Database Final Project

Group 2

李文堯、陳怡婕

Outline

1. Background of research topics
2. Methods
3. Data source and analysis pipeline
4. Construction of a database system
5. E-R model
6. Web interface

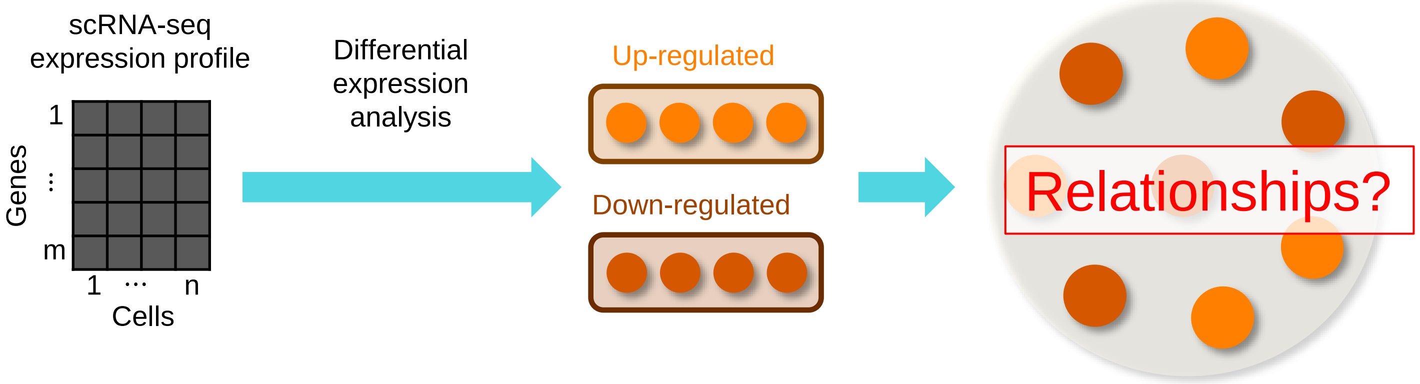
Introduction

- Cell-type identification heavily relies on an optimal clustering result, which is highly subjective due to the lack of ground-truth labels.
- The labeling of cell types of a scRNA-seq dataset require comprehensive prior knowledge of marker genes for each cell type.[1]
- Most of methods regard gene expression as the input feature and rarely take the relationships among genes into consideration.

[1]A comparison of single-cell trajectory inference methods. Nat Biotechnol(2019) : 547-554

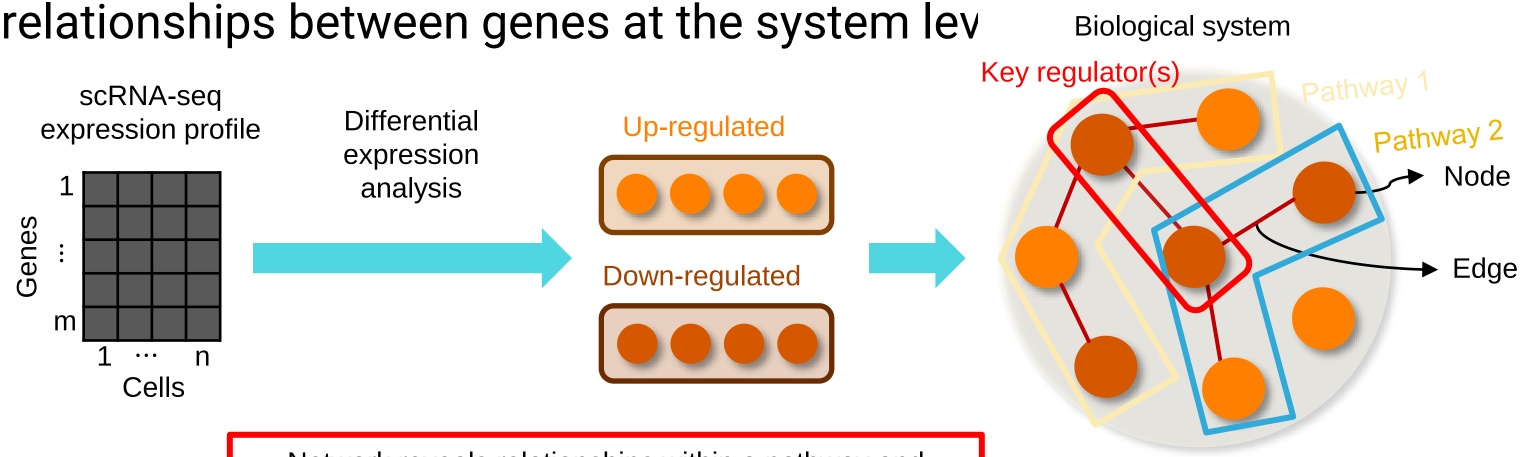
Network Biology helps understand the relationships between genes

- Traditional methods on scRNA-seq data are limited to “point” changes.



Network Biology helps understand the relationships between genes

- Traditional methods on scRNA-seq data are limited to “point” changes.
- Network Biology provides a comprehensive understanding of the relationships between genes at the system lev



Network reveals relationships within a pathway and cross-talk between pathways.

Node: Gene
Edge: Relationship between two genes (a gene pair)

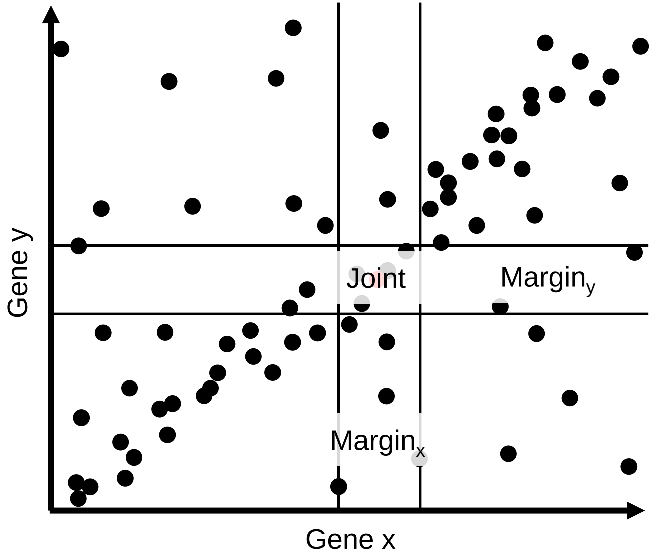
[1] Barabási, AL. et al. *Nat Rev Genet* 5, 101–113 [2] Cha, J. et al. *Exp Mol Med* 52, 1798–1808 (2020).

Aims

- Preserving the information contained in scRNA-seq data through networks allows the observation of interactions between genes.
- Take advantage of prior gene network to get a more meaningful low-dimensional representation of genes.
- Identify the cell type of individual cells base on single cell network.

SCN inference methods – CSN

- Cell-Specific Network (CSN) [1] constructs one network for one cell by quantifying *statistical independence*.



Under independence:

$$Pr(Joint) = Pr(Margin_x) \times Pr(Margin_y)$$

hence,

$$\rho = Pr(Joint) - Pr(Margin_x) \times Pr(Margin_y) = 0$$

When two genes are dependent:

$$Pr(Joint) > Pr(Margin_x) \times Pr(Margin_y)$$

hence,

$$\rho = Pr(Joint) - Pr(Margin_x) \times Pr(Margin_y) > 0$$

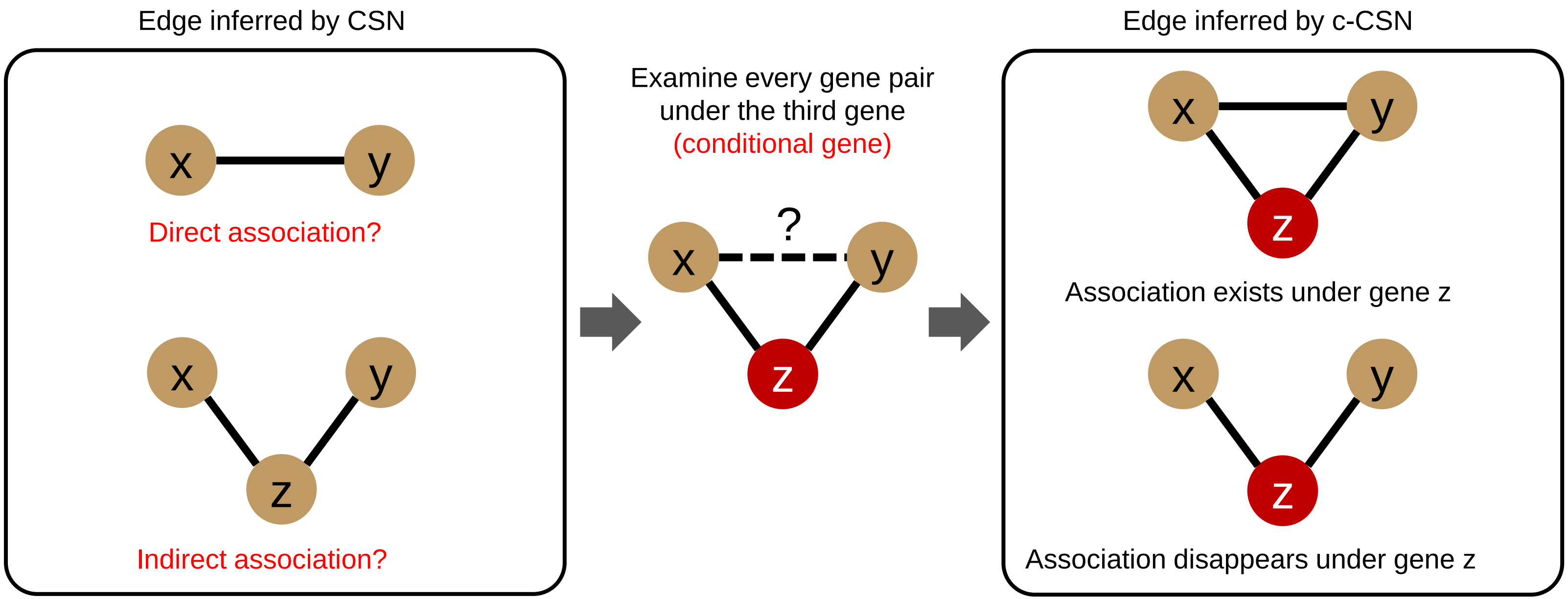
ρ serves as the statistics in CSN model.

$\rho \sim Normal, n \to \infty$

[1] Hao Dai et al. *Nucleic Acids Res*. 47(11):e62

SCN inference methods – c-CSN

- CSN suffers from overestimation of edges [1, 2].
- Conditional Cell-Specific Network (c-CSN) eliminates indirect associations by conditional independence [3].



Edge inferred by CSN

Direct association?

Indirect association?

Examine every gene pair under the third gene (conditional gene)

Edge inferred by c-CSN

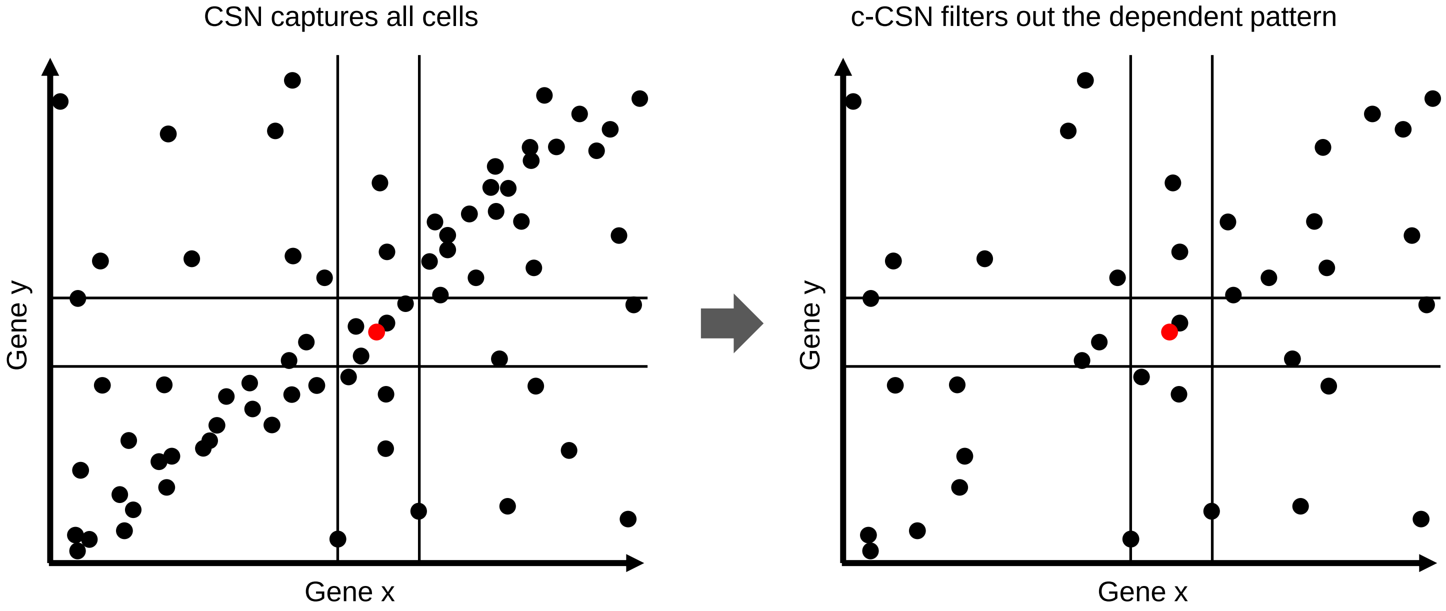
Association exists under gene z

Association disappears under gene z

[1] Li et al. *Genom Proteom Bioinf*. 19.2, 319-329 (2021).
[2] Hao Dai et al. *Zool Res*. 41(6): 599–604 (2020).
[3] Li et al. *Genom Proteom Bioinf*. 19.2, 319-329 (2021).

SCN inference methods – c-CSN

- Conditional Cell-Specific Network (c-CSN) eliminates indirect associations by conditional independence [1].



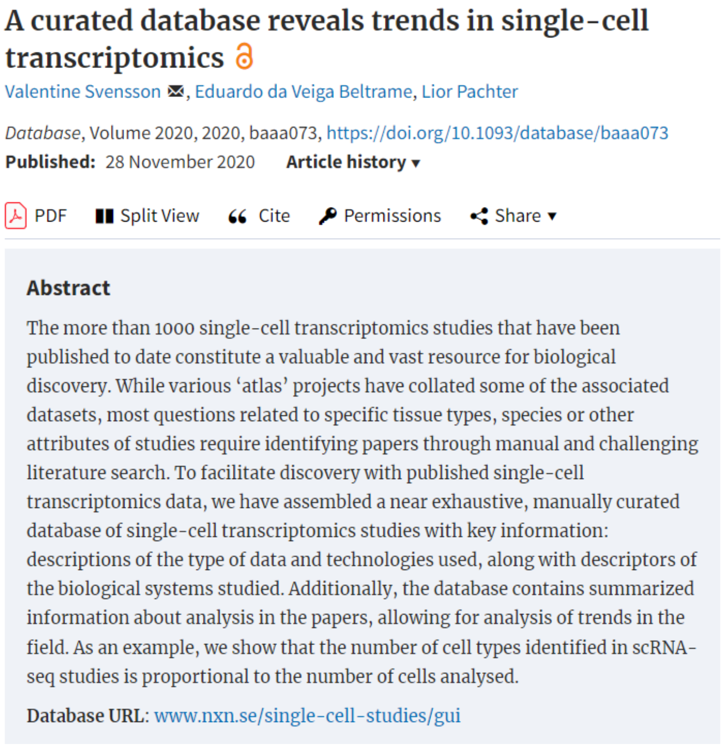
CSN captures all cells

c-CSN filters out the dependent pattern

[1] Li et al. *Genom Proteom Bioinf*. 19.2, 319-329 (2021).

Data

- **GEM Dataset**
  - Kim
  - Guo
  - Yan
  - LiNormal
- **Download from**: https://www.nxn.se/single-cell-studies
- **Data preprocessing**
  - Applying logarithm transformation (2 based)
  - Removing genes expressed in less than 10 cells



A curated database reveals trends in single-cell transcriptomics

Valentine Svensson, Eduardo da Veiga Beltrame, Lior Pachter

Database, Volume 2020, 2020, baaa073, https://doi.org/10.1093/database/baaa073

Published: 28 November 2020 **Article history** ▾

Abstract

The more than 1000 single-cell transcriptomics studies that have been published to date constitute a valuable and vast resource for biological discovery. While various 'atlas' projects have collated some of the associated datasets, most questions related to specific tissue types, species or other attributes of studies require identifying papers through manual and challenging literature search. To facilitate discovery with published single-cell transcriptomics data, we have assembled a near exhaustive, manually curated database of single-cell transcriptomics studies with key information: descriptions of the type of data and technologies used, along with descriptors of the biological systems studied. Additionally, the database contains summarized information about analysis in the papers, allowing for analysis of trends in the field. As an example, we show that the number of cell types identified in scRNA-seq studies is proportional to the number of cells analysed.

Database URL: www.nxn.se/single-cell-studies/gui

Identification of the cell types

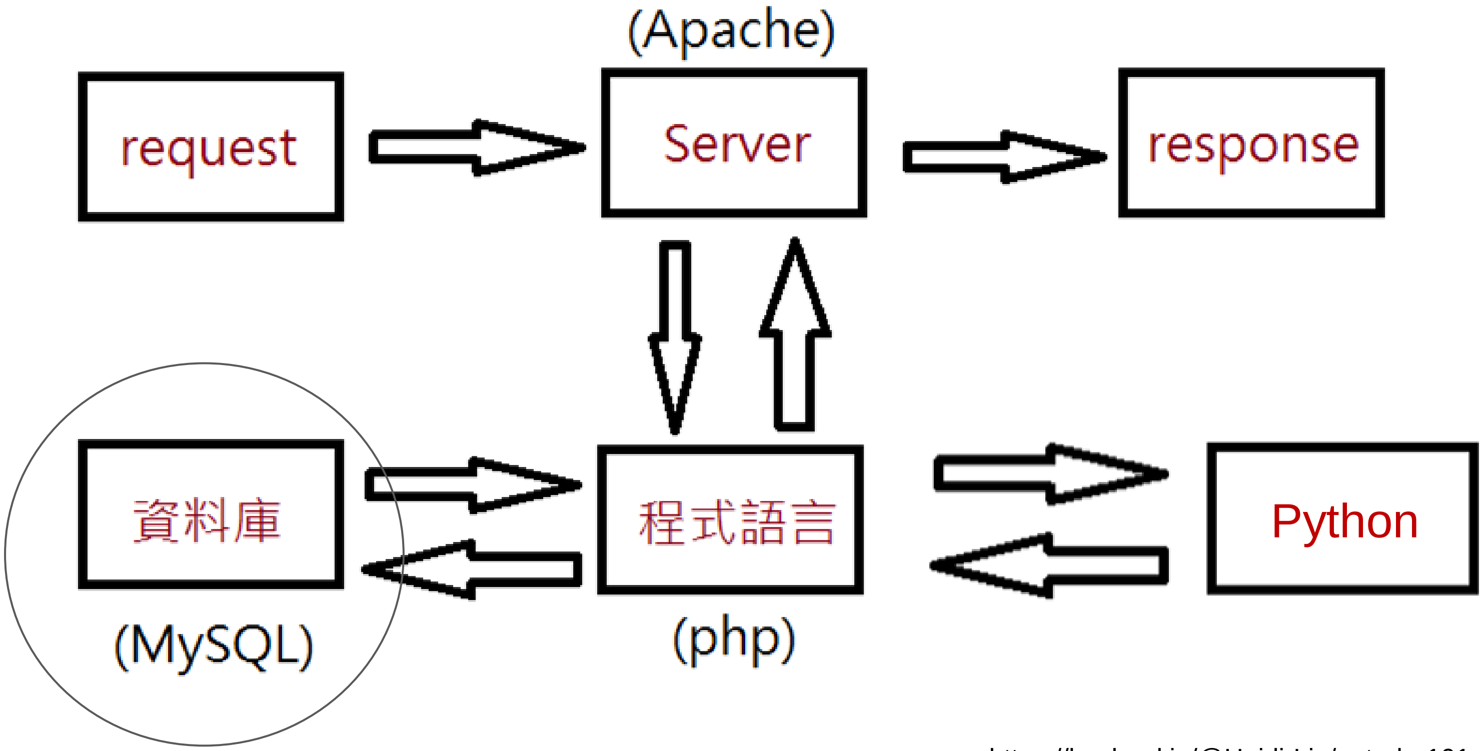
- Input data:
  - Gene expression matrix
  - Network degree matrix
- Dimension reduction:
  1. UMAP [1, 2]
- Clustering algorithm: Louvain clustering [3]
- Compare with Three difference clustering output
  - Paper base cell types
  - GEM base cell types
  - Network degree base cell types

[1] McInnes, Leland, John Healy, and James Melville. *arXiv preprint arXiv:1802.03426* (2018).

[2] Becht, Etienne, et al. *Nature biotechnology* 37.1 (2019): 38-44.

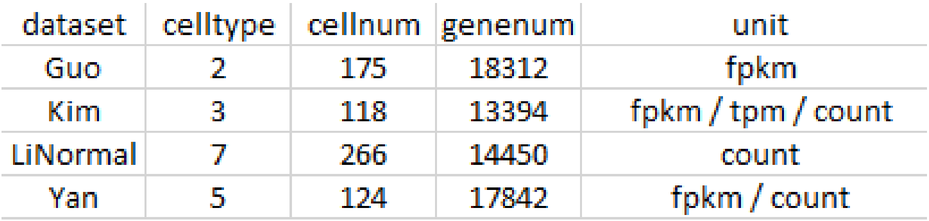
[3] Blondel, Vincent D., et al. *Journal of statistical mechanics: theory and experiment* 2008.10 (2008): P10008.

Construction of a Database System

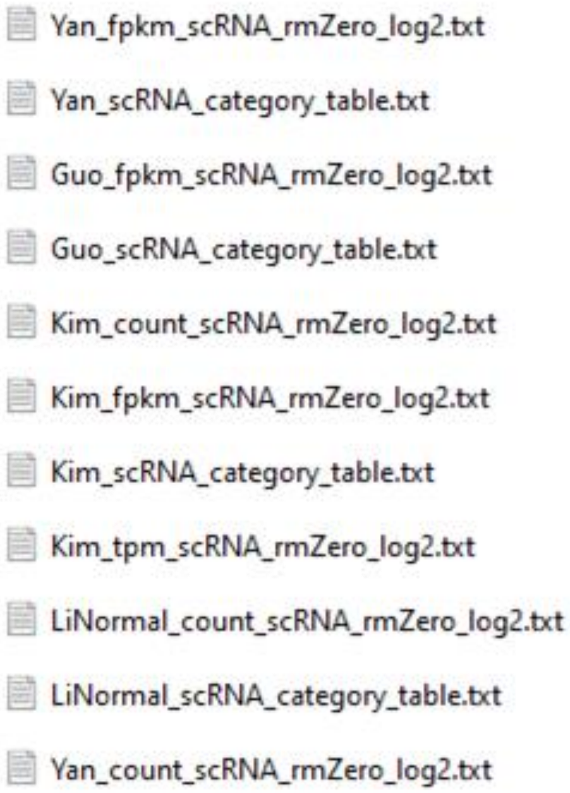


https://hackmd.io/@Heidi-Liu/note-be101-php-and-mysql

Construction of a Database System

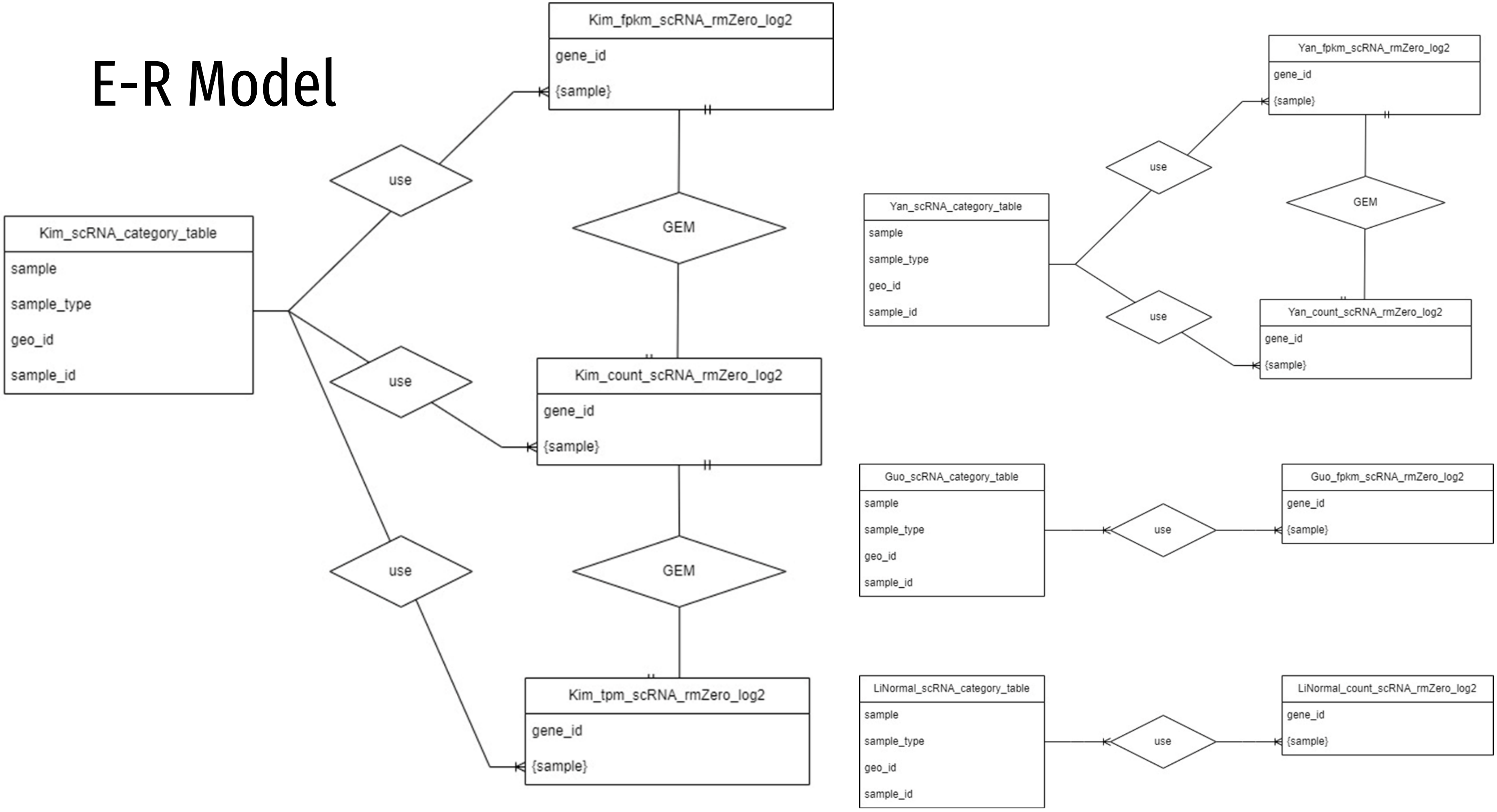


dataset	celltype	cellnum	genenum	unit
Guo	2	175	18312	fpkm
Kim	3	118	13394	fpkm / tpm / count
LiNormal	7	266	14450	count
Yan	5	124	17842	fpkm / count



Yan_fpkm_scRNA_rmZero_log2.txt
Yan_scRNA_category_table.txt
Guo_fpkm_scRNA_rmZero_log2.txt
Guo_scRNA_category_table.txt
Kim_count_scRNA_rmZero_log2.txt
Kim_fpkm_scRNA_rmZero_log2.txt
Kim_scRNA_category_table.txt
Kim_tpm_scRNA_rmZero_log2.txt
LiNormal_count_scRNA_rmZero_log2.txt
LiNormal_scRNA_category_table.txt
Yan_count_scRNA_rmZero_log2.txt

E-R Model



Kim_scRNA_category_table
sample
sample_type
geo_id
sample_id

use

Kim_fpkm_scRNA_rmZero_log2
gene_id
{sample}

GEM

use

Kim_count_scRNA_rmZero_log2
gene_id
{sample}

use

GEM

Kim_tpm_scRNA_rmZero_log2
gene_id
{sample}

Yan_scRNA_category_table
sample
sample_type
geo_id
sample_id

use

Yan_fpkm_scRNA_rmZero_log2
gene_id
{sample}

GEM

use

Yan_count_scRNA_rmZero_log2
gene_id
{sample}

Guo_scRNA_category_table
sample
sample_type
geo_id
sample_id

use

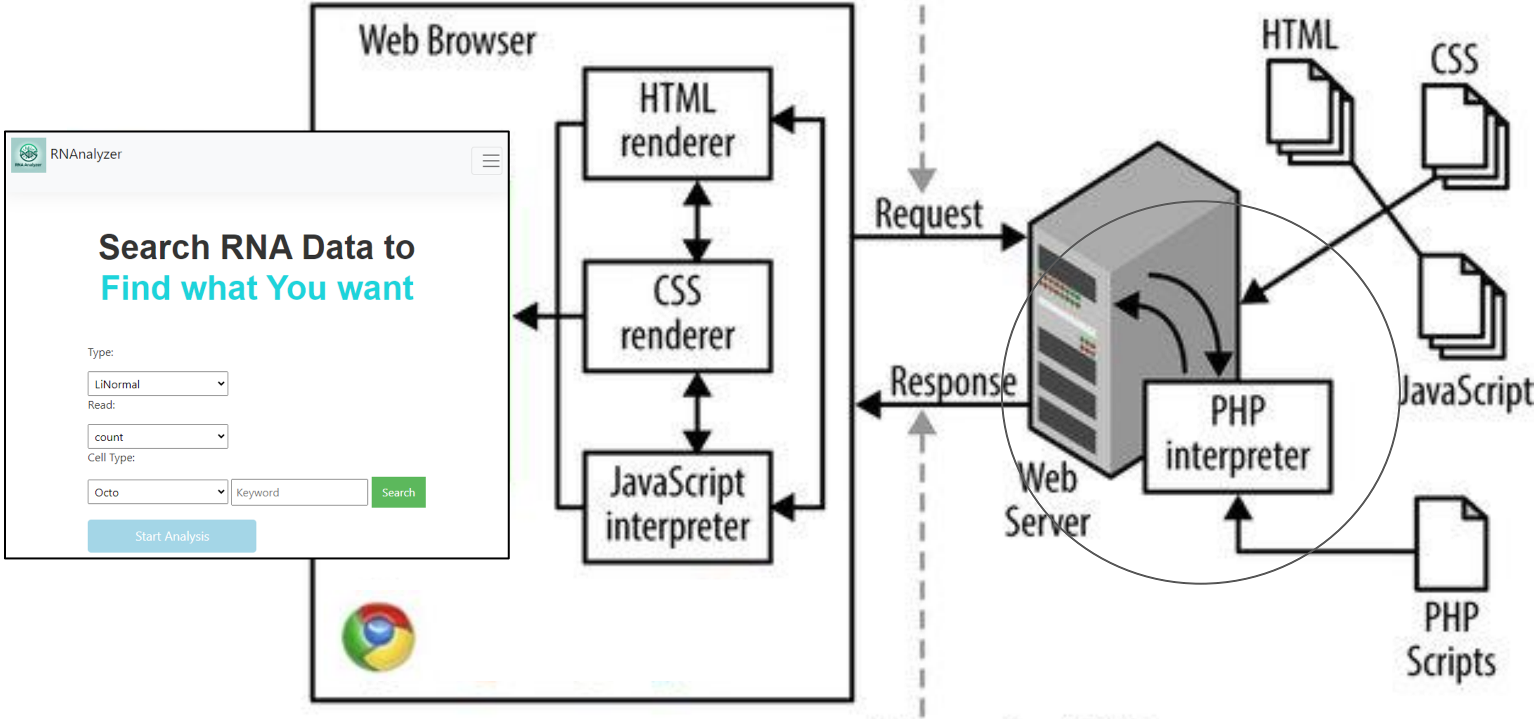
Guo_fpkm_scRNA_rmZero_log2
gene_id
{sample}

LiNormal_scRNA_category_table
sample
sample_type
geo_id
sample_id

use

LiNormal_count_scRNA_rmZero_log2
gene_id
{sample}

Could be for HTML, CSS, PHP or a combination.



RNAnalyzer

Search RNA Data to Find what You want

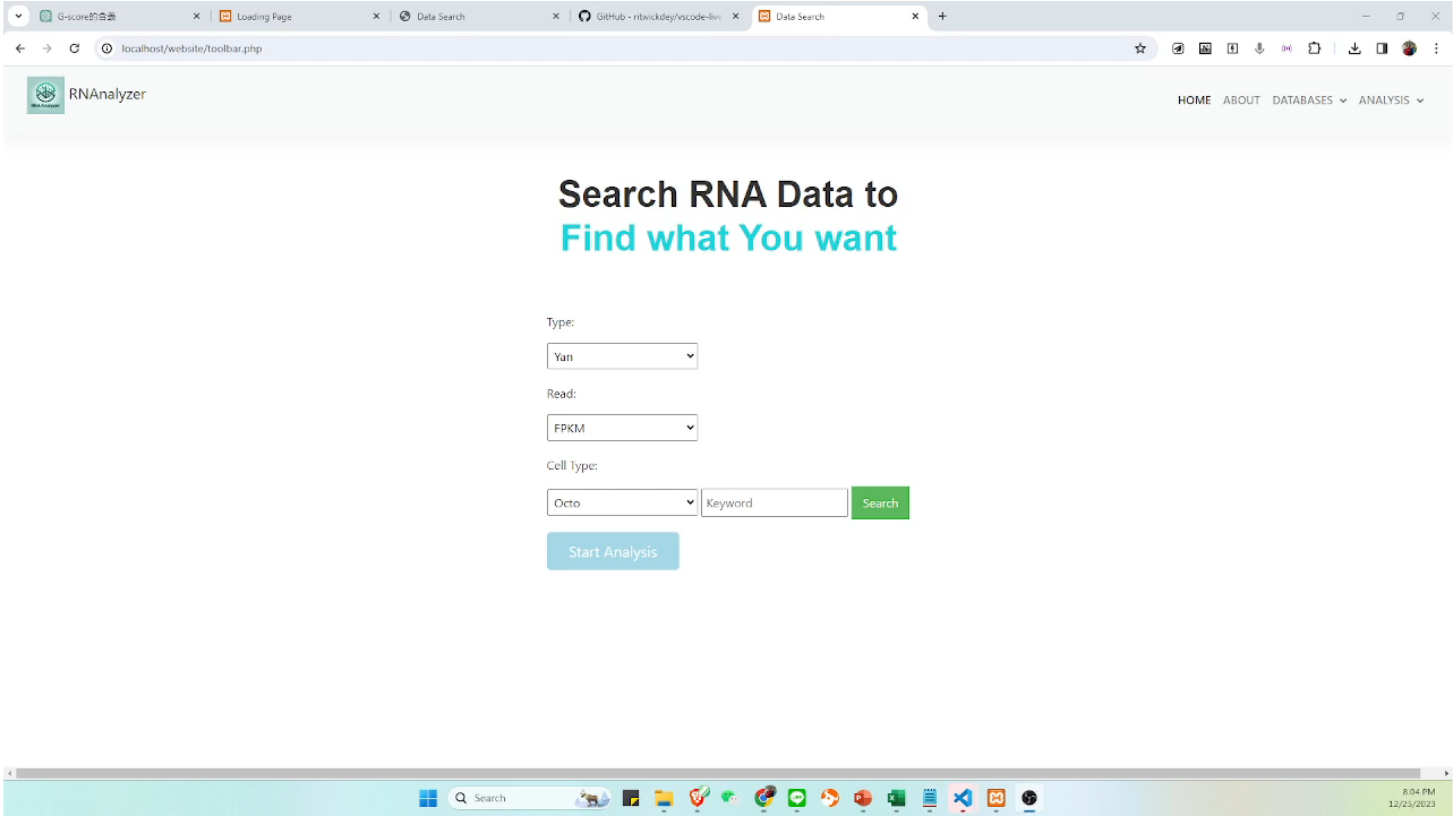
Type: LiNormal

Read: count

Cell Type: Octo

Response is not PHP, but the result of interpreting PHP, usually more HTML and CSS.

Video Demo



localhost/website/toolbar.php

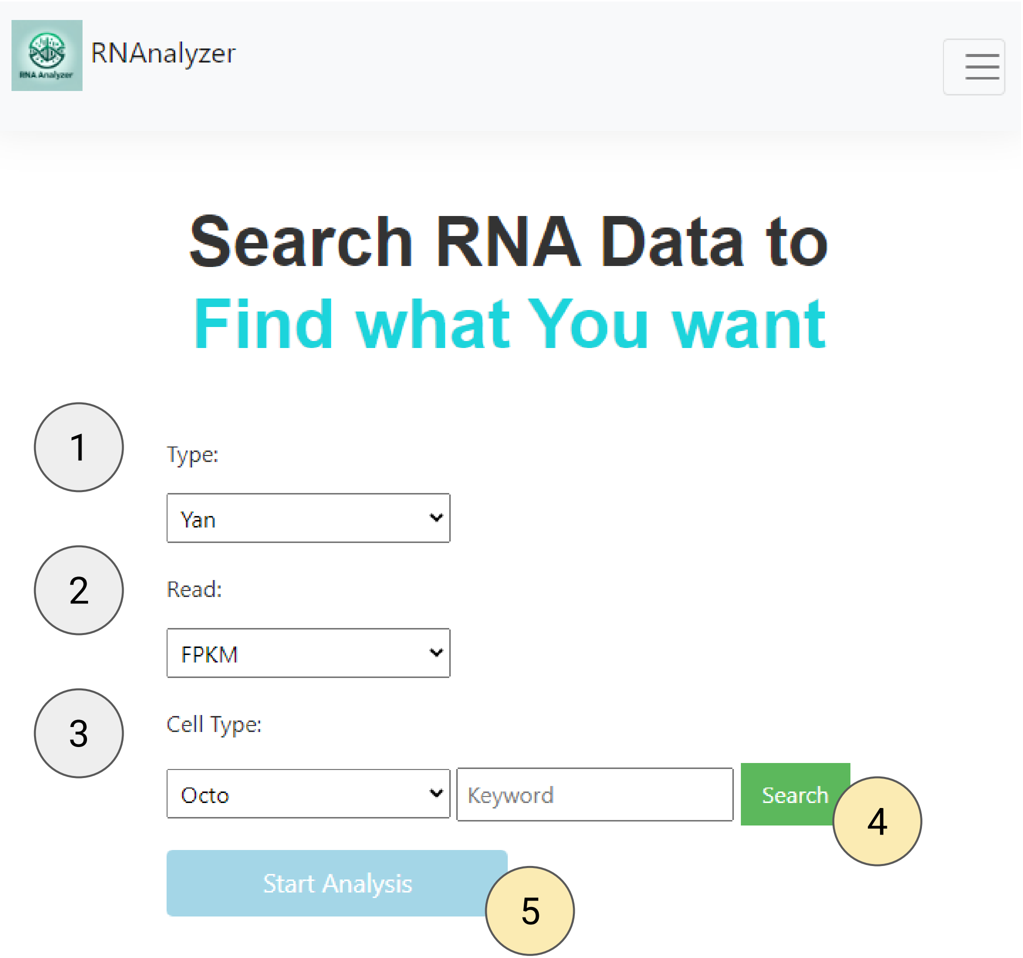
Search RNA Data to
Find what You want

Type: Yan

Read: FPKM

Cell Type: Octo

Search Page



RNAnalyzer

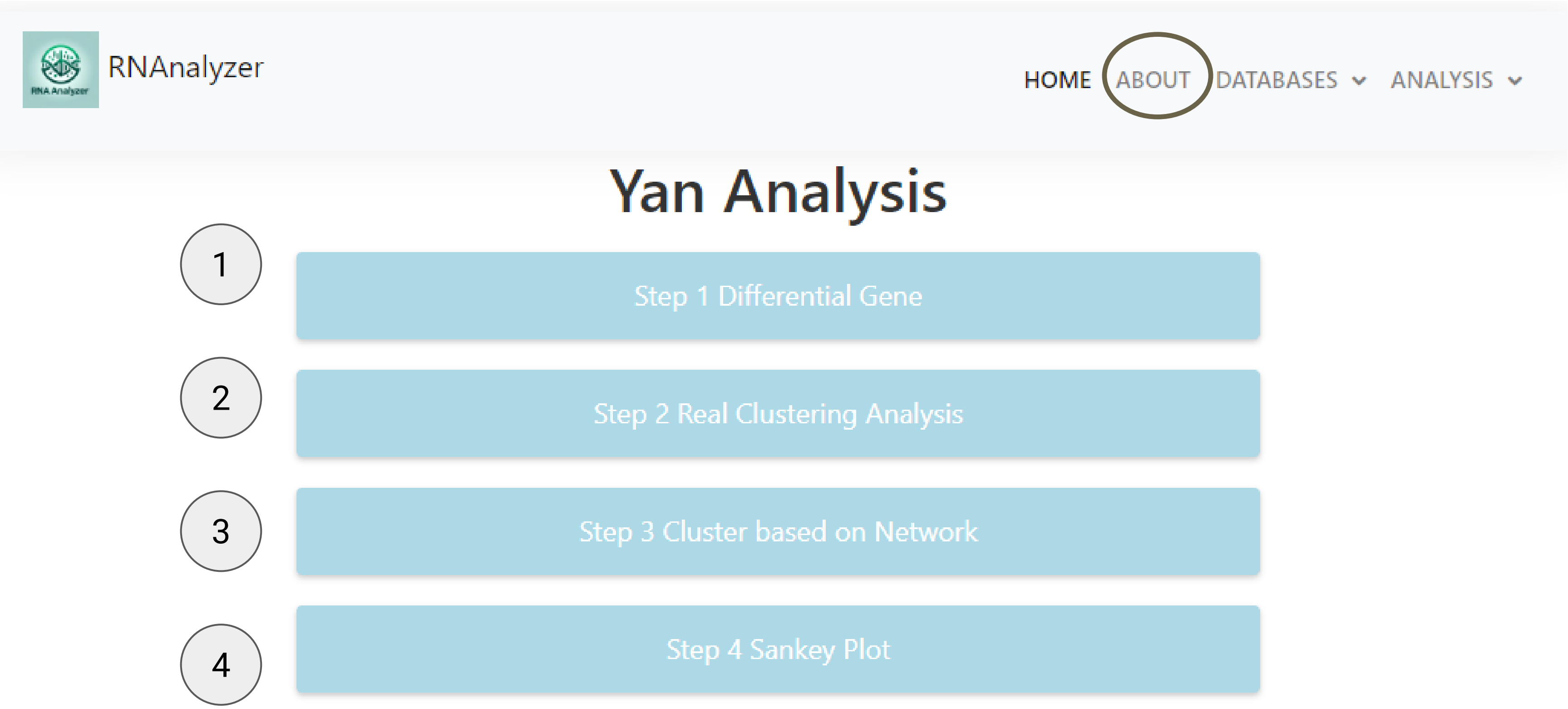
Search RNA Data to
Find what You want

Type: Yan

Read: FPKM

Cell Type: Octo

Analysis Page



Yan Analysis

1
Step 1 Differential Gene
2
Step 2 Real Clustering Analysis
3
Step 3 Cluster based on Network
4
Step 4 Sankey Plot

Thanks for your listening!