Comparing De Novo Single-Cell Mutational Signatures to COSMIC References

Introduction

- Cancer is a result of somatic mutations in a cell's DNA that cause rapid reproduction of that cell [1].
- Mutations can be categorized based on single-base substitutions (SBS).
- Collecting all mutation types of a sample results in a mutational profile vector.
- Mutational processes leave characteristic patterns in somatic mutations called mutational signatures.
- Mutational signatures can be estimated by factorizing many mutational profiles into signatures and mutations caused using nonnegative matrix factorization [2].
- Extracting new signatures from scratch is called de novo extraction.
- The Catalogue of Somatic Mutations in Cancer (COSMIC) contains a collection of known mutational signatures.
- Extraction of known signatures is done based on bulk data, where all the mutations in the tumor are aggregated [4].
- Single-cell data captures mutational heterogeneity between cells.

Research question

- How can we compare single-cell de novo mutational signatures to mutational signatures fitted from the COSMIC library?

Sub questions

- What are the differences between single-cell de novo data and bulk tumor data?
- Can the single-cell data be explained by existing signatures found in the COSMIC library?

Methods

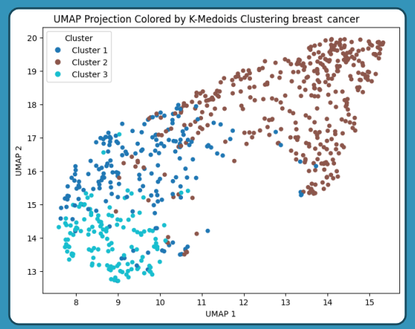
- Generate mutational profiles from single-cell data.
- Cluster, subsample, and visualize mutational profiles with k-medoids and UMAP [3].
- Fit COSMIC signatures to samples.
- Extract de novo signatures using NMF
- Compare de novo to COSMIC signatures:
  - Directly compare cosine similarity
  - Decompose de novo signatures into COSMIC signatures

Results data source

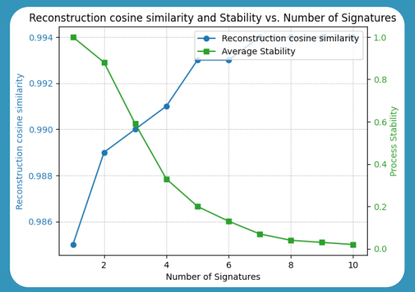
- Breast cancer dataset with 688 samples

Results

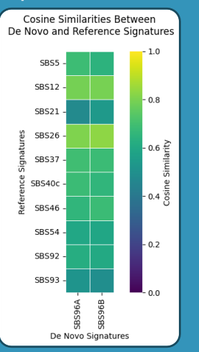
- Mutational profiles show low inter-cluster variability (cosine similarity between medoids > 0.97).
- Suggests a few distinct mutational processes active within tumors.

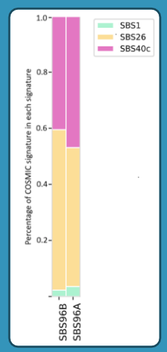


- Two de novo signatures extracted per dataset (based on stability > 0.8).
- De novo signatures are highly similar within each dataset (cosine > 0.96).
- Overfitting possible: even 1 signature gives cosine > 0.98 reconstruction similarity.



- No direct 1:1 match to known COSMIC signatures (all cosine < 0.85).
- Composite decomposition reveals matches to COSMIC SBS1, SBS26, SBS40c.
- Decomposition cosine similarity: all > 0.85; therefore, signatures are likely composite.





Conclusion

- Single-cell datasets have low diversity in terms of mutational processes.
- Low mutation counts result in possible large stochastic variations in mutational profiles of various cells.
- Multiple stable signatures can be found, possibly a result of overfitting.
- Future studies with more datasets are necessary

TU Delft

Fedde de Haas
Supervisors: Sara Costa & Ivan Stresec
Responsible Professor: Joana de Pinho Gonçalves

Reference:
[1] Stratton, Michael R, Peter J Campbell, and P Andrew Futreal (2009). "The cancer genome". In: Nature 458.7239, pp. 719–724.
[2] Alexandrov, Ludmil B, Serena Nik-Zainal, et al. (2013). "Deciphering signatures of mutational processes operative in human cancer". In: Cell reports 3.1, pp. 246–259
[3] Becht, Etienne et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP". In: Nature biotechnology 37.1, pp. 38–44.
[4] Alexandrov, Ludmil B, Jaegil Kim, et al. (2020). "The repertoire of mutational signatures in human cancer". In: Nature 578.7793, pp. 94–101.