

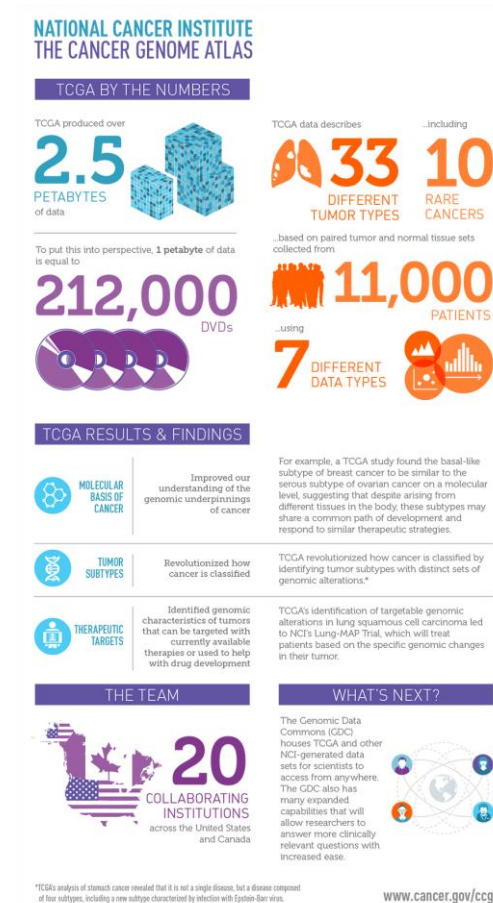
Differential expression & coexpression analysis of clear cell renal carcinoma vs normal cells

Dataset downloaded from The Cancer Genome Atlas

By: Frank M. Jenkins

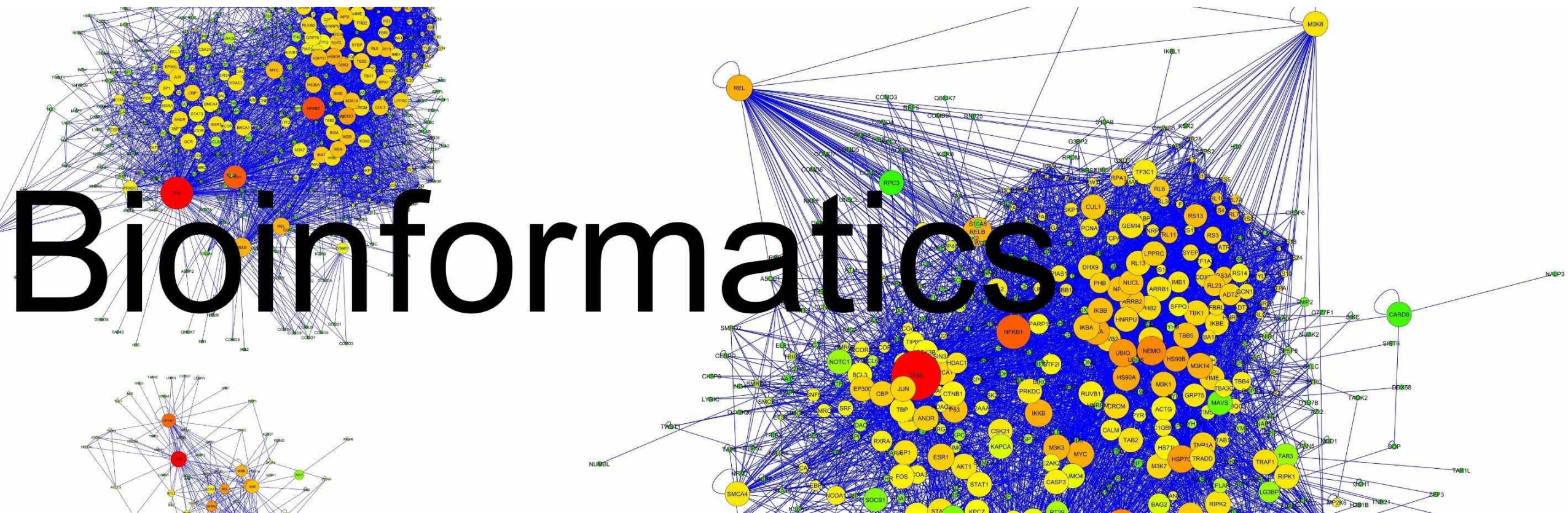
Springboard – Data Analysis with Python

The Cancer Genome Atlas (TCGA)



Clear cell renal carcinoma

- Clear cell renal carcinoma is a form of kidney cancer which primarily afflicts men in their sixties and seventies
- It is called 'clear cell' because the cells generally have a clear cytoplasm surrounded by a distinct cell membrane containing round and uniform nuclei
- In general, renal cell carcinomas are a kidney cancer that originates in the lining of the proximal convoluted tubule, a part of the network of small tubes in the kidney responsible for transporting urine
- Initial treatment involves partial or complete removal of the affected kidney, and absent metastasis, the five year survival rate post-surgery is 65 – 90%
- The greatest risk factors for RCC are lifestyle related e.g. obesity, smoking, and hypertension



Problem identification and target audience

- Novice bioinformaticians need accessible resources to learn programmatic techniques for genomic analysis
- There are many different approaches for solving common genomic problems
- In this analysis, we want to identify differences in gene expression between cancer and normal cells, in order to identify the genes involved in cancer formation and metastasis
- Identification of those genes allows us to develop strategies to inhibit the deleterious functions of oncogenes and oncoproteins

Workflow

- Find and import dataset
- Data wrangling
- Filtering
- Exploratory data analysis
- Correlation analysis
- Differential expression analysis
- Unsupervised learning
- Hypothesis testing
- Interpretation of results

Find and import data

- The data, gene expression values for cancer (clear cell renal carcinoma) and normal cells, was downloaded from TCGA as a csv file
- First, the working directory was changed to reflect the path where the dataset is stored
- Then, the dataset was assigned to a filename object
- Then, the dataset was imported into a Jupyter Notebook using the `pandas read_csv` command
- Finally, the columns containing the gene names was set as the index
- The dataset can be found at:
 - <https://tcga.xenahubs.net/download/TCGA.KIRC.sampleMap/HiSeqV2.gz>

Out[22]:

	TCGA- BP- 4162- 01	TCGA- CJ- 5677-11	TCGA- DV- 5566- 01	TCGA- BP- 5191- 01	TCGA- BP- 5200- 01	TCGA- BP- 4347- 01	TCGA- BP- 4770- 01	TCGA- B0- 5696-11	TCGA- BP- 4762- 01	TCGA- BP- 4158- 01	...	TCGA- B0- 5104- 01	TCGA- A3- 3313- 01	TCGA- B2- 5633- 01	TCGA- CJ- 4872- 01	TCGA- CJ- 5684- 01	TCGA- CJ- 4886- 01
sample																	
ARHGEF10L	10.5030	10.8969	10.7612	10.2063	10.0616	10.0193	8.4364	11.1427	9.9150	10.7621	...	10.5586	9.5556	10.2652	9.4981	10.2751	9.9521
HIF3A	5.5283	6.4943	5.7842	5.0063	5.3326	7.6102	6.0422	5.9789	3.7802	7.2039	...	6.2061	4.8253	5.1680	8.6195	6.2080	6.3111
RNF17	3.8036	0.0000	0.0000	0.0000	0.0000	0.3386	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RNF10	11.4379	12.2130	11.5478	12.3439	11.5149	11.4065	12.1963	12.3946	11.3734	11.3251	...	11.7338	11.7709	11.4968	11.9561	11.7518	11.5691
RNF11	11.4180	11.8248	11.3190	10.3413	11.2923	10.9971	11.8936	11.7217	12.2329	11.2993	...	11.7629	10.2373	11.4637	11.4391	11.3992	11.4937

5 rows × 606 columns



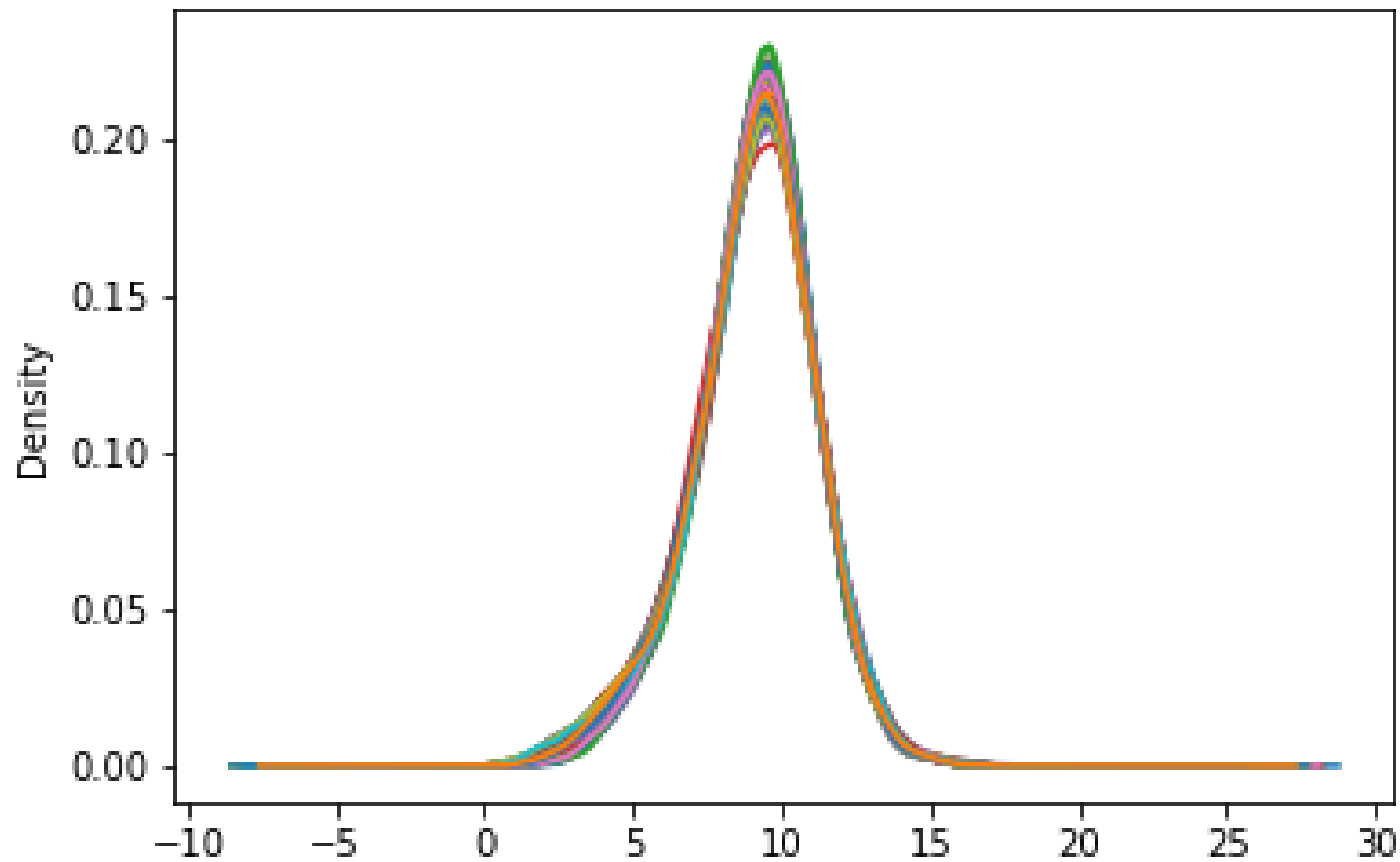
Data Preparation

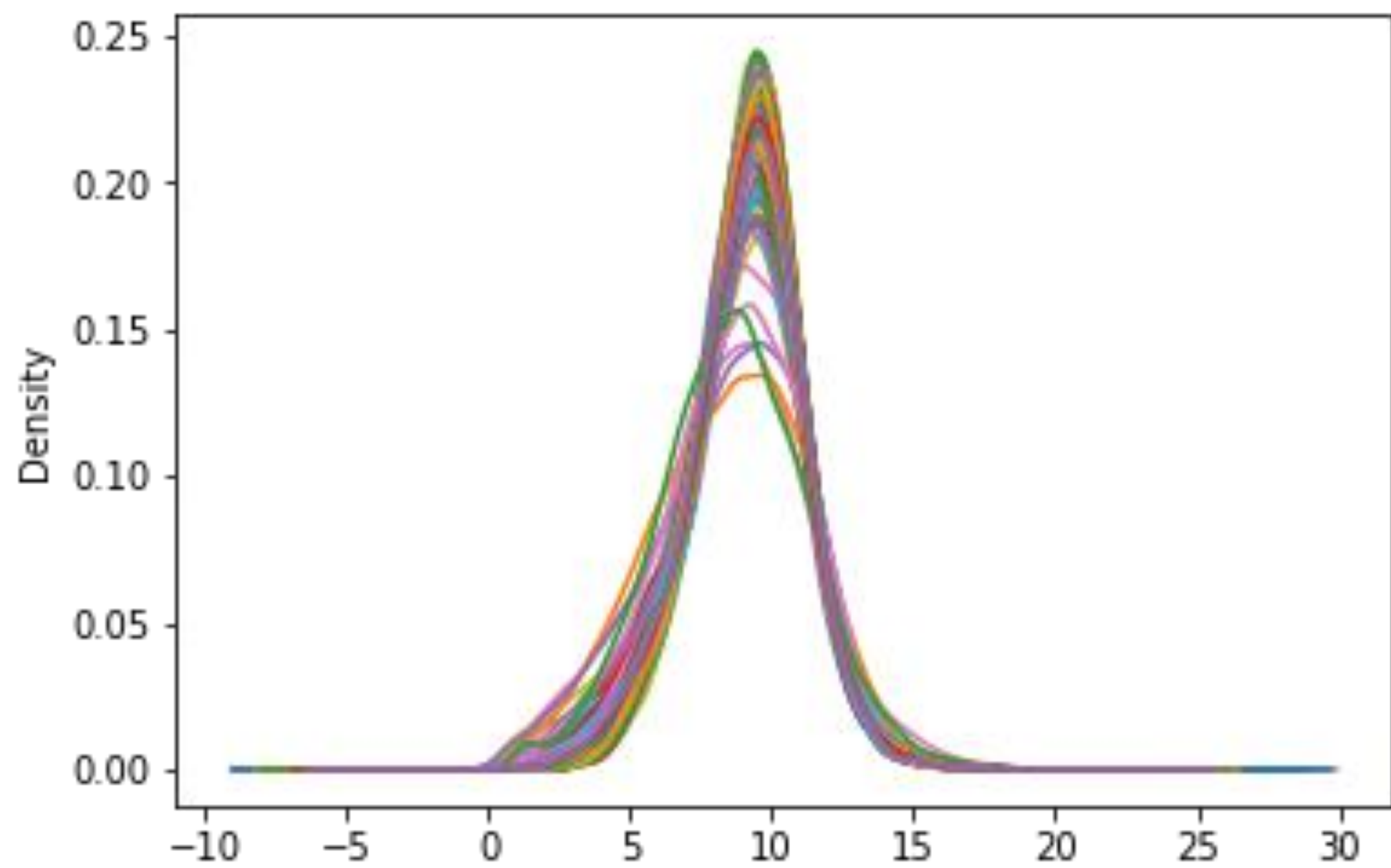


Data wrangling and filtering

- All rows with zero values were dropped.
- The dataset was split into two Pandas dataframes:
 - The first dataframe contains normal cell gene expression values
 - The second dataframe contains cancer cell gene expression values
- The original dataset has been normalized and log2 transformed.

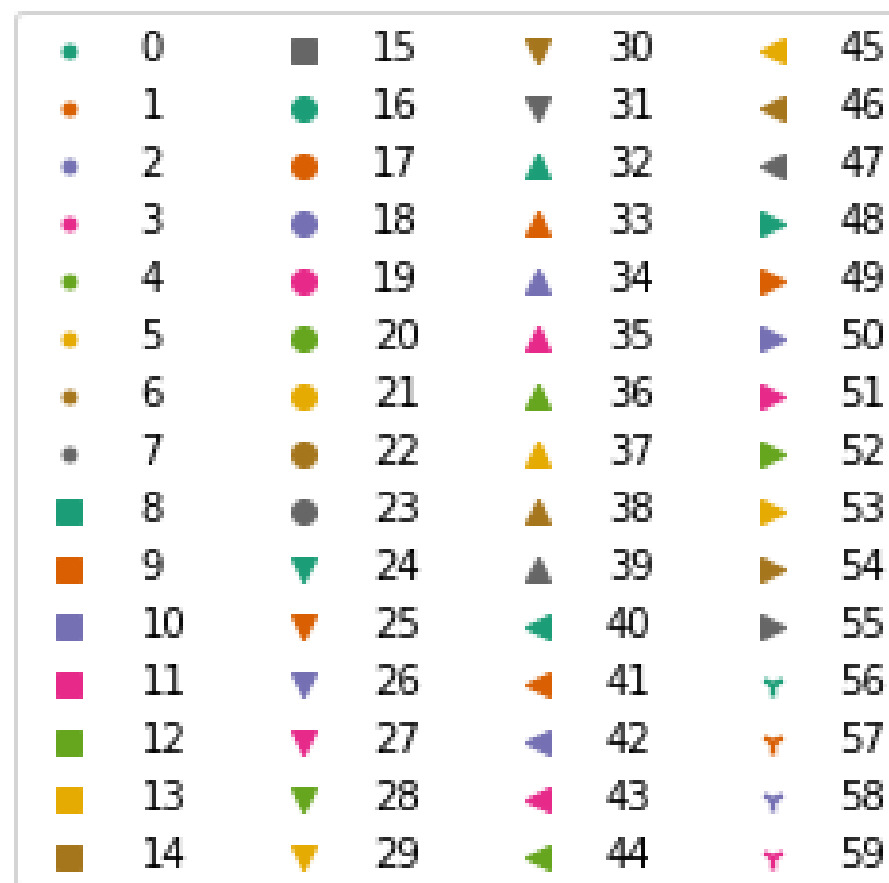
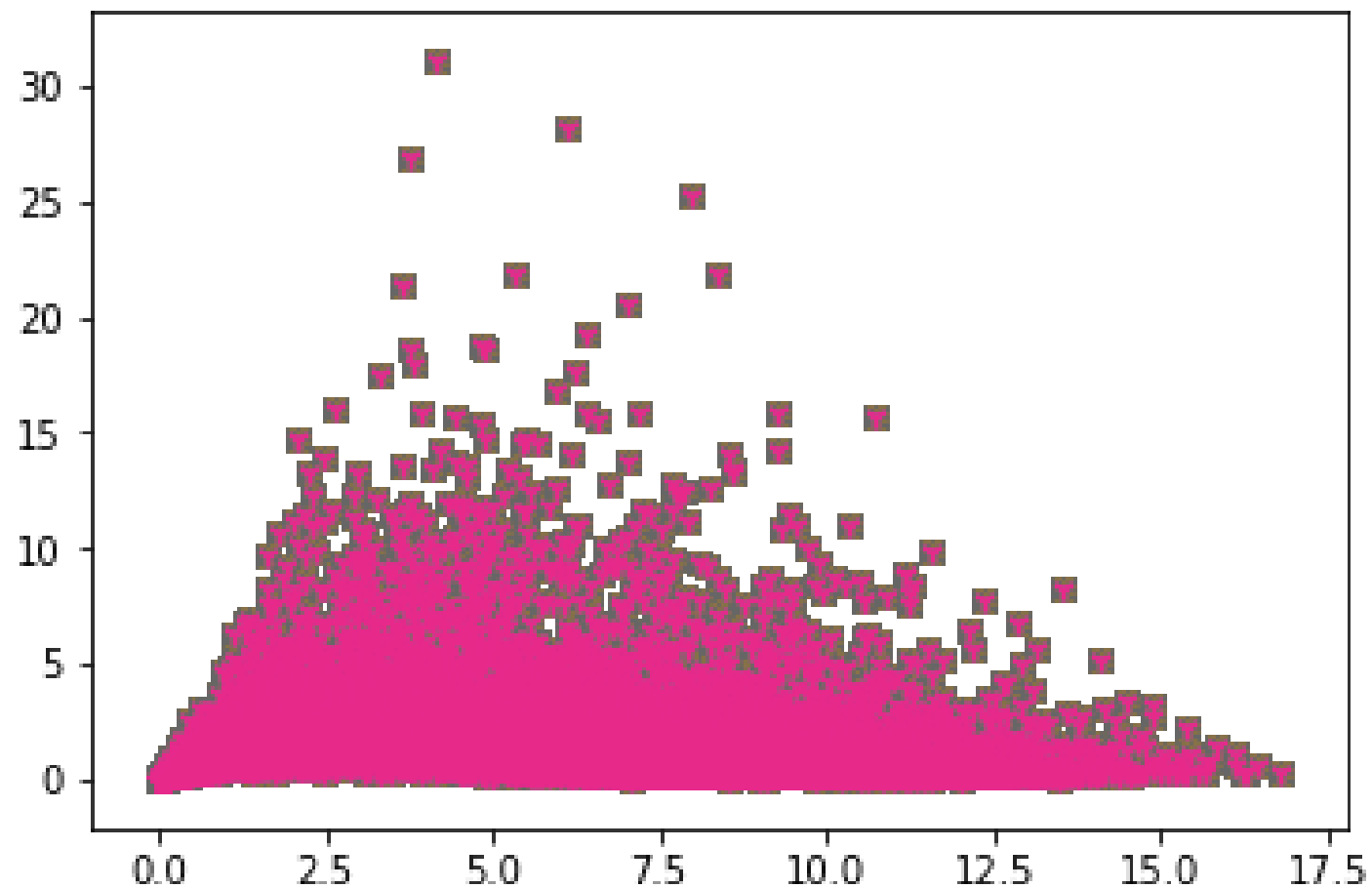
Exploratory Data Analysis

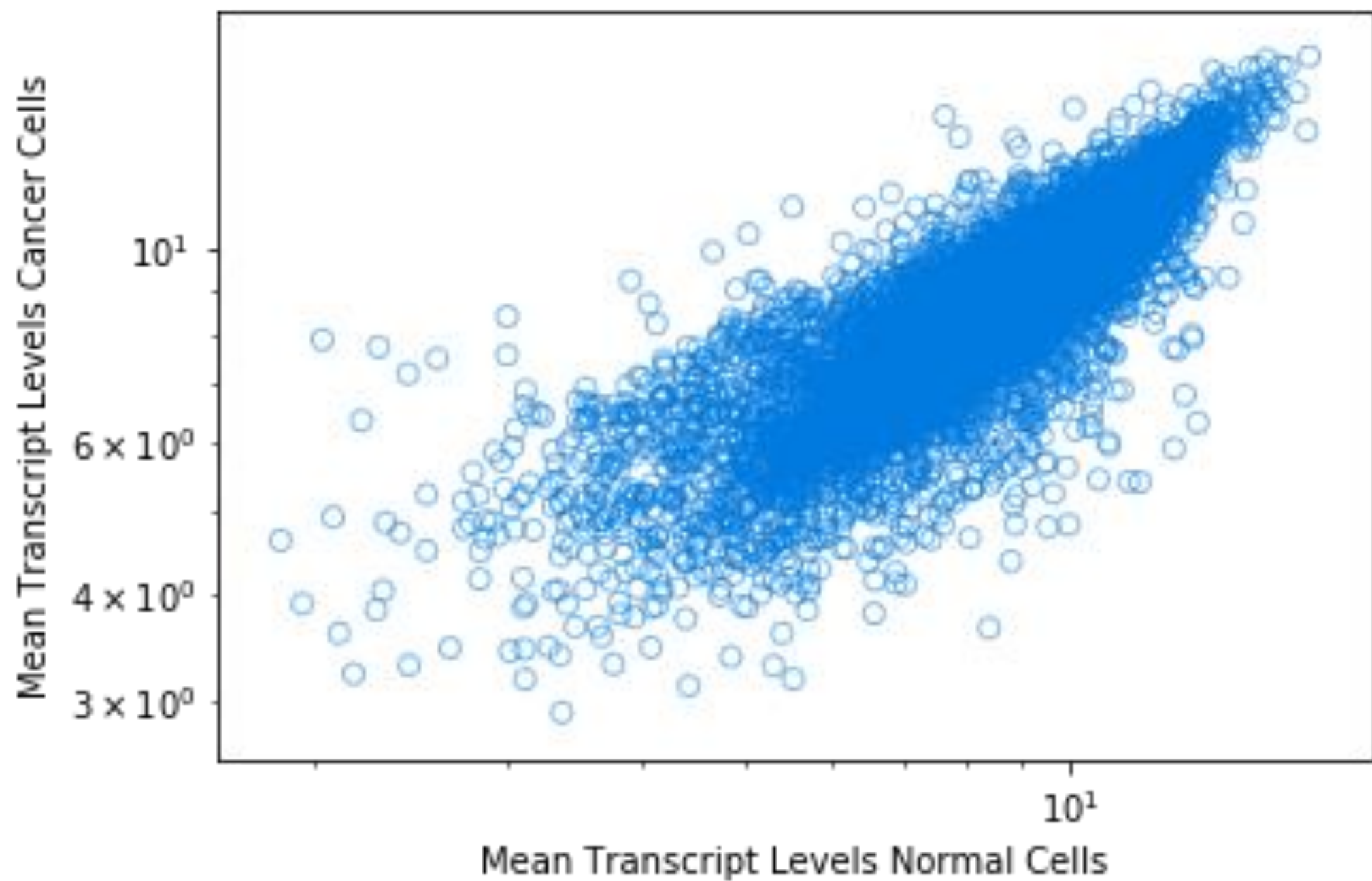




Confirm the shape of the distribution

- The two plots shown above are kde (density) plots
- We can use kde plots to check the normality of our data
- The first plot plots normal cell expression values
- The second plot plots cancer cell expression values
- As shown, the data appears to be normally distributed





Confirm that true biological differences exist between cancer and normal cell gene expression

- The first scatter plot (above) plots the mean against the variance of the entire dataset
- The second scatter plot plots mean expression levels for cancer genes against normal genes
- These plots suggest real biological differences between cancer and normal gene expression levels

Differential expression analysis

The t-test method was used for this analysis

- There are a number of different methods which can be used for differential expression analysis
- They include the t-test method and other methods such as empirical Bayes
- In this case, because the number of samples is large and the data is normally distributed, the t-test method was appropriate
- The two images below show the t-test formula and the top 25 differentially expressed genes

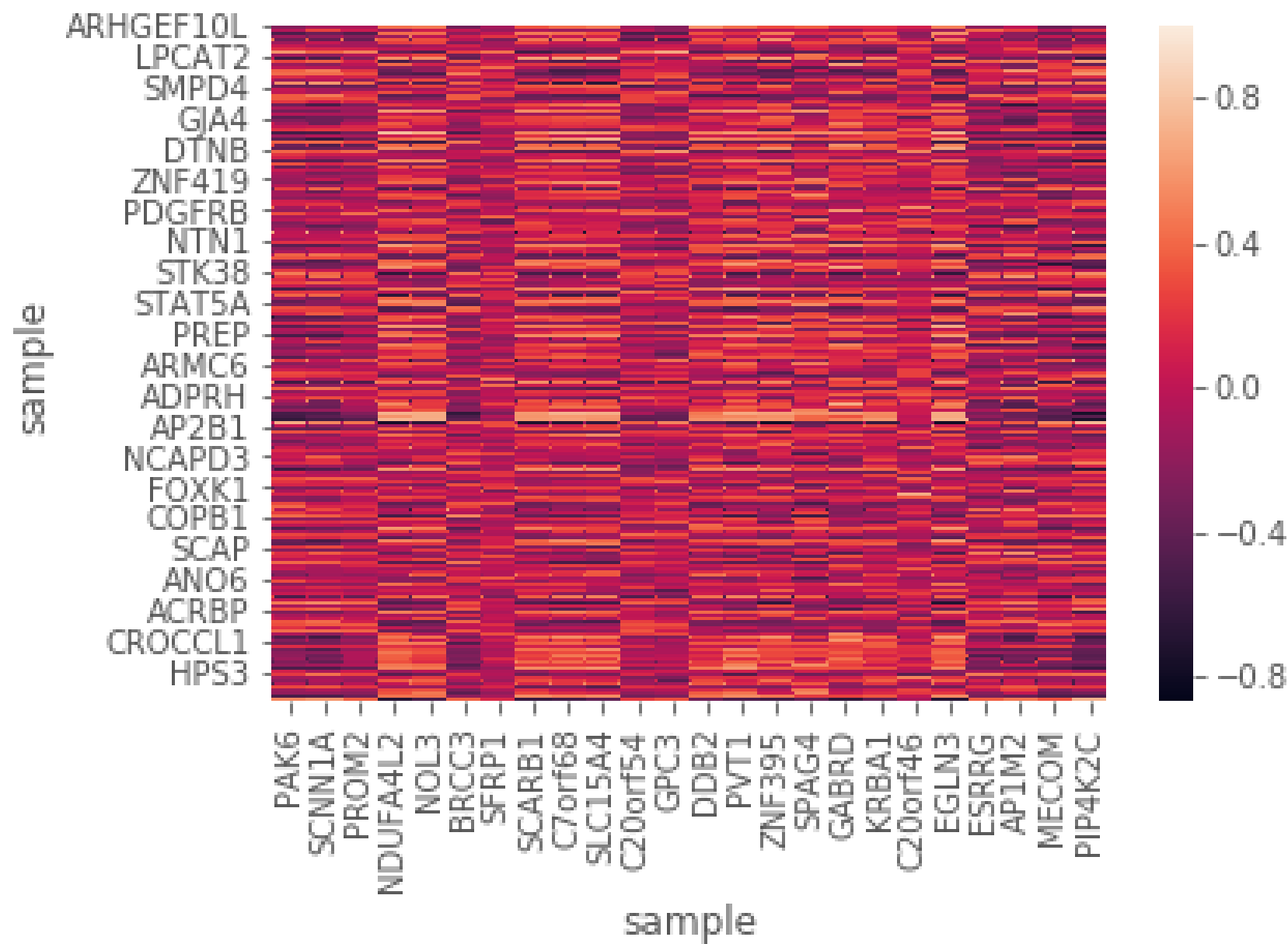
$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}}$$

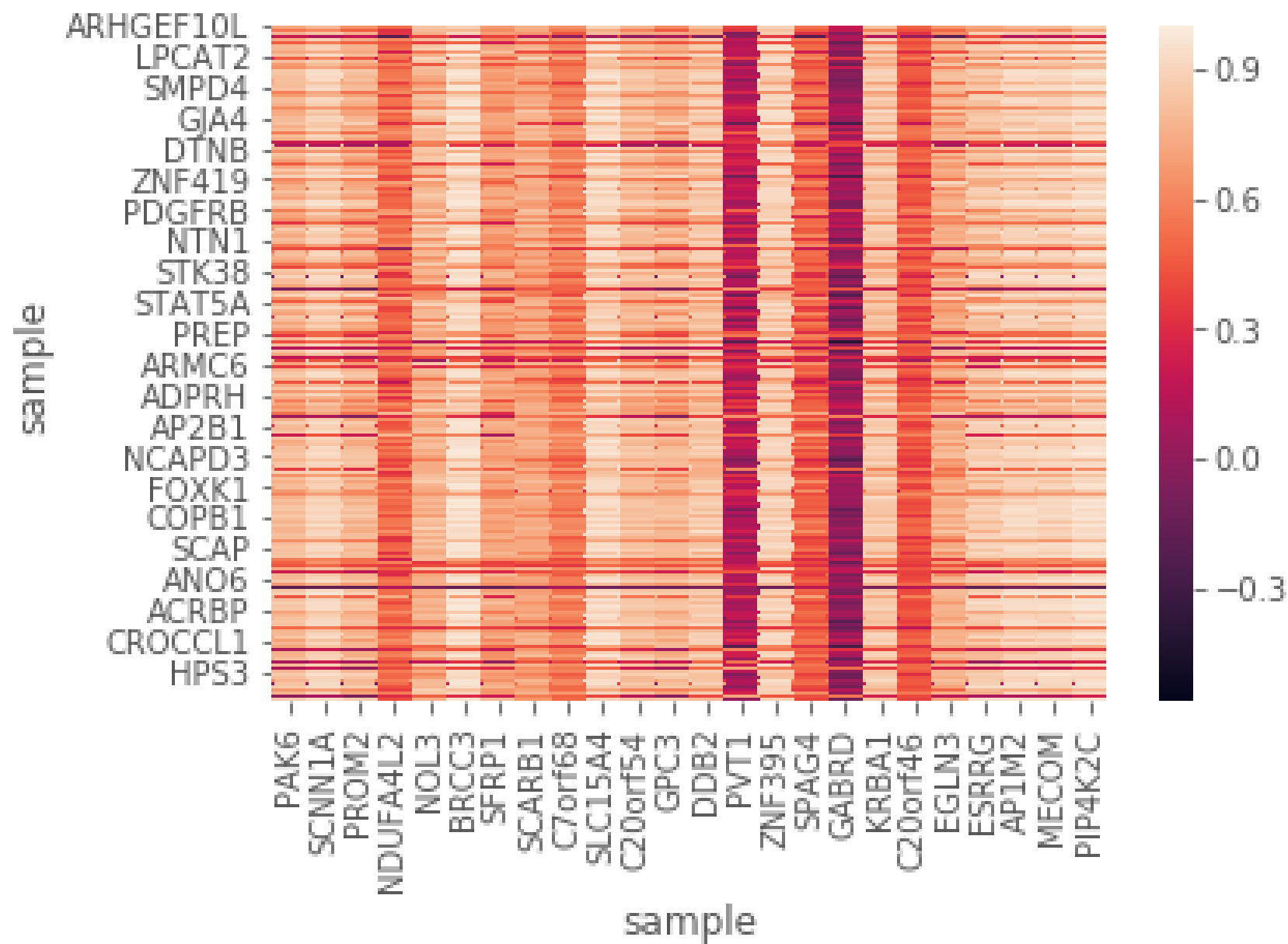
```
In [59]: results = diffexp.iloc[1:25]
results
```

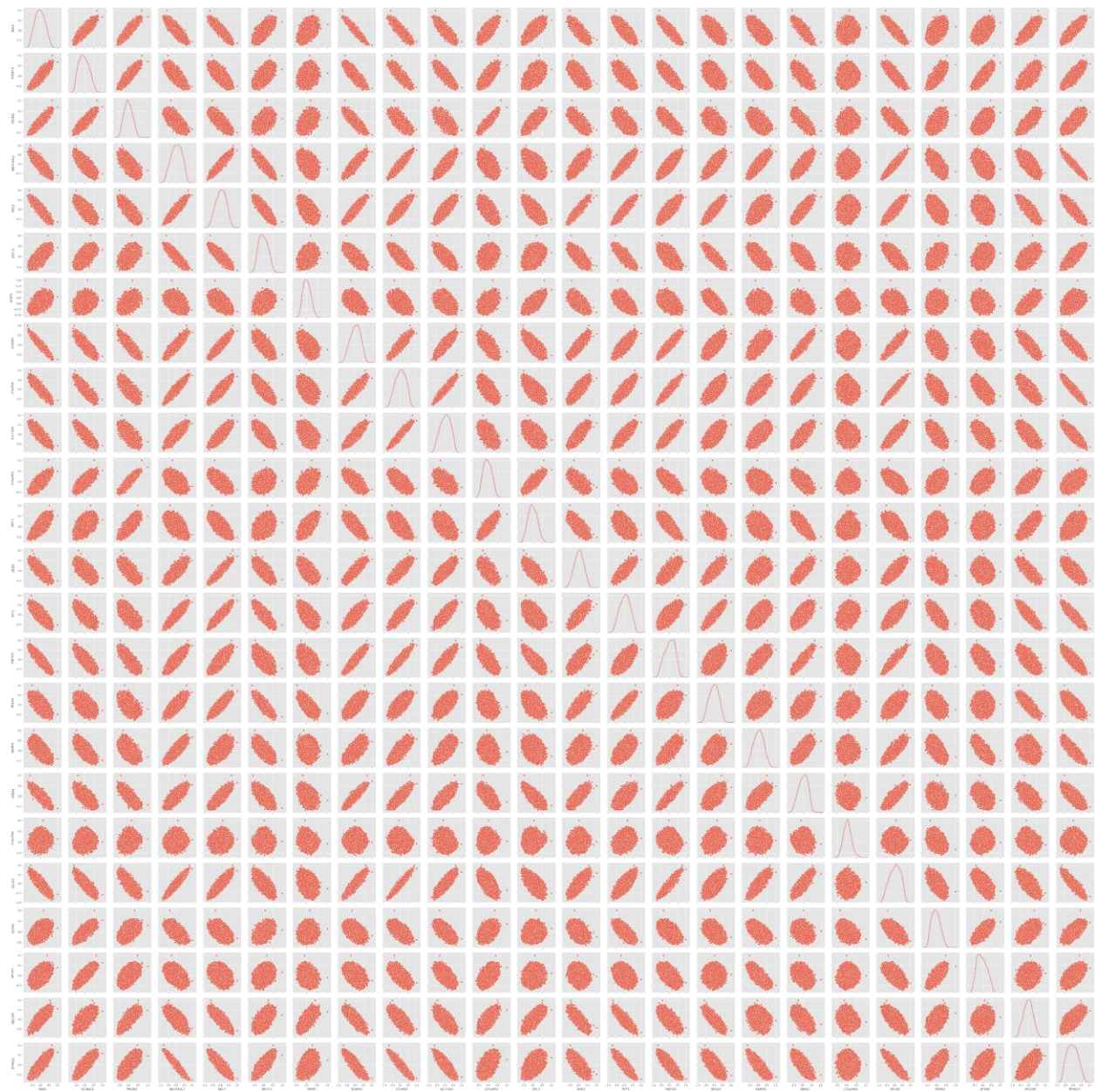
```
Out[59]: sample
PAK6      45.037406
SCNN1A    41.761799
PROM2     41.630070
NDUFA4L2  41.419890
NOL3      41.102853
BRCC3     39.666073
SFRP1     39.578176
SCARB1    39.336044
C7orf68   39.328517
SLC15A4   38.738762
C20orf54  37.866042
GPC3      37.633499
DDB2      37.573465
PVT1      37.449893
ZNF395    36.747181
SPAG4     36.731714
GABRD     36.592500
KRBA1     36.530676
C20orf46  36.410145
EGLN3     36.152210
ESRRG     36.070852
AP1M2     36.035638
MECOM     35.674698
PIP4K2C   35.659240
Name: ttest, dtype: float64
```

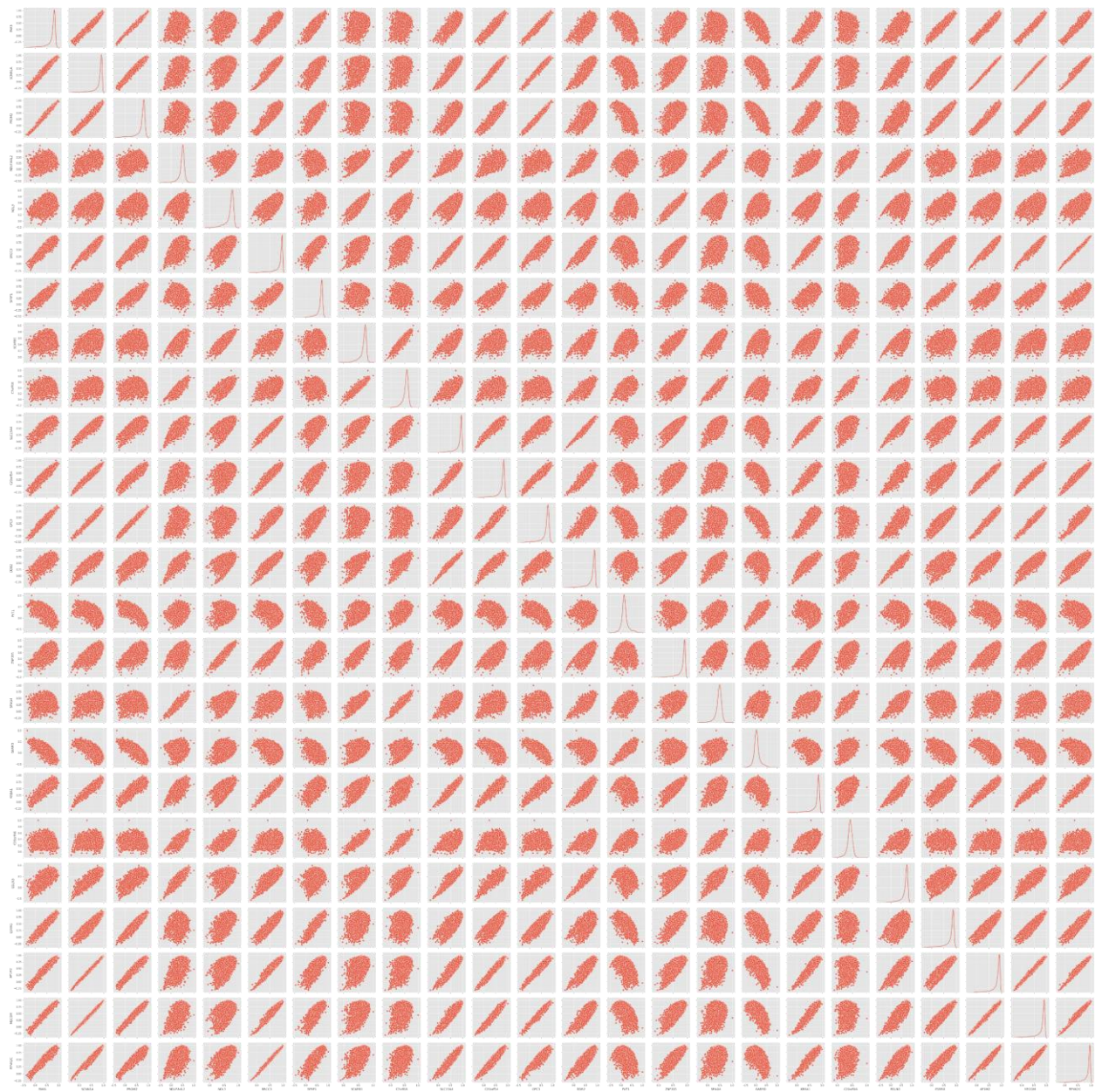
Correlation analysis

- Correlation analysis was done using the Pandas `corr()` method
- Heatmaps were generated showing correlation with our differentially expressed genes
- We notice a large number of negatively correlated genes in the cancer expression dataframe
- Conversely, we see a large number of positively correlated genes in the normal cell expression dataframe
- This is not surprising, since oncogenes produce deleterious proteins which often inhibit important cellular functions such as cell cycle control and apoptosis
- Pairs plots were also generated to further demonstrate the trends which appear in the heatmaps









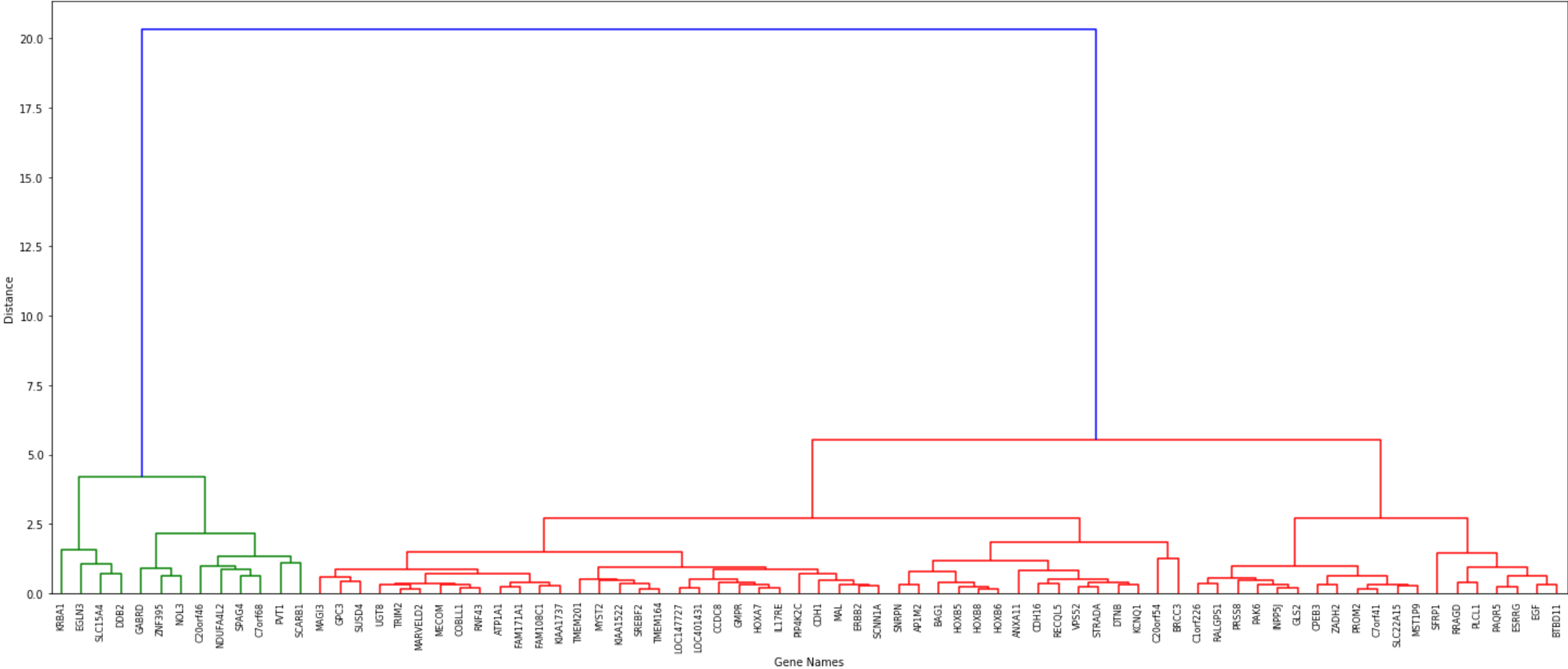
Top correlated genes

- Next, a dataframe was created containing correlated values for the top 25 differentially expressed genes
- In order to perform the analysis, the number of replicates in the cancer gene expression dataframe was reduced to 74
- Then, a correlation analysis was done which identified genes with expression values closely correlated with the top 25 differentially expressed genes
- This gene set was then used in hierarchical clustering

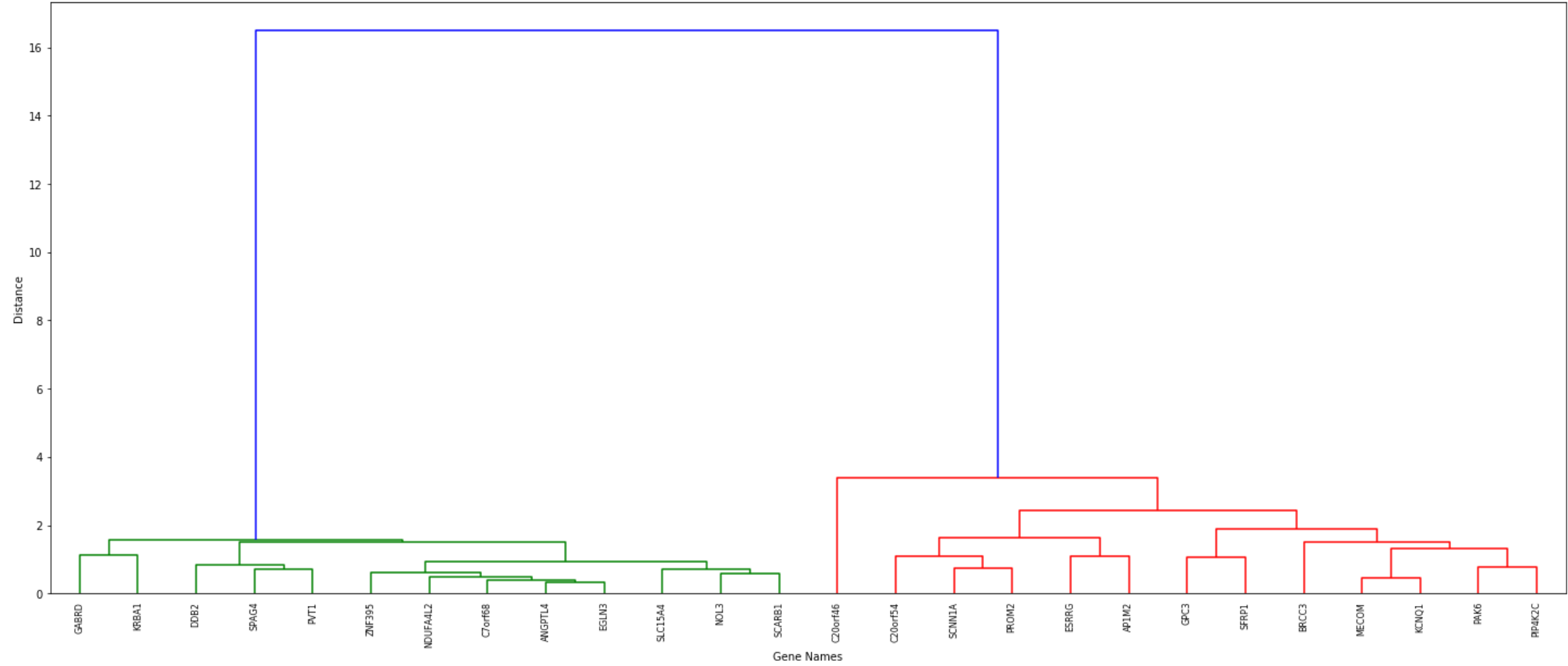
Unsupervised learning

- As detailed above, hierarchical clustering was done on genes closely correlated with the top 25 differentially expressed genes
- This operation was done to identify a gene candidate with a unique expression profile
- The gene ANGPTL4 was identified
- A survey of the literature was done to elucidate on the role ANGPTL4 plays in tumorigenesis

Hierarchical Clustering Dendrogram



Hierarchical Clustering Dendrogram



NULL Hypothesis: ANGPTL4 is not overexpressed in clear cell renal carcinoma tumor cells relative to expression levels in normal cells

Alternate Hypothesis: ANGPTL4 is overexpressed in clear cell renal carcinoma tumor cells relative to expression levels in normal cells

```
In [14]: df1.loc['ANGPTL4'].mean(axis=0)
```

```
Out[14]: 7.894029166666665
```

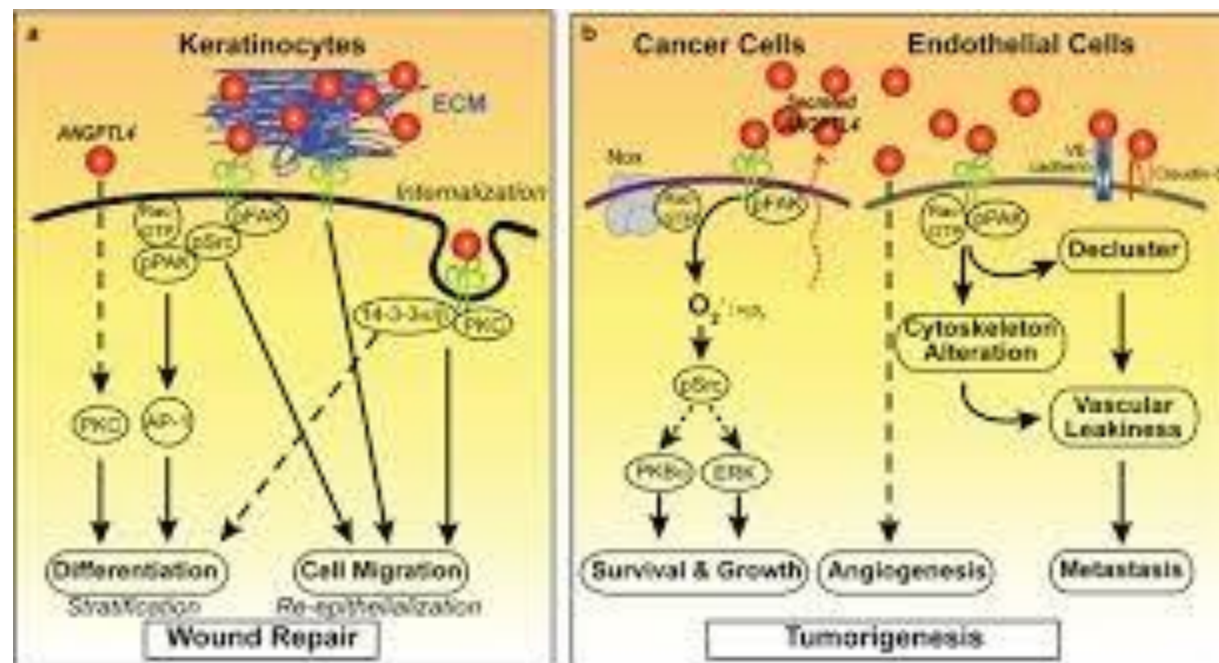
```
In [15]: df2.loc['ANGPTL4'].mean(axis=0)
```

```
Out[15]: 13.540759626168226
```

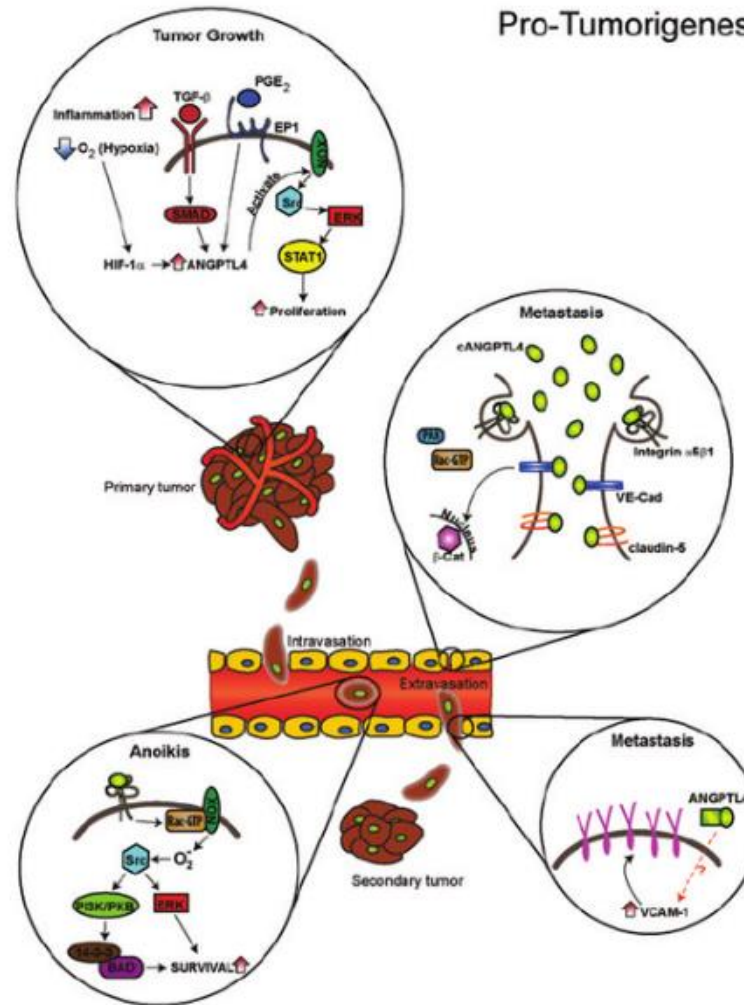
```
In [16]: from scipy import stats  
stats.ttest_ind(df1.loc['ANGPTL4'], df2.loc['ANGPTL4'])
```

```
Out[16]: Ttest_indResult(statistic=-23.8749431619333, pvalue=2.8860803388857175e-89)
```

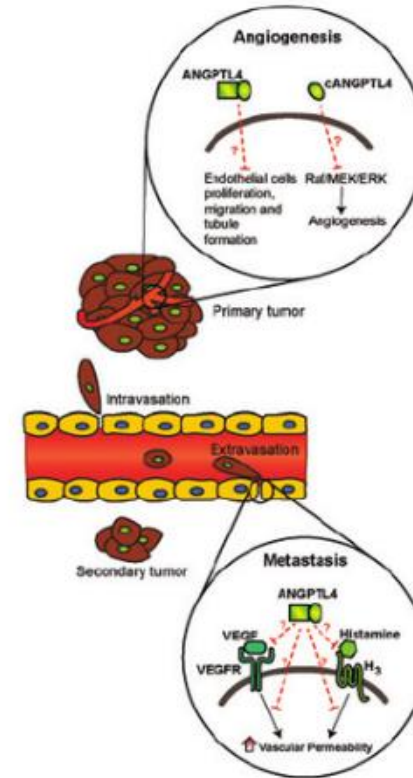
- Our p-value is < 0.05 , providing strong evidence against the NULL hypothesis
- Therefore, we can reject the NULL hypothesis
- This shows that ANGPTL4 is overexpressed in clear cell renal carcinoma tumor cells compared to expression levels in normal cells



Pro-Tumorigenesis



Anti-Tumorigenesis



Conclusion

- As shown, our p-value is < 0.05 , therefore we reject the NULL hypothesis, and we can assert that ANGPTL4 is overexpressed in clear cell renal carcinoma cells
- It is important to note, although there has been extensive research elucidating on the role ANGPTL4 plays in tumorigenesis and important biological functions such as glucose homeostasis, we were able to identify ANGPTL4 with no prior knowledge
- This project demonstrates basic techniques for gene expression analysis, and shows how open ended data analysis can spearhead biological research and hypothesis development

References

- The Cancer Genome Atlas: About TCGA, retrieved July 31, 2018, available at, <https://cancergenome.nih.gov/abouttcga>
- Clear Cell Renal Cell Carcinoma, Wikipedia, last edited Nov. 28, 2017, available at, https://en.wikipedia.org/wiki/Clear_cell_renal_cell_carcinoma
- Renal Cell Carcinoma, Wikipedia, last edited June 14, 2018, available at, https://en.wikipedia.org/wiki/Renal_cell_carcinoma
- Jie Tan, Ming & Teo, Ziqiang & Sng, Ming Keat & Zhu, Pengcheng & Tan, Nguan Soon. (2012). Emerging Roles of Angiopoietin-like 4 in Human Cancer. Molecular cancer research : MCR. 10. 677-88. 10.1158/1541-7786.MCR-11-0519
- Z. Pengcheng et al., A decade of research, Bioscience Reports (Dec. 22, 2011)