TECHNICAL NOTE

Gene Signature Comparison on Polly

OVERVIEW

Identifying datasets that match a particular Gene signature is a powerful strategy deployed by biopharma R&D teams. This can be used for many applications, for example, identifying potential new uses for existing drugs. At its core, this technique involves comparing the differential gene expression profiles with a user-defined gene signature. This gene signature can then be used to search public databases of gene expression data for other drugs or compounds that can revert the disease signature, indicating a potential therapeutic effect.

Extracting relevant signatures from public databases, however, can be a challenging task due to the varying processing pipelines, syntaxes, schemas, and metadata annotations used at the source. We address these challenges through Polly's RNA-Seq Omixatlas, the world's most comprehensive and curated set of more than 45,000 transcriptomics datasets

In this technical note, we discuss how users can perform signature comparisons using Polly's RNA-Seq Atlas.

HOW POLLY HELPS?

Polly's RNA-Seq OmixAtlas (OA) contains 45,000 highly curated gene expression studies collected from GEO (Gene Expression Omnibus). This richly curated resource provides a useful base for researchers looking to find datasets with similar transcriptional profiles to their gene sets of interest. All datasets on Polly are:

Consistently Processed

End to end data processing (identifier mapping, QC, normalization and alignment) is orchestrated through the Kallisto pipeline. Consistent processing on the entire Atlas allows samples to be reliably combined into cohorts and used to develop RNA-Seq signatures.

Enriched with Metadata

All datasets are enriched with over 21 searchable metadata fields (disease, gene, tissue, drug, control etc.) at the dataset, sample and feature level. This means that users can easily run SQL queries to find datasets with normal to disease comparisons and/or define cohorts of their choice.

OUR APPROACH

GENERATE QUERY SIGNATURE(S)

Users can employ their own methodology for arriving at a gene signature. A query signature will comprise of the following:

- A list of genes that were significantly differentially expressed in the experiment with Log Fold Change, p-values and adjusted p-values
- Query Methodology: Our experts work with your scientists to capture transcriptome profile and generate queries (gene signature vectors) that will be run on Polly's signature database.

  Example of a query: Given an input of gene set and Log Fold Change values, search for all datasets that show maximum cosine similarity scores with the input genes and their differential expression results

CREATING A SIGNATURE DATABASE DERIVED FROM DATA ON POLLY

- Experiment designs of all RNA-Seq datasets on Polly are evaluated. Datasets containing control and perturbation samples are then extracted from this collection.
- A differential expression analysis is performed on these cohorts of control and perturbed samples
- The resulting differential computation includes a distinct ID for the Differential Comparison, Gene Names, and their values for Log Fold Change, p-Value, and adjusted p Values.
- These results, along with metadata such as perturbations, controls, disease, drug, or genotype, are indexed on Polly in queryable. GCT files.
- Simultaneously a database of gene signatures vectors is created based on your choice of thresholds for Log Fold Change and adjusted p-value cut-offs.

IDENTIFYING DATASETS SIMILAR TO THE QUERY SIGNATURE

This signature database can now be queried to identify datasets with similar transcriptional profiles to the Query Signature. For instance, users can run complex SQL queries to identify:

- Datasets where diseased samples are compared to normal and are similar to the query signature
- Datasets where a particular disease is treated with some drug and shows a reverse profile to the query signature.

RANKING THE OUTPUT

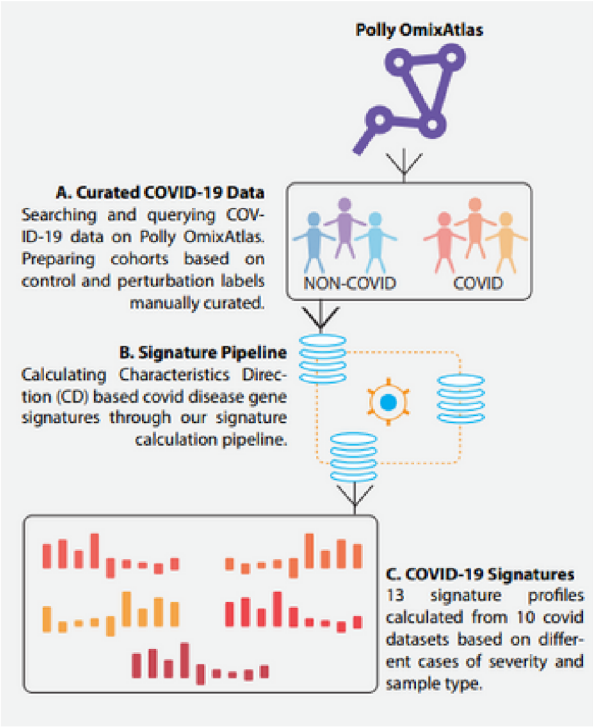
- Our experts work with your team to identify your preferred method for finding similar or dissimilar transcriptome profile from the database and rank these. We can employ any of the popularly known similarity scores such as Jaccard index, cosine similarity, concordance/discordance ratio etc.
- Additionally we also create a random query gene signature with same number of differentially expressed genes and obtain the distribution of similarity scores to serve as a background distribution. These help in identifying significant similarity scores for the given query signature.

CASE STUDY

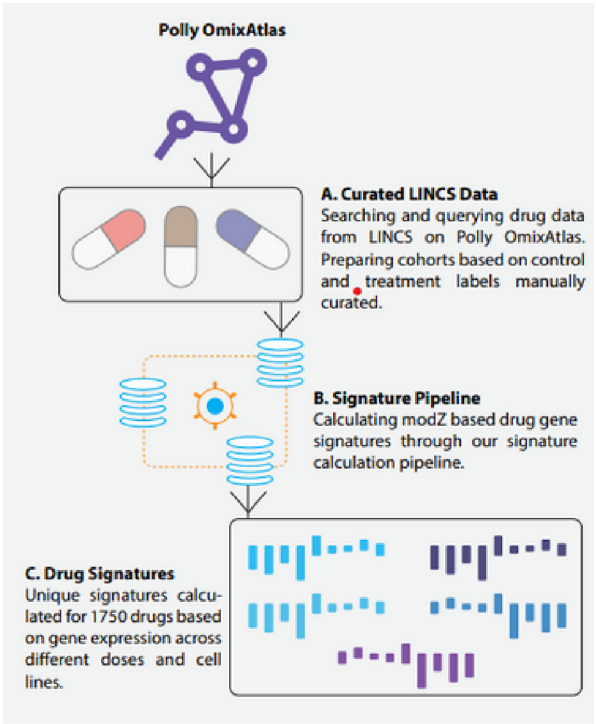
PREDICTING SYNERGISTIC DRUG COMBINATIONS FOR COVID-19

METHODLOGY

We used signature reversal and multivariate gene expression signatures to identify potential drug combinations for COVID-19. To do this, publicly available transcriptomics data from COVID-19 studies and drug signatures from LINCS were compiled, processed and curated. All datasets were ingested through Polly's proprietary curation pipeline, enriched with ontology-backed metadata and engineered to a queryable .gct format.



Generating Covid 19 Query Signatures



Creating a Database of Drug Signatures

RESULTS

- 37 reference drug candidates based on similarity between drugs and disease profile were identified on Polly. Drugs with low drug-disease similarity, across most disease profiles were shortlisted.
- Drug combinations were evaluated based on similarity to reference drugs and disease signature reversal. 28 combinations with low reference drug similarity and high disease signature reversal were prioritized.

ABOUT US

Polly delivers ML-ready biomedical molecular data that is curated to accelerate drug discovery. Hosting a rich repository of more than 45,000 RNA-Seq datasets, Polly is a customizable platform that assists with the comprehensive analysis of integrated biomedical data.
For help with performing gene signature comparisons, get in touch with us at sabu.george@elucidata.io