Hide&Seq Cornerstone Meeting/S

Goal/ Client Requirements

- Drug for a rare genetic disease, personlized therapeutics.
- RNAseq of the patient will act as the input. Optimum drugs are the required output
- Main database: LC1000, CMAP etc (stores drug activity information, among other stuff)

Other things to keep in mind:

- $Business perspective$- Is it novel enough for us to commercialise this tool and monitize it?

(……no, already out there. And if we want to patent a service, will take more than a month)

- Timeline- Can it be done by August for the client? If not then, when? If not for them, then?

Envisioned Outline

1. RNA Seq Data Analysis: Collect RNASeq Data from patients, DEG with a target expression. Aim: Identity DEG from the sample, dysregulated compared to the target. Tools that can be used: edgeR, DESeq2, or limma. Data PreProcessing
2. GSEA: Score and ranking of of DEG

▼ GSEA?

- It helps to identify whether a pre-defined set of genes or gene sets show statistically significant differences between different experimental conditions or groups.
- Gene sets: combination of genes required or present in a particular biological pathway. Gene sets can represent biological pathways, molecular functions, cellular components, or disease-associated signatures.
- Data: RNASeq- quantification of the RNA transcription
- Ranked Gene List: Rank the most differential expressed gene, creates a set
- Enrichment Score calculation: Score of over/under expression of the ranked gene list.
- Bottom line: Some statistical tests on gene seq/s that ranks them at the end of it.

3. Drug Target Database: Links the drugs for the dysregulated genes: This is where things begin to get complicated. Query DBs
4. Rank for the Query (drug affinity rank)
5. Output

Lets say in the case of hypercholesterolemia, the defective genes would be- LDLR (Low- Density Lipoprotein Receptor), PCSK9, APOB, LDLRAP1, APOE. We can consider these as the list of DEGs, in general any defect to these genes/ genes set would lead to the onset of the disease.

```
Pseudo how it might look.

# Step 1: Data Preprocessing
preprocess_data()

# Step 2: Differential Gene Expression Analysis
differential_expression_analysis()

# Step 3: Query CMAP and L1000 Databases
L1000_query_results = queryl1000_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(1000_query_results)

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

Steps Breakdown

▼ 1) RNA Seq Data Analysis/ Data PreProcessing

DEG: Genes or RNA transcripts that echibit different expression levels between two or more groups. In our context: patient and control (defect vs healthy). This gives us an insight into the molecular mechanisms inderlying the disease or conditions.

To make things easier, transcriptomics == RNASeq study

Steps involved:

1) Data preprocessing- Quality control, filtering, normalization of the raw genetic or RNA data to remove noise.

2) Quantification- Abundnace of RNA transcripts. Mapping the sequenced reads to a ref genome. Read counts or estimated expression values to each gene or transcripts

3) Statistical analysis- edgeR, DESeq2, or limma can be used to identify the DEGs. Reads the count data from the quantification, factors include sample size, variablility b/w samples etc to determine the DEGs.

4) Enrichment analysis- *see above block*

Tools for DEG: DESeq2, edgeR, limma, voom, Cuffdiff, NOISeq. All R packages (the most used are DESeq2, edgeR, and limma)

Show example of differential gene expression code from sepsis

Steps involved:

- install packages, libs and dependencies.
- Build the count matrix. AnnotationDbi package

- Load the count matrix- gene names/ IDS as the row naes and sample counts as the columns
- limma package to fit a linear model to the gene expression data.
- Create the DGE_list obj
- Normalize the DGE
- topTags() function to obtain list of the DEG from the LRT(Likelihood Ratio Test using glmLRT())

Doubts: How do we batch this? Script has to be run for every RNASeq input? Automation? Script iteration for everyrun

What else? Galaxy Servers- No novelty, no script..

▼ 2) DB Connections

- Google Cloud Connection
- Maayan Lab Cloud- https://maayanlab.cloud/sigcom-lincs/#/SignatureSearch/Set

Other method

Pre-Made tools

Concluding Remarks

- Every tool is already out there and can take anyone a few hours of digging around to get to the required step. No real novelty that can be commericalised

- If a one click/ submission end to end tool is to made it would defiantely take more than a month and a lot more intellectual prowers on the table. Example: One submission get the result.

This would mean building pipeline, infrastrcture, and compute that takes care of end to end processing. Can this be felxible enough to be a one time build that can work for every/ any sample?

Guides/ Links

Data Portal

https://maayanlab.cloud/sigcom-lincs/#/SignatureSearch/UpDown

GCP Connection

https://lincsportal.ccs.miami.edu/signatures/bigquery

GCP Doc

https://docs.google.com/document/d/1Bddq9cNGzrfEWSRlMy36JC3yD6c8-BH-6K-Qvs3__M0/edit

---

Pages

Client Engagement