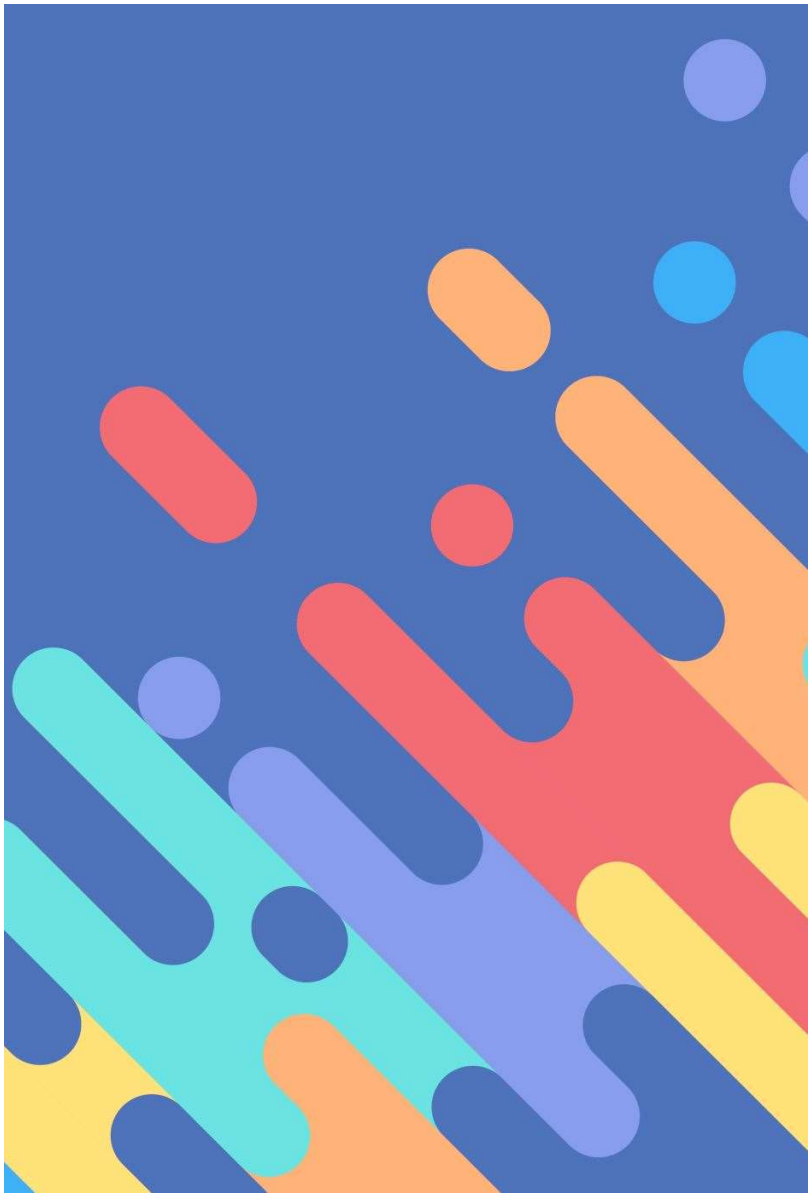
Use of an XGBoost classifier to explain tumor resistance in bulk RNA seq from scRNAseq

Marco A. Tello Palencia

CPSC-545 Final project presentation



RESEARCH ARTICLES

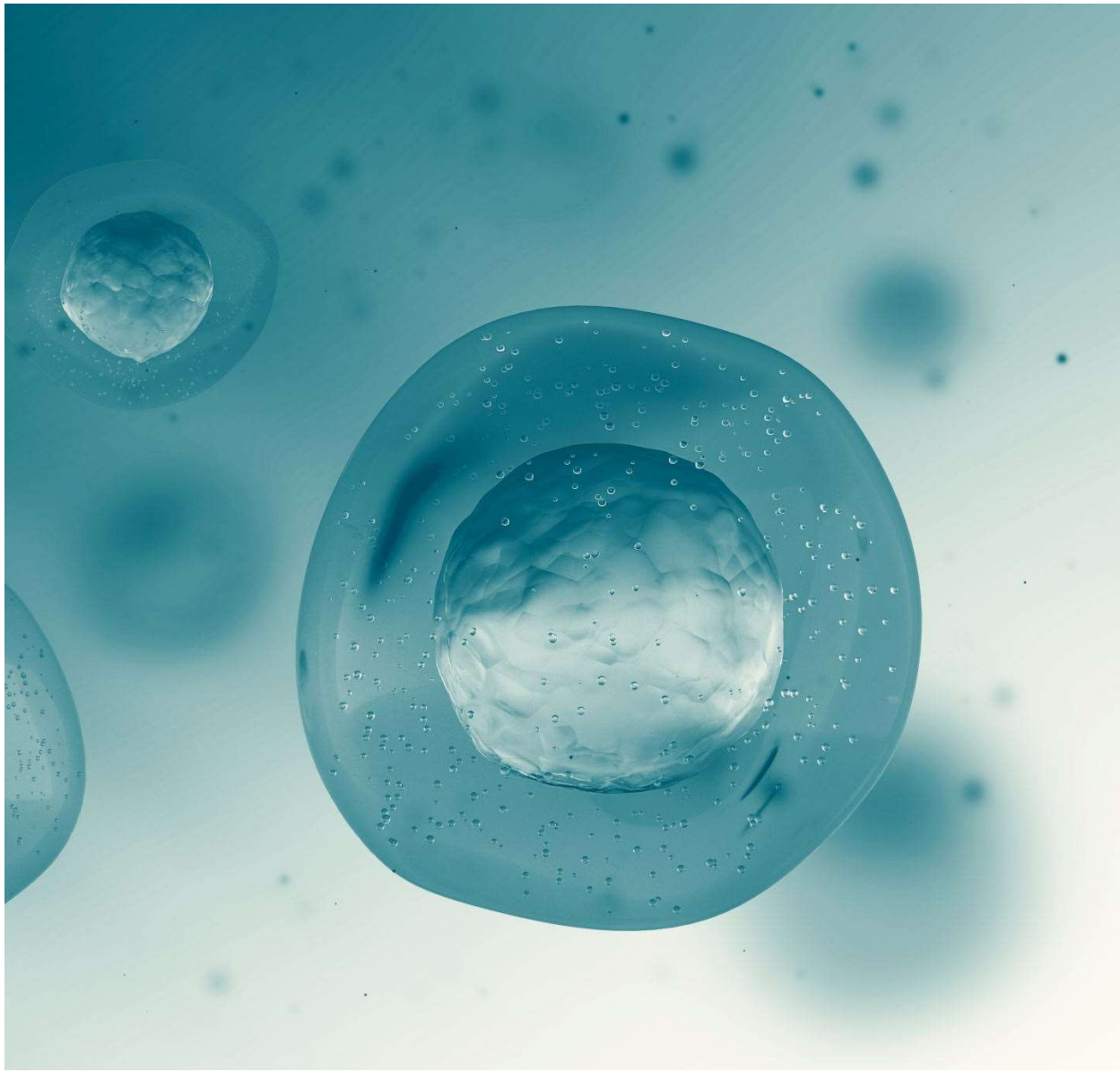
CANCER GENOMICS

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

Itay Tirosh,[1*] Benjamin Izar,[1,2,3*†‡] Sanjay M. Prakadan,[1,4,5,6] Marc H. Wadsworth II,[1,4,5,6] Daniel Treacy,[1] John J. Trombetta,[1] Asaf Rotem,[1,2,3]

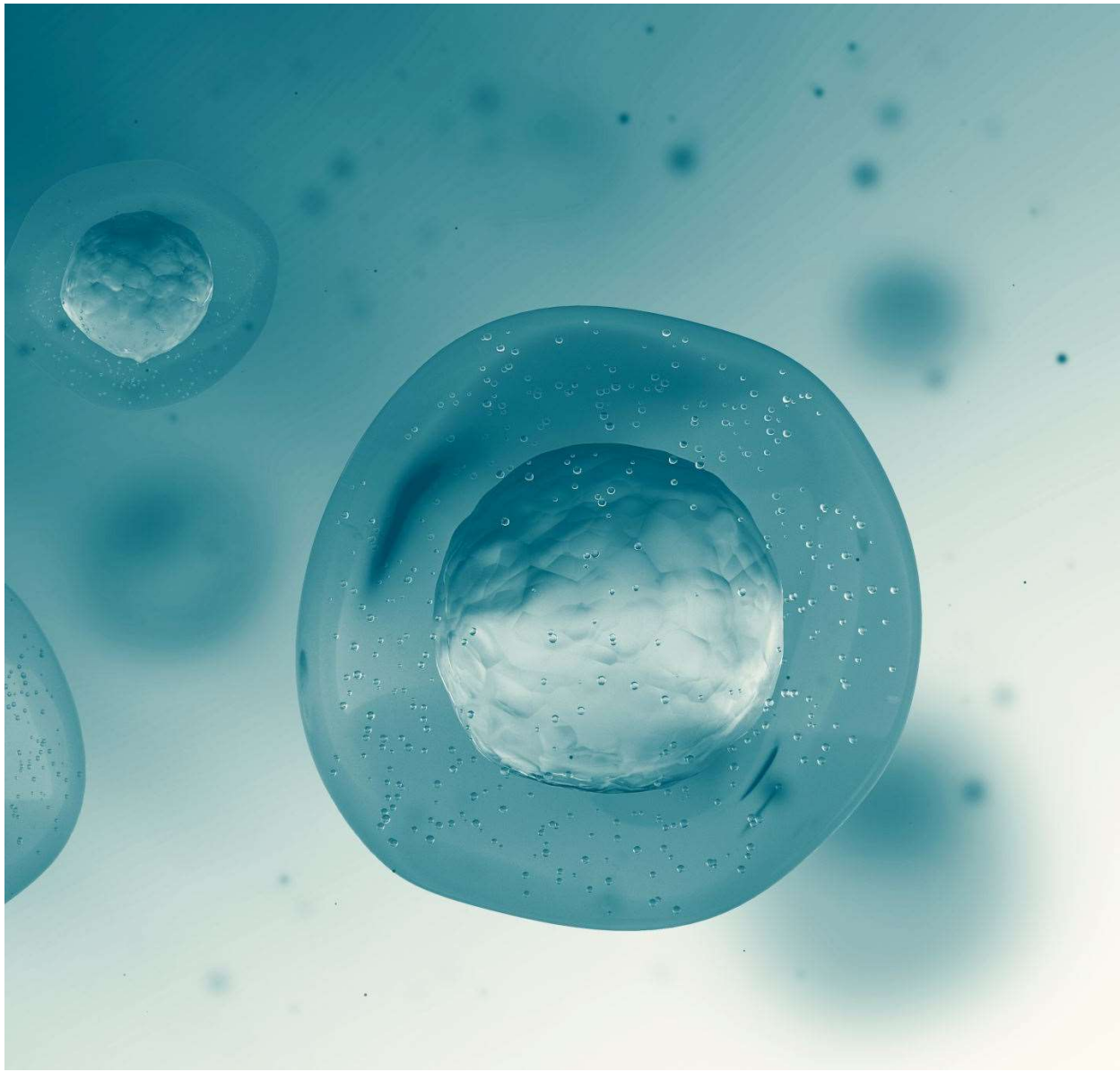
Tumors are complex ecosystems

- Tumors, intricate ecosystems shaped by diverse cell types.
- Interactions among malignant, immune, and stromal cells crucial for cancer development (1).
- Cellular composition and interplay's pivotal roles in tumor behavior (2).



Melanoma cell composition and treatment resistance

- It is possible that subsets of malignant cells and the microenvironment play essential roles in the response to treatments (3).
- Melanomas with the BRAF V600E mutation have a defined treatment however most tumors with this mutation develop resistance (4,5)



The dataset

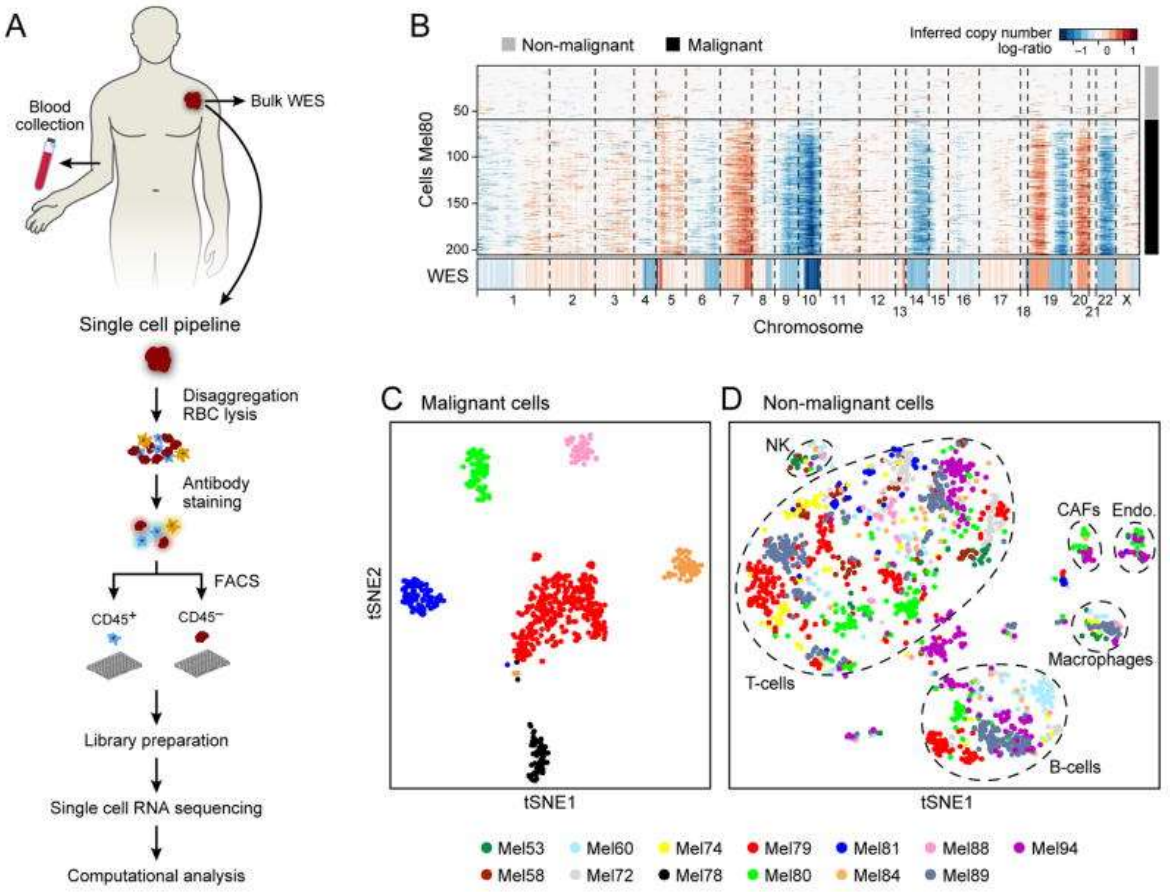
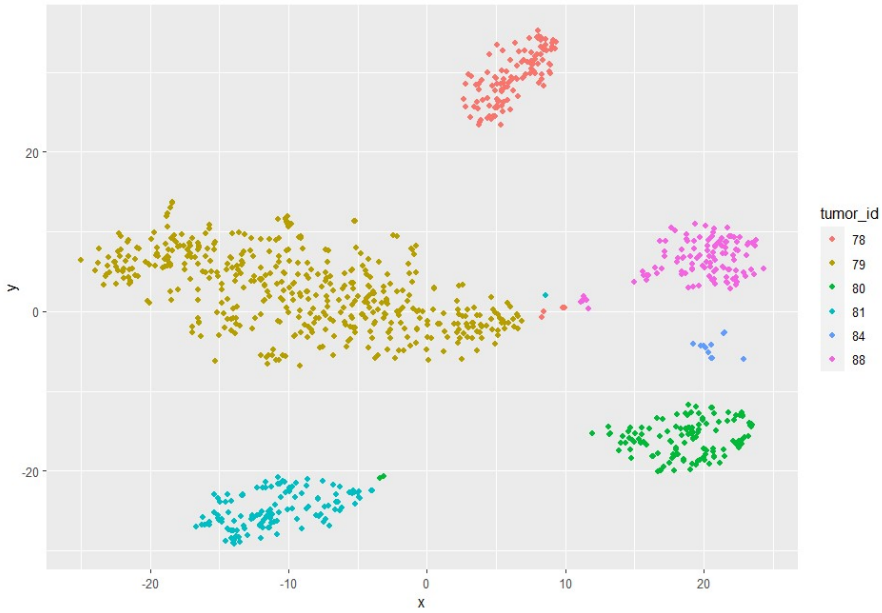
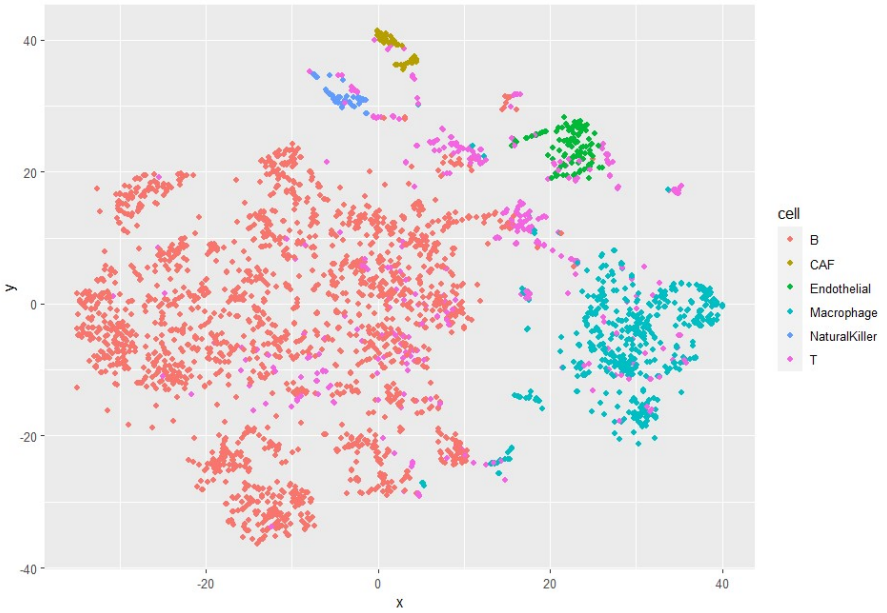


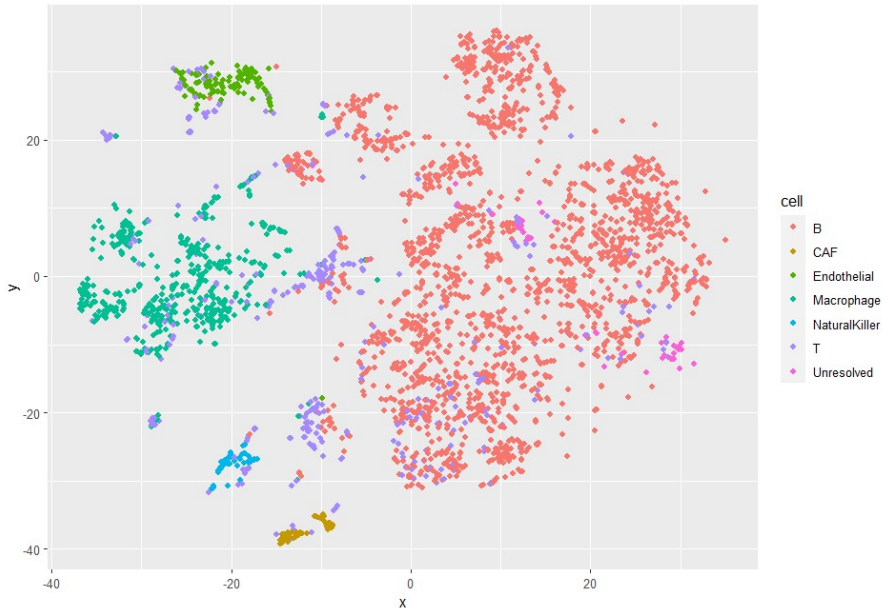
Figure 1. Tirosh, et al 2016

Exploratory data analysis – reproduce plots

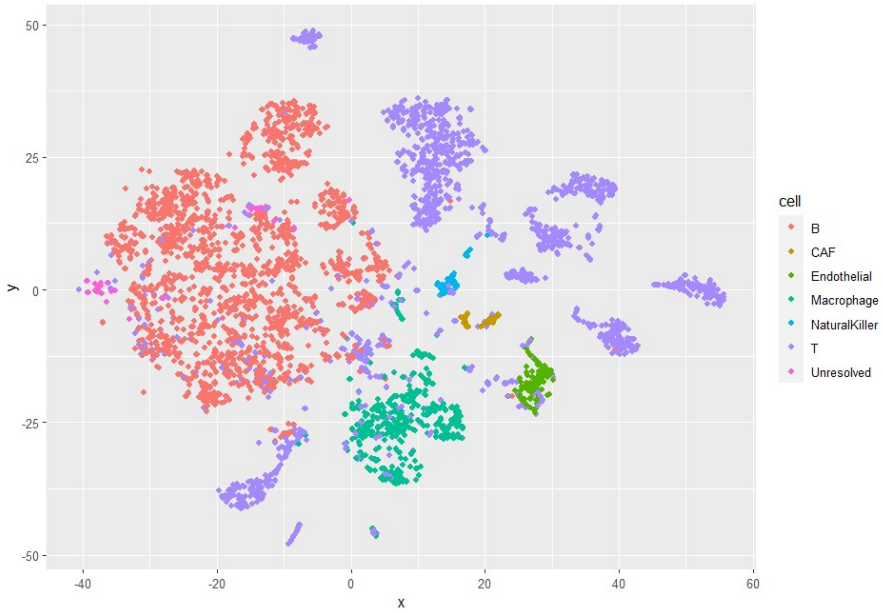


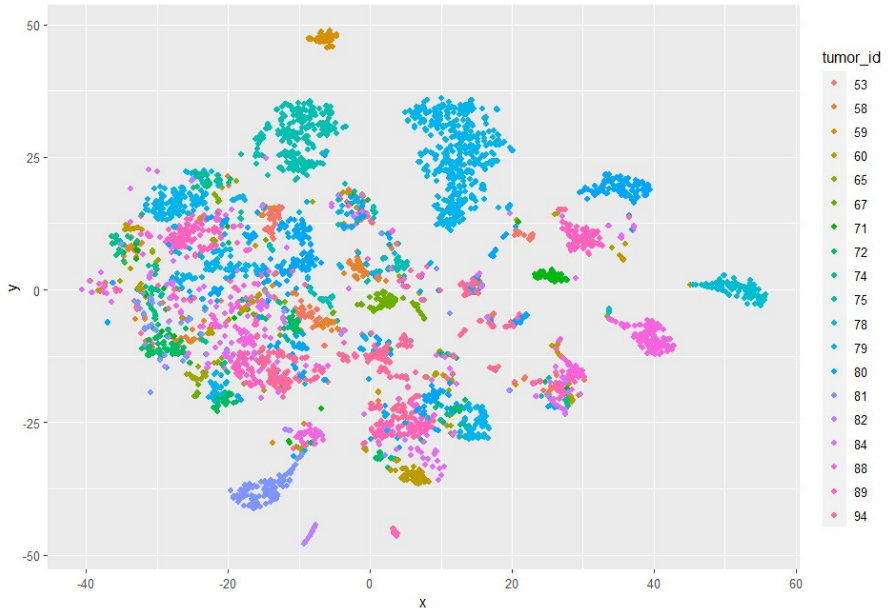


Exploratory data analysis – using the full data

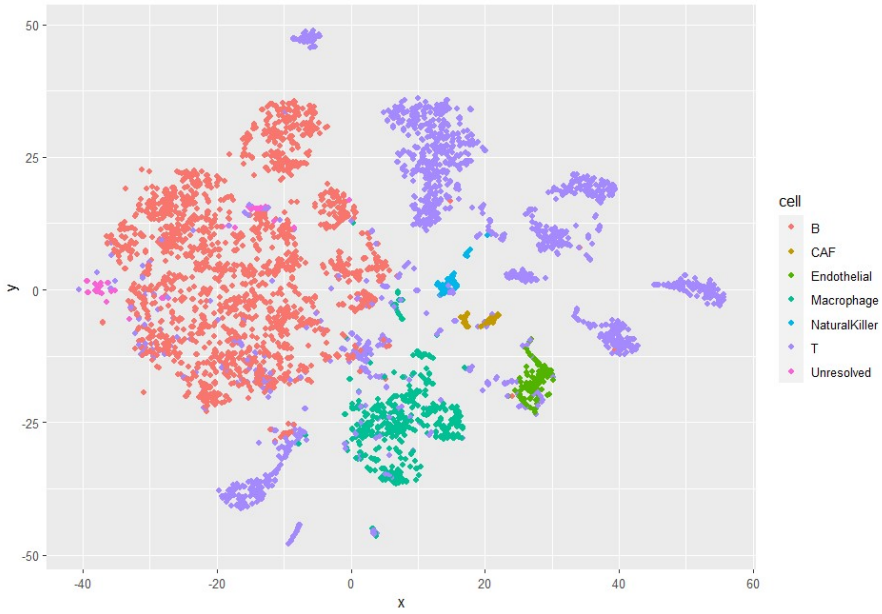


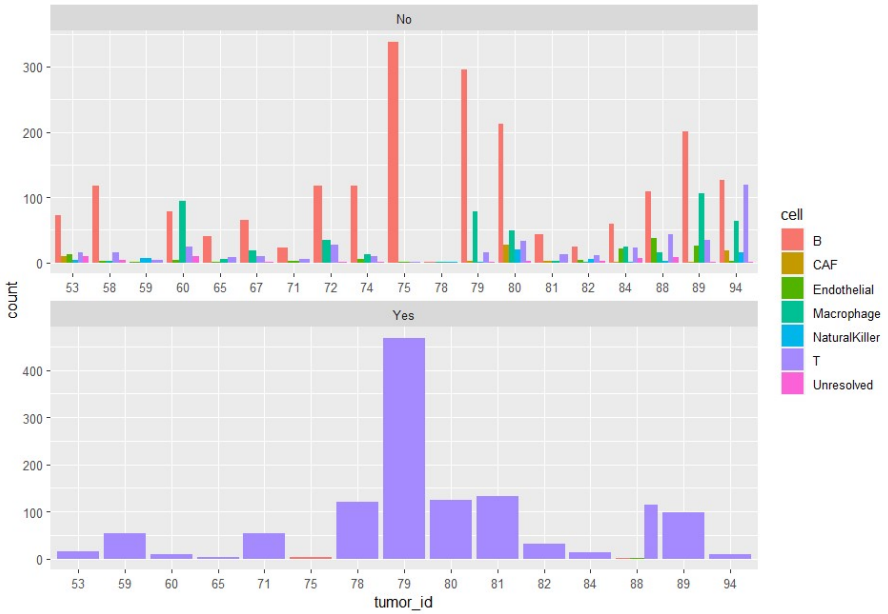
Initial plan – Predict malignancy using cell type expression profiles





The problem… malignancy and cell type are associated





Complementary data from bulk-RNAseq

Samples from donors before and after development of resistance

Identified transcriptomic profiles associated with this change

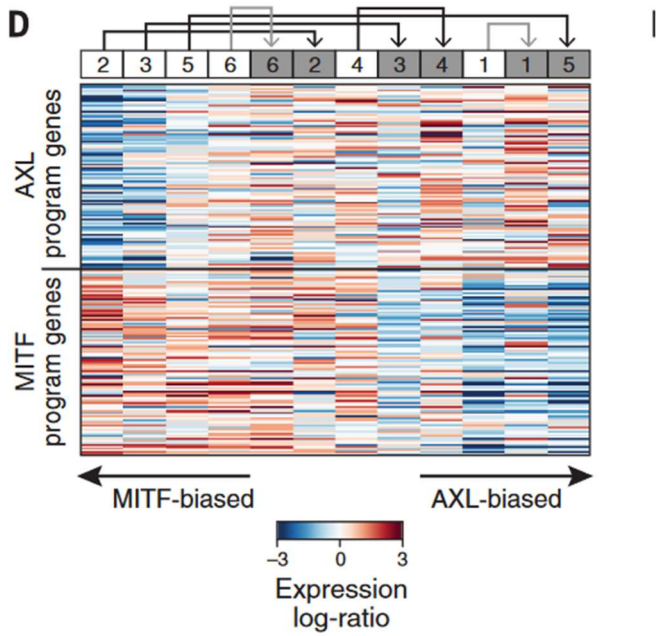
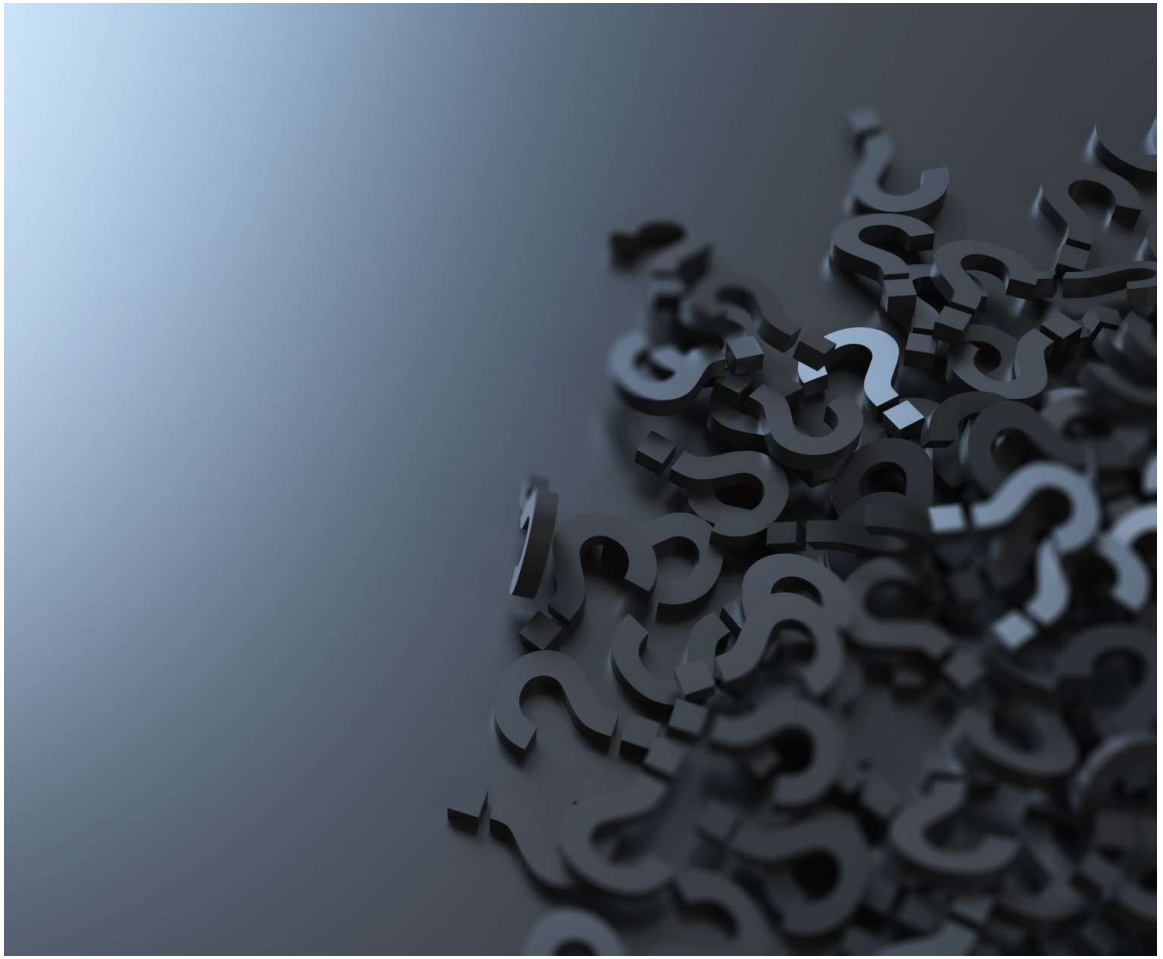


Figure 3D. Tirosh, et al 2016

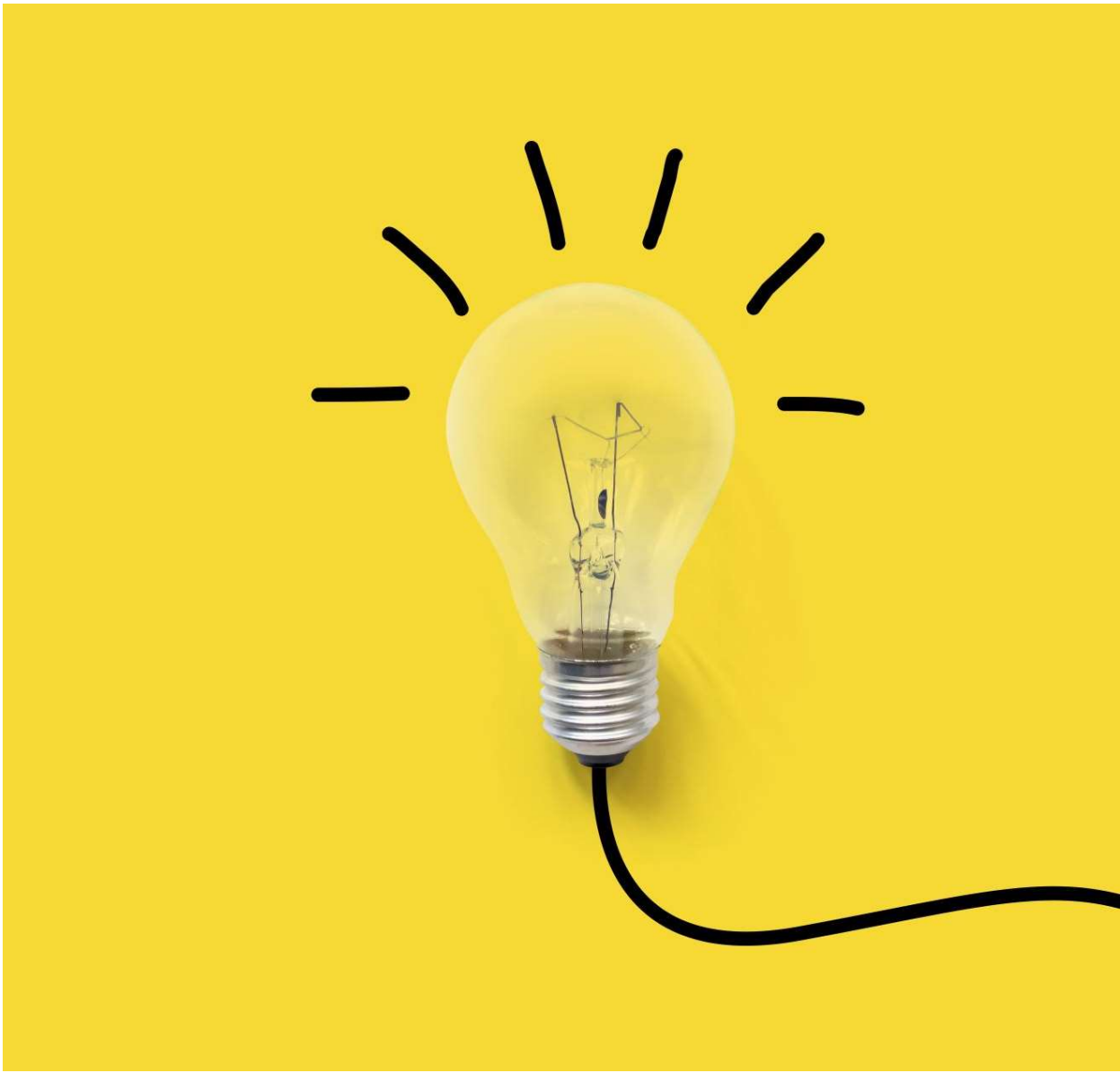
Change of question…

- Is it possible that these differences can be captured in the scRNAseq?
- If so, is it possible to transfer the knowledge from sc-RNAseq to bulk samples?
    - Predict tumor resistance on bulk based on scRNA-seq profiles



New plan

- Use an XGBoost classifier trained on sc-RNAseq to predict resistance.
- Use local interpretability of Shapley Additive explanations (SHAP) to explain what genes contribute most to predict resistance



The feature table design

To reduce high-dimensionality

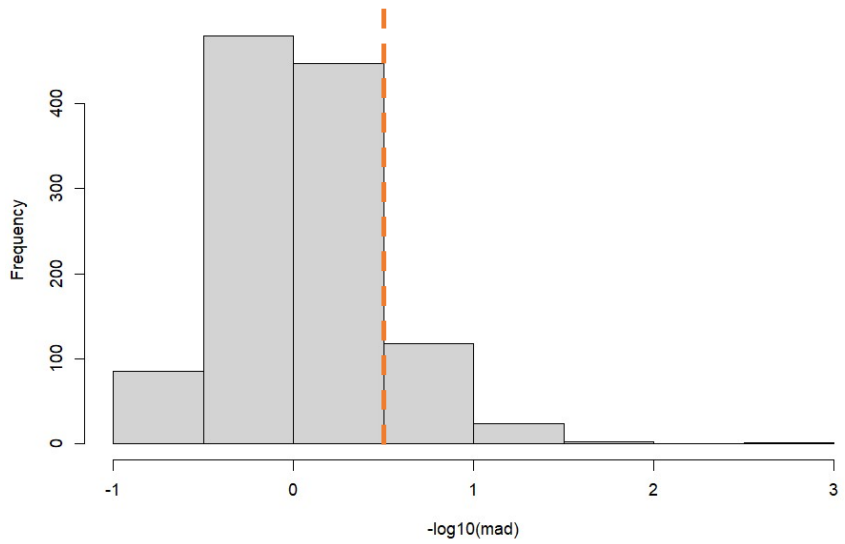
- Conserve top 10% most variable genes in single cell (median absolute deviation)

Genes shared with validation dataset.

- Final subset of 558 genes

Expression standardized within cells

- Capture relative variability within cell
- Reduce impact of cell-cell variability



Supervised XGBoost model results

- Classification of cells into resistant or not (16% of cells positive)
- Train split: 70% of cells
- 10-fold CV for hyper-parameter tunning in train split
  - Optimized for average precision
  - Random search 500 iterations

- 10-fold evaluation in test split:
  - F1 Score: 0.945 (±0.07)
  - AUROC: 0.98 (±0.05)
  - AUPR: 0.97 (±0.07)

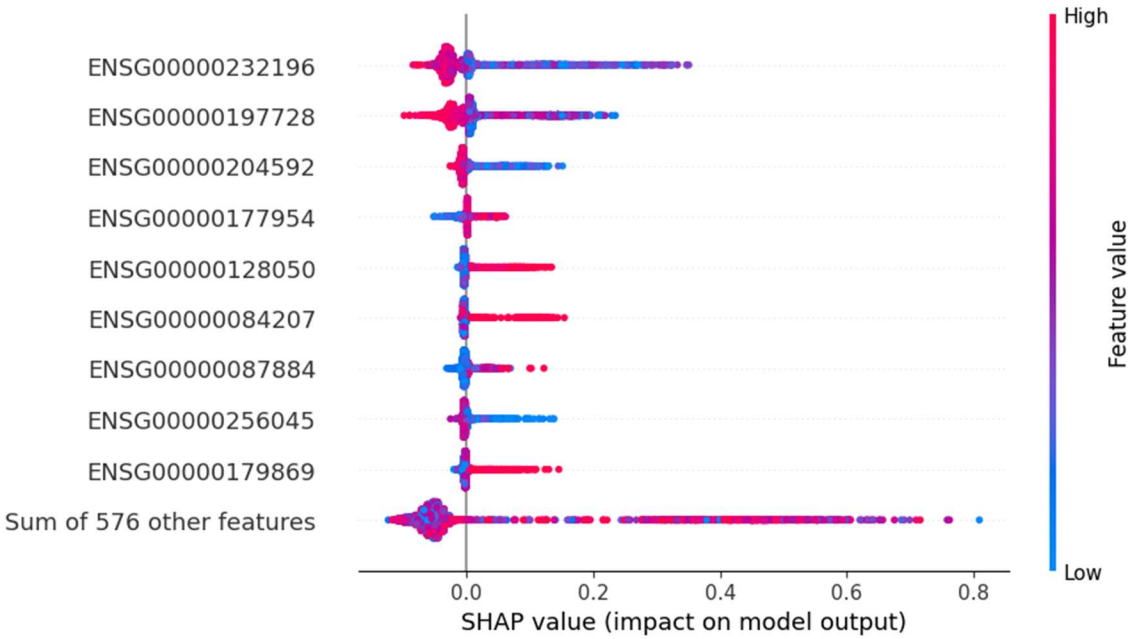
Results from the validation dataset

- Two donors out of five were accurately classified.
- Influenced by the imbalanced training dataset
- What genes primarily influenced the model prediction for these samples?

Donor	Time	Prob(No)	Pro(Yes)	Resistant
1	Pre	0.899	0.101	No
1	Post	0.947	0.053	Yes
2	Pre	0.91	0.09	No
2	Post	0.919	0.081	Yes
3	Pre	0.747	0.253	No
3	Post	0.403	0.597	Yes
4	Pre	0.952	0.048	No
4	Post	0.945	0.055	No
5	Pre	0.825	0.175	No
5	Post	0.969	0.031	Yes

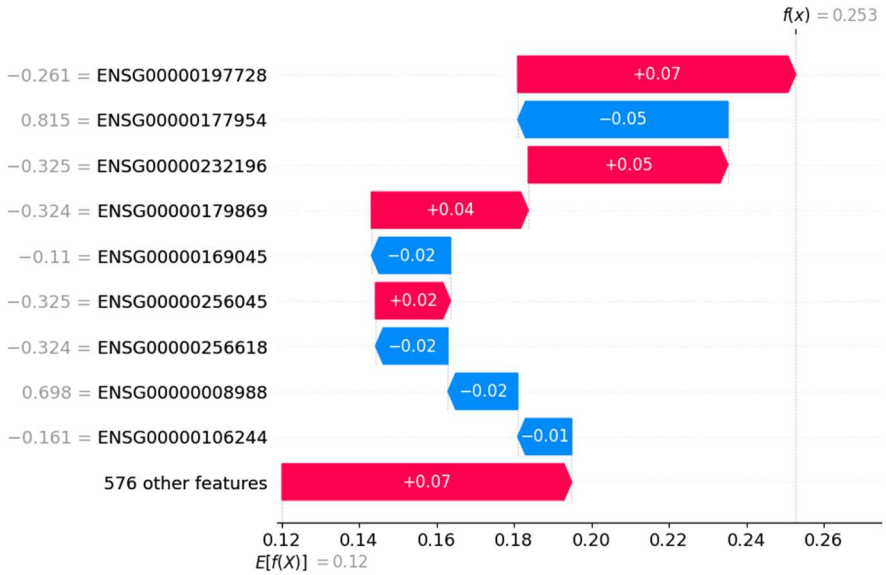
Overview of SHAP results – Full model

- **MTRNR2L4**
  - Antiapoptopic
  - Predicted to be extracellular
- **RPS26**
  - Ribosomal protein
- **HLA-E**
  - Involved in immune self-nonself discrimination
- **RPS27**
  - Ribosomal protein

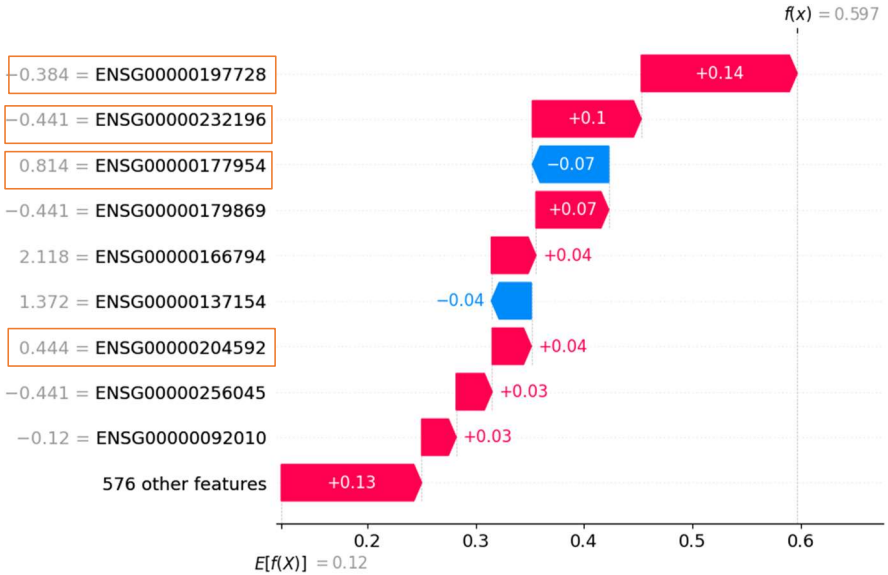


Donor which transcriptomic profile changed when developed a resistance

Before treatment



After treatment



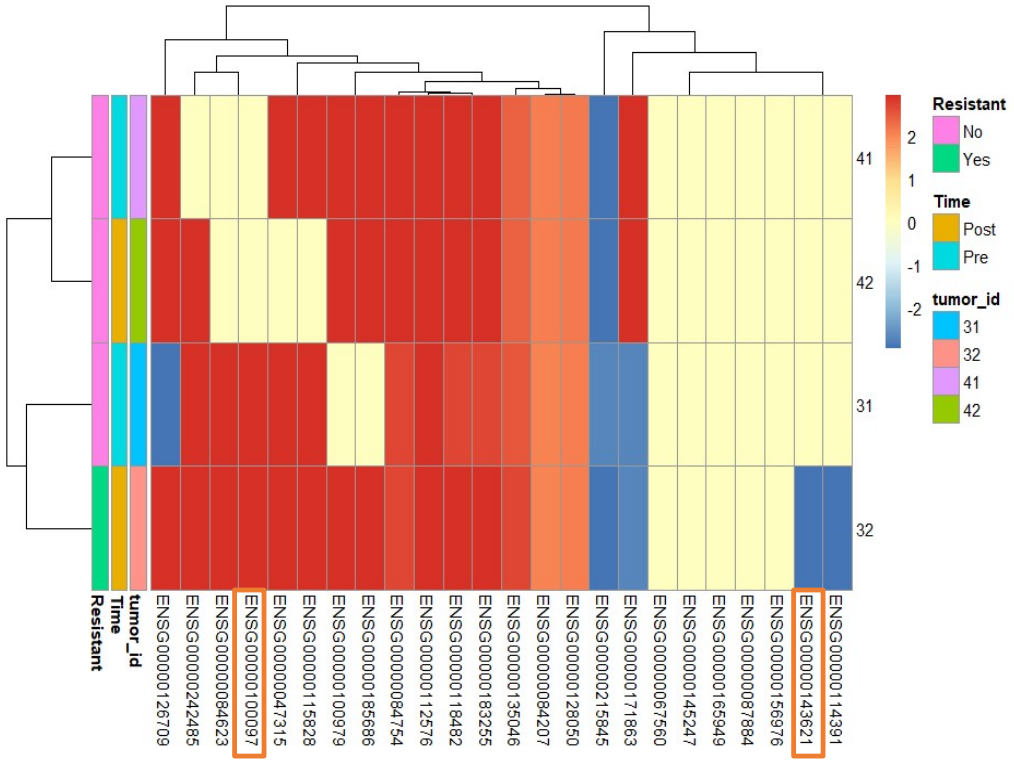
Comparison of model relevance for samples accurately predicted

ENSG00000143621 (ILF2)

Transcription factor required for T-cell expression of the interleukin 2 gene

ENSG00000100097 (Galectin)

Extracellular matrix, regulating cell proliferation.



Conclusions

- Identified several genes which were relevant for the prediction of resistance in sc-RNAseq
- Translation of knowledge from scRNAseq to bulk is probably influenced by the imbalanced dataset.
- Recovered many genes in bulk that drove predictions in scRNA-seq
- Identified new genes with a differential contribution to prediction of resistance in bulk-RNAseq

References

1. D. Hanahan, R. A. Weinberg, Cell 144, 646–674 (2011).
2. C. E. Meacham, S. J. Morrison, Nature 501, 328–337 (2013).
3. A. Snyder et al., N. Engl. J. Med. 371, 2189–2199 (2014).
4. N. Wagle et al., J. Clin. Oncol. 29, 3085–3096 (2011).
5. E. M. Van Allen et al., Cancer Discov. 4,94–109 (2014).