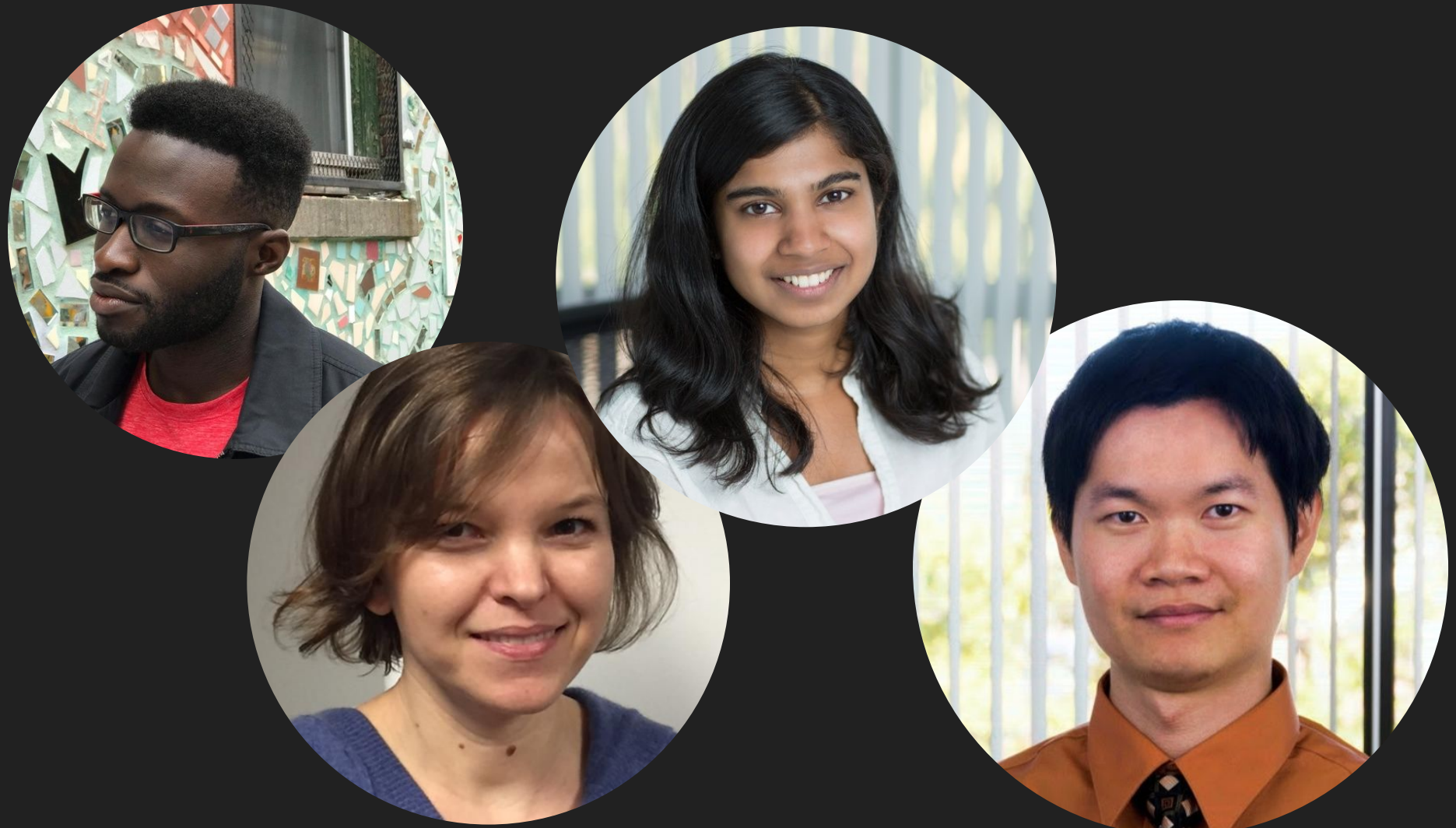
Team

Steve “Mr. TATA” Osazuwa

Ola “commandlinegirl” Zalcman

Naina “Remote” Thangaraj

Arkarachai
“Chai”
Fungtammasan



Problem

Doctor/Researcher/Scientist

What are genes to investigate?

What are genes to investigate?

What are the patients habits?

What are genes to investigate?

Smoker?

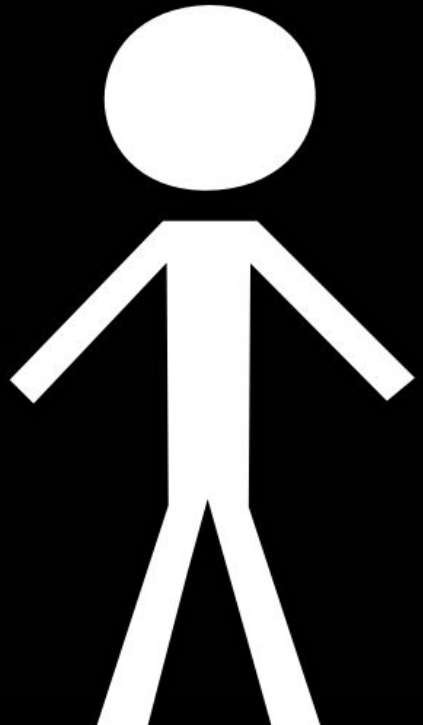
What are the patients habits?

What are genes to investigate?

Smoker?

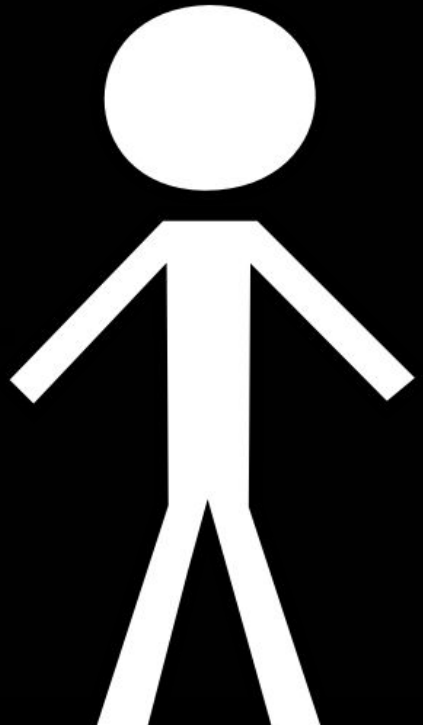
Partier?

What are the patients habits?



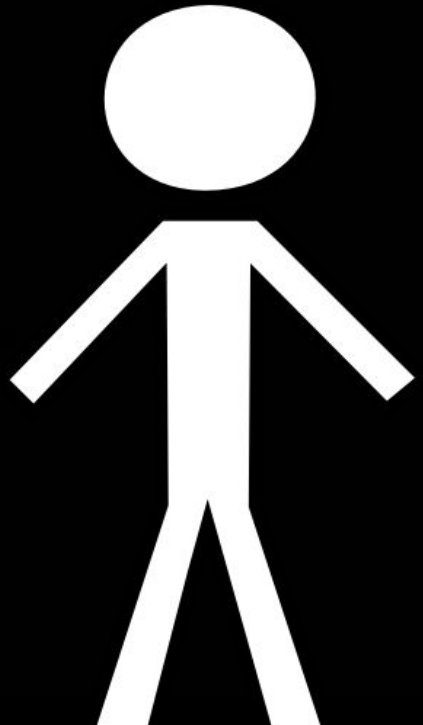
Smoker?

Nah

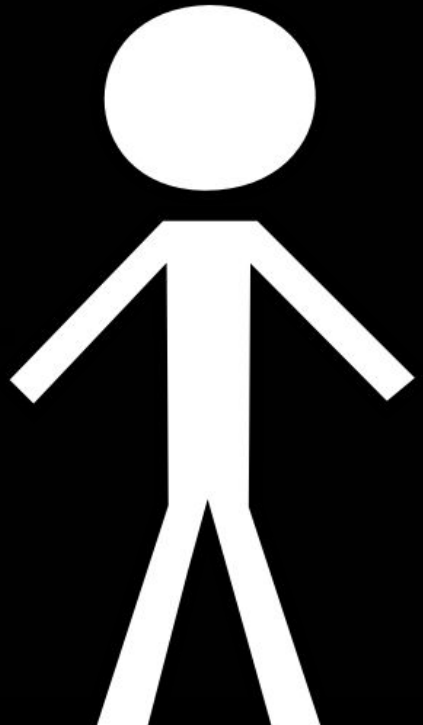


Partier?

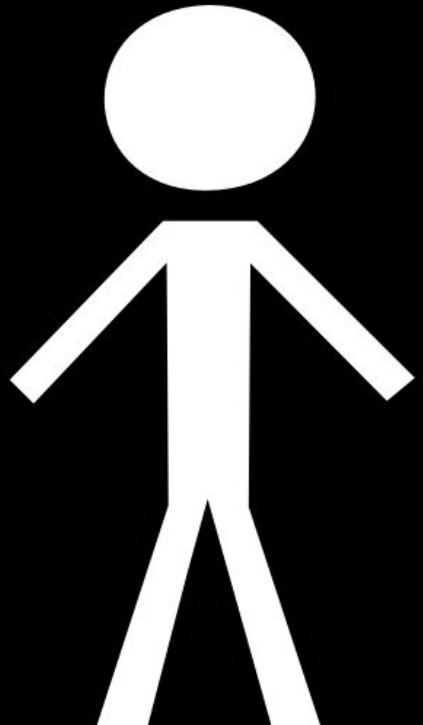
Yeah!



I know know what genes to investigate and likely causal variants



AI +



AI +

Given some information about the patient can we infer phenotypic or even genotypic data

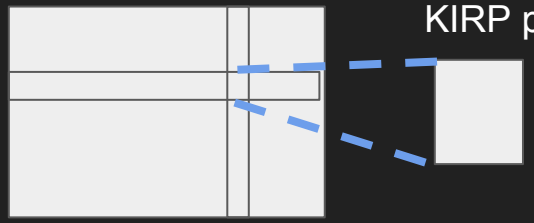
Methods

Patient or Gene grouping based on expression profile

Data

Capture critical features

TCGA GEM

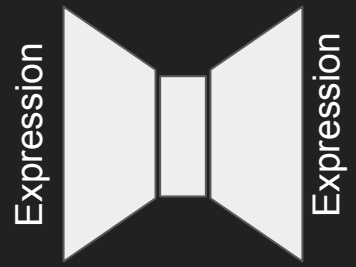


KIRP patients

KIRP related genes

Variational autoencoder

https://jmetzen.github.io/



Genotypic RNAseq FPKM

(Clemson's PanTCGA Expression Data)

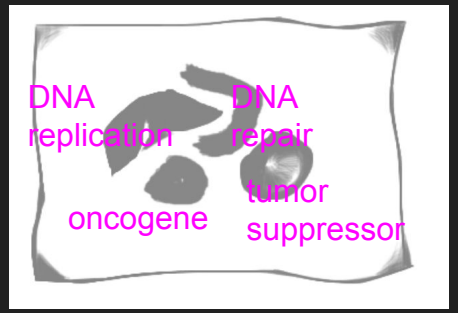
Phenotypic Clinical observations

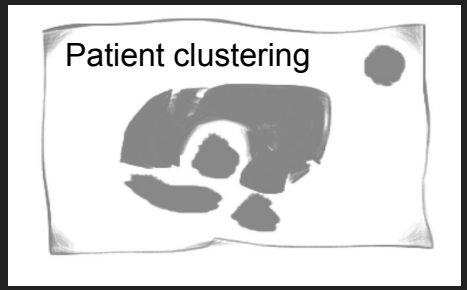
Patient or Gene grouping based on expression profile

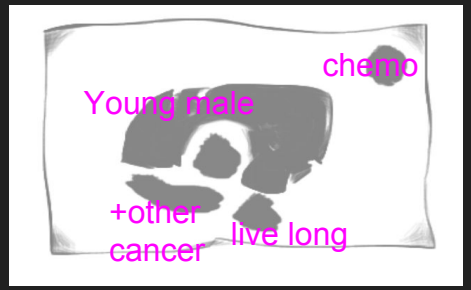
Expected clustering

Label by phenotype

t-SNE dimensional reduction





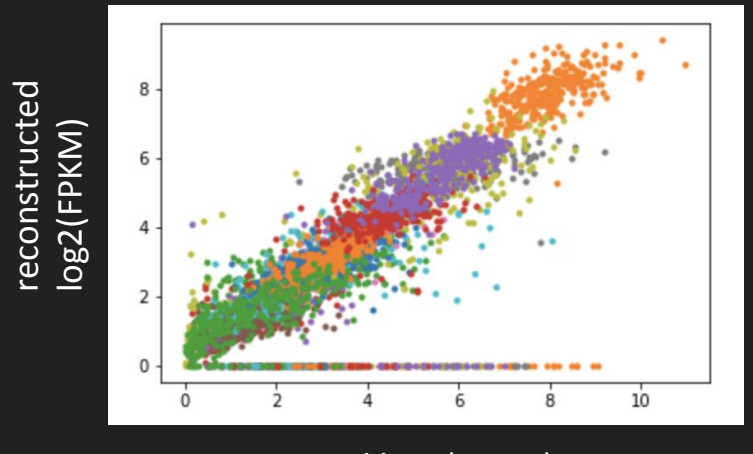


?

Results

Reconstruct gene expression using autoencoder

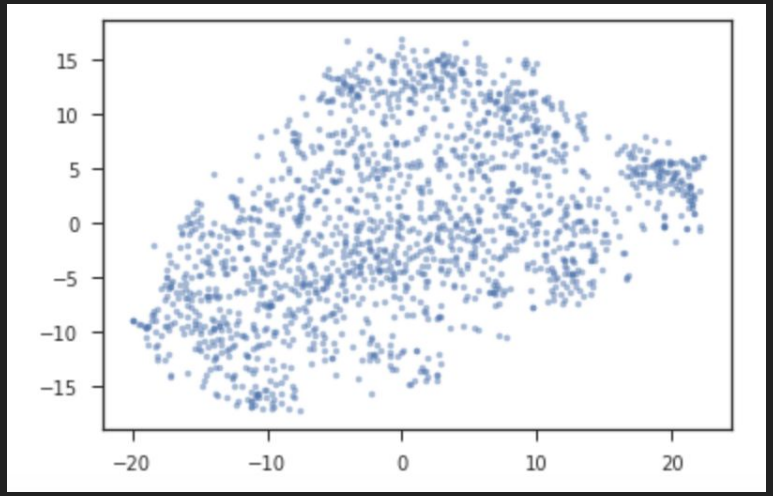
Correlation between the original and reconstructed gene expression
for randomly selected genes



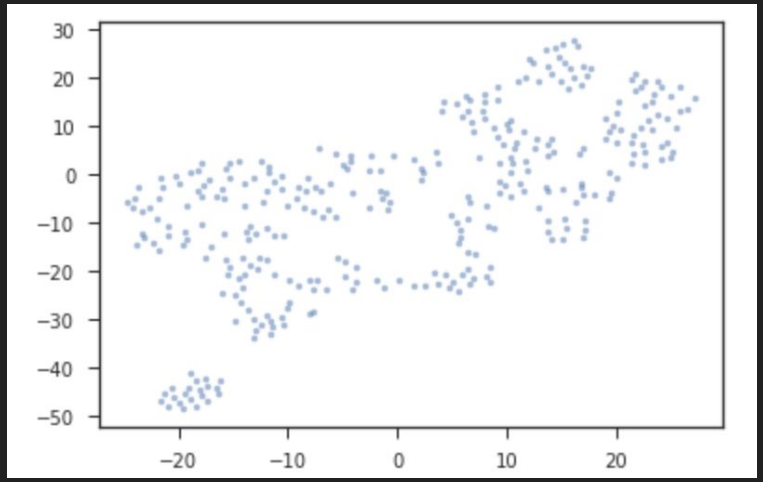
The variational autoencoder reconstructs fairly faithfully gene expression values from the latent space

Clustering results (Realty)

Gene cluster

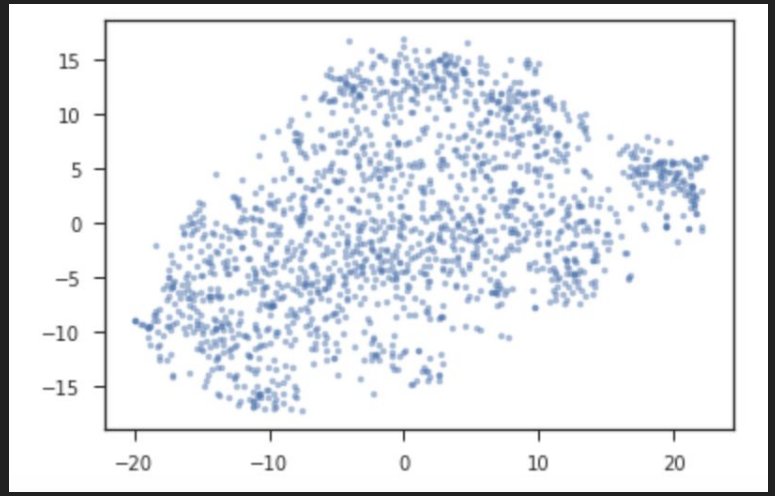


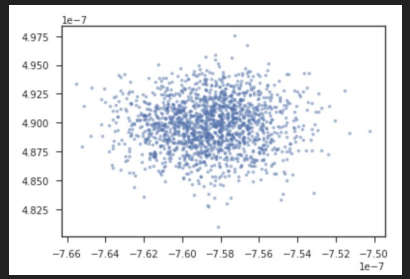
Patient cluster

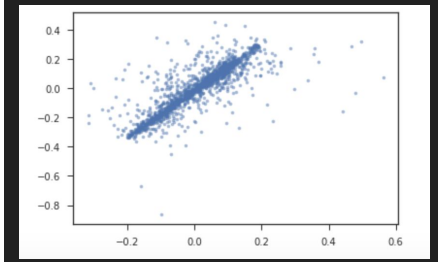


Clustering results (Realty)

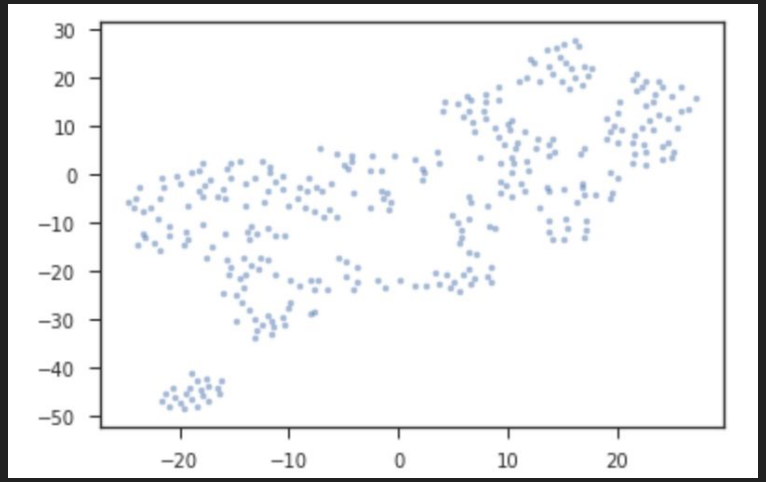
Gene cluster

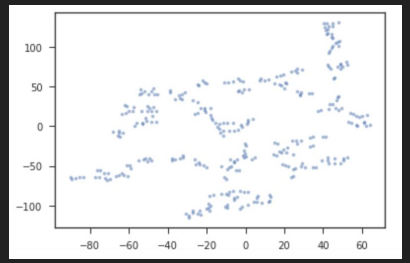


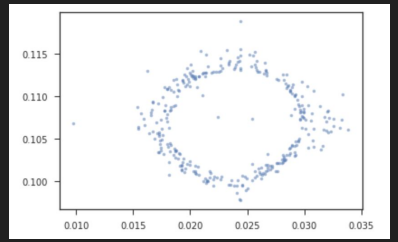




Patient cluster







Insights

VAE and t-SNE clustering

VAE

- Great with “-omic” data where we often lack truth/labels (unsupervised learning method for capturing meaningful features of the data)
- Latent encoding provides discoverable data structure
- Structure not always present in latent vector. Additional tuning required
- Random sampling approach requires more care with Cost function definition

t-SNE

Powerful method for embedding high-dimensional data in a low-dimensional space of two or three dimensions

- Efficient visualization of multi-dimensional space
- Very sensitive to certain parameters, e.g. perplexity, which is related to the number of nearest neighbors of other methods
- Requires careful evaluation and tuning