EXAMENSARBETE INOM MEDICINSK TEKNIK,
AVANCERAD NIVÅ, 30 HP

STOCKHOLM, SVERIGE 2020

Computational evalutation of pathway activity on the progression of a set of cancers

PATRIK BJÄRTEROT

KTH
SKOLAN FÖR KEMI, BIOTEKNOLOGI OCH HÄLSA

Pathway analysis of subtypes in breast cancer

PATRIK BJÄRTEROT

Master in Molecular Techniques in Life science
Date: May 3, 2020
Supervisor: Lukas Käll, Gustavo Jeuken
Examiner:
KI, KTH, SU

Abstract

This is the abstract.

Contents

1 Introduction **1**
- 1.1 Research Question 1
- 1.2 Breast Cancer 1
- 1.3 Cancer Subtyping 2
- 1.4 Cancer Treatments 2
- 1.5 The bigger picture of cancer (this should be renamed) 2
- 1.6 2

2 Methods **3**
- 2.1 PCA 3
- 2.2 PCA based pathway analysis 3
- 2.3 Failures of null 4
- 2.4 Sunburst plotting 4
- 2.5 GSEA based pathway analysis 4

3 Results **6**
- 3.1 PCA based pathway analysis 6
- 3.2 Sunburst Plotting 6
- 3.3 GSEA Validation 6

4 Discussion **8**
- 4.1 PCA 8
- 4.2 PCA based pathway analysis 8
- 4.3 Sunburst Plotting 8

5 Conclusions **9**

6 Future Work **10**

7 Ethical reflections **11**

8 Acknowledgements **12**

Chapter 1

Introduction

1.1 Research Question

This project was aimed to evaluate a method for pathway activity and to test it in the context of breast cancer. As gene expression by itself is often biased and noisy, a hypothesis was made that simply reducing the noise and looking at the points of highest variance would lower bias and give higher resolution results in the context of pathway analysis

1.2 Breast Cancer

Breast cancer is the most common cancer among women (WHO), accounting for nearly 15% of all cancer deaths among women. While the mortality rate has slowly decreased in the last 30 years(https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/mortalityheading-Two), one will agree that major advances are needed in order to continue the trend. Furthermore, the field of oncology is not only about reducing mortality, it is also very important to ensure disease-free survival, as well as the overall physical and mental health of the patient and their family. Granted, most of these aspects can be mitigated by advances made in the disease treatment/prevention

Breast cancer has been studied for a long time, with the earliest paper recorded in PubMed from 1789 to almost 25 000 papers in 2019. The sad truth is though giant leaps in detection and treatments have been made since 1789, we still cannot call breast cancer a cured disease.

1.3 Cancer Subtyping

Currently, the primary source of cancer phasing regards subtypes. Receptor subtyping has proven very helpful in breast cancer and has been a major player in improving the overall survival of the disease, as more targeted treatments such as tamoxifen can be used for patients that have tumors with estrogen receptors, and need not be used on receptor negative patients.

A very common subtyping method used in breast cancer is PAM50, a tumor profiling test that determines whether ER+/HER2- cancers are likely to metastasize. It is known that the PAM50 'Luminal A' subtype in breast cancer has a higher hazard than do other subtypes [1].

Subtyping provides a great basis personalized treatments, however, could it also be used as a basis for targeted new treatments? Since cancer is a very heterogenic disease, subtyping allows us to phase biologically similar patients together to decide for treatment, prognosis and overall care.

1.4 Cancer Treatments

One area where subtyping has proven very important is in cancer treatments. Treatments have been shown to affect different subgroups differently and have thus

Maybe one of the reasons for the sluggish development of treatments for different cancers is that they are tested on very dissimilar patients. Taking Tamoxifen as an example, if it were tested on estrogen receptor positive as well as negative patients, the effect would be diminished if the subtypes were not identified.

1.5 The bigger picture of cancer (this should be renamed)

1.6

Chapter 2

Methods

Since the project did not require any wet lab, the materials are all computer based. The project is mostly based on the python programming language, with common packages numpy, pandas and scikit-learn being the most used. As for the methods, the project centered around PCA analysis and information mining from said analysis.

2.1 PCA

To give some background to PCA, it is a dimensionality reduction method which simplifies high-dimensional data while keeping as much variability as possible. High-dimensional data, while common, can lead to misleading information as some dimensions can lead to noisy datasets that might give a type 2 error. However, too few dimensions can lead to loss of information, the extent of which is evaluated in this project

PCA analysis performs orthogonal transformation to create a linear function where the first component shows the largest possible variance between the samples. Each following component then shows the highest possible variance provided it is orthogonal to the previous component. Principal components are displayed as eigenvectors, which are explained by their eigenvalues

2.2 PCA based pathway analysis

Following the PCA transformation, pathways were defined using the reactome database and the available genes. Subsequently, comparisons between patient groups were made using Mann-whitney U tests for each pathway in a high-

throughput manner. Beginning with the Intclusts defined in the metabric paper, each cluster was compared to all other clusters, in order to determine what separates them on a pathway basis. This was also done using benign breast tissue samples as an additional cluster, comparing each intclust to the normal samples to identify differences not outlined by the other malignant samples.

Additionaly, this comparison was also performed comparing hormone receptor subtypes, as well as comparing triple negative cancers with non-triple negative patients.

Lastly, there was an interest and a hypothesis that mutation data would give interesting results, and this was analysed as well. However, since there were more than 170 different mutations, and sunburst plotting is not yet optimized for a great number of

2.3 Failures of null

One observed effect of the multiple hypothesis testing was that the p-values did not follow a null distribution when converted to Z-values. This suggests that there were confounders that needed to be accounted for. One way of doing this was explained by (((SOURCE))) wherein a truncated gaussian distribution would be fitted to the Z-values and deriving new p-values from that. This was done using the scipy modules norm and truncnorm.

2.4 Sunburst plotting

Due to the large number of pathways, simply looking at tables of pathways ranked by q-value was not quite sufficient to get a good idea of the relationship between groups. Therefore, a way of visually displaying the pathways in a clearer manner had to be used. The sunburst plots were an excellent way of doing so. The plots were colored by q-value and the size of each cell is proportional to the number of genes in the pathway.

2.5 GSEA based pathway analysis

As a way of benchmarking this new method of pathway analysis, GSEA was performed. Primarily, one downside of GSEA, and in particular the python version of gseapy, is that it was relatively difficult to get to work. Aside from the particularity of the input dataframe, which reminds one of trying to get a cat into a bath, it was impossible to run in the windows and linux operating

system. It was possible, however, to run on the OSX operating system. In hindsight, the difficulty of running it on linux and windows is most likely due to some errors regarding multiprocessing, and would not work even though one specified one process.

Chapter 3

Results

3.1 PCA based pathway analysis

Results from the pathway analysis can be seen in figure 3.1

3.2 Sunburst Plotting

3.3 GSEA Validation

Table 3.1: Results of the pathway analysis from looking at the sunburst plot

Cluster	Sunburst
1	Cell Cycle, Base Excision Repair, DNA Double Strand Break Repair, RNA Pol I transcription, Gene Silencing by RNA, Senescence, Rho GTPase Signaling, ESR Signaling, TCF complex
2	mTORC1, Iron Uptake and Transport, Insulin Receptor
3	Cell Cycle, DNA Replication, DNA Double Strand Break Repair, Mismatch Repair, DNA Damage Bypass, SUMOylation, Megakaryocyte Production
4ER+	Cell Cycle, DNA Double Strand Break Repair, Rho GTPase Signaling
4ER-	SLC Mediated Transport, Immune System, Neuronal System, Metabolism of Vitamins and Cofactors, DAG/IP3 Signaling, CaM Pathway
5	Cell Cycle, Metabolism of Amine Derived Hormones, Tryptophan Catabolism, Rho gtpases
6	IP Metabolism, Wnt
7	CaM Pathway, NMDA Receptors, Cilium Assembly, Metabolism of Vitamins/Cofactors, Metabolism of Nucleotides, Ion Channel Transport
8	Immune System, Transport of Small Molecules, Sphingolipid Metabolism, GPCR Signaling, DAG/IP3, Negative Regulation of Akt Network, Hemostasis
9	Cell Cycle, KSHRP/TTP/TRF1, Riboflavin Metabolism, Glutamine/Glutamate Metabolism
10	Cell Cycle, DNA Replication, Rho GTPases, Sphingolipid Metabolism, Nucleotide Metabolism

Table 3.1: Results of the pathway analysis from looking at the sunburst plot

Chapter 4

Discussion

4.1 PCA

4.2 PCA based pathway analysis

4.3 Sunburst Plotting

Sunburst plotting together with the q-value analysis described above became an excellent combination of analyses as the table provides very specific information of the pathways and the sunburst plots provide a broader picture of the pathway relationships of the tumor cells.

One aspect where the advantage of using sunburst plotting became undeniable was when analysing parent child pathway relationships. As can be seen in ((((((figure)))))) the activity of ((((parent)))) differs from the activity of the (((((child))))). This thus becomes somewhat of a barcoding analysis where a straight line of parent child pathways with low q-value is quite convincing evidence for a difference in pathway activity.

This relatively novel method of analysis has resulted in a few points of benchmarking and what the underlying biology might be.

- A parent pathway with relatively medium q-value and several child-pathways where at least one of them has very low q-value is a good indicator that the parent pathway q-value might be affected by the significant child pathway.
- A significant pathway with no children should be an interesting pathway to study as it has been broken down as much as possible and still poses a significant difference in activity compared to other clusters.

Chapter 5

Conclusions

Chapter 6

Future Work

Chapter 7

Ethical reflections

One concern of cancer research, and especially in this project, is the scarcity of longitudinal data. As tumors are instantly removed upon discovery, any genetic testing occurs only at one point in time. It would be scientifically beneficial to have several points of genetic testing to analyse the genetic changes in the tumor. Ethically, however, this is vital problem. Tumors must be removed as soon as they are discovered, especially if malignancy is likely. This removes the ability to study tumors longitudinally.

However, there is one way to somewhat mitigate this by looking at recurring cases, meaning cases where the tumor has been removed, yet new malignancies occur. This allows for two longitudinal datapoints to be studied.

är detta kanske lite utanför thesis eftersom jag inte använder mer än en datapunkt?

Chapter 8

Acknowledgements

References

[1] M. Pu et al. “Research-based PAM50 signature and long-term breast cancer survival”. In: *Breast Cancer Res Treat* 179.1 (2020), pp. 197–206. ISSN: 0167-6806 (Print) 0167-6806. DOI: 10.1007/s10549-019-05446-y.