

# Technical Training #3: Paper Overview

Presented by: Annie, Erin, Alex, Isha, Yves, Goral



# How to Read & Understand a Scientific Paper

## General Tips:

- Different process than reading an article / blog / newspaper
- Read sections in a different order
- Should take notes (have a highlighter & pen ready)
- Read it multiple times
- Should follow up with more research
- May take longer than you think

# How to Read & Understand a Scientific Paper

## Steps:

1. **SKIM** the abstract
2. Read the **INTRODUCTION**
3. Summarize the background
  - a. Identify the BIG QUESTION(S) / purpose
4. Read the **RESULTS**
  - a. Focus on figures → you can get a lot of information from them
5. Read the **CONCLUSION / discussion**
6. Read the materials & methods
7. NOW read the abstract
8. Follow-up by reading similar papers

# What is the main purpose of the research?

- To find specific genetic alterations that likely cause colorectal cancer
  - Prognostic and predictive markers
    - **Prognostic-** identifies alterations and determines likelihood of developing disease, or a recurrence or progression of it.
    - **Predictive-** used in studying and determining appropriate therapeutic interventions; identifies who would have favorable or unfavorable outcomes if given a specific therapy based on biomarkers that are or are not found in others.
  - This paper focuses on prognostic markers for colorectal cancer

# State of current research in the field

- [2019 study using stool samples for noninvasive testing of biomarkers for CRC](#)
  - Choi HH, Cho YS, Choi JH, Kim HK, Kim SS, Chae HS. Stool-Based miR-92a and miR-144\* as Noninvasive Biomarkers for Colorectal Cancer Screening. *Oncology*. 2019;97(3):173-179. doi:10.1159/000500639
- [Using ceRNA networks to find markers for Triple-Negative breast cancer](#)
  - Liu Z, Mi M, Li X, Zheng X, Wu G, Zhang L. lncRNA OSTN-AS1 May Represent a Novel Immune-Related Prognostic Marker for Triple-Negative Breast Cancer Based on Integrated Analysis of a ceRNA Network. *Front Genet*. 2019;10:850. Published 2019 Sep 13. doi:10.3389/fgene.2019.00850
- [Finding lncRNAs \(long noncoding RNA\) associated with CRC tumorigenesis using ceRNA networks to develop outcome/risk scores for CRC](#)
  - Yang ZD, Kang H. Exploring prognostic potential of long noncoding RNAs in colorectal cancer based on a competing endogenous RNA network. *World J Gastroenterol*. 2020;26(12):1298-1316. doi:10.3748/wjg.v26.i12.1298

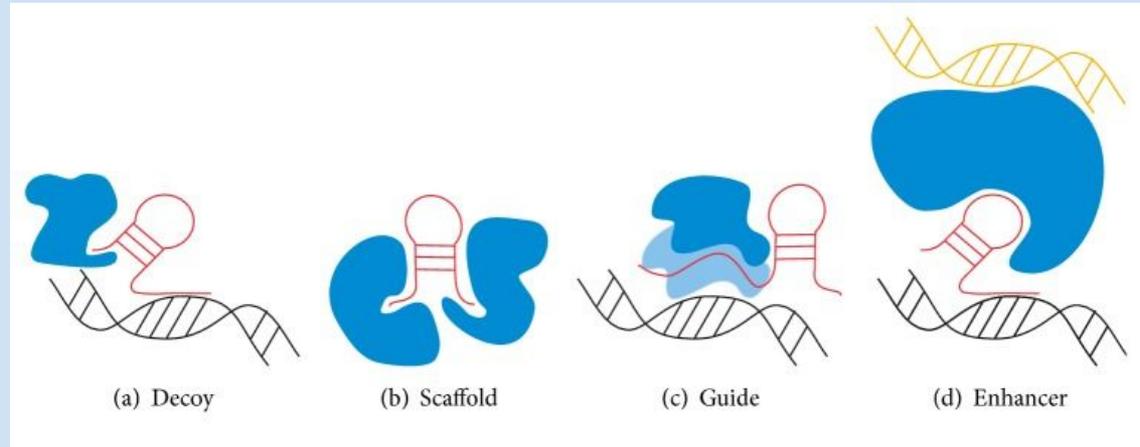
Terminology +  
Scientific  
Background

# lncRNA

## Definition:

relatively well-characterized class of **noncoding RNA (ncRNA) molecules**, involved in the regulation of various cell processes, including transcription, intracellular trafficking, and chromosome remodeling

**Context:** Work in breakout rooms!



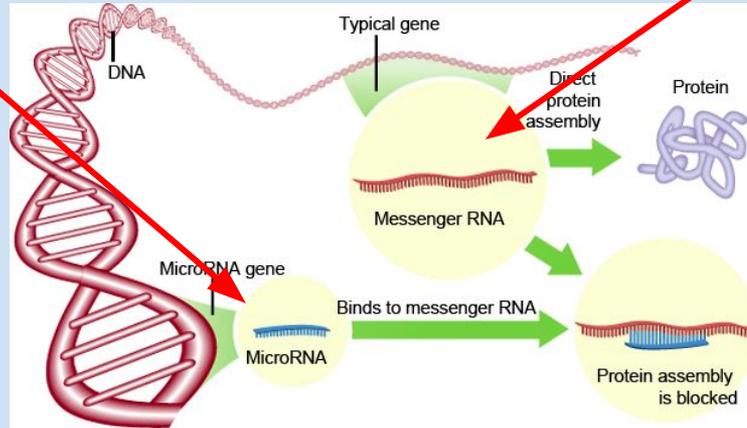
# miRNA and mRNA

## miRNA

### Definition:

**MicroRNAs** (miRNAs) are small noncoding RNAs that function as major players of posttranscriptional gene regulation in diverse species.

**Context:** Discuss in breakout rooms!



## mRNA

### Definition:

**Messenger RNA (mRNA)** is a single-stranded RNA molecule complementary to DNA strands of gene -- DNA of gene can be transcribed into an mRNA molecule that will end up making one specific protein.

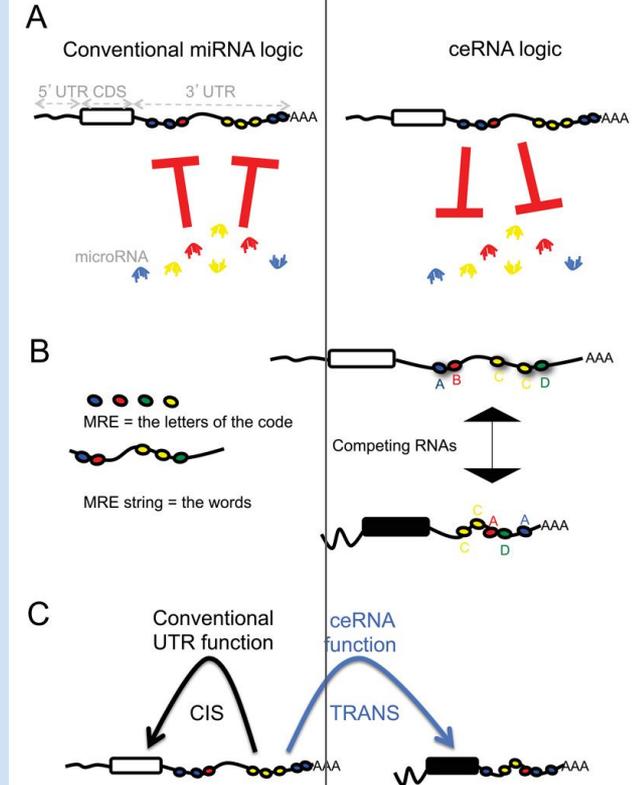
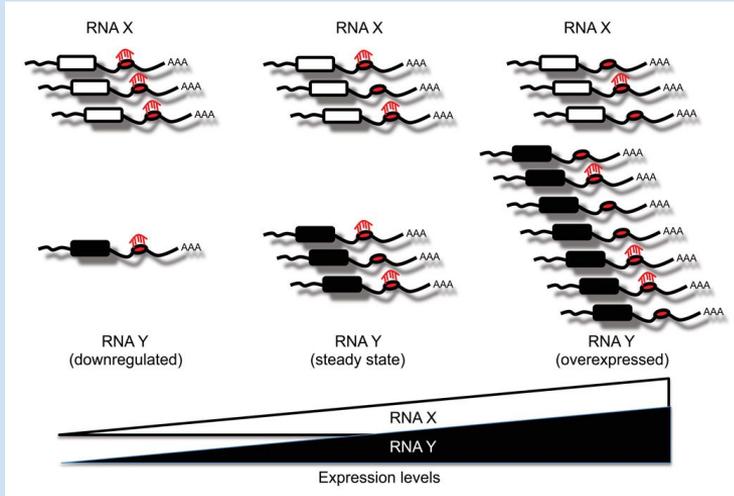
**Question:** Based on the paper, how do you think mRNA-miRNA interactions may indicate a prognostic marker?

# ceRNA and ceRNA Networks

## Definition:

**competing endogenous RNAs** (**ceRNAs**) regulate other RNA transcripts by competing for shared **microRNAs**. Changes in the expression of one or multiple miRNA targets alter the number of unbound miRNAs and lead to observable changes in miRNA activity

**ceRNA Networks:** the network of different ceRNAs and how they interconnect

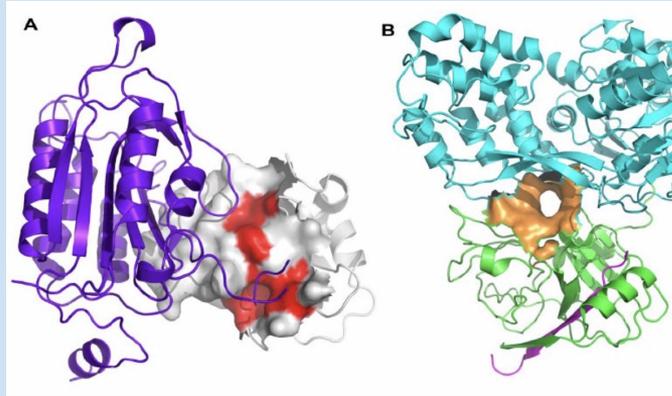


# PPI Network

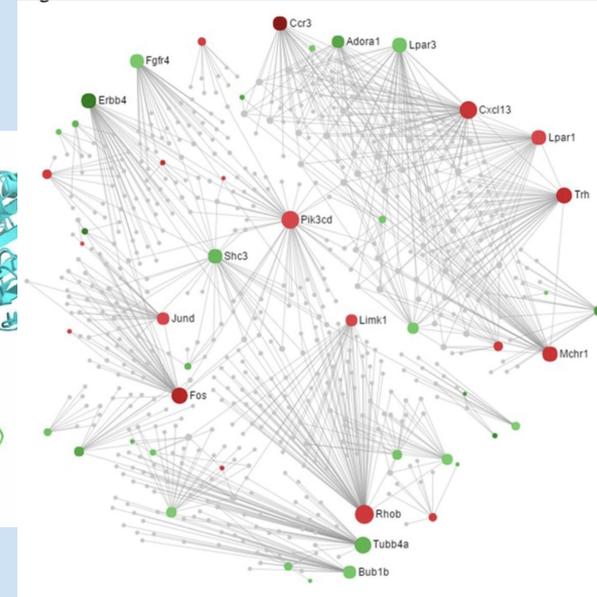
## Definition:

**PPI** = protein-protein interaction, which are the physical contacts of high specificity established between 2+ protein molecules due to biochemical events driven by electrostatic forces, H-bonding, and hydrophobic effects

**Context:** Discuss in breakout rooms!



PPI



PPI Network Analysis of mRNA Expression Profile

# Prognostic Marker

## Definition:

Biomarkers used to measure the progress of a disease in a patient, and response to a therapeutic solution. Helpful for stratifying patients into groups. .

**Context:** Discuss in breakout rooms!

Table1: DNA prognostic markers for common cancer types<sup>[3]</sup>

Cancer	DNA markers
Thyroid cancer	RET-PTC, NTRK1, PTEN, TP53, PI3K, AKT, CTNNB1, PAX8, RAS, BRAF, TSHR
Bladder cancer	FGFR3, TERT, STAG2, AURKA
Ovarian cancer	TP53, WT1, Ki67, Topo-II, BRCA1, BRCA2
Cervical cancer	Ki67, MYC, p16INK4a, PTEN, Bm-3a
Breast cancer	BRCA1, BRCA2, HER-2, TP53, EGFR
Prostate cancer	GSTP1, MYC, PTEN, APC, PCA3, PSMA, AMACR, BRCA1, BRCA2
Colorectal cancer	KRAS, BRAF, PIK3CA, TP53, APC, SFRP2, ITGA4, GATA4, GATA5, OSMR

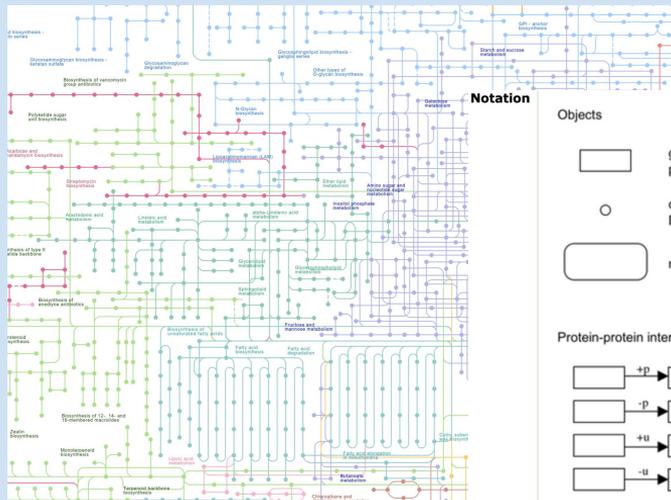
# KEGG Pathway

## Definition:

**KEGG** = Kyoto Encyclopedia of Genes and Genomes.

The KEGG PATHWAY database is a collection of manually drawn graphical diagrams (KEGG pathway maps), representing different molecular pathways for metabolism, genetic information processing, etc.

**Context:** Discuss in breakout rooms!



### Notation

#### Objects

- gene product, mostly protein but including RNA
- chemical compound, DNA and other molecule
- map

#### Arrows

- molecular interaction or relation
- link to/from another map
- indirect link or unknown reaction
- missing interaction (eg., by mutation)
- drug structure link or pointer used to add legend

#### Protein-protein interactions

- phosphorylation
- dephosphorylation
- ubiquitination
- deubiquitination
- glycosylation
- methylation
- activation
- inhibition
- indirect effect or state change
- binding / association
- dissociation
- complex

#### Gene expression relations

- expression
- repression
- expression
- repression

#### Enzyme-enzyme relations

- two successive reaction steps

# Colorectal Cancer

- 1/21 men and 1/23 women will develop it in the USA
- 2nd leading cause of cancer death in women; 3rd in men
- Death rate is decreasing with advances in screening techniques and improvements in treatment
  - Treatment includes surgery, radiotherapy, and chemo
  - It is dependent on size, location, stage of cancer, and age/overall health of patient



# Colorectal Cancer

Risk factors include

- older age
- a diet that is high in animal protein, saturated fats, and calories or low in fiber
- high alcohol consumption
- breast, ovary, or uterine cancer
- a family history of colorectal cancer
- Crohn's disease, or irritable bowel disease (IBD)
- overweight and obesity
- smoking
- a lack of physical activity



# Data Sources

## TCGA

- Catalogue genetic mutations responsible for cancer.
- Improve our ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease.

For CRC:

- over 500 individuals
- 480 tumor and 41 normal samples

**NOTE: As advised by the mentors, we won't be working with TCGA**

The screenshot shows the NCI website page for the Cancer Genome Atlas Program. The header includes the NIH logo and the text "NATIONAL CANCER INSTITUTE". Navigation links include "1-800-4-CANCER", "Live Chat", "Publications", and "Dictionary". A secondary navigation bar contains "ABOUT CANCER", "CANCER TYPES", "RESEARCH", "GRANTS & TRAINING", "NEWS & EVENTS", "ABOUT NCI", and a search bar. The breadcrumb trail reads "Home > About NCI > NCI Organization > CCG > Research > Structural Genomics". A sidebar menu for "TCGA" includes "Program History", "TCGA Cancers Selected for Study", "Publications by TCGA", "Using TCGA", and "Contact". The main content area features the title "The Cancer Genome Atlas Program" and a paragraph describing the program as a landmark cancer genomics program characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. It mentions the joint effort between the NCI and the National Human Genome Research Institute, starting in 2006. A second paragraph states that over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data, which will remain publicly available. At the bottom, there are two images: one showing a group of diverse human silhouettes and another showing a DNA double helix with the word "Patterns" overlaid.

# Data Sources

- The largest fully public repository for high-throughput molecular abundance data
- The database has a flexible and open design that allows the submission, storage and retrieval of many data types.

**GSE32323** : 17 normal and 17 tumor samples

**GSE21510** : 44 normal and 104 tumor samples

**GSE8671** : 32 normal and 32 tumor samples

NCBI Resources How To Sign in to NCBI

GEO Home Documentation Query & Browse Email GEO

COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.  
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

## Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession

### Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

### Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

### Browse Content

Repository Browser	
DataSets:	4348
Series:	130860
Platforms:	20991
Samples:	3632114

### Information for Submitters

Login to Submit	Submission Guidelines	MIAME Standards
	Update Guidelines	Citing and Linking to GEO
		Guidelines for Reviewers
		GEO Publications

# Example file Images

The screenshot shows the NCBI GEO website interface. At the top left is the NCBI logo. In the center is the GEO logo with the text "Gene Expression Omnibus". A prominent red banner contains a COVID-19 notice: "COVID-19 is an emerging, rapidly evolving situation. Get the latest public health information from CDC: https://www.coronavirus.gov. Get the latest research from NIH: https://www.nih.gov/coronavirus. Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/." Below the banner is a navigation bar with links for HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. The main content area shows the accession "GSE32323" and a "GO" button. A "Series GSE32323" section is visible, with a link to "Query DataSets for GSE32323". Below this, a table lists details for the series: Status (Public on Mar 20, 2012), Title (Screening for Epigenetically Masked Genes in Colorectal Cancer using 5-aza-2'-deoxycytidine treatment, Microarray and Gene Expression Profile), Organism (Homo sapiens), Experiment type (Expression profiling by array), and Summary (Unearthing of silenced genes in colorectal cancer (CRC) is of great importance).

Platforms (1) [GPL570](#) [HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (44) [GSM800742](#) patient 006b, normal, homogenized  
[GSM800743](#) patient 011b, normal, homogenized  
[GSM800744](#) patient 024b, normal, homogenized

#### Relations

BioProject [PRJNA147273](#)

Analyze with GEO2R

#### Download family

[SOFT formatted family file\(s\)](#)  
[MINiML formatted family file\(s\)](#)  
[Series Matrix File\(s\)](#)

#### Format

[SOFT](#) [?](#)  
[MINiML](#) [?](#)  
[TXT](#) [?](#)

Supplementary file	Size	Download	File type/resource
<a href="#">GSE32323_RAW.tar</a>	211.3 Mb	<a href="#">(http)(custom)</a>	TAR (of CEL)

Raw data provided as supplementary file

Processed data included within Sample table

# MATERIALS AND METHODS

## Screening Potential Hub Genes

### DAVID

(The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery)

It provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes

**Input:** You can either load a gene list from a file or paste a gene list to the text box.

**Output:** It depends on the type of analysis you're doing. In most of the cases your result is going to be a table.

### STRING database

It is a database of known and predicted protein-protein interactions.

The interactions include direct (physical) and indirect (functional) associations.

**Input:** You can either load a gene list from a file or paste a gene list to the text box.

**Output:** You can export graphs as images, simple tabular text files or even protein sequence.

# MATERIALS AND METHODS

## Potential Prognostic Values of Candidate Genes

### GEPIA database

- Interactive web server for analyzing the RNA sequencing expression data
- Facilitate data mining in wide research areas, scientific discussion and the therapeutic discovery process.

**Input:** gene symbols or Ensemble ID

**Output:** It depends on the type of analysis you're doing. In most of the cases you're result is going to be a table.

### StarBase database

- Facilitate the comprehensive exploration of miRNA-target interaction maps
- Useful to assigning miRNAs to their regulatory target genes

**Input:** Nucleotides of a mature sequence or a mature miRNA sequence

**Output:** Information about the target gene and visual sequence alignments matched to a specific CLIP-Seq

# MATERIALS AND METHODS

## Screening Related miRNAs and lncRNAs Based on the Hub mRNAs

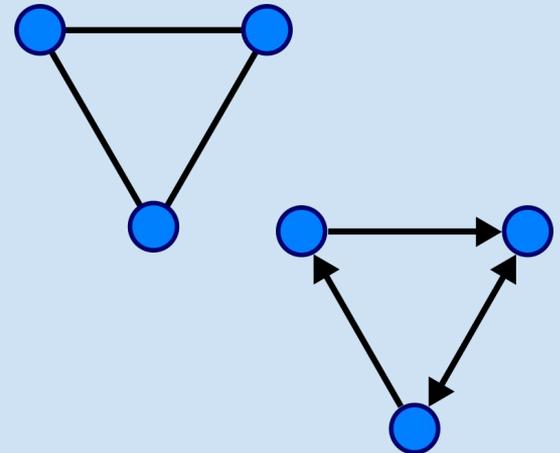
### MiRNet

An easy-to-use web-based tool designed for creation, customization, visual exploration and functional interpretation of miRNA-target interaction networks.

**Input:** a list of miRNAs, genes, small molecules, etc.; or a data table microarray or RNAseq

**Output:** a graph → What is a graph?

graphs are mathematical structures used to model pairwise relations between objects.



# MATERIALS AND METHODS

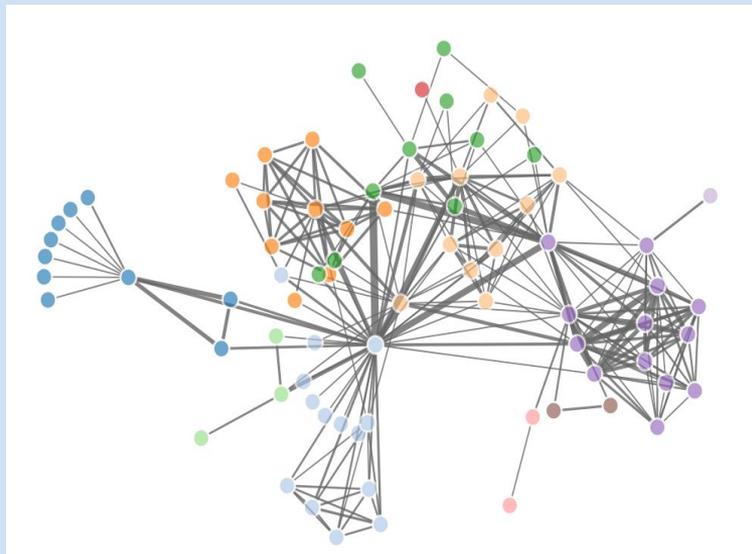
## Construction and In-Depth Analysis of the ceRNA Network

### Network D3

It makes easy to create these network graphs from R. It is designed to take a simple data frame that has two columns specifying the *sources* and *targets* of the nodes in a network and turn it into a graph. You can easily customise the look and feel of the graph.

**Input:** Dataframe

**Output:** Graph



# MATERIALS AND METHODS

Statistical Analysis and Network Visualization

