# An Investigation on the Possibilities of Targeting Transcription Factor Cascades for Precision Oncology

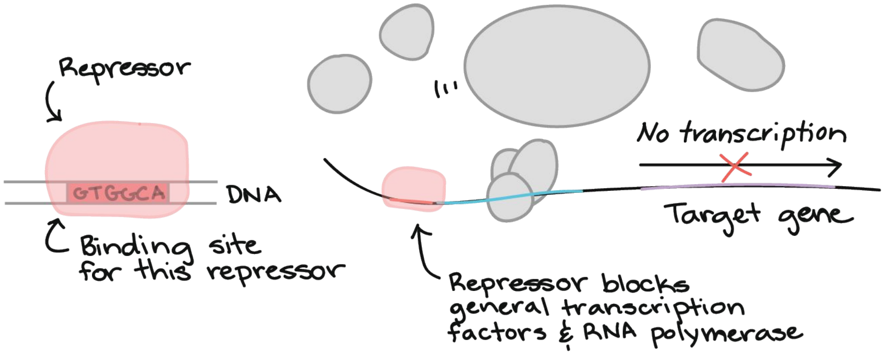
- Sonish Sivarajkumar

# Transcription factor

Transcription factors are proteins involved in the process of converting, or transcribing, DNA into RNA.

Transcription factors include a wide number of proteins, excluding RNA polymerase, that initiate and regulate the transcription of genes.

There are 8107 Transcription Factors in Humans.

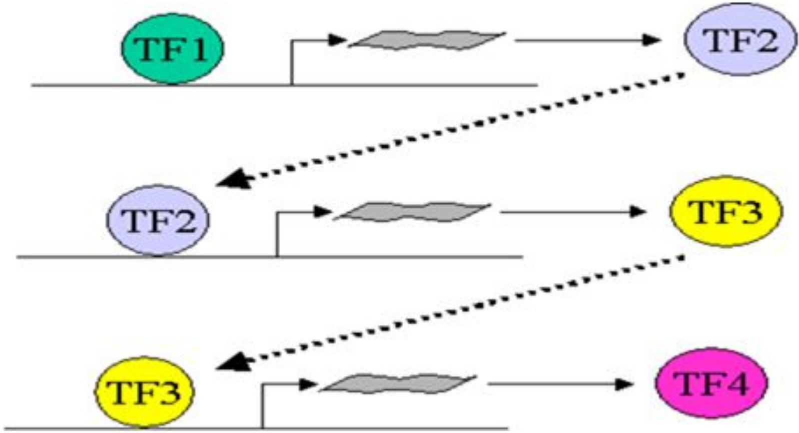

Repressor
GTGGCA
DNA
Binding site for this repressor
No transcription
Target gene
Repressor blocks general transcription factors & RNA polymerase

# Transcription Factor Cascade

Transcription factors are encoded by genes.

Regulation of their expression also needs transcription factors. So there are regulatory relationships among transcription factors.

After the primary transcription factors are expressed, they will activate the transcription of the second transcription factor genes or other target genes. The second transcription factors will activate the transcription of the third transcription factor genes or other target genes.



# Significance of TF Cascades dataset

The lack of availability of  Transcription Factor cascades is the one of the underlying cause of slow cancer research.

Transcription Factor is one of the core participant in the cause death in a patient as it causes the **domino** effect within the gene .

One Transcription Factor mutation leads to further mutation of other Transcription Factors .

# Dataset Creation

- Step 1 : downloaded the interaction network from **string** database
- Step 2: Converted the known TF genes into string ID format
- Step 3: filtered out the interactions that are present in the TF gene list
- Step 4: Python script to make the network

Biomart is a website which converts IDs from one form to another . Eg Ensp (string protein ID ) can be to HGNC gene ID (Hugo Gene ID)
BioMart (ensembl.org)

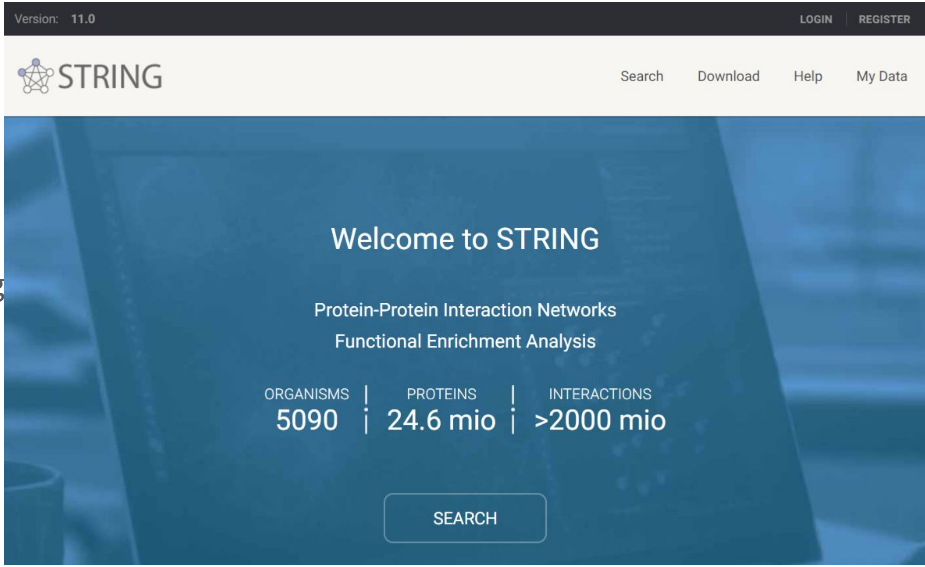
# String DB

STRING is a database  of known and predicted protein–protein interactions.

Contains 4 million lines

contains information from numerous sources, including experimental data, computational prediction methods and public text collections.

[STRING: functional protein association networks (string-db.org)](string-db.org)



Version: 11.0

Welcome to STRING

Protein-Protein Interaction Networks

Functional Enrichment Analysis

ORGANISMS 5090 | PROTEINS 24.6 mio | INTERACTIONS >2000 mio

# Python to create network graph for every chain individually

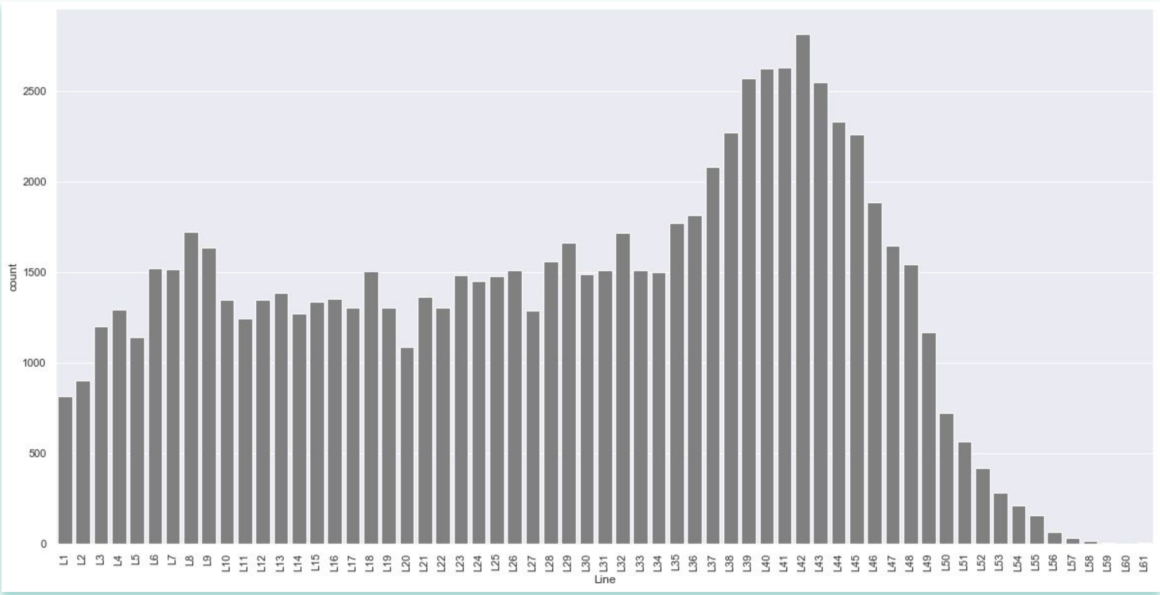
Using Python package Stringdb and igraph we are able to get network of every individual chain we created earlier

A report file for created for every chain to keep tabs of connectivity and scores as well as many different details of the interactions.

# Creating Chains(Cascades) of genes

Using Python, the chains  or cascades of TFs were formed.

81488 unique chains were formed , and the longest chain was of length 62



# Analysis & Investigation

Exploratory Data Analysis

Graph Analytics

Enrichment Analysis

# Analysis & Investigation

The network is too difficult to understand with more than 400 unique nodes and over 9000 interactions

First step was to re-index the data.

We performed the following analysis on the generated dataset to  find some key insights

1. **Exploratory Data Analysis** - analyzing data set  to summarize their main characteristics, using statistical graphics and other data visualization methods

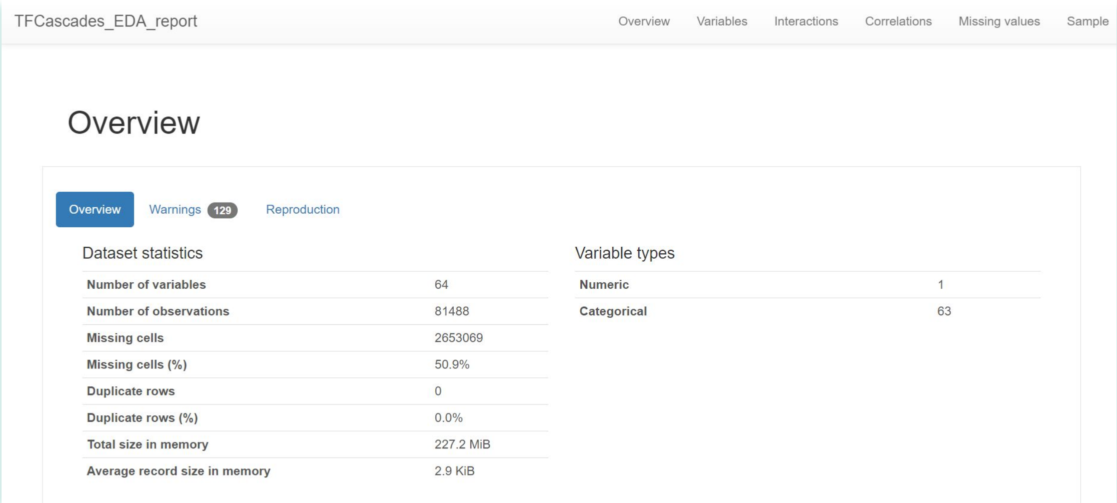
2. **Graph Analytics** - Build network or graphs out of the data and find the high-scoring nodes

3. **Enrichment Analysis** - Functional exploration of  high-scoring nodes including their biological pathways

# EDA and Report

Using Python, we did Exploratory Data Analysis on the TF Cascades dataset

Created a detailed report of the  EDA as an interactive  HTML webpage



## Overview

Dataset statistics

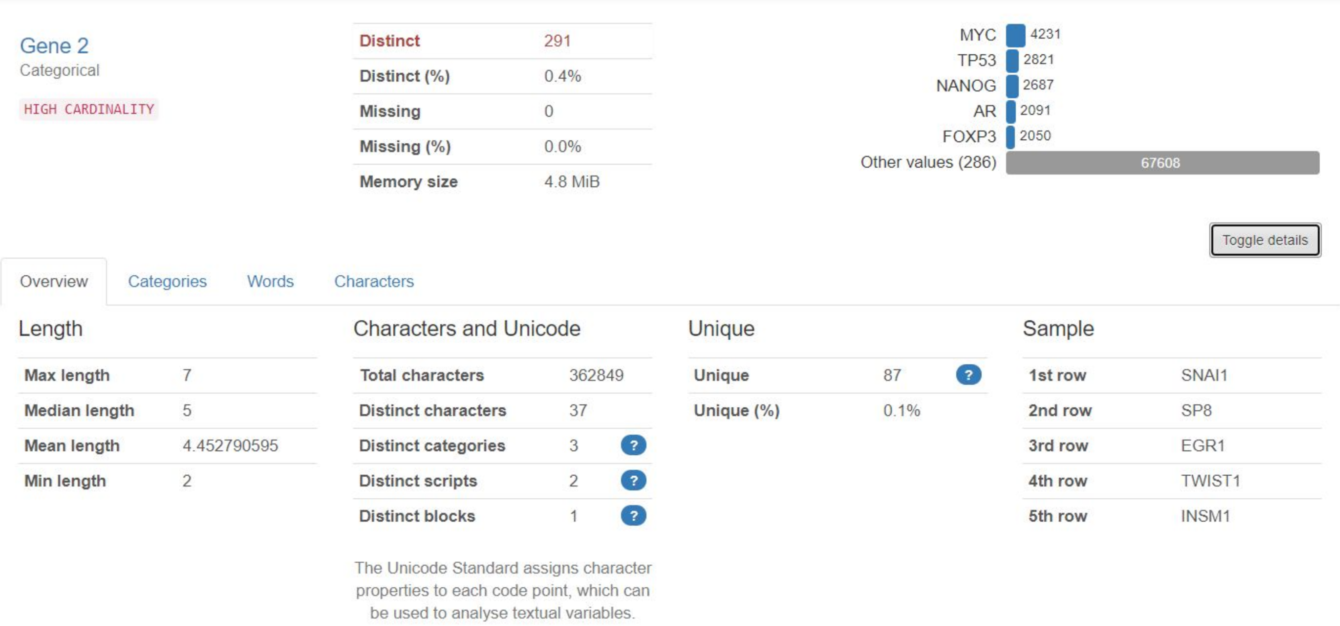
| | |
|---|---|
| **Number of variables** | 64 |
| **Number of observations** | 81488 |
| **Missing cells** | 2653069 |
| **Missing cells (%)** | 50.9% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0.0% |
| **Total size in memory** | 227.2 MiB |
| **Average record size in memory** | 2.9 KiB |

Variable types

| | |
|---|---|
| **Numeric** | 1 |
| **Categorical** | 63 |

# EDA - Summary

Similarly all 62  columns are analysed



Gene 2

Categorical

HIGH CARDINALITY

| | |
|---|---|
| Distinct | 291 |
| Distinct (%) | 0.4% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 4.8 MiB |

| | |
|---|---|
| MYC | 4231 |
| TP53 | 2821 |
| NANOG | 2687 |
| AR | 2091 |
| FOXP3 | 2050 |
| Other values (286) | 67608 |

Toggle details

Overview Categories Words Characters

### Length

| | |
|---|---|
| Max length | 7 |
| Median length | 5 |
| Mean length | 4.452790595 |
| Min length | 2 |

### Characters and Unicode

| | |
|---|---|
| Total characters | 362849 |
| Distinct characters | 37 |
| Distinct categories | 3 ? |
| Distinct scripts | 2 ? |
| Distinct blocks | 1 ? |

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

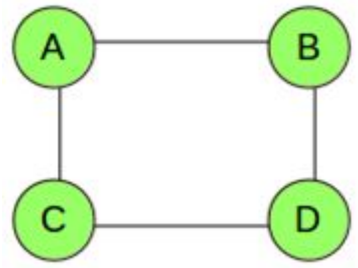
### Unique

| | |
|---|---|
| Unique | 87 ? |
| Unique (%) | 0.1% |

### Sample

| | |
|---|---|
| 1st row | SNAI1 |
| 2nd row | SP8 |
| 3rd row | EGR1 |
| 4th row | TWIST1 |
| 5th row | INSM1 |

# Graph Theory



In statistics, graph theory is the study of graphs, which are mathematical structures used to model pairwise relations between objects. A graph is a combination of vertices (nodes) and edges.

G = (V, E) where V represents the set of all vertices and E represents the set of all edges of the graph.

Here we are trying to answer the following **questions** using graph theory,

- Among all 62  Lines together, which TFs  have higher influence?
- In each individual Lines or cascades, which TFs  have higher influence?
- which TFs are  least influential - in each Lines?
- which TFs are least influential- in all Lines together?

We can **study the most influential TFs for drug discovery and targeted therapy**

# Complex graph theory and network analysis :

To assign roles and make categories between individuals we will calculate mathematical indicators from the theory of complex graphs:

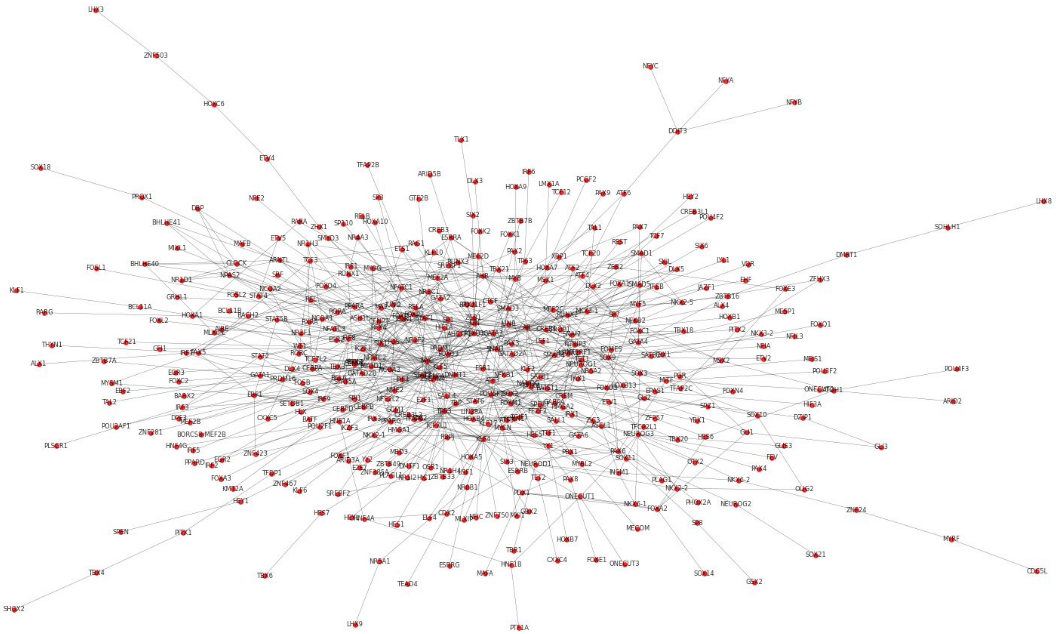
**Betweenness centrality**: This indicator can detect individuals who influence the transfer of information.

**Centrality of proximity**: This indicator makes it possible to detect the individuals who have a significant power on the transfer of information. $C(x) = \frac{N}{\sum_y d(y,x)}$ where d(y,x) is the distance between vertices x and y, and N is the number of nodes.

**Eigenvector centrality**: The individuals having a high spectral centralized are the individuals who have the most relation in the network, they are central and have influence in a general way on the network. For a given graph G:=(V,E) , with |V| vertices, let A=(av,t) be the adjacency matrix, i.e.(av,t)=1, if vertex v is linked to vertex t ,and (av,t)=0 otherwise. The relative centrality score of vertex v can be defined as:

$$x_v = \frac{1}{\lambda}\sum_{t\in M(v)} x_t = \frac{1}{\lambda}\sum_{t\in G} a_{v,t}x_t$$

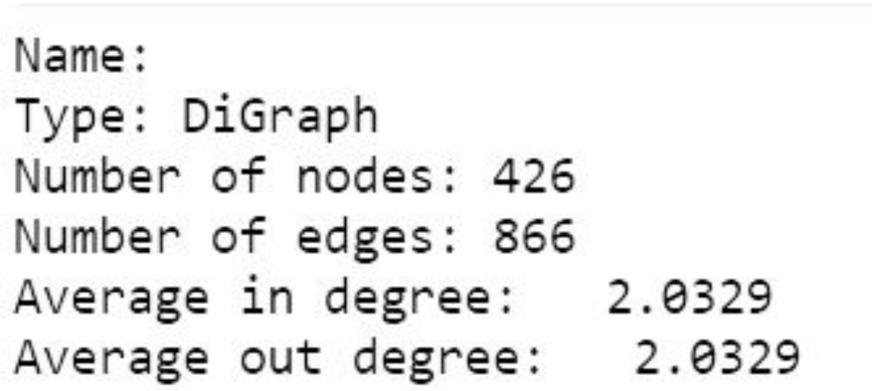
# Graph Analytics Summary - Python



All TF interactions

Using NetworkX python library

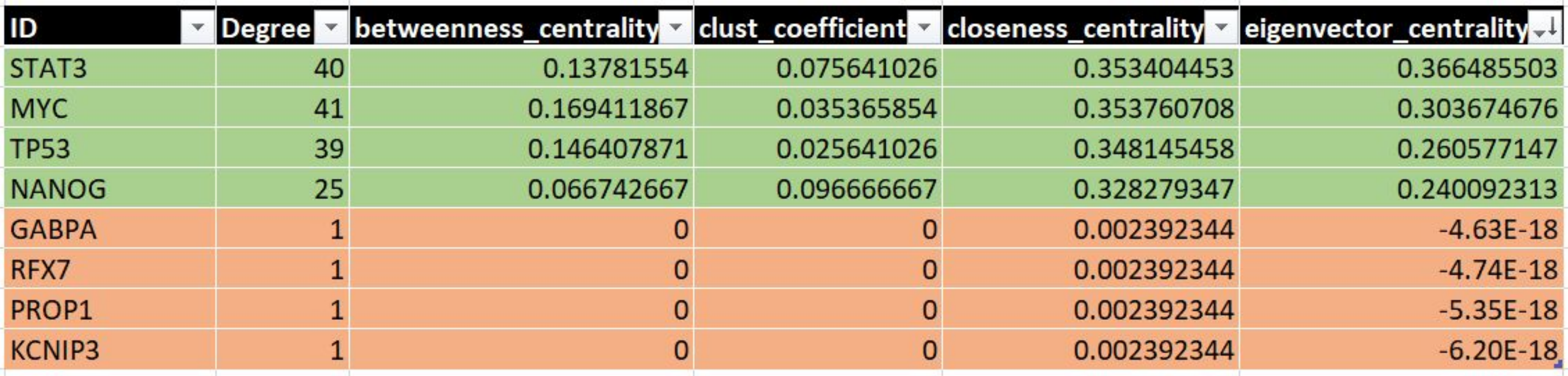
# Graph Analytics Summary - Python



```
Name:
Type: DiGraph
Number of nodes: 426
Number of edges: 866
Average in degree:   2.0329
Average out degree:   2.0329
```

# Graph Analytics Summary - Python

Green - Top 4 most influential TFs - STAT3, MYC, TP53, NANOG
Red - Least 4 influential TFs - GABPA, RFX7, PROP1, KCNIP3



| ID | Degree | betweenness_centrality | clust_coefficient | closeness_centrality | eigenvector_centrality |
|---|---|---|---|---|---|
| STAT3 | 40 | 0.13781554 | 0.075641026 | 0.353404453 | 0.366485503 |
| MYC | 41 | 0.169411867 | 0.035365854 | 0.353760708 | 0.303674676 |
| TP53 | 39 | 0.146407871 | 0.025641026 | 0.348145458 | 0.260577147 |
| NANOG | 25 | 0.066742667 | 0.096666667 | 0.328279347 | 0.240092313 |
| GABPA | 1 | 0 | 0 | 0.002392344 | -4.63E-18 |
| RFX7 | 1 | 0 | 0 | 0.002392344 | -4.74E-18 |
| PROP1 | 1 | 0 | 0 | 0.002392344 | -5.35E-18 |
| KCNIP3 | 1 | 0 | 0 | 0.002392344 | -6.20E-18 |

# Enrichment Pathway Analysis

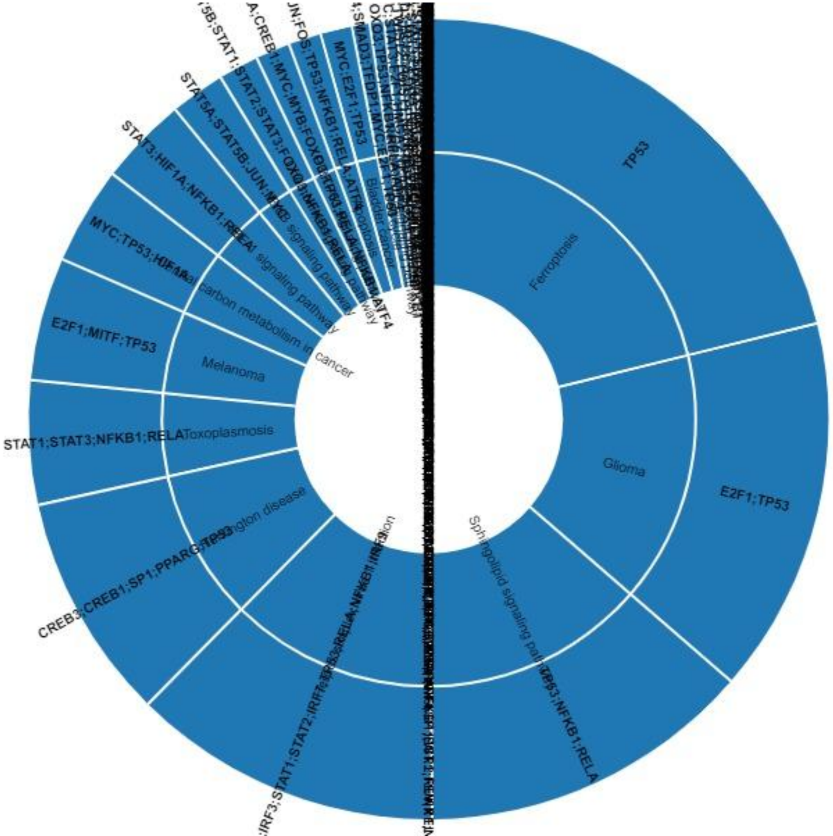
Filtered TF interactions of the top 4 most influential TFs , ie. STAT3, MYC, TP53, NANOG

Questions:

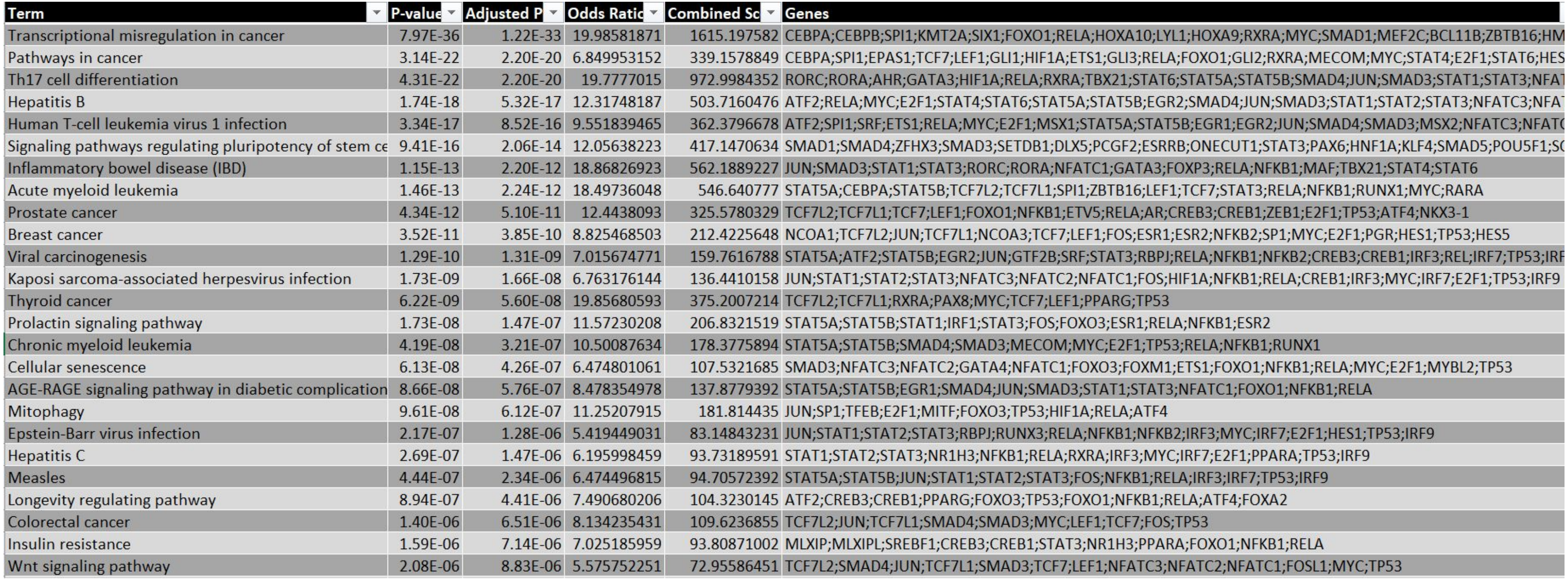
- What does this TF do?
- How and where does it do it?
- Does it make sense to see it on this list?
- Does it interact with other TFs?
- Does its behaviour change during disease, disorder or therapy?

Performed Enrichment Pathway analysis on all interacting TFs

Visualized the Adjusted P-value using sunburst diagram



# Enrichment Analysis Summary



| Term | P-value | Adjusted P | Odds Ratic | Combined Sc | Genes |
|---|---|---|---|---|---|
| Transcriptional misregulation in cancer | 7.97E-36 | 1.22E-33 | 19.98581871 | 1615.197582 | CEBPA;CEBPB;SPI1;KMT2A;SIX1;FOXO1;RELA;HOXA10;LYL1;HOXA9;RXRA;MYC;SMAD1;MEF2C;BCL11B;ZBTB16;HM |
| Pathways in cancer | 3.14E-22 | 2.20E-20 | 6.849953152 | 339.1578849 | CEBPA;SPI1;EPAS1;TCF7;LEF1;GLI1;HIF1A;ETS1;GLI3;RELA;FOXO1;GLI2;RXRA;MECOM;MYC;STAT4;E2F1;STAT6;HES |
| Th17 cell differentiation | 4.31E-22 | 2.20E-20 | 19.7777015 | 972.9984352 | RORC;RORA;AHR;GATA3;HIF1A;RELA;RXRA;TBX21;STAT6;STAT5A;STAT5B;SMAD4;JUN;SMAD3;STAT1;STAT3;NFAT |
| Hepatitis B | 1.74E-18 | 5.32E-17 | 12.31748187 | 503.7160476 | ATF2;RELA;MYC;E2F1;STAT4;STAT6;STAT5A;STAT5B;EGR2;SMAD4;JUN;SMAD3;STAT1;STAT2;STAT3;NFATC3;NFAT |
| Human T-cell leukemia virus 1 infection | 3.34E-17 | 8.52E-16 | 9.551839465 | 362.3796678 | ATF2;SPI1;SRF;ETS1;RELA;MYC;E2F1;MSX1;STAT5A;STAT5B;EGR1;EGR2;JUN;SMAD4;SMAD3;MSX2;NFATC3;NFATC |
| Signaling pathways regulating pluripotency of stem ce | 9.41E-16 | 2.06E-14 | 12.05638223 | 417.1470634 | SMAD1;SMAD4;ZFHX3;SMAD3;SETDB1;DLX5;PCGF2;ESRRB;ONECUT1;STAT3;PAX6;HNF1A;KLF4;SMAD5;POU5F1;S |
| Inflammatory bowel disease (IBD) | 1.15E-13 | 2.20E-12 | 18.86826923 | 562.1889227 | JUN;SMAD3;STAT1;STAT3;RORC;RORA;NFATC1;GATA3;FOXP3;RELA;NFKB1;MAF;TBX21;STAT4;STAT6 |
| Acute myeloid leukemia | 1.46E-13 | 2.24E-12 | 18.49736048 | 546.640777 | STAT5A;CEBPA;STAT5B;TCF7L2;TCF7L1;SPI1;ZBTB16;LEF1;TCF7;STAT3;RELA;NFKB1;RUNX1;MYC;RARA |
| Prostate cancer | 4.34E-12 | 5.10E-11 | 12.4438093 | 325.5780329 | TCF7L2;TCF7L1;TCF7;LEF1;FOXO1;NFKB1;ETV5;RELA;AR;CREB3;CREB1;ZEB1;E2F1;TP53;ATF4;NKX3-1 |
| Breast cancer | 3.52E-11 | 3.85E-10 | 8.825468503 | 212.4225648 | NCOA1;TCF7L2;JUN;TCF7L1;NCOA3;TCF7;LEF1;FOS;ESR1;ESR2;NFKB2;SP1;MYC;E2F1;PGR;HES1;TP53;HES5 |
| Viral carcinogenesis | 1.29E-10 | 1.31E-09 | 7.015674771 | 159.7616788 | STAT5A;ATF2;STAT5B;EGR2;JUN;GTF2B;SRF;STAT3;RBPJ;RELA;NFKB1;NFKB2;CREB3;CREB1;IRF3;REL;IRF7;TP53;IRF |
| Kaposi sarcoma-associated herpesvirus infection | 1.73E-09 | 1.66E-08 | 6.763176144 | 136.4410158 | JUN;STAT1;STAT2;STAT3;NFATC3;NFATC2;NFATC1;FOS;HIF1A;NFKB1;RELA;CREB1;IRF3;MYC;IRF7;E2F1;TP53;IRF9 |
| Thyroid cancer | 6.22E-09 | 5.60E-08 | 19.85680593 | 375.2007214 | TCF7L2;TCF7L1;RXRA;PAX8;MYC;TCF7;LEF1;PPARG;TP53 |
| Prolactin signaling pathway | 1.73E-08 | 1.47E-07 | 11.57230208 | 206.8321519 | STAT5A;STAT5B;STAT1;IRF1;STAT3;FOS;FOXO3;ESR1;RELA;NFKB1;ESR2 |
| Chronic myeloid leukemia | 4.19E-08 | 3.21E-07 | 10.50087634 | 178.3775894 | STAT5A;STAT5B;SMAD4;SMAD3;MECOM;MYC;E2F1;TP53;RELA;NFKB1;RUNX1 |
| Cellular senescence | 6.13E-08 | 4.26E-07 | 6.474801061 | 107.5321685 | SMAD3;NFATC3;NFATC2;GATA4;NFATC1;FOXO3;FOXM1;ETS1;FOXO1;NFKB1;RELA;MYC;E2F1;MYBL2;TP53 |
| AGE-RAGE signaling pathway in diabetic complication | 8.66E-08 | 5.76E-07 | 8.478354978 | 137.8779392 | STAT5A;STAT5B;EGR1;SMAD4;JUN;SMAD3;STAT1;STAT3;NFATC1;FOXO1;NFKB1;RELA |
| Mitophagy | 9.61E-08 | 6.12E-07 | 11.25207915 | 181.814435 | JUN;SP1;TFEB;E2F1;MITF;FOXO3;TP53;HIF1A;RELA;ATF4 |
| Epstein-Barr virus infection | 2.17E-07 | 1.28E-06 | 5.419449031 | 83.14843231 | JUN;STAT1;STAT2;STAT3;RBPJ;RUNX3;RELA;NFKB1;NFKB2;IRF3;MYC;IRF7;E2F1;HES1;TP53;IRF9 |
| Hepatitis C | 2.69E-07 | 1.47E-06 | 6.195998459 | 93.73189591 | STAT1;STAT2;STAT3;NR1H3;NFKB1;RELA;RXRA;IRF3;MYC;IRF7;E2F1;PPARA;TP53;IRF9 |
| Measles | 4.44E-07 | 2.34E-06 | 6.474496815 | 94.70572392 | STAT5A;STAT5B;JUN;STAT1;STAT2;STAT3;FOS;NFKB1;RELA;IRF3;IRF7;TP53;IRF9 |
| Longevity regulating pathway | 8.94E-07 | 4.41E-06 | 7.490680206 | 104.3230145 | ATF2;CREB3;CREB1;PPARG;FOXO3;TP53;FOXO1;NFKB1;RELA;ATF4;FOXA2 |
| Colorectal cancer | 1.40E-06 | 6.51E-06 | 8.134235431 | 109.6236855 | TCF7L2;JUN;TCF7L1;SMAD4;SMAD3;MYC;LEF1;TCF7;FOS;TP53 |
| Insulin resistance | 1.59E-06 | 7.14E-06 | 7.025185959 | 93.80871002 | MLXIP;MLXIPL;SREBF1;CREB3;CREB1;STAT3;NR1H3;PPARA;FOXO1;NFKB1;RELA |
| Wnt signaling pathway | 2.08E-06 | 8.83E-06 | 5.575752251 | 72.95586451 | TCF7L2;SMAD4;JUN;TCF7L1;SMAD3;TCF7;LEF1;NFATC3;NFATC2;NFATC1;FOSL1;MYC;TP53 |

# Enrichment Pathway Analysis - on each cascade

We had around 80,000 cascades

Each cascade was individually analysed

On a average, we obtained 25 pathways for each cascade

For 80,000 cascades, we got **2 million pathways**

A new database was created with all 2 million pathways, their interacting TF genes and statistical measurements like P-values

# Summary of the Project - To be edited

In this study, we have  developed a compendium of TF-cascades encoded in the human genome as an **TFCascades database**.

We have performed **exploratory data analysis(EDA)** and  done an extensive exploration of the dataset

Applied **Graph analytics**   TF-cascade network to identify and prioritize important TFs in cascades as drug targets, which will be useful in developing precision medicine.

Performed **Enrichment  pathway analysis** to understand the underlying biological processes out of the study

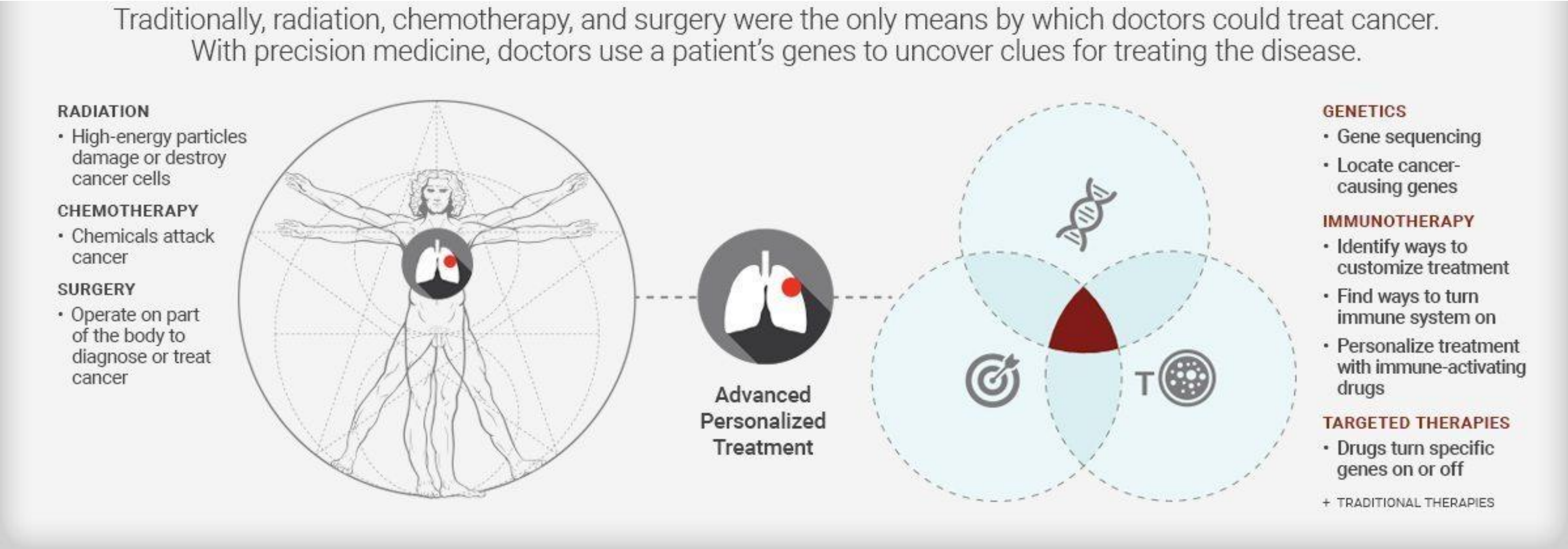
# Conclusion

TF-cascade network to **identify and prioritize important TFs** in cascades as drug targets, which will be useful in developing precision medicine.

Graph Analytics results showed that **STAT3, MYC, TP53, NANOG** are the most influencing TFs for cancer formation

Enrichment analysis results showed the most correlation pathways for study in cancer genomics

# Traditional Medicine vs Precision Medicine



Traditionally, radiation, chemotherapy, and surgery were the only means by which doctors could treat cancer. With precision medicine, doctors use a patient's genes to uncover clues for treating the disease.

**RADIATION**

- High-energy particles damage or destroy cancer cells

**CHEMOTHERAPY**

- Chemicals attack cancer

**SURGERY**

- Operate on part of the body to diagnose or treat cancer

**GENETICS**

- Gene sequencing
- Locate cancer-causing genes

**IMMUNOTHERAPY**

- Identify ways to customize treatment
- Find ways to turn immune system on
- Personalize treatment with immune-activating drugs

**TARGETED THERAPIES**

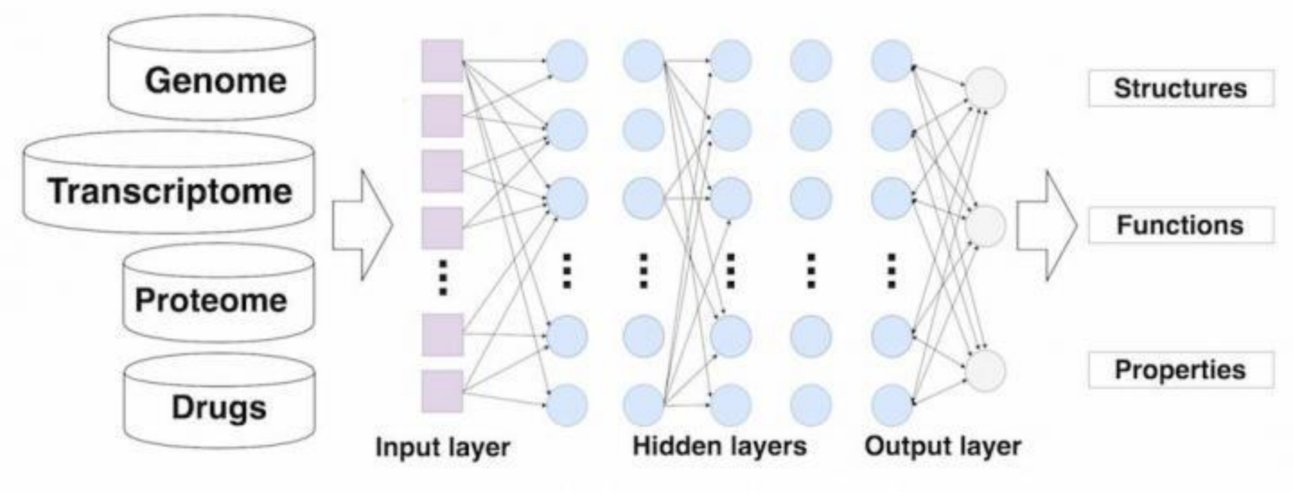
- Drugs turn specific genes on or off

+ TRADITIONAL THERAPIES

# Future Plans

Use Artificial Intelligence and this project results  for drug discovery

Expand the work to targeted therapy and precision medicine


Genome
Transcriptome
Proteome
Drugs
Input layer
Hidden layers
Output layer
Structures
Functions
Properties