami.edu/.
2. Access the LINCS Query interface: On the LINCS Data Portal homepage, click on the "LINCS Query" tab located in the top navigation menu.
3. Enter sig_id in the search field: In the LINCS Query interface, there should be a search field where you can enter your sig_id value. Input the sig_id associated with your gene expression profile or perturbagen of interest.
4. Retrieve perturbagen information: Submit the query, and the LINCS Query interface will fetch the information related to the provided sig_id. This includes details about

the perturbagen, such as the name of the drug or compound used in the experiment.

5. Explore perturbagen details: Review the retrieved information to gather insights into the perturbagen, its properties, mechanisms of action, and any available information on its therapeutic indications or associations with the dysregulated genes or pathways.

By following these steps, you should be able to access the list of drugs or perturbagens associated with the sig_id in the LINCS L1000 database. Remember that the sig_id uniquely identifies the gene expression profile or perturbation experiment and can be used to retrieve relevant information about the corresponding drugs or compounds used in those experiments.
