Survival analysis of pathway activity as a prognostic determinant in breast cancer

Gustavo S. Jeuken
Science for Life Laboratory,
School of Engineering Sciences in Chemistry, Biotechnology and Health,
KTH – Royal Institute of Technology,
Box 1031, 171 21 Solna, Sweden

Nicholas P. Tobin,
Department of Oncology and Pathology,
Karolinska Institutet and University Hospital,
Stockholm, Sweden

Lukas Käll*
Science for Life Laboratory,
School of Engineering Sciences in Chemistry, Biotechnology and Health,
KTH – Royal Institute of Technology,
Box 1031, 171 21 Solna, Sweden
* *To whom correspondence should be addressed.*

May 18, 2021

Abstract

Motivation: Most activities in a cell are dependent on not individual, but rather sets of biomolecules. Such sets are often referred to as pathways. Pathways are useful to study as their activities are frequently more directly related to the behavior of the cell than their components. Traditional pathway analysis gives significance to differences in the pathway components' concentrations between sample groups. Here we instead propose a PCA-based method for estimating individual samples pathway activity that allows us to investigate more advanced statistical models. Specifically, we investigate the pathway activities' association with patients' survival time based on the transcription profiles of the METABRIC dataset.

Results: Our implementation shows that pathway activities are better prognostic markers for survival time in METABRIC than the individual transcripts. We also demonstrate that we can regress out the effect of individual pathways on other pathways, which allows us to estimate the other pathways' residual pathway activity on survival.

Availability: Python package is available from `https://github.com/statisticalbiotechnology/porch`, and the code generating the plots available at `https://github.com/statisticalbiotechnology/metabric-pathway-survival`. **Contact:** `lukask@kth.se`

1 Introduction

Modern high throughput techniques in molecular biology enable the measurement of thousands or even millions of analytes in any sample of biological origin. Such wealth of data makes it possible to describe samples very accurately, with a precision that opens the door to new understandings of the mechanisms governing biological and medical processes.

Although this abundance of data indicates a possibility for acquiring knowledge, the sheer number of dimensions in such measurements also presents us with many challenges: when analyzing high-dimensional vectors with statistical methods one easily faces the curse of dimensionality, i.e. that the sample space grows exponentially with each added dimension [1]. This becomes a problem when the number of samples is lower than the number of measured analytes. Further, in e.g. differential expression analysis, which tests the concentration differences of each measurement separately, we to some degree run into decreased sensitivity due to multiple hypothesis testing.

One of the more promising ways of alleviating these problems is through pathway analysis [17, 25]. Proteins frequently operate in an orchestrated manner and phenotypes are frequently the consequence of sets of proteins and not just of single-proteins. Metabolic pathways, or other aggregated biological knowledge that groups analytes, provide a model-driven way of combining molecular information and thus also the measurements from a high throughput experiment in a biologically meaningful way.

Cell proliferation, as an example, has been established as one of the hallmarks of cancer [13, 14]. It is a complex, systemic process that involves many different mechanisms, in different parts of the cell, as well as signaling pathways that regulate them. In breast cancers, it has been shown that one can separate Luminal A from Luminal B subtypes based on the level of proliferation [22]. Thus, when investigating how molecular profiles of tumors affect patients, one may want to look at proliferation as a whole, instead of focusing on individual analytes.

Traditional pathway analysis first determines the quantitative differences in gene expression between patient groups, and subsequently either tests the significantly differential genes annotation for enrichment in the tested pathway [26, 6, 12], or tests if the genes belonging to a pathway have more extreme differences than other genes, using so-called gene-set enrichment analysis [25]. Both types of analysis leave the user with a significance value for the analytes in a pathway annotation being differentially abundant under the different conditions. Yet this feels unsatisfying, as embedding the pathway analysis in the statistical test significantly limits the types of statistical tests that can be applied in the analysis.

An alternative is offered by single sample pathway analysis, which promises activity scores for each sample and pathway. The method ASSESS fits two mixture models to data to quantify pathways [8], however, the supervised learning step of the procedure makes further statistical testing challenging. The Pathifier method fits principal curves and outputs binary pathway states (activated or not) [7]. The method ssGSEA, produces enrichment scores, i.e. internal comparisons of the ranking of the genes in a pathway within each sample [15].

Another critique is that most pathway analysis methods derive statistical significance by comparing the behavior of genes belonging to a pathway with those that do not. Goeman & Bühlmann [10] argue in favor of self-contained tests, in which the significance of a pathway relates only to the expression of said pathway's genes, resulting in a more restrictive null hypothesis that leads to a higher statistical power.

Here, we propose a method for pathway summarization technique that is done distinctively from the statistical analysis. We investigate its ability to perform pathway-level survival analysis on the METABRIC [4] breast cancer set.

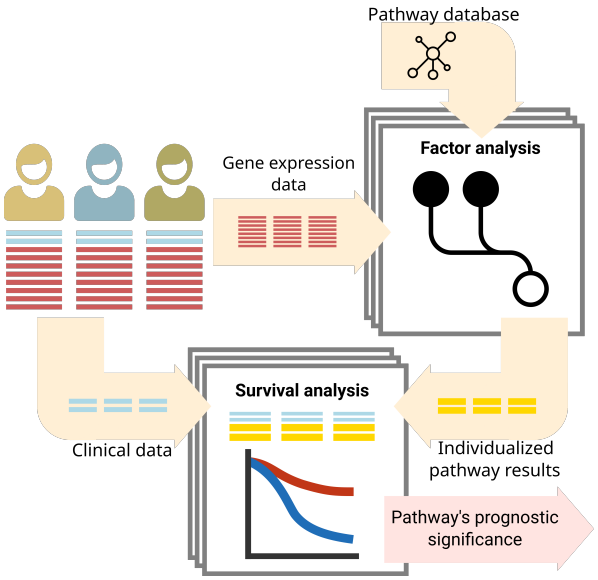

Figure 1: **A method to analyze the coupling of pathway activity on patient survival**. We used factor analysis to give, for each sample and pathway, an individualized pathway readout. We then combine those readouts with the survival information to perform survival analysis on a pathway level.

2 Approach

2.1 A pathway-level survival analysis

We implemented a method for evaluating a pathway activity's influence on survival. The method is generic for any censored data, however, here we demonstrate its efficiency on the METABRIC breast cancer dataset. The transcription profiles of the 1980 breast cancers in the METABRIC data were downloaded and grouped according to the Reactome database into pathway groups. After the operation, each pathway group contained the transcripts corresponding to proteins in each pathway, and as there are overlaps between pathways in Reactome, this means that each transcript could appear in multiple pathways. The expression matrix of each pathway was factorized into eigensamples (left-singular vectors) and eigengenes (right-singular vectors) using the first eigenvector of a Singular Value Decomposition (SVD) following Wall *et al.* [28]. The eigensample represents the linear combination of genes and the eigengene the linear combination of samples that best explains the variance of the expression matrix of the pathway. This slightly backward naming convention stems from the idea that an eigengene models the variation of samples of a typical gene, while an eigensample models the variation of the genes in a typical sample.

We then project each sample into this eigengene and use this result as a representation of the sample's pathway activity. These pathway activities together with the survival information for each patient were fed into a Cox model for proportional hazards, which regresses these values against the survival information (survival time and disease-specific death) to obtain both the regression coefficient and the statistical significance for the effect of each pathway activity on patient survival. The significance, first obtained as p values were multiple testing corrected into q values [24]. Figure 1 gives an overview of the procedure.

Before we investigate the output of our model, we would like to motivate the choice of SVD as a means to capture pathway activity. As an example, let's consider the pathway 'Metabolism of folate and pterines" consisting of 26 proteins. When investigating the covariation matrix of the transcripts and the survival time for 907 samples from patients that were deceased in the METABRIC cohort,

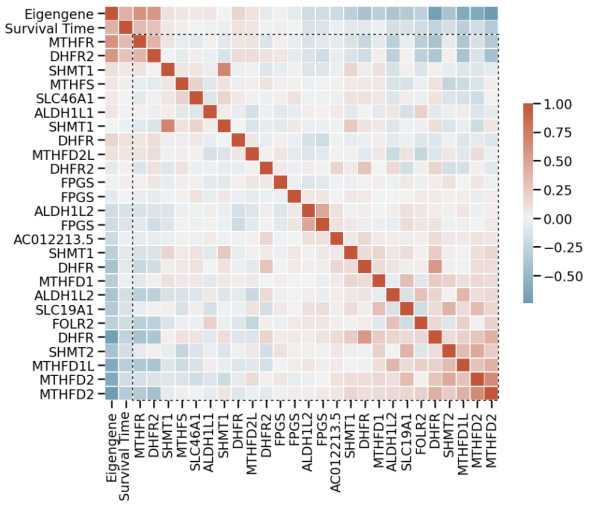

Figure 2: **Heatmap of the coefficients of the Pearson correlation matrix between the variables survival time, the eigengene of the gene expression values, and the gene expression values within the pathway 'Metabolism of folate and pterines" for samples from diseased patients**. We sorted the genes in the expression matrix by their contribution to the eigensample and encompassed them with a dotted line. The patients' survival time both correlates and anti-correlates with the different genes. However, the eigengene captures the covariational trend within the expression data and correlates well with the patients' survival time.

we found both examples of transcript levels that correlate positively and negatively with the survival time (Figure 2). By having the probe reads ranked by their contribution to the eigensample, we see that the eigensample captures information from genes that have both a positive and negative influence on survival. We also see that probes that have a negative contribution to the eigensample also have a negative influence on survival, while the opposite is also true. It is the linear combination of these expressions, obtained in an unsupervised manner, that is used for later survival analysis.

3 Methods

3.1 METABRIC Transcription profiles, and their annotation

Normalized gene expression and clinical annotation in the METABRIC dataset were downloaded from the European Genome-phenome Archive. This data consists of microarray reads from 1992 breast cancer specimens, primarily fresh frozen, together with the clinical annotation, including survival information, of the respective patients. There are 12 duplicate samples from this dataset that were removed from our analysis.

Pathway annotations were retrieved from the Reactome database [16], version 76, with annotations as Ensembl gene IDs. The pathways' gene ID annotations were converted to Illumina probe IDs (HT_12_v4), through BioMart [23], by assuming that any transcript associated with each Reactome protein's underlying gene was associated with the pathway.

The clinical endpoint for this study was breast cancer-specific survival (BCSS) defined as patients who have not died from breast cancer in the study period from the date of surgery to the end of follow-up.

3.2 Estimation of Pathway Activities

Let $\mathcal{G} = \{a_1, a_2, ..., a_g\}$ be the set of all measurements we want to study, the Illumina probes in our case, and let $\mathcal{P} = \{b_1, b_2, ..., b_p\}$, $\mathcal{P} \subset \mathcal{G}$ be the subset of measurements that are associated with our pathway of interest. If we define $\mathbf{A}_{g\times m}$ as the matrix of log-transformed and standardized measurements of our samples (where m is the number of samples), then we can define $\mathbf{B}_{p\times m}$ as a matrix constructed by using only the rows in $\mathbf{X}$ that are present in $\mathcal{P}$. We then decompose $\mathbf{B}$ using a Truncated Singular Value Decomposition (SVD):

$$\mathbf{B}_{p\times m} \approx \mathbf{U}_{p\times 1}\mathbf{S}_{1\times 1}\mathbf{V}^T_{m\times 1}$$

Here we name the pathway's left singular vector $\mathbf{U}$ as an eigensample, and the right singular vector $\mathbf{V}$ as an eigengene, following the nomenclature in Wall *et al.* [28]. Note that the eigengene contains one vector element per sample, and we will use these vector elements as a measure of pathway activity. Practically we here make use of the *scikit-learn* python package [21].

3.3 Proportional Hazards Model

Cox's proportional hazards model [3, 18] relates the survival function $S_i(t)$ of patient i to the value of any of its covariates X_i through the following:

$$\lambda(t|X_i) = \lambda_0(t)\exp(X_i \cdot \beta)$$

Where $\lambda(t)$ is the hazard function and is defined as the rate of mortality at time t, for patients that have survived up to that time, and $\lambda_0(t)$ is the baseline hazard defined as $\lambda(t|0)$. The hazard function is then associated with the survival function by $\lambda(t) = -S'(t)/S(t)$

Here we used the pathway activities derived previously as independent variables X for a Cox regression, doing so one pathway at a time, to study the connection between each of the samples' pathway activities and the survival of patients. This regression gave us both a coefficient β showing the magnitude of the effect of a pathway, and a p value representing the statistical significance of its coefficient. The p values were subsequently corrected for multiple testing into q values [24]. In here we make use of the *lifelines* python package [5].

3.4 Concordance Index

The concordance index (or C-index) is a generalization of the area under the curve (AUC) classifier performance that can take into account censored data. It represents the model's accuracy in ranking the survival times[27]. It can be calculated as,

$$c = \frac{\sum_{i\neq j} \mathbb{1}_{t_i<t_j}\mathbb{1}_{\eta_i>\eta_j} d_j}{\sum_{i\neq j} \mathbb{1}_{t_i<t_j} d_j}$$

Here we have the observed survival time, t_i, and survival status, d_i, for every patient, i, and we associate that information to $\eta_i = X_i \cdot \beta$, calculated with 5-fold cross-validation, to prevent over-fitting β to X_i. Just as for AUC, a concordance index of 0.5 corresponds to a null prediction, and 1 to a perfect prediction.

3.5 Availability of code

A python package for running the analysis are downloadable under an Apache 2.0 license at `https://github.com/statisticalbiotechnology/porch`, and the code generating the plots for this paper is available at `https://github.com/statisticalbiotechnology/metabric-pathway-survival`.

4 Results

This section describes the results of our method when executed on the METABRIC dataset. We will subsequently investigate the robustness of the derived regression models, by permutation tests. We will then expand on the relationship between cell proliferation and those results and how our methods are specifically suited to deal with strong signals such as this one. Finally, we present a tool to visualize pathway results intuitively.

4.1 Analysis of breast cancer data set

The transcription profiles of the 1980 breast cancers in the METABRIC data were downloaded, and they were analyzed with our method. Supplementary Table S1 list the pathways and their prognostic significance, and the results are also available as an interactive plot (`https://statisticalbiotechnology.github.io/metabric-pathway-survival/results.html`). We found 1030 pathways (out of the 2214 pathways in Reactome) that were associated with patient survival with q¡0.05. We note that most pathways involved with the Cell Cycle are indeed associated with survival, as one would expect, as well as pathways relating to DNA replication and DNA repair.

Cell proliferation is a driver of cancer [13] that affects most other aspects of cancer cell activity. We regressed out the influence of proliferation on other pathway activities, by adding an independent variable for cell proliferation in our Cox's regression model for each pathway. This operation intends to highlight associations that otherwise are drowned by the more trivial background of cell proliferation. Traditionally proliferation is quantified by the transcription of marker genes, like MKI67 [20]. However, here we instead used the pathway activity of the "Diseases of the mitotic cell cycle", calculated by our model, as a proxy for abnormal cell proliferation.

After regressing out our measure of proliferation the number of significant (q ¡ 0.05) pathways, as expected, is reduced to 264 (Supplementary Table S2, (`https://statisticalbiotechnology.github.io/metabric-pathway-survival/results_proliferation.html`). We see that the operation removes the significance of most pathways relating to the cell cycle and DNA replication. Also, several other pathways significance diminishes, e.g. the "Metabolism of folate and pterines" pathway became insignificant, indicating that its prognostic power was mostly driven by cell proliferation (when removing that confounding effect of cell proliferation, its effect on the survival of the patients is removed, from $q = 9 \cdot 10^{-13}$ to $q = 0.17$). However, pathways relating to cell pH regulation due to respiratory oxidation [11] (e.g. "Bicarbonate transporters", from $q = 2 \cdot 10^{-10}$ to $q = 7 \cdot 10^{-5}$) and HER2-signaling [19] (e.g. "GRB7 events in ERBB2 signaling", from $q = 3 \cdot 10^{-8}$ to $q = 7 \cdot 10^{-5}$) remain highly significant after the operation. This demonstrates an advantage of calculating individual measurements of pathway activity for each sample, as we can subtract effects of known confounders.

4.2 Pathway's eigensamples appear stable and their estimates appear well calibrated

For the results to be meaningful, the eigensample should be stable, i.e. if the decomposition is repeated for similar samples, a similar eigensample should result. This is especially important for datasets that are smaller than the ones used here. To simulate this, we randomly subsampled 20% (396 samples) of the tumors 100 times, each time performing the same decomposition, and made pairwise comparisons of the resulting eigensamples using the cosine distance (Supplementary Figure S1). We found that most eigensamples were indeed stable, and most importantly, the most significant pathways in respect to survival prognosis were all stable.

We also performed a permutation test to check the calibration of the p values coming from the Cox regression model. We randomly selected 100 pathways by their regression p value, by first sampling a uniformly in the range $[-5, 1]$ and selecting the pathway with the p value closest to 10^{-a}, then for each pathway the association between the gene expression values and the survival information was permuted $100/n$ times, where n is the regression's p value. The fraction of permutations with a more

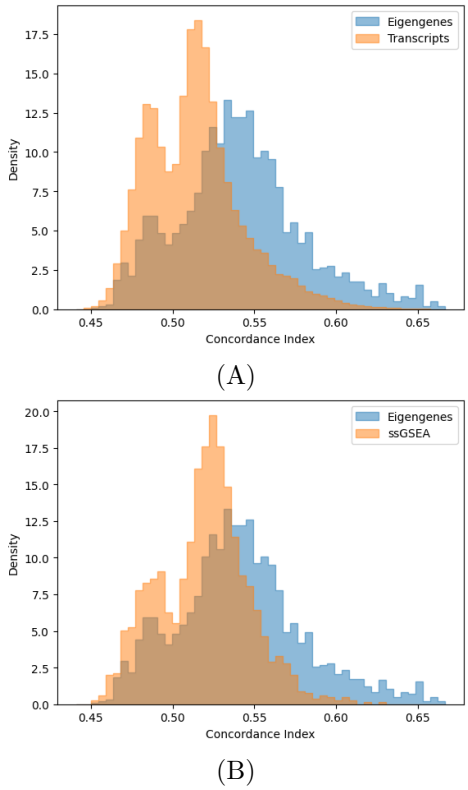

Figure 3: **Comparison of how well eigengenes predicts survival.** A histogram of the concordance indices calculated with a 5-fold cross-validation of Cox regressions based on eigengenes compared to (A) individual transcripts and (B) ssGSEA enrichment scores.

extreme outcome was then compared to the original Cox model's p value. The results (Supplementary Figure S2) indicate that our model is indeed well-calibrated for the null hypothesis that there is no association between gene expression in the form of the pathway activation and patient survival.

Finally, to test whether the signal we see comes from a meaningful set of transcripts, as opposed to any set of transcripts, we performed a gene set permutation analysis (Supplementary Note S1), which confirmed the relevance of the manually curated sets.

4.3 Pathways are more predictive of survival than individual transcripts

We also tested how well eigengenes can be used to predict the survival of each patient. We do this by, for each pathway, performing 5-fold cross-validation using the Cox model to predict survival on the holdout data, measuring the success of the prediction using the Concordance Index. Figure 3A show the distribution of the Concordance Index after performing the cross-validation on all pathways, and contrasts it to the distribution of the concordance index obtained by doing the same process with individual transcripts. We see that there is a general gain in predictive power when combining the information contained in individual transcripts using our method.

4.4 Eigengenes are more predictive of survival than enrichment scores

We compared our method to the alternative method of using the enrichment scores of single-sample geneset enrichment analysis (ssGSEA)[2] as an independent variable in a Cox-regression. To obtain the enrichment scores, we used ssGSEA as implemented in `https://github.com/broadinstitute/`

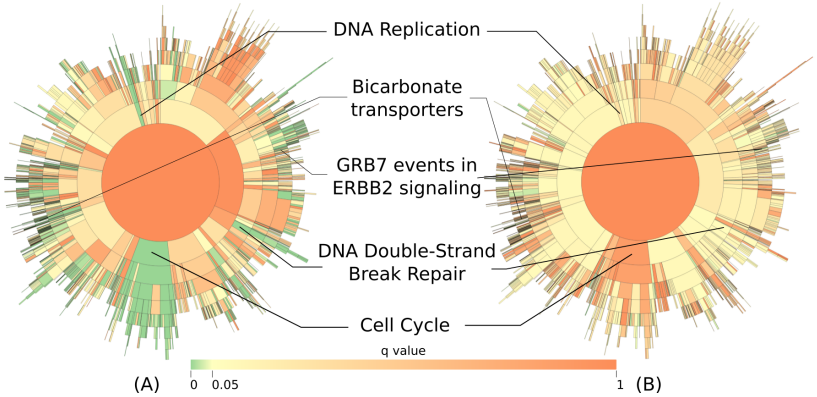

Figure 4: **Sunburst diagrams offer an exploratory visualization of the influence of the pathways in the Reactome hierarchy on the survival of patients.** We use an interactive sunburst chart to visualize and explore the result. This allows us to navigate the pathway hierarchy and see the relationships in a more natural way than e.g. in a static table. The sunburst (A) without and (B) with the pathway "Diseases of mitotic cell cycle" regressed out. Interactive versions of the same figures are available from https://statisticalbiotechnology.github.io/metabric-pathway-survival.

`ssGSEA2.0` with its default parameters. We found our eigengenes more sensitive than ssGSA, with 1030 vs 352 reported pathways with q values below 5%. We also found that the concordance indices of our model outperformed the ones produced by the ssGSEA-fed model (Figure 3B).

4.5 An interactive visualization of pathway-level results

We know that biological pathways, as described by pathway databases, are not independent, i.e. genes can belong to multiple pathways at the same time. Moreover, the relationship between pathways also has a hierarchical nature, and in many pathway databases, large pathways are composed of smaller, connected pathways. This information is lost when we present a table with the results of our analysis, and for this reason, we developed an interactive visualization of the results based on a sunburst chart that highlights not only the numerical results but also the hierarchy and relationship between pathways. Figure 4 shows a static view of our interactive figure, the full plot is available at `https://statisticalbiotechnology.github.io/metabric-pathway-survival/`

5 Discussion and Conclusions

Here we have presented a method for deriving a quantitative value of pathway activity, and we demonstrated its usefulness when associating pathway activity to patient survival. We do so by singular value decomposition of the expression of genes inside a pathway, using the resulting eigensample to ascribe a pathway activity to each sample, and testing its association to the survival of patients using a Cox Proportional Hazards model.

Over the years databases of metabolomic pathways have been collected, condensing knowledge of molecular biology into causative relations. We here have made use of such relations, or at least their ability to group molecules in functional relevant subsets. Our idea is that the correlation of the concentration differences of such a subset of gene products can be seen as a manifestation of the same biological phenomenon. Thus, we argue that pathway-level statistics give a clearer view of how biological processes govern clinical outcomes. Another welcome property of pathway-level statistics is that we test fewer hypotheses than at the gene level, which increases the statistical power of the method.

The results of our analysis showed many pathways that are involved with cell proliferation as the most significantly associated with the survival of the patients. Although this is an obvious result, it is also a welcome find that validates our method. Furthermore, our strategy to extract individual measurements of pathway activity for each sample allows us to elaborate on the statistical model of survival. By regressing out the effects of proliferation, the confounding effect of this major perturbation of our samples was removed, allowing us to see other features of our data more clearly [9].

At the heart of the method is the unsupervised dimensionality reduction that produces a pathway activation score for each sample that can be used later for survival analysis. Because of this general first step, our method could be modified to be suitable for application on any analysis where readouts are used to derive statistical significance.

Finally, another possible extension of our method is to include different modalities of data such as proteomics or metabolomics. As long as the novel data can be grouped into pathways, and the assumption of covariation within and between the different modalities hold, they can be included in the same factor analysis step and their information used for integrating pathway activity in an even broader sense.

Supplementary information

The supplementary information includes:

- Note S1: A null model based on set permutation.
- Supplementary Figure S1: Test of stability of pathways eigengene vectors
- Supplementary Figure S2: Investigation of the statistical calibration of p values from the Cox regression model
- Table S1: Significance of Cox-regression of Reactome's pathways
- Table S2: Significance of Cox-regression of Reactome's pathways, when regressing out pathway R-HSA-9675126

Acknowledgements

This study makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium, which was funded by Cancer Research UK and the British Columbia Cancer Agency Branch. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at HPC2N partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Funding

This work has been supported by a grant from the Swedish Foundation for Strategic Research (BD15-0043).

References

[1] Naomi Altman and Martin Krzywinski. The curse (s) of dimensionality. *Nat Methods*, 15(6):399–400, 2018.

[2] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 2009.

[3] David R Cox. Regression models and life-tables (with discussion). *JR Stat. Soc. Ser. B*, 34:187–220, 1972.

[4] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.

[5] Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Sean Reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel WIlson, Gabriel, Luis Moneda, Arturo Moncada-Torres, Kyle Stark, Harsh Gadgil, Jona, Karthikeyan Singaravelan, Lilian Besson, Miguel Sancho Peña, Steven Anton, Andreas Klintberg, GrowthJeff, Javad Noorbakhsh, Matthew Begun, Ravin Kumar, Sean Hussey, Skipper Seabold, and Dave Golland. Camdavidsonpilon/lifelines: v0.25.8, January 2021.

[6] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*, 4(9):R60, 2003.

[7] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

[8] Elena Edelman, Alessandro Porrello, Justin Guinney, Bala Balakumaran, Andrea Bild, Phillip G Febbo, and Sayan Mukherjee. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, 22(14):e108–e116, 2006.

[9] Jeroen Gilis, Steff Taelman, Lucas Davey, Lennart Martens, and Lieven Clement. Pitfalls in reanalysis of observational omics studies: a post-mortem of the human pathology atlas. *BioRxiv*, 2020.

[10] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[11] Andrej Gorbatenko, Christina Wilkens Olesen, Ebbe Boedtkjer, and Stine Falsig Pedersen. Regulation and roles of bicarbonate transport in cancer. *Frontiers in physiology*, 5:130, 2014.

[12] Johannes Griss, Guilherme Viteri, Konstantinos Sidiropoulos, Vy Nguyen, Antonio Fabregat, and Henning Hermjakob. ReactomeGSA-efficient multi-omics comparative pathway analysis. *Molecular & Cellular Proteomics*, 19(12):2115–2125, 2020.

[13] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[14] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[15] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14(1):1–15, 2013.

[16] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl_1):D428–D432, 2005.

[17] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 2012.

[18] David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

[19] Mark M Moasser. The oncogene her2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene*, 26(45):6469–6487, 2007.

[20] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Aleix Prat and Charles M Perou. Deconstructing the molecular portraits of breast cancer. *Molecular oncology*, 5(1):5–23, 2011.

[23] Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research*, 43(W1):W589–W598, 2015.

[24] John D. Storey. The positive false discovery rate: a bayesian interpretation and the q -value. *Ann. Statist.*, 31(6):2013–2035, 12 2003.

[25] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[26] Saeed Tavazoie, Jason D Hughes, Michael J Campbell, Raymond J Cho, and George M Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281, 1999.

[27] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[28] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pp. 91–109. Springer, 2003.