DECIPHERING CANCER HETEROGENEITY WITH MACHINE LEARNING

SIGNATURE FITTING ANALYSIS ON SINGLE CELLS IN RELATION TO PSEUDO-BULK DATA

AUTHOR

Rotar Mircea-Raul

SUPERVISORS

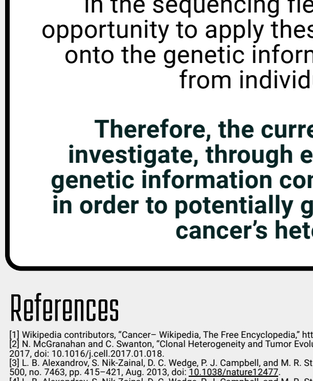
Joana Gonçalves

Sara Costa

Ivan Stresec

INTRODUCTION

"Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body" [1]



The heterogenous character of cancer is one of the main reasons to why it is so difficult to treat [2]. Understanding mutagens such as tobacco or UV-light, that give rise to this heterogeneity, has been essential in the pursuit of finding more specialized treatment solutions[3]. Mathematical models and frameworks have been able to uncover specific patterns of mutations left behind by these processes, modifications in the DNA material referred as mutational signatures [4].

Bypassing previous limitations, recent advances in the sequencing field have provided the opportunity to apply these mathematical models onto the genetic information coming directly from individual cells [5].

Therefore, the current project aims to investigate, through existing methods, the genetic information coming from single-cells in order to potentially gain more insights into cancer's heterogeneity.

RESEARCH QUESTION

What is the effect of performing mutational signature fitting for single-cell by relating it to pseudo-bulk?

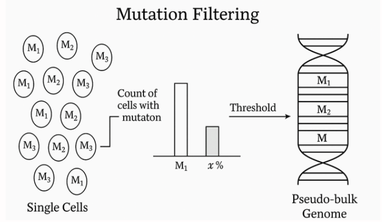
- Do single-cells achieve a better accuracy in the reconstruction of the mutational profile in comparison to pseudo-bulk across several metrics?
- If single cells can find more active mutational mutational signatures in comparison to pseudo-bulk, can we cluster these single-cells in a way that would explain the way in which the pseudo-bulk data was fitted?
- Do pseudo-bulk samples generated from subpopulations of cells differ from the one that was generated from the entire population?

METHODOLOGY

Data: 688 points sampled from a breast cancer tumor

SigProfilerAssignment library

Generation of pseudo-bulk data from single cell



Multiple percentages were chosen as the threshold to simulate the noise in the variant calling pipeline process - 0, 3, 5, 8, 10 and 15 %.

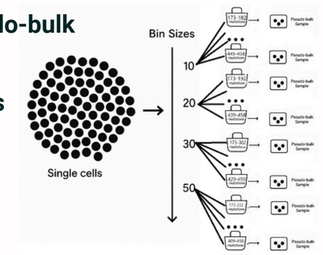
Q1. Accuracy of reconstruction of mutational profile between single-cells and pseudo-bulk

- Cosine similarity
- Pearson's correlation coefficient

Q2. Clustering single-cells

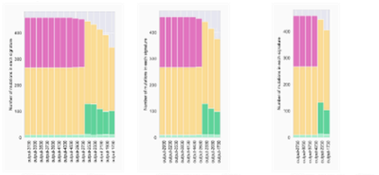
- K-medoids clustering with cosine similarity
- Elbow method for selecting number of clusters
- UMAP for visualization

Q3. Generation of pseudo-bulk samples based on subpopulation of cells



RESULTS

- There are 3 clusters in which we can group our cells based on their associated reconstructed mutational profiles.
- The cluster with the highest relative cell count has the majority of cells with the same set of active signatures as the pseudo-bulk.



(a) Bucket size of 20 (b) Bucket size of 30 (c) Bucket size of 50

Figure 5: VCF Ranges for Buckets

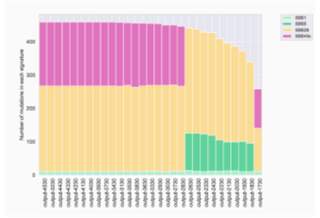
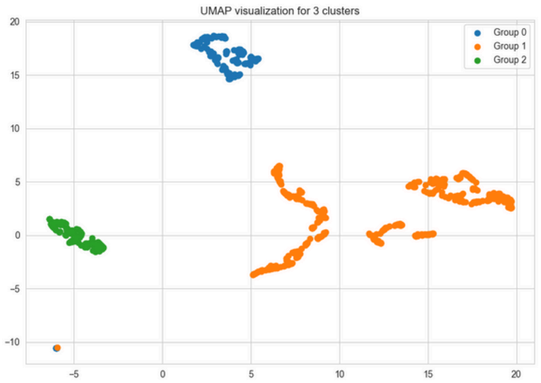


Figure 6: Bucket size of 10

- There are 75 cells which achieve a better accuracy of the reconstruction of mutational profile than the pseudo-bulk, across all metrics.
- The points in the graphs represent the values for the single-cells, while the red lines represent the values of the pseudo-bulk, for the corresponding metrics.
- The blue cells are performing worse, red are ones performing better on the current metric, while green are cells performing better across both metrics, in relation to pseudo-bulk



- Subpopulations of cells, with a relatively smaller number of mutations, generate pseudo-bulk samples that generally differ in the set of active mutational signatures identified in relation to the one generated from the entire population.
- By contrast, subpopulations of cells with a higher count of mutations generate pseudo-bulk samples having the same set of active signatures as the entire population one.
- There seems to be a divergence between the signatures fitted within subpopulations of cells with a relatively smaller mutation count and the signatures identified in the corresponding pseudo-bulk samples.

LIMITATIONS AND CONCLUSIONS

- Small number of mutations describing the single-cells we worked with
- No ground-truth bulk data describing the genome from which the single-cells were sampled from
- Future research should focus on performing single cell analysis as this has the potential to enhance our knowledge in the field of oncology

References

[1] Wikipedia contributors, "Cancer– Wikipedia, The Free Encyclopedia," https://en. wikipedia.org/wiki/Cancer, accessed: 2025-06-20.
[2] N. McGranahan and C. Swanton, "Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future," *Cell*, vol. 168, no. 4, pp. 613–628, Feb. 2017, doi: 10.1016/j.cell.2017.01.018.
[3] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013, doi: 10.1038/nature12477.
[4] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering signatures of mutational processes operative in human cancer," *Cell Reports*, vol. 3, no. 1, pp. 246–259, 2013, doi: 10.1016/j.celrep.2012.12.008.
[5] C. Ziegenhain *et al.*, "Comparative Analysis of Single-Cell RNA Sequencing Methods," *Mol. Cell*, vol. 65, no. 4, pp. 631-643.e4, Feb. 2017, doi: 10.1016/j.molcel.2017.01.023.