Generation of Benchmarking Data Set for Quantitative Comparison of Pathway Scoring Methods and Application of Pathway-based Analysis

Michelle Meier
mimeier@student.ethz.ch
Msc Systems Biology, D-BIOL, ETH-Zurich

Supervisor: Dr. Natalie Davidson

1 ABSTRACT

While pathway scoring methods are gaining more traction, there is still lack of a comprehensive benchmarking data set that allows a quantitative comparison of these methods. Evaluation of pathway scoring methods is crucial for pathway-based phenotype prediction and analysis, which builds the foundation for many studies in cancer research.
In this study, I generated a benchmarking data set for six curated cancer pathways (hypoxia, tumor protein 53 (p53), phosphoinositide 3 kinase (pi3k), notch, metastasis and cell cycle) and evaluated five different pathway scoring methods. The singscore method was then used to compute pathway scores for all further analyses. Firstly, the pathway scoring method was used for the analysis of pathway dysregulation in palbociclib resistant cells. While pathway analysis of palbociclib resistant cells suggested a novel significant dysregulation of notch in resistant cell, it could not confirm previously observed disturbances in pi3k. Secondly, pathway scores were calculated for The Cancer Genome Atlas (TCGA) cohort. Linear regression analysis of early stage kidney renal papillary cell carcinoma (KIRP) showed a significant relationship between gender, age, race and the pathways cell cycle and pi3k. Also, in early breast cancer, the model suggests not only a relationship between racial background and p53 but also between age and metastasis and cell cycle. Lastly, prior malignancy and ethnicity significantly contribute to various pathway scores in early stage hepatocellular carcinoma. Many of the results observed from linear regression analysis could be confirmed by findings from prior studies which emphasizes the importance of these models. However, it was discovered that some pathways, such as hypoxia, and cancer types, such as thymoma, are more suitable for this type of analysis than others.

2 INTRODUCTION

With the emergence of next generation sequencing methods over the past two decades [1], transcriptomic data has grown exponentially. While this development was of utter importance, it also came with the challenge of interpreting high-volume outputs of these methods. As genes never act alone but in networks, a promising approach to face this challenge is pathway scoring methods. These methods integrate expression data and curated biological pathways as gene sets in order to determine pathway regulation in a sample. Moreover, prediction models based on pathway scores have been shown to be more robust than models based on expressions of single genes [2]. Various pathway scoring methods are already available. However, it is currently not possible to quantitatively compare these methods as there is no benchmarking data set that combines different tissue types, perturbation agents and the targeted pathway. There is a series of international platforms that archive high-throughput transcriptomic data, such as gene expression omnibus (GEO), which provides easy access to data sets of over 130’000 distinct experiments [3]. Hence, now is the perfect time to generate such a benchmarking data set.
Pathway regulation in cancer has been discussed at length [4]–[6] and it is therefore not surprising that pathway scoring methods have gained traction in that field. The Cancer Genome Atlas (TCGA) research network is a collection of over 11’000 samples and clinical data from roughly 20 different cancer types. This data availability allows us to study pathway dysregulation in specific cancer subtypes, within age groups, ethnicities or gender, which can be of great importance in regard to treatment and therapy [7].
As cancer cells show an aberrant growth phenotype, cell cycle is one of the most commonly dysregulated pathways in cancer[8], [9]. As an example, common treatment for advanced

estrogen hormone receptor positive and negative breast cancer patients is endocrine therapy coupled with cell cycle inhibitors [10]. More specifically, these cell cycle inhibitors target cyclin-dependent kinases 4 and 6 (CDK4/6), which are crucial for the Gap1 (G1) and synthesis (S) – phase checkpoint [11]. Currently, there are three CDK4/6 inhibitors approved by the US Food and Drug Administration (FDA)[12]: palbociclib, ribociclib, and abemaciclib. Unfortunately, treatment with palbociclib is often accompanied by resistance [13]. While many different mechanisms of resistance have been discussed [14]–[16], there is still no conclusive answer to what the main driver of resistance is. Pathway analysis of resistant and sensitive cells could provide a better insight to what pathways are relevant for resistance

In this study, the aim was to generate a comprehensive benchmarking data set for six curated cancer pathways (hypoxia[17], tumor protein 53 (p53)[4], phosphoinositide 3 kinase (pi3k) [4], notch signaling [4], cell cycle [4] and metastasis[18]) and apply five different pathway scoring methods: gene set variant enrichment (GSVA)[19], single sample gene set enrichment analysis (ssGSEA)[20], zscore [21], pathway level analysis of gene expression (PLAGE)[22] and single sample scoring (singscore)[23]. Further, I compared the performance of the different methods on my benchmarking data set using a rank distribution metric. I then used the best methods to compute pathway scores for palbociclib resistant samples in order to study how dysregulation of pathways contribute to the resistance phenotype. Lastly, I used linear regression analysis to investigate pathway regulation in specific cancer types from the TCGA cohort, namely kidney renal papillary cell carcinoma (KIRP), glioblastoma, thymoma and hepatocellular carcinoma, with respect to different demographic features. These findings will then be compared to previous studies in order to confirm the result

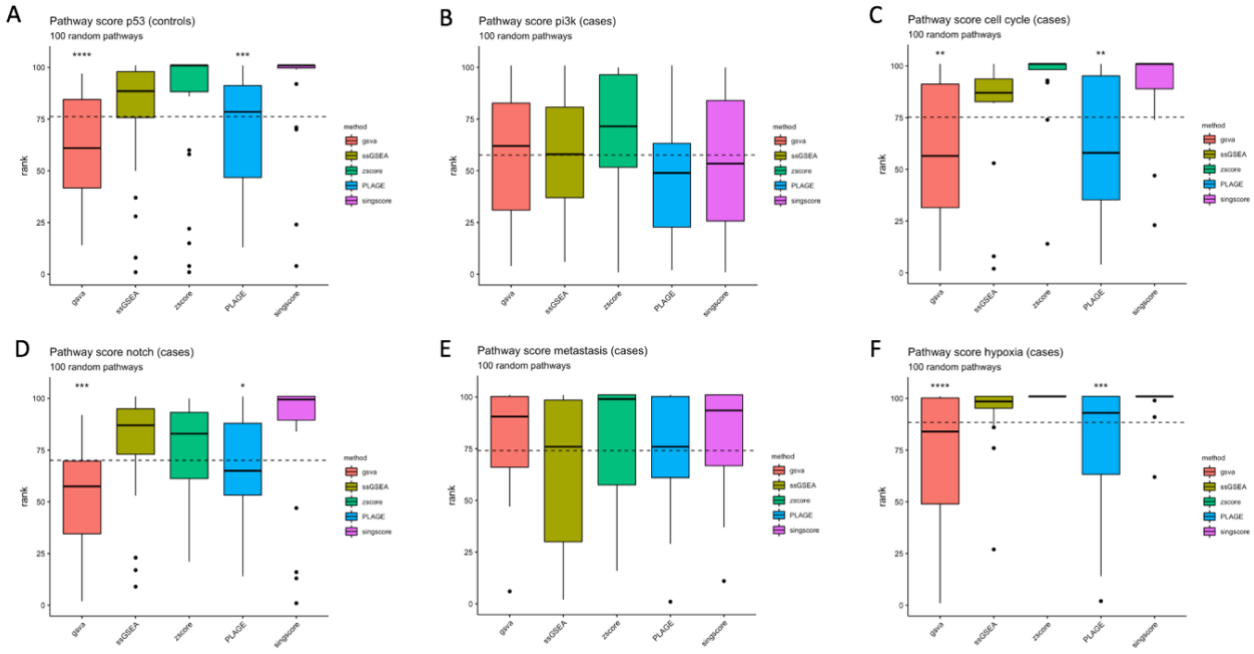


Figure 1: I used a rank distribution metric to evaluate five different pathway scoring methods (GSVA, ssGSEA, zscore, PLAGE, singscore, from left to right in each plot) by comparing their performance in identifying the true pathway out of 100 randomly generated gene sets in each sample. The higher the rank, the better the performance. Each figure **A-F** is dedicated to a pathway studied. The title of each plot describes the pathway and specifies in brackets if case or control samples were enriched for the true gene set in quality control. The dashed line indicates the mean rank over all methods. For each method, the rank for all samples is summarized in a boxplot showing the mean (black centered line), upper and lower quantiles and whiskers extending to 1.5x the interquantile range. Points beyond the whiskers show outliers. The performance of all methods was compared to the singscore method using a student's t test. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > *** > 0.0001 > **** > 0. The singscore method outcompetes GSVA and PLAGE in two third of the studied pathway. No significant difference in performance was observed for metastasis and pi3k.

3 RESULTS

3.1 BENCHMARKING DATA SET

I was able to identify enough relevant samples for each pathway to generate a sufficiently large benchmarking data set. The final benchmarked data set contains a total of 6 pathways and 1'382 samples. The largest set was found for metastasis (N=628) while the sets for all other pathways consist of roughly 250 samples (Table S 2).

3.1.1 DATA AVAILABILITY

The generated benchmarking data set for the pathways analyzed in this study are publicly available at https://github.com/michellemeier27/Semesterproject/tree/master/benchmarking_data_set. For each pathway, it contains all relevant samples along with the following meta information for each sample: The GEO series and sample accession ID, the pathway of interest, whether the sample is case or control, a Boolean describing if there are case and control samples in the corresponding series and finally the direction of gene set enrichment using ssGSEA.

3.2 PATHWAY SCORING EVALUATION

Five different pathway scoring methods (GSVA, ssGSEA, zscore, PLAGE and singscore) were applied to each sample of every pathway specific subset in order to determine which

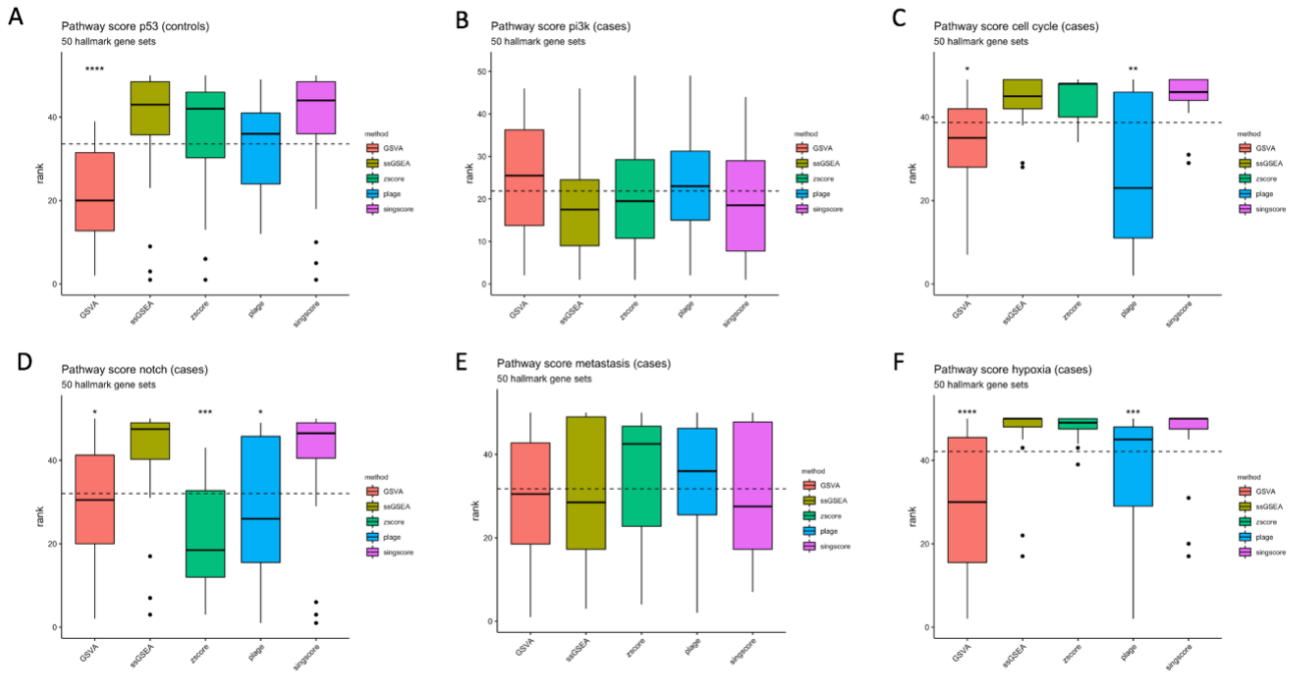


Figure 2: I used a rank distribution metric to evaluate five different pathway scoring methods (GSVA, ssGSEA, zscore, PLAGE, singscore, from left to right in each plot) by comparing their performance in identifying the true pathway out of 50 defined hallmark gene sets in each sample. The higher the rank, the better the performance. Each figure **A-F** is dedicated to a pathway studied. The title of each plot describes the pathway and specifies in brackets if case or control samples were enriched for the true gene set in quality control. The dashed line indicates the mean rank over all methods. For each method, the rank for all samples is summarized in a boxplot showing the mean (black centered line), upper and lower quantiles and whiskers extending to 1.5x the interquantile range. Points beyond the whiskers show outliers. The performance of all methods was compared to the singscore method using a student's t test. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > *** > 0.0001 > **** > 0. The singscore method outcompetes at least one other method in two third of the studied pathways, while no difference in performance is observed for the other third.

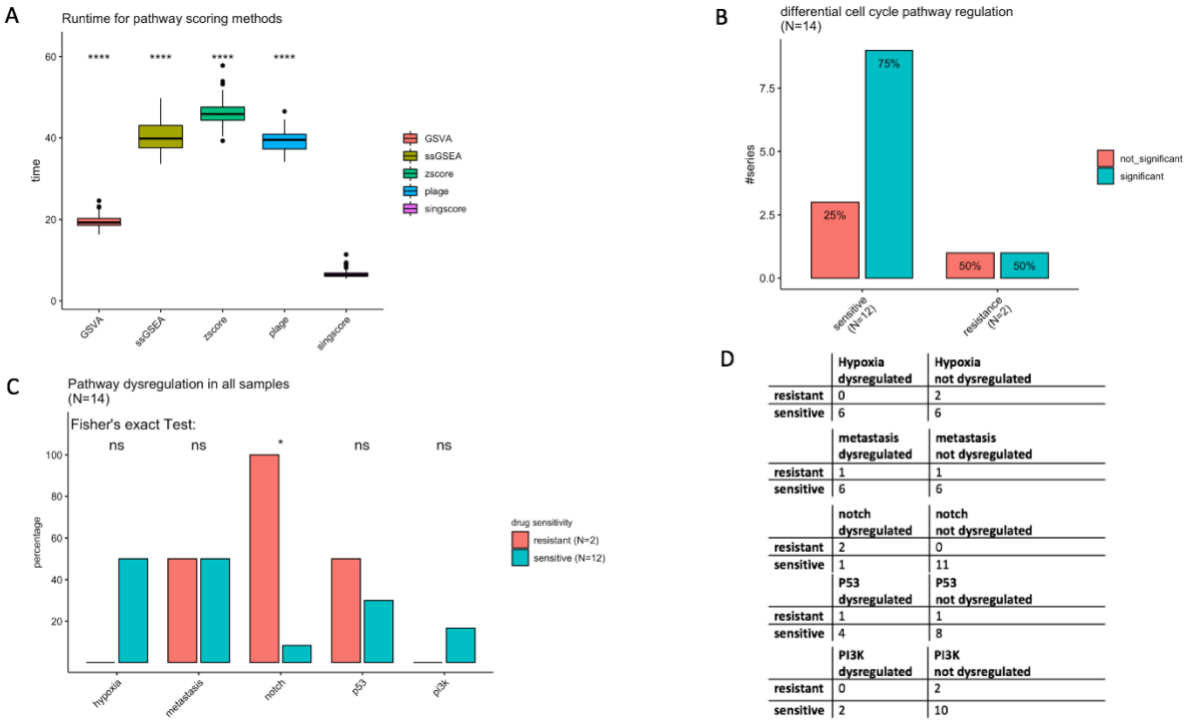
method most reliably identified alterations in the pathway of question. In order to further compare the performance of the singscore method to the other methods, a student's *t*-test was performed.

3.2.1 IDENTIFICATION OF TRUE PATHWAY OUT OF 100 RANDOM PATHWAYS

The methods were compared using a rank distribution metric as described in section 5.2.3. The highest rank is 101 and indicates perfect identification of the wanted pathway out of 100 randomly generated gene sets by the corresponding scoring method. Figure 1 summarizes the results of the pathway scoring evaluation. With exception of pi3k, the mean of singscore is always above the mean over all methods (dashed line in each figure). As can be seen in Figure 1 A,C,D,F, in two third of the studied pathways, singscore performed significantly better than PLAGE and GSVA while no difference in performance was noted for ssGSEA and zscore. For the metastasis pathway (Figure 1 E), no difference in performance was observed across all methods.

3.2.2 IDENTIFICATION OF TRUE PATHWAY OUT OF 50 HALLMARK GENE SETS

The methods were compared using a rank distribution metric as described in section 5.2.4. The highest rank is 50 and indicates perfect identification of the wanted pathway out of 50 hallmark gene sets gene sets by the corresponding scoring method. Figure 2 summarizes the results of the pathway scoring evaluation. The singscore methods outcompetes at least one other method in two third of the studied pathways (Figure 2 A,C,D,F). While there is no significant



	Hypoxia dysregulated	Hypoxia not dysregulated
resistant	0	2
sensitive	6	6

	metastasis dysregulated	metastasis not dysregulated
resistant	1	1
sensitive	6	6

	notch dysregulated	notch not dysregulated
resistant	2	0
sensitive	1	11

	P53 dysregulated	P53 not dysregulated
resistant	1	1
sensitive	4	8

	PI3K dysregulated	PI3K not dysregulated
resistant	0	2
sensitive	2	10

Figure 3: **A**: The runtime was measured for each method when computing the rank distribution for pi3k identification out of 50 hallmark gene sets. Each method was run 60 times and the resulting time measurments shown in a Tukey boxplot. The boxplot shows the mean (black centered line), upper and lower quantiles and whiskers extending to 1.5x the interquantile range. Points beyond the whiskers show outliers. The runtime of all methods was compared to the singscore method using a student's t test. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > *** > 0.0001 > **** > 0. The singscore methods clearly outcompetes all other methods with a mean runtime of 6.68 seconds. **B**: Regulation of cell cycle pathway was compared between cell cycle inhibitor resistant (right) and sensitive (left) cells upon treatment. Sensitive samples were expected to show dysregulation, while in resistant samples cell cycle should not be significantly dysregulated compared to control samples. Out of 12 sensitive samples, 8 samples showed differential cell cycle regulation. Out of 2 resistant samples, only 1 sample showed differential cell cycle regulation. **C**: The relative number of samples with differentially regulated pathways are compared between resistant (salmon) and sensitive (blue) samples. Significant notch dysregulation in drug resistant compared to sensitive samples was confirmed using a Fisher's exact test (p = 0.03). Also, hypoxia is dysregulated in half of the sensitive samples while dysregulation did not occur in any resistant sample (p = 0.49). **D**: Occurance of significant differential pathway regulation was counted by hand after comparing treated to untreated for both resistant and sensitive samples and plugged into the contingency tables. These were then used to compute the Fisher's exact test.

performance difference in the metastasis and pi3k pathway (Figure 2 B,E), the pi3k pathway seems to have a rather low mean (dashed line) across all methods. Moreover, this low mean does not result from some methods being outliers and performing much worse than others; all methods perform poorly for this pathway.
Overall, the rank distribution when detecting the true pathway out of random pathways (Figure 1) showed a similar pattern to detecting the true pathway out of defined hallmark gene sets (Figure 2).

3.2.3 SCORING METHODS RUNTIME

To not only compare the methods by performance, runtime of each method was measured when computing the rank distribution for pi3k identification out of 50 hallmark gene sets. The result can be seen in the Tukey boxplot in Figure 3 A. The methods were compared to the runtime of singscore and evaluated using a student's t test. Singscore significantly outperformed all other methods with a mean running time of 6.68 seconds. GSVA came to a mean of 19.51, ssGSEA 40.19, zscore 46.27 and PLAGE 39.32 seconds.

3.3 CELL CYCLE INHIBITORS

Pathway scores for all relevant samples were computed and the scores compared between palbociclib treated and untreated for resistant and sensitive samples. The plots in Figure 3 B and C were generated from manually counting the occurrences of significantly differentially regulated pathways between case and control samples. Firstly, regulation of the cell cycle pathway was compared between cell cycle inhibitor resistant and sensitive cells upon treatment (Figure 3 B). Of the 12 reportedly sensitive samples, only one fourth (N=3) samples did not show significant cell cycle dysregulation. However, out of 2 reportedly resistant samples half of the samples (N=1) showed cell cycle dysregulation. This result was not statistically significant which most probably is due to the small sample size of this study (N=14).
Further, regulation of all other pathways was investigated for both sensitive and resistant samples separately. The results for the resistant samples can be viewed in Figure S 3 A. Both resistant samples showed significantly enriched notch pathway, one sample enrichment in metastasis and one in p53. No enrichment was found in the pi3k and hypoxia pathway for both samples. The results for the sensitive samples are depicted in Figure S 3 B. In half of the samples (N=6), the pathways metastasis and hypoxia were dysregulated. While one third (N=4) of the sensitive samples studied showed dysregulation in p53, only one and two samples were dysregulated in pi3k and notch, respectively. Lastly, I also directly compared pathway dysregulation between resistant and sensitive samples and calculated statistical significance using a Fisher's exact test. The contingency tables for the calculations can be found in Figure 3 D. The plot in Figure 3 C compares the relative number of samples that showed dysregulation for the corresponding pathway for sensitive (blue) and resistant (salmon) samples. The only pathway that showed significant (P=0.03) differential regulation between resistant and sensitive samples was notch. However, it can be observed that all resistant samples do not show any dysregulation in hypoxia, while half of the sensitive samples do (P=0.49).

3.4 TCGA PATHWAY ANALYSIS

Overall, when looking at all models for all cancer subtypes, one can see that the adjusted R^2 is consistently lower than 0.5 for all p53 models. On the contrary, adjusted R^2 is high in hypoxia for all models. An overview of all adjusted R^2 can be found in Figure 4 C.
Also, not all weight estimates are shown in this report. For a full comprehensive list of all factors, estimates, p values, standard deviations and adjusted R^2 for all models see https://github.com/michellemeier27/Semesterproject/tree/master/results/linear_regression%20.

3.4.1 GENDER BIAS IN MOLECULAR SIGNATURES IN TCGA PATIENTS

I performed a linear regression analysis on early stage KIRP (N=77) and all stages glioblastoma (N=65) data available in TCGA (see **Error! Reference source not found.**). Figure 4A and B show forest plots of the weight estimates for the factors age, race and gender in the linear regression models. For the KIRP data, which can be seen in Figure 4 A and B, I observed a statistically significant relationship between gender of a patient and pathway score in two models (2 out of 6 pathways): For pi3k (adjusted R^2 = 0.53, intercept = 0.12) and cell cycle (adjusted R^2 = 0.49, intercept = 0.02). Additionally, both models also showed a significant relationship between being African or African American, age and the pathway score.
For the glioblastoma data, there was no significant relationship between gender, age or racial background for all pathway scores. Notably, despite varying with the factors used to build the model and removing outliers, the optimized linear models for pathway scores for pi3k and

C

	pi3k	p53	metastasis	notch	hypoxia	cell cycle
early BRCA	0.30	0.34	0.45	0.29	0.54	0.37
late BRCA	0.50	0.22	0.36	0.28	0.61	0.40
GBM	0.00	0.49	0.45	-0.03	0.49	0.25
KIRP	0.53	0.31	0.50	0.37	0.45	0.49
THYM	0.90	0.27	0.74	0.77	0.79	0.93
early LIHC	0.60	0.37	0.53	0.39	0.64	0.62
late LIHC	0.42	0.16	0.78	0.80	0.56	0.46

adjusted *R squared* | < 0.35 | < 50 | < 70 | < 100

Figure 4: **A,B**: These forest plots show estimated weights of two linear models (see Table S 4 for detailed model) for the pathway score pi3k (adjusted R_2 = 0.53, intercept = 0.12) and cell cycle (adjusted R_2 = 0.49, intercept = 0.02) which were fitted to KIRP patient data (N=77). Only estimates for the factors age, race and gender are shown. The defaults assumed by the model for the shown factors are female (N= 20) for gender and American Indian or Alaskan native (N= 1) for race. Information on race was not available for 3 patients. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > ***. The white dot shows the mean of the weight estimate distribution, the blue box spans to a confidence interval of 90%. Whiskers extend to a confidence interval of 95%. Both models identified a significant relationship between racial background, gender, age and the corresponding pathway score. **C**: This table presents an overview of all adjusted R_2 in all linear regression models. The R_2 value shows, how much of the variance in the dependent variable can be explained by the model. Adjusted R_2 values were colored in according to the legend at the bottom of the table. All models for the thymoma data show a high adjusted R_2, while the adjusted R_2 for most p53 models are very low. Legend: BRCA = breast cancer, GBM = glioblastoma, LIHC = hepatocellular carcinoma, THYM = thymoma, KIRP = kidney renal papillary cell carcinoma.

notch rendered a very low adjusted R_2 around 0. The other pathway models showed an adjusted R_2 ranging from 0.25 to 0.5 (Figure 4 C).

3.4.2 LINEAR REGRESSION FOR BREAST CANCER, THYMOMA AND HEPATOCELLULAR CARCINOMA

I performed a linear regression analysis to study the relationship between pathway scores and demographic features for breast cancer (N=337 early stages and N=115 in late stages), thymoma (N=38) and hepatocellular carcinoma (N=91 early stages and N=36 in late stages) (see 5.4.4). Using a linear model for the pathway score of p53 (adjusted R_2= 0.3, intercept = 0.12), a significant relationship between being white and pathway scores p53 was observed in early stage breast cancer patients. Also, the results suggest a relationship between pathway score and age of the patient in the model for cell cycle (adjusted R_2= 0.37, intercept = 0.09) and metastasis (adjusted R_2= 0.45, intercept = 0.2). These results are visualized using a forest plot in Figure 5 A-C. However, the linear regression model did not suggest a relationship between racial background and hypoxia pathway score in early (p = 0.16) and in late stages (p = 0.66).

Next, I built linear regression models for the thymoma data. While no demographic factors are suggested to contribute to the pathway score by the model, the adjusted R_2 were strikingly high for all pathway models but p53 (adjusted R_2 = 0.27). Moreover, the only significant factors are the pathway scores of the other pathways. Figure 5E and D show the forest plots for the pathway factors for the pi3k (adjusted R_2= 0.9, intercept = 0.16) and cell cycle model (adjusted R_2= 0.93, intercept = 0.16).

Lastly, the model for early stage hepatocellular carcinoma showed a significant relationship between having had a prior malignancy and the pathway score in the following models:

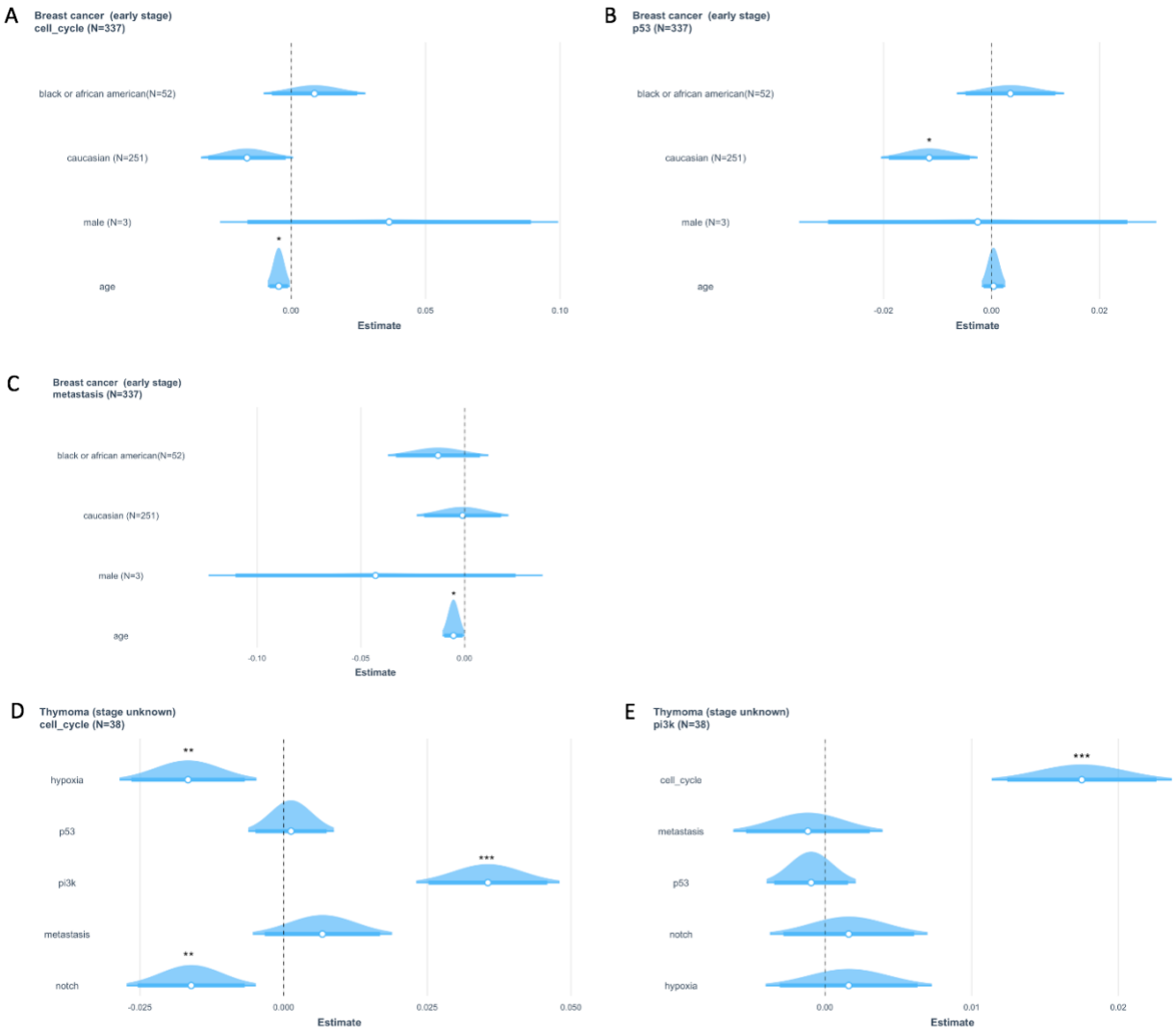


Figure 5: **A-C:** These forest plots show estimated weights of three linear models (see Table S 4 for detailed model) for the pathway score cell cycle (adjusted R_2 = 0.37, intercept = 0.09), p53 (adjusted R_2 = 0.3, intercept = 0.12) and metastasis adjusted R_2 = 0.45, intercept = 0.2) which were fitted to early stage breast cancer patient data (N=337). Only estimates for the factors age, race and gender are shown. The defaults assumed by the model for the shown factors are female (N=334) for gender and Asian (N=15) for race. Information on race was not available for 19 patients. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > ***. The white dot shows the mean of the weight estimate distribution, the blue box spans to a confidence interval of 90%. Whiskers extend to a confidence interval of 95%. **A,C**: Both models identified a significant relationship between age and the corresponding pathway score. **B**: The linear model suggests a relationship between being Caucasian (N=251) and the pathway score for p53.
C-D: These forest plots show estimated weights of two linear models (see Table S 4 for detailed model) for the pathway score cell cycle (adjusted R_2 = 0.93, intercept = 0.16) and pi3k (adjusted R_2 = 0.9, intercept = 0.16) which were fitted to thymoma patient data (N=38). These two models were chosen as they have a very high adjusted R_2 value. Only estimates for the other pathway scores studied are shown. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > ***. The white dot shows the mean of the weight estimate distribution, the blue box spans to a confidence interval of 90%. Whiskers extend to a confidence interval of 95%. **D:** The model suggests contribution of the pathway scores hypoxia, notch, pi3k to the cell cycle score. **E**: The model suggests contribution of the pathway score of cell cycle to the pi3k score.

Hypoxia (adjusted R_2 = 0.64, intercept = 0.14), metastasis (adjusted R_2 = 0.53, intercept = 0.03) and notch (adjusted R_2 = 0.39, intercept = 0.07) (Figure 6 C,A,B). Additionally, the linear model for hypoxia score suggests a significant contribution of age and not being Hispanic or Latino to the pathway score (Figure 6 C). I also observed a relationship between ethnicity, age and the pathway score in the p53 model (adjusted R_2 = 0.37, intercept = 0.08) (Figure 6 D). These results are visualized using a forest plot in Figure 6.

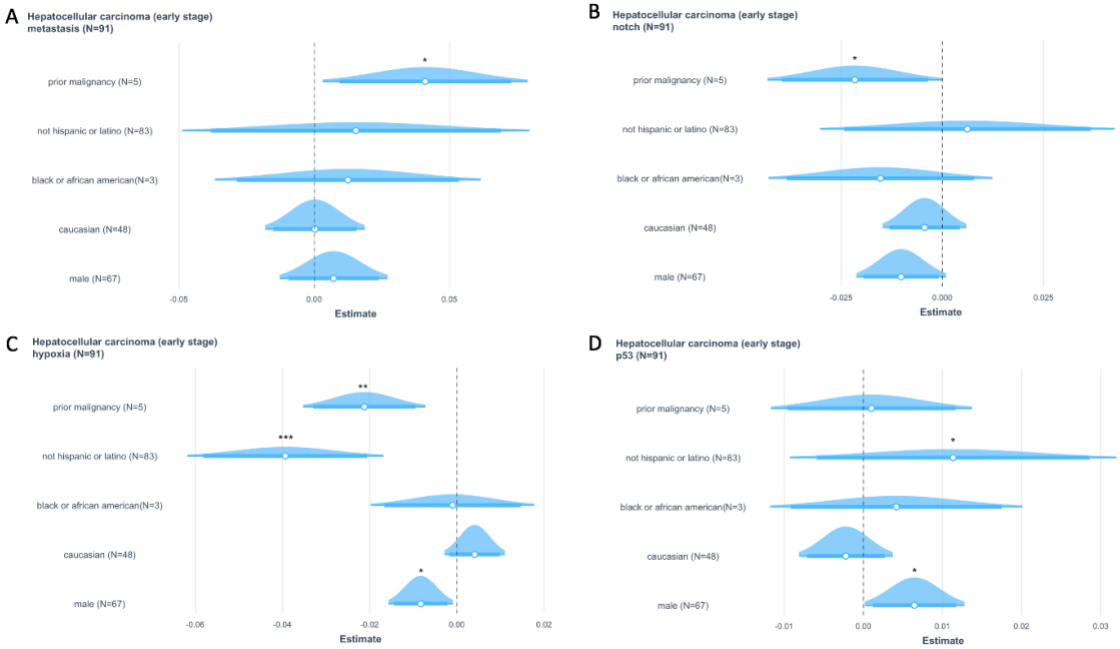


Figure 6: These forest plots show estimated weights of four linear models (see Table S 4 for detailed model) for the pathway score hypoxia (adjusted R_2 = 0.64, intercept = 0.14), p53 (adjusted R_2 = 0.37, intercept = 0.08), notch (adjusted R_2 = 0.39, intercept = 0.07) and metastasis (adjusted R_2 = 0.53, intercept = 0.03) which were fitted to early stage hepatocellular carcinoma patient data (N=91). Only estimates for the factors age, race, ethnicity (Hispanic or Latino), prior malignancies and gender are shown. The defaults assumed by the model for the shown factors are female (N= 24) for gender, Hispanic or Latino (N=3) for ethnicity, no prior malignancies (N=86) for prior malignancy and Asian (N= 35) for race. Information on race and/or ethnicity was not available for 5 patients. Significant p values are represented by asterisks as follows: 0.05 > * > 0.01 > ** > 0.001 > ***. The white dot shows the mean of the weight estimate distribution, the blue box spans to a confidence interval of 90%. Whiskers extend to a confidence interval of 95%. **A, B**: Both models identify significant relationships between having had prior malignancies and the corresponding pathway score. **C:** The model suggests a significant relationship between ethnicity, gender, having had prior malignancies and the pathway score. **D**: The model suggests a significant relationship between ethnicity, gender and the pathway score.

4 DISCUSSION

4.1 BENCHMARKING DATA SET

The benchmarking data set generated in this study is the first of its kind: It combines data across different tissues, perturbation agents and targeted pathways and makes this heterogeneous data easily available. While much care was taken in order to keep the benchmarking data set as pure as possible, some points need to be considered. Firstly, the benchmarking data set only contains six different pathways and they each vary in sample sizes. Additionally, some pathways, like metastasis, are rather difficult to define as they may integrate many different pathway responses [24][25]. Secondly, sample selection was based only on the all RNA-Sequencing (RNA-Seq) and chromatin immunoprecipitation sequencing (ChiP-seq) sample and signature search (ARCHS4) database, which integrates many distinct experiments but is by all means not complete. During my study, I have realized that in many experiments the samples listed on GEO do not coincide with the samples found on ARCHS4. Moreover, the ARCHS4 data base realigns all reads found in the experiments to the most recent human reference genome and provides the raw integer reads for genes on a gene symbol basis. While I can fully appreciate that aligning the reads of all experiment to a common reference genome benefits the handling of heterogeneous data, a great number of gene symbols could not be assigned to an ensemble ID. Therefore, I was not able to normalize for gene length. However, as the data in this study was normalized for library size and only a defined set of genes was considered for each pathway, this should not have affected the outcome of the analysis.

4.2 PATHWAY SCORING ANALYSIS

In order to evaluate the different pathway scoring methods used in this study, they were compared based on their performance and runtime. As mentioned, singscore significantly outperforms GSVA and PLAGE in identifying the true pathway among 100 randomly generated gene sets in four out of six pathways studied. Singscore, zscore and ssGSEA commonly have similar scores and thus the runtime of these methods was compared. The runtime of singscore was one fourth of the runtime of ssGSEA and even one eighth for zscore. Therefore, based on its good performance across all pathways and excellent runtime, singscore was used to calculate pathway scores in all further analyses.
The same methods were also assessed in identifying the true pathway out of 50 hallmark gene sets. Generally, as pathways are highly interconnected, it was expected that the methods perform worse compared to identification of the true pathway out of random gene sets. However, this was only true for pi3k. Interestingly, the overall rank distribution was strikingly similar to the distributions observed for identification out of 100 random pathways. For example, the mean rank across all methods for pi3k is lower compared to the other pathways in both performance analyses. On the other hand, hypoxia scores are exceptionally high in both analyses. This finding suggests that the properties of the pathway might have an effect on the performance of pathway scoring methods, which makes some pathways inherently easier to detect with these methods than others. Many previous studies have discussed the performance of pathway scoring methods [26], [27], but none of them consider the underlying pathway's properties. Future studies could provide insight into how pathway structure affects pathway scoring and if and how it must be incorporated into pathway scoring methods.

4.3 CELL CYCLE INHIBITOR PATHWAY ANALYSIS

Firstly, to address a fundamental issue of this analysis, I want to underline that this study was conducted with only 12 sensitive and 2 resistant samples. This is due to the lack of relevant samples found in the ARCHS4 database. Nevertheless, the results from this study are discussed in more detail as follows: Palbociclib resistant samples were hypothesized to exhibit similar cell cycle activity to untreated samples, while sensitive samples were expected to show significant downregulation of cell cycle upon treatment. For the sensitive samples, these expectations were met as 75% of the samples analyzed showed significant dysregulation in cell cycle. As only two resistant samples were studied and one sample showed significant dysregulation of cell cycle, the hypothesis on cell cycle regulation in resistant cells could neither be confirmed nor rejected. Further, I compared the significantly dysregulated pathways in resistant and sensitive samples. Previous studies have identified Phosphoinositide 3-kinase 3C (PIK3CA, in pi3k) and (retinoblastoma protein 1) Rb1 as potential driver mutations for *de novo* resistance [14], [15], [28]. Additionally, *acquired* resistance has been shown through non-canonical cell cycle activation induced by pi3k signaling [15]. Lastly, a recent study also associated dysregulation in DNA repair and statin 3 with palbociclib resistance [29]. Unfortunately, my results do not confirm dysregulation of the pi3k pathway. The contribution of Rb1 loss or mutation to resistance could not be assessed as this gene was considered part of the cell cycle pathway. Finally, the dysregulation of DNA repair and statin 3 could not be quantified as these pathways were not included in the benchmark data set. Interestingly, the analysis suggests a never-seen dysregulation of notch. While one should keep in mind that the sample size for this analysis was very small, future studies with larger sample sizes could further study the contribution of notch to the resistance phenotype.

4.4 TCGA PATHWAY ANALYSIS

4.4.1 GENDER BIAS IN MOLECULAR SIGNATURES IN TCGA PATIENTS

This study was conducted based on results from Yuan *et al* [30].They discuss gender-biases in various different cancer types. This study was set up in order to confirm the observed gender-bias in KIRP and lack thereof in glioblastoma. As described, the linear regression model suggests a significant relationship between gender and pi3k in early stage KIRP. This coincides nicely with the findings of Yuan *et al*, as they report significant gender-specific somatic copy number alterations in various genes of the pi3k pathway. The results of this study can also confirm the reported lack of gender-bias in glioblastoma. To my knowledge, no previous studies have investigated the relationship between gender and cell cycle. Both cell cycle and pi3k were also associated with age and racial background. A study from 2017 [31] reports a potential contribution of single nucleotide polymorphisms (SNPs) in genes associated with pi3k to the risk factor in prostate cancer, especially in specific age groups, smoking status, body mass index and racial background. Future pan cancer studies could evaluate, if this contribution of SNPs to risk is specific to prostate cancer and whether gene polymorphisms could also influence pi3k regulation after diagnosis.

4.4.2 LINEAR REGRESSION FOR BREAST CANCER, THYMOMA AND HEPATOCELLULAR CARCINOMA

As mentioned, in this study a linear relationship between racial background and p53 was found for early breast cancer. Previous studies [32], [33] have also identified an influence of racial background on p53 status. This is discussed in the context of prognosis and survival and is therefore highly relevant for treatment. Also, in this study age at diagnosis was found to

contribute to cell cycle and metastasis. Purushotham *et al.* [34] report an inverse relationship between age at diagnosis and distant metastasis in breast cancer, which confirms my findings. To my knowledge, no studies discuss the potential contribution of age to cell cycle. Lastly, I want to discuss a previously reported relationship between being African American and hypoxia by Bhandari *et al* [34]. The results from the linear regression model built in this study does not suggest any dependence of hypoxia on racial background. However, there are a few fundamental differences in the set-up of the two studies that must be considered: Firstly, in this study I used the hypoxia hallmark gene set to compute the pathway scores, while Bhandari *et al.* based their scores on enrichment of the Buffa mRNA signature [35]. Also, their sample size is roughly three times as large (N=997 vs N=337), which results from the reduced TCGA cohort used in this study. Lastly, but most importantly, the findings of this study only suggest that there is no linear relationship and does not exclude possible contribution in a non-linear fashion.

The most common risk factor for hepatocellular carcinoma is cirrhosis, often caused by a chronic hepatitis B/C infection [36]. However, diabetes, alcohol and drug abuse, obesity and other prior malignancies are also considered risk factors [37]. As described, the linear regression analysis suggests a relationship between having had prior malignancies and metastasis, notch and hypoxia, which is half of the pathways studied. Unfortunately, the National Cancer Institute does not provide detailed information on what kind of prior malignancy the patients suffered from. As prior malignancy plays such an important role in the development of hepatocellular carcinoma, this relationship and its potential importance in treatment should be investigated in further studies. In this study, I also report significant contribution of ethnicity to hypoxia and p53. Prior studies have not only discussed variation in incidence rates among different ethnicities in the US [38], but also variation in diagnosis and curative treatment receipt [39]. This delay in treatment could explain the differential pathway regulation.

The most surprising result of the linear regression analysis of thymoma was the strikingly high adjusted R^2 values for almost all pathway models; Apart from p53, all models explained more than 73 % of the variance in the data. Interestingly, none of the demographic features were significant, the only significant factors contributing to the pathway scores were the scores from other pathways. The most extreme observation was made for pi3k: 90 % of the variance was explained by only one other pathway (cell cycle) as significant factor. While it is known that the main effector of pi3k, mTOR, is associated with cell growth [40] and cell cycle is a significant factor in almost all pi3k models studied, it remains unknown why this strong linear dependence is unique to this cancer type.

Quite contrary results were observed for glioblastoma. In these models, adjusted R^2 never exceeded 0.5 and two pathways had incredibly low values. These results suggest that pathway regulation in some cancer types might be more suitable for linear regression analysis than others. Also, this seems to hold true for specific pathways: While most model for hypoxia have a high adjusted R^2, p53 models cannot explain large portions of the variance for almost all cancer types studied. These findings emphasize the complexity of not only pathway dysregulation in cancer but also the innate complexity of pathway networks. Despite being able to support many of the results obtained from this study with previous findings, some results remain unexplained or contradict literature. Nevertheless, considering the linear regression models used in this study are rather simple it is remarkable how many results they managed to reproduce.

5 METHODS

5.1 GENERATING BENCHMARKING DATA SET

5.1.1 RNA-SEQ DATA ACQUISITION

Human RNA-Seq gene counts and the corresponding meta data were downloaded using the web application ARCHS4 [41] and are publicly available at https://amp.pharm.mssm.edu/archs4/download.html. Gene counts were generated by aligning raw reads to the human reference genome (GRCh38) using Kallisto [42]. All gene counts are on gene-level and rounded to integer levels. The meta data was then queried in RStudio using the Bioconductor "rhdf5" (v 2.32.0) package [43]. In order to identify suitable samples for the benchmarking data set, I used a list of search terms (Table S 1) dedicated to each of the pathways of interest. The query result was then further curated manually to ensure that only relevant samples were kept for further analysis. Finally, the samples for each pathway were completed with relevant meta data. Meta data included are the GEO accession IDs for each sample, the GEO accession ID for the corresponding series, a reference whether the sample was diseased (case) or healthy (control), a note on whether there are both cases and controls in the series and the pathway of interest. Some series were subdivided into subseries as either only a few samples were relevant for the study or there were too many conditions in one series, e.g. different cell lines. RNA-Seq data was then extracted for all identified samples for each pathway. For technical replicates, the mean of the expression values was used to merge them into one sample.

5.1.2 DATA PROCESSING AND FILTERING

Unfortunately, the gene counts could not be transformed to fragments per kilobase per million mapped reads (FPKM) [44] or transcript per million (TPM) [45] as the gene symbol annotation on ARCHS4 is incomplete and the gene length for a substantial number of genes could not be determined. For this reason, further analysis was performed with raw integer gene counts. Lowly expressed genes across all samples were removed by applying a filter from the Bioconductor package "edgeR" (v 3.30.0)[46] to the log transformed (raw expression +1) values. Most parameters were kept at their default value. As an exception, the *min.count* parameter was adjusted in order to get approximately 12'000 genes in each pathway (see Table S 2 for a more detailed overview). Next, I normalized the library sizes within series using upper-quantile (UQ) normalization[47] and the resulting normalized libraries were compared to the unnormalized libraries using a Tukey boxplot for confirmation. Lastly, the normalized gene counts were scaled and centered across all samples within a pathway.

5.1.3 DATA ANALYSIS AND FURTHER FILTERING (QUALITY CONTROL)

T-distributed stochastic neighbor embedding (*t*-SNE) plots were generated over all series in each pathway using the CRAN package "Rtsne" (v 0.15) [48]. Due to small sample sizes within series the perplexity parameter was adjusted to 5.
Further, in order to ensure differential expression between case and control samples, the following analysis were performed for each individual series:

- *T*-SNE plots were generated and clustering of case and control samples analyzed.
- Principal component analysis (PCA) [49] was performed and the results visualized

using the CRAN package "factoextra" (v 1.0.7) [50]

- ssGSEA [20] was performed using the Bioconductor package "GSVA" (v 1.36.0) [19] using publicly available gene sets from the Molecular Signature Database (MSigDB) (v7.1). The gene sets applied for each pathway are listed in Table S 3 in more detail. The enrichment result was then plotted using the package "ggbiplot" (v 0.55) [51]. Additionally, to confirm the enrichment of the gene set in the case or control samples a Tukey boxplot comparing the enrichment scores was created. The meta data for each sample was completed by adding a reference, denoted as "direction", for which sample group the enrichment score was higher.

Series that did not show sufficient differential expression or clustering were removed from the benchmarking data set.

5.2 PATHWAY SCORING METHODS

5.2.1 OVERVIEW OF PATHWAY SCORING METHODS

The following table (Table 1) provides an overview over the different pathway scoring methods used and very briefly describes the rationale behind them. All methods are suitable for analysis of single samples.

Table 1: Summary of the five different pathway scoring methods used. For each method, a brief description and reference for more detail is provided

Method	Description	Reference
GSVA	Estimates the distribution of gene expression across all samples and then calculates Kolmogorov-Smirnov rank statistics.	Hänzelmann *et al.* [19]
ssGSEA	Calculates enrichment score for a single sample for a gene set by comparing the distribution of gene expression ranks inside and outside the gene set.	Barbie *et al.* [20]
zscore	Enrichment score for a gene set made up of added individual gene activity zscores. Standardizes gene activity across all samples. Assumes normally distributed expression data and gene independence in gene sets.	Lee *et al.* [21]
PLAGE	Dimension reduction by decomposing variance in expression for each gene set using singular value decomposition. Standardizes across all samples. Assumes normally distributed expression data.	Tomfohr *et al.* [22]
singscore	Genes are ranked according to mRNA abundance and enrichment scores computed based on normalized mean percentile ranks.	Foroutan *et al.* [23]

5.2.2 COMPUTATION OF PATHWAY SCORES

The Bioconductor package "GSVA" (v 1.36.0) [19] offers a selection of four different scoring methods: GSVA, ssGSEA, z-score and PLAGE. Additionally, the singscore method available in the Bioconductor package "singscore" (v 1.8.0) [23] was used as a fifth scoring method. The expression data was normalized and filtered as described above for all five methods. An additional filtering step which removed samples with less than two samples in either control or case group was applied to the benchmark data set.

5.2.3 *IDENTIFICATION OF TRUE PATHWAY OUT OF 100 RANDOM PATHWAYS*

The benchmarked data set containing a total of 6 pathways and 1'382 samples was then used to compare various pathway scoring methods available in R. Pathway scoring methods were applied to each pathway individually. In order to compare the different methods, the score difference for the true pathway between case and control samples in each series were compared to the pathway score differences between case and control for 100 randomly generated pathways. The methods were run on each pathway separately (see Figure S 1 for more detail). The gene sets corresponding to the true pathway for each pathway can be obtained from Table S 3. The score differences were then sorted, and the rank of the true pathway noted. The rank distribution then served as an evaluation metric. A student's t test was performed to determine statistically significant differences in rank distributions between singscore and all other methods using the "stats" package (v 3.6.2) in R.

5.2.4 *IDENTIFICATION OF TRUE PATHWAY OUT OF ALL HALLMARK GENE SETS*

To investigate the potential of the pathway scoring methods to detect the true pathway out of non-random gene sets, I also performed an analysis using the 50 hallmark gene sets available on MSigDB at https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=H. The true pathways for each pathway were defined as mentioned (Table S 3) except for the cell cycle pathway as this true pathway is not a hallmark gene set. The hallmark G2M checkpoint gene set was used instead. The pathway scoring methods GSVA, ssGSEA, z-score, PLAGE and singscore were applied to each hallmark pathway and the score difference between case and control samples were noted for each pathway. Each method was run on all 50 pathways (see Figure S 2 for more details). Similar as mentioned above, the rank distribution of the true pathway was used as an evaluation metric.

5.2.5 *RUNTIME*

In order to compare the computational efficiency of the different methods, the runtime was measured for true pi3k pathway identification out of all hallmark genes. Time measurements were repeated 60 times for all methods and a student's t test performed using the "stats" package (v 3.6.2) in R.

5.3 PATHWAY ANALYSIS CELLS TREATED WITH CELL CYCLE INHIBITORS

5.3.1 *IDENTIFICATION OF RELEVANT SAMPLES*

For this analysis, samples treated with CDK4/6 inhibitors palbociclib, abemaciclib and ribociclib and their corresponding placebo samples were retrieved from the ARCHS4 database (search terms can be found in Table S 1). Also, I searched for palbociclib resistant samples and included them in the study. The resulting samples were manually curated to ensure that only relevant samples were kept for further analysis. Some experimental series were also subdivided into subseries. Additionally, each sample was annotated according to which drug they were treated with and whether they were resistant or sensitive to it. Lastly, the RNA-Seq data for those samples was processed and filtered as described for the benchmarking data set above.

5.3.2 *PATHWAY SCORES ANALYSIS*

The pathway scores for each pathway were calculated using the singscore method for hypoxia, metastasis, p53, notch and cell cycle. The score for pi3k was calculated using the z-score method. Then, the pathway scores were compared between samples that had been treated with

different cell cycle inhibitors and placebo samples using a Tukey box plot. Statistical significance to confirm differential pathway signature was calculated using a student's t test using the "stats" package (v 3.6.2) in R. First, cell cycle dysregulation was analyzed by counting how many series showed significant cell cycle dysregulation in resistant and sensitive samples. The result was then visualized using a Tukey box plot. Finally, pathway regulation was compared between sensitive and resistant samples by counting how many series showed differential pathway regulation. This was done for all pathways. A Fisher's exact test calculated using the "stats" package (v 3.6.2) in order to confirm statistical significance.

5.4 TCGA PATHWAY ANALYSIS

5.4.1 DATA ACQUISITION

As the circumstances did not allow me to compute pathway scores myself, pathway scores for all pathways for roughly one third of the TCGA samples (N=3753) were kindly provided by Natalie Davidson and can be found at https://github.com/michellemeier27/Semesterproject/tree/master/TCGA. The scores were calculated using the singscore method. The corresponding meta data was then obtained from the National Cancer Institute which is publicly available at https://portal.gdc.cancer.gov/repository. The meta data includes but is not limited to age at diagnosis, gender, ethnicity, race, treatment, pathological stage, primary diagnosis and TCGA project ID.

5.4.2 PATHWAY ANALYSIS OVER ALL CANCER TYPES

The samples were then divided into early and late pathological stages as defined by the American Joint Committee of Cancer (AJCC). The stages I, IA, IB, II, IIA, IIB, IIC and IIS were denoted as early and III, IIIA, IIIB, IIIC, IV, IVA, IVB and IVC were denoted as late stages. Also, the following analysis was performed on individual cancer types. The cancer types were defined by their TCGA project ID, each of which is dedicated to a specific cancer.

5.4.3 GENDER BIAS IN MOLECULAR SIGNATURES IN TCGA PATIENTS

Samples corresponding to KIRP and glioblastoma were searched using the TCGA project IDs KIRP and GBM, respectively. While only early stage KIRP samples were analyzed, all glioblastoma samples were considered due to the lack of pathological stage information. For both cancers, linear regression analysis was performed separately for each pathway score using the "stats" package (v 3.6.2) in R. The starting point of each model was the full model containing all possible variables available for that set of samples and only removed if necessary. Reasons for variable exclusion were insufficient factor levels or too many missing data entries. If the adjusted R^2 was sufficiently high, the full model was used for further analysis. Full models with low adjusted R^2 were adapted by removing variables to maximize the adjusted R^2. Additionally, the fit of the model was evaluated using the residuals and QQ-plot and potential outliers identified using the residuals/leverage plot generated by the linear regression function in R. Outliers were only removed if the adjusted R^2 was close to zero. This was true for the pi3k and notch models in glioblastoma. However, as removal of the outliers did not affect the model parameters, they were kept for further analysis. For a more detailed overview of the formulas used in this linear regression analysis, see Table S 4. The results were then visualized in a forest plot using the CRAN package "jtools" in R[52] for the factors of interest.

5.4.4 Linear Regression Breast cancer, Thymoma and Hepatocellular carcinoma

Samples corresponding to breast cancer, thymoma and hepatocellular carcinoma were searched using the TCGA project IDs BRCA, THYM and LIHC, respectively. For breast cancer and hepatocellular carcinoma samples, early and late stage samples were analyzed separately. As there was no pathological stage information available for the thymoma samples, all samples were analyzed. Linear regression analysis was performed as described above.

5.5 Data availability

The codes used to perform all analyses in this study are available at https://github.com/michellemeier27/Semesterproject, along with all plots in this report plus the linear regression analysis forest plots for late stage breast cancer and hepatocellular carcinoma not shown in this report. Also, the weight estimates for all linear regression models and the singscore pathway scores for the TCGA samples received are also available online.

6 REFERENCES

[1] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, 2016, doi: 10.1038/nrg.2016.49.

[2] M. P. Segura-Lepe, H. C. Keun, and T. M. D. Ebbels, "Predictive modelling using pathway scores: Robustness and significance of pathway collections," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–11, 2019, doi: 10.1186/s12859-019-3163-0.

[3] T. Barrett *et al.*, "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, Nov. 2012, doi: 10.1093/nar/gks1193.

[4] F. Sanchez-Vega *et al.*, "Oncogenic Signaling Pathways in The Cancer Genome Atlas," *Cell*, vol. 173, no. 2, pp. 321-337.e10, 2018, doi: 10.1016/j.cell.2018.03.035.

[5] S. M. MacNeil, W. E. Johnson, D. Y. Li, S. R. Piccolo, and A. H. Bild, "Inferring pathway dysregulation in cancers from multiple types of omic data," *Genome Med.*, vol. 7, no. 1, p. 61, Jun. 2015, doi: 10.1186/s13073-015-0189-4.

[6] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, no. 1, p. 126, 2012, doi: 10.1186/1471-2105-13-126.

[7] R.-F. Sun, Q.-Q. Yu, and K. H. Young, "Critically dysregulated signaling pathways and clinical utility of the pathway biomarkers in lymphoid malignancies," *Chronic Dis. Transl. Med.*, vol. 4, no. 1, pp. 29–44, 2018, doi: https://doi.org/10.1016/j.cdtm.2018.02.001.

[8] D. Hanahan and R. A. Weinberg, "The Hallmarks of Cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000, doi: 10.1016/S0092-8674(00)81683-9.

[9] Z. A. Stewart, M. D. Westfall, and J. A. Pietenpol, "Cell-cycle dysregulation and anticancer therapy," *Trends Pharmacol. Sci.*, vol. 24, no. 3, pp. 139–145, Mar. 2003, doi: 10.1016/S0165-6147(03)00026-9.

[10] N. C. Turner, P. Neven, S. Loibl, and F. Andre, "Advances in the treatment of advanced oestrogen-receptor-positive breast cancer," *Lancet*, vol. 389, no. 10087, pp. 2403–2414, Jun. 2017, doi: 10.1016/S0140-6736(16)32419-9.

[11] B. O'Leary, R. S. Finn, and N. C. Turner, "Treating cancer with selective CDK4/6 inhibitors," *Nat. Rev. Clin. Oncol.*, vol. 13, no. 7, pp. 417–430, 2016, doi: 10.1038/nrclinonc.2016.26.

[12] M. Shah, M. R. Nunes, and V. Stearns, "CDK4/6 Inhibitors: Game Changers in the Management of Hormone Receptor–Positive Advanced Breast Cancer?," *Oncology (Williston Park).*, vol. 32, no. 5, pp. 216–222, May 2018.

[13] A. A. Mohammed, H. Rashied, and F. M. Elsayed, "CDK4/6 inhibitors in advanced breast cancer, what is beyond?," *Oncol. Rev.*, vol. 13, no. 2, p. 416, Jul. 2019, doi: 10.4081/oncol.2019.416.

[14] B. O'leary *et al.*, "The genetic landscape and clonal evolution of breast cancer resistance to palbociclib plus fulvestrant in the PALOMA-3 trial," *Cancer Discov.*, vol. 8, no. 11, pp. 1390–1403, 2018, doi: 10.1158/2159-8290.CD-18-0264.

[15] C. Guarducci *et al.*, "Mechanisms of Resistance to CDK4/6 Inhibitors in Breast Cancer and Potential Biomarkers of Response," *Breast Care*, vol. 12, no. 5, pp. 304–308, 2017, doi: 10.1159/000484167.

[16] A. McCartney *et al.*, "Mechanisms of Resistance to CDK4/6 Inhibitors: Potential Implications and Biomarkers for Clinical Practice," *Front. Oncol.*, vol. 9, no. July, pp.

2–9, 2019, doi: 10.3389/fonc.2019.00666.
[17] B. Muz, P. de la Puente, F. Azab, and A. K. Azab, "The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy," *Hypoxia (Auckland, N.Z.)*, vol. 3, pp. 83–92, Dec. 2015, doi: 10.2147/HP.S93413.
[18] T. N. Seyfried and L. C. Huysentruyt, "On the origin of cancer metastasis," *Crit. Rev. Oncog.*, vol. 18, no. 1–2, pp. 43–73, 2013, doi: 10.1615/critrevoncog.v18.i1-2.40.
[19] S. Hänzelmann, R. Castelo, and J. Guinney, "GSVA: Gene set variation analysis for microarray and RNA-Seq data," *BMC Bioinformatics*, vol. 14, pp. 1–21, 2013, doi: 10.1186/1471-2105-14-7.
[20] D. A. Barbie *et al.*, "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1," *Nature*, vol. 462, no. 7269, pp. 108–112, 2009, doi: 10.1038/nature08460.
[21] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring Pathway Activity toward Precise Disease Classification," *PLOS Comput. Biol.*, vol. 4, no. 11, p. e1000217, Nov. 2008.
[22] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, no. 1, p. 225, 2005, doi: 10.1186/1471-2105-6-225.
[23] M. Foroutan, D. D. Bhuva, R. Lyu, K. Horan, J. Cursons, and M. J. Davis, "Single sample scoring of molecular phenotypes," *BMC Bioinformatics*, vol. 19, no. 1, p. 404, 2018, doi: 10.1186/s12859-018-2435-4.
[24] X. Guan, "Cancer metastases: challenges and opportunities," *Acta Pharm. Sin. B*, vol. 5, no. 5, pp. 402–418, 2015, doi: https://doi.org/10.1016/j.apsb.2015.07.005.
[25] K. A. H. Nwabo *et al.*, "Developmental pathways associated with cancer metastasis: Notch, Wnt, and Hedgehog," *Cancer Biol. Med.*, vol. 14, no. 2, p. 109, 2017, doi: 10.20892/j.issn.2095-3941.2016.0032.
[26] A. L. Tarca, G. Bhatti, and R. Romero, "A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity," *PLoS One*, vol. 8, no. 11, p. e79217, Nov. 2013.
[27] C. Yu, H. J. Woo, X. Yu, T. Oyama, A. Wallqvist, and J. Reifman, "A strategy for evaluating pathway analysis methods," *BMC Bioinformatics*, vol. 18, no. 1, p. 453, Oct. 2017, doi: 10.1186/s12859-017-1866-7.
[28] N. S. Zainal *et al.*, "Effects of palbociclib on oral squamous cell carcinoma and the role of PIK3CA in conferring resistance," *Cancer Biol. Med.*, vol. 16, no. 2, pp. 264–275, May 2019, doi: 10.20892/j.issn.2095-3941.2018.0257.
[29] N. M. Kettner *et al.*, "Combined Inhibition of STAT3 and DNA Repair in Palbociclib-Resistant ER-Positive Breast Cancer," *Clin. Cancer Res.*, vol. 25, no. 13, pp. 3996 LP – 4013, Jul. 2019, doi: 10.1158/1078-0432.CCR-18-3274.
[30] Y. Yuan *et al.*, "Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients," *Cancer Cell*, vol. 29, no. 5, pp. 711–722, 2016, doi: 10.1016/j.ccell.2016.04.001.
[31] T. Liu *et al.*, "Gene polymorphisms in the PI3K/AKT/mTOR signaling pathway contribute to prostate cancer susceptibility in Chinese men," *Oncotarget*, vol. 8, no. 37, pp. 61305–61317, May 2017, doi: 10.18632/oncotarget.18064.
[32] K. A. Dookeran, J. J. Dignam, K. Ferrer, M. Sekosan, W. McCaskill-Stevens, and S. Gehlert, "p53 as a Marker of Prognosis in African-American Women with Breast Cancer," *Ann. Surg. Oncol.*, vol. 17, no. 5, pp. 1398–1405, 2010, doi: 10.1245/s10434-009-0889-3.
[33] K. A. Dookeran *et al.*, "Race and the prognostic influence of p53 in women with breast cancer," *Ann. Surg. Oncol.*, vol. 19, no. 7, pp. 2334–2344, Jul. 2012, doi:

10.1245/s10434-011-1934-6.
[34] A. Purushotham *et al.*, “Age at diagnosis and distant metastasis in breast cancer- A surprising inverse relationship,” *Eur. J. Cancer*, vol. 50, no. 10, pp. 1697–1705, Jul. 2014, doi: 10.1016/j.ejca.2014.04.002.
[35] F. M. Buffa, A. L. Harris, C. M. West, and C. J. Miller, “Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene,” *Br. J. Cancer*, vol. 102, no. 2, pp. 428–435, Jan. 2010, doi: 10.1038/sj.bjc.6605450.
[36] G. Ramakrishna, A. Rastogi, N. Trehanpati, B. Sen, R. Khosla, and S. K. Sarin, “From cirrhosis to hepatocellular carcinoma: new molecular insights on inflammation and cellular senescence,” *Liver cancer*, vol. 2, no. 3–4, pp. 367–383, Aug. 2013, doi: 10.1159/000343852.
[37] G. Fattovich, T. Stroffolini, I. Zagni, and F. Donato, “Hepatocellular carcinoma in cirrhosis: Incidence and risk factors,” *Gastroenterology*, vol. 127, no. 5, pp. S35–S50, Nov. 2004, doi: 10.1053/j.gastro.2004.09.014.
[38] R. Wong and D. A. Corley, “Racial and Ethnic Variations in Hepatocellular Carcinoma Incidence within the United States,” *Am. J. Med.*, vol. 121, no. 6, pp. 525–531, Jun. 2008, doi: 10.1016/j.amjmed.2008.03.005.
[39] N. E. Rich *et al.*, “Racial and Ethnic Differences in Presentation and Outcomes of Hepatocellular Carcinoma,” *Clin. Gastroenterol. Hepatol.*, vol. 17, no. 3, pp. 551-559.e1, Feb. 2019, doi: 10.1016/j.cgh.2018.05.039.
[40] R. J. Shaw and L. C. Cantley, “Ras, PI(3)K and mTOR signalling controls tumour cell growth,” *Nature*, vol. 441, no. 7092, pp. 424–430, 2006, doi: 10.1038/nature04869.
[41] A. Lachmann *et al.*, “Massive mining of publicly available RNA-seq data from human and mouse,” *Nat. Commun.*, vol. 9, no. 1, p. 1366, 2018, doi: 10.1038/s41467-018-03751-6.
[42] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nat. Biotechnol.*, vol. 34, no. 5, pp. 525–527, 2016, doi: 10.1038/nbt.3519.
[43] B. F. and G. P. and M. Smith, “R Interface to HDF5.” 2019.
[44] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, vol. 5, no. 7, pp. 621–628, 2008, doi: 10.1038/nmeth.1226.
[45] G. P. Wagner, K. Kin, and V. J. Lynch, “Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples,” *Theory Biosci.*, vol. 131, no. 4, pp. 281–285, 2012, doi: 10.1007/s12064-012-0162-3.
[46] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.
[47] X. Li, N. G. F. Cooper, T. E. O’Toole, and E. C. Rouchka, “Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies,” *BMC Genomics*, vol. 21, no. 1, p. 75, 2020, doi: 10.1186/s12864-020-6502-7.
[48] J. H. Krijthe, “Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation.” 2015.
[49] M. Ringnér, “What is principal component analysis?,” *Nat. Biotechnol.*, vol. 26, no. 3, pp. 303–304, 2008, doi: 10.1038/nbt0308-303.
[50] A. K. and F. Mundt, “factoextra: Extract and Visualize the Results of Multivariate Data Analyses.” 2020.
[51] V. Q. Vu, “ggbiplot: A ggplot2 based biplot.” 2011.
[52] J. A. Long, “jtools: Analysis and Presentation of Social Scientific Data.” 2019.

[53] G. L. Semenza, “Hypoxia-inducible factor 1: master regulator of O2 homeostasis,” *Curr. Opin. Genet. Dev.*, vol. 8, no. 5, pp. 588–594, 1998, doi: https://doi.org/10.1016/S0959-437X(98)80016-6.
[54] L. Gossage, T. Eisen, and E. R. Maher, “VHL, the story of a tumour suppressor gene,” *Nat. Rev. Cancer*, vol. 15, no. 1, pp. 55–64, 2015, doi: 10.1038/nrc3844.
[55] Q. Li and G. Lozano, “Molecular pathways: targeting Mdm2 and Mdm4 in cancer therapy,” *Clin. Cancer Res.*, vol. 19, no. 1, pp. 34–41, Jan. 2013, doi: 10.1158/1078-0432.CCR-12-0053.
[56] S. Nag, J. Qin, K. S. Srivenugopal, M. Wang, and R. Zhang, “The MDM2-p53 pathway revisited,” *J. Biomed. Res.*, vol. 27, no. 4, pp. 254–271, Jul. 2013, doi: 10.7555/JBR.27.20130030.
[57] H. Furukawa *et al.*, “PRIMA-1 induces p53-mediated apoptosis by upregulating Noxa in esophageal squamous cell carcinoma with TP53 missense mutation,” *Cancer Sci.*, vol. 109, no. 2, pp. 412–421, Feb. 2018, doi: 10.1111/cas.13454.
[58] V. J. N. Bykov *et al.*, “Reactivation of mutant p53 and induction of apoptosis in human tumor cells by maleimide analogs.,” *J. Biol. Chem.*, vol. 280, no. 34, pp. 30384–91, Aug. 2005, doi: 10.1074/jbc.M501664200.
[59] J. Friedman *et al.*, “Inhibition of WEE1 kinase and cell cycle checkpoint activation sensitizes head and neck cancers to natural killer cell therapies,” *J. Immunother. cancer*, vol. 6, no. 1, p. 59, Jun. 2018, doi: 10.1186/s40425-018-0374-2.
[60] A. K. Brenner, H. Reikvam, A. Lavecchia, and Ø. Bruserud, *Therapeutic targeting the cell division cycle 25 (CDC25) phosphatases in human acute myeloid Leukemia - The possibility to target several kinases through inhibition of the various CDC25 isoforms*, vol. 19, no. 11. 2014.
[61] Y. Lu *et al.*, “Upregulated cyclins may be novel genes for triple-negative breast cancer based on bioinformatic analysis,” *Breast Cancer*, 2020, doi: 10.1007/s12282-020-01086-z.
[62] S. Aref, M. El Agdar, O. Salama, T. A. Zeid, and M. Sabry, “Significance of NOTCH1 mutations détections in T-acute lymphoblastic leukemia patients,” *Cancer Biomarkers*, vol. 27, pp. 157–162, 2020, doi: 10.3233/CBM-190967.
[63] J. R. Göthert, R. L. Brake, M. Smeets, U. Dührsen, C. G. Begley, and D. J. Izon, “NOTCH1 pathway activation is an early hallmark of SCL T leukemogenesis,” *Blood*, vol. 110, no. 10, pp. 3753–3762, Nov. 2007, doi: 10.1182/blood-2006-12-063644.
[64] J. Reichrath and S. Reichrath, “A Snapshot of the Molecular Biology of Notch Signaling: Challenges and Promises BT - Notch Signaling in Embryology and Cancer: Molecular Biology of Notch Signaling,” J. Reichrath and S. Reichrath, Eds. Cham: Springer International Publishing, 2020, pp. 1–7.
[65] K. S. Albain *et al.*, “S1-5: Modulation of Cancer and Stem Cell Biomarkers by the Notch Inhibitor MK-0752 Added to Endocrine Therapy for Early Stage ER+ Breast Cancer.,” *Cancer Res.*, vol. 71, no. 24 Supplement, pp. S1-5 LP-S1-5, Dec. 2011, doi: 10.1158/0008-5472.SABCS11-S1-5.
[66] I. Espinoza and L. Miele, “Notch inhibitors for cancer treatment,” *Pharmacol. Ther.*, vol. 139, no. 2, pp. 95–110, Aug. 2013, doi: 10.1016/j.pharmthera.2013.02.003.
[67] T. Trimarchi *et al.*, “Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia,” *Cell*, vol. 158, no. 3, pp. 593–606, Jul. 2014, doi: 10.1016/j.cell.2014.05.049.
[68] F. Huang, J. Chen, Z. Wang, R. Lan, L. Fu, and L. Zhang, “δ-Catenin promotes tumorigenesis and metastasis of lung adenocarcinoma,” *Oncol. Rep.*, vol. 39, no. 2, pp. 809–817, 2018, doi: 10.3892/or.2017.6140.
[69] Y.-T. Tee, G.-D. Chen, L.-Y. Lin, J.-L. Ko, and P.-H. Wang, “NM23-H1: a Metastasis-

Associated Gene," *Taiwan. J. Obstet. Gynecol.*, vol. 45, no. 2, pp. 107–113, 2006, doi: https://doi.org/10.1016/S1028-4559(09)60206-0.

[70] R. Qiu *et al.*, "BRMS1 coordinates with LSD1 and suppresses breast cancer cell metastasis," *Am. J. Cancer Res.*, vol. 8, no. 10, pp. 2030–2045, Oct. 2018.

[71] T. D. Zeng, B. Zheng, W. Zheng, and C. Chen, "CD82/KAI1 inhibits invasion and metastasis of esophageal squamous cell carcinoma via TGF-β1," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 22, no. 18, pp. 5928–5937, 2018, doi: 10.26355/eurrev_201809_15922.

[72] S. Zhou, X. Tang, and F. Tang, "Krüppel-like factor 17, a novel tumor suppressor: its low expression is involved in cancer metastasis," *Tumour Biol.*, vol. 37, no. 2, pp. 1505–1513, Feb. 2016, doi: 10.1007/s13277-015-4588-3.

[73] Q. Li *et al.*, "Gas1 Inhibits Metastatic and Metabolic Phenotypes in Colorectal Carcinoma," *Mol. Cancer Res.*, vol. 14, no. 9, pp. 830 LP – 840, Sep. 2016, doi: 10.1158/1541-7786.MCR-16-0032.

[74] S. Ozturk *et al.*, "SDPR functions as a metastasis suppressor in breast cancer by promoting apoptosis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 3, pp. 638–643, Jan. 2016, doi: 10.1073/pnas.1514663113.

[75] C. Corno and P. Perego, "KiSS1 in regulation of metastasis and response to antitumor drugs," *Drug Resist. Updat.*, vol. 42, pp. 12–21, 2019, doi: https://doi.org/10.1016/j.drup.2019.02.001.

[76] M. Canel, A. Serrels, M. C. Frame, and V. G. Brunton, "E-cadherin–integrin crosstalk in cancer invasion and metastasis," *J. Cell Sci.*, vol. 126, no. 2, pp. 393 LP – 401, Jan. 2013, doi: 10.1242/jcs.100115.

[77] Y. Wang *et al.*, "PI3K inhibitor LY294002, as opposed to wortmannin, enhances AKT phosphorylation in gemcitabine-resistant pancreatic cancer cells," *Int. J. Oncol.*, vol. 50, no. 2, pp. 606–612, 2017, doi: 10.3892/ijo.2016.3804.

[78] G. Petroni, S. C. Formenti, S. Chen-Kiang, and L. Galluzzi, "Immunomodulation by anticancer cell cycle inhibitors," *Nat. Rev. Immunol.*, pp. 1–11, 2020, doi: 10.1038/s41577-020-0300-y.

7 SUPPLEMENTARY INFORMATION

7.1 ABBREVIATIONS

HIF1: Hypoxia-inducible factor 1;VHL: Von Hippel Lindau suppressor gene; TP5: tumor protein 53; PRIMA-1: p53 re-activation and induction of massive apoptosis; MIRA-1: maleimide-derived molecule; CDK: cyclin dependent kinase; CDKN: cyclin dependent kinase inhibitor; RB1: retinoblastoma protein 1; E2F: E2 factor; p16: tumor protein 16; CCN: cyclin; CDC: cell divison cycle; CREB: cAMP response element binding protein; EP300: E1A binding protein ; JAG: jagged ; CNTN6: contactin 6; DKM5a: Lysine-specific demethylase 5A; FBXW7: F-box and WD repeat domain containing 7; T-ALL: T-cell acute lymphoblastic leukaemia; DLL: delta-like-ligand; LUNAR: Leukemia-induced non-coding activator RNA; EMT:epithelial mesenchymal transition ; BRMS1: Breast cancer metastasis-suppressor 1; KAI1: Kangai 1; KLF17: Krüppel-like-factor 17; GAS-1: Growth arrest–specific 1; SDPR: serum deprivation response; PTEN: Phosphatase and tensin homolog; PIK3: Phosphoinositide 3-kinase; INPP4b: Inositol polyphosphate-4-phosphatase; TSC: tuberous sclerosis complex; STK11: Serine/Threonine kinase 11; RHEB: Ras homolog enriched in brain; PPP2R1A: Protein Phosphatase 2 Scaffold Subunit Aalpha; RICTOR: Rapamycin-insensitive companion of mammalian target of rapamycin; mTOR: mammalian target of rapamycin; AKT: protein kinase B

7.2 METHODS

Table S 1:Search terms used for each pathway to query the data downloaded from the ARCHS4 website.

Pathway	Search Terms
hypoxia	hypoxia OR hif1[53] OR hypoxic OR vhl [54]
p53	p53 OR tp53 OR mdm2[55], [56] OR mdm4 [55] OR prima-1[57] OR mira-1[58]
cell cycle	cell cycle OR cdkn[4] OR ccnd[4] OR ccne1 [4] OR cdk2[4] OR cdk4[4] OR cdk6[4] OR rb1[4] OR OR e2f[4] OR p16[4] OR wee1[59] OR cdc20 OR cdc25[60] OR ccna[61]
notch	notch OR creb[4] OR ep300[4] OR jag[4] OR cntn6[4] OR kdm5a[4] OR fbxw7[4] OR t-all[62][63] OR dll[64] OR mk0752[65] [66] OR ro4929097[66] OR mrk-003[66] OR mk-0752[66] OR pf03084014[66] OR lunar[67]
metastasis (EMT)	metastasis OR catenin[68] OR nm23[69] OR brms1[70] OR kai1[71] OR klf17[72] OR gas1[73] OR sdpr[74] OR kiss1[75] OR neoangiogenesis[25] OR emt OR e-cadherin[76] OR metastases
pi3k	pi3k OR pten[4] OR pik3[4] OR inpp4b[4] OR tsc[4] OR stk11[4] OR rheb[4] OR ppp2r1a[4] OR rictor[4] OR mtor[4] OR wortmannin[77] OR ly294002[77] OR akt[4]

cell cycle inhibitors	palbociclib[78] OR abemaciclib[78]OR ribociclib[78] OR cdk4[78] OR cdk6[78]

Table S 2: Summary of sample and series size, gene size, normalisation strategy and filtering cutoff for each pathway.

pathway	min.counts	genes after filtering	Library normalisation	# sample	# series
hypoxia	8	11634	UQ	295	42
p53	6	11566	UQ	216	33
cell cycle	7	11038	UQ	242	31
notch	7	11665	UQ	203	31
metastasis (EMT)	10	11133	UQ	628	25
pi3k	7	12617	UQ	245	34

Table S 3: Summary of all gene sets used in ssGSEA analysis for each pathway. Up to three pathways were used for the biplot analysis and the most accurate pathway was then used to compare enrichment scores between case and control samples using a Tukey boxplot. Gene set names are equivalent to the names listed on the MSigDB website.

pathway	gene sets for biplot	gene set for boxplot (true pathway)
hypoxia	HALLMARK_HYPOXIA, HALLMARK_TFG_BETA_SIGNALLING	HALLMARK_HYPOXIA
p53	HALLMARK_P53_PATHWAY, HALLMARK_DNA_REPAIR, HALLMARK_APOPTOSIS	HALLMARK_P53_PATHWAY
cell cycle	KEGG_CELL_CYCLE, HALLMARK_G2M_CHECKPOINT, HALLMARK_MITOTIC SPINDLE	KEGG_CELL_CYCLE
notch	HALLMARK_NOTCH_SIGNALING, HALLMARK_APOPTOSIS, GO_CELL_GROWTH	HALLMARK_NOTCH_SIGNALING
metastasis (EMT)	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION, KEGG_FOCAL_ADHESION, KEGG_ADHERENS JUNCTIONS	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
pi3k	HALLMARK_PI3K_AKT_MTOR_SIGNALING, GO_CELL_GROWTH, KEGG_MTOR_SIGNALING_PATHWAY	HALLMARK_PI3K_AKT_MTOR_SIGNALING

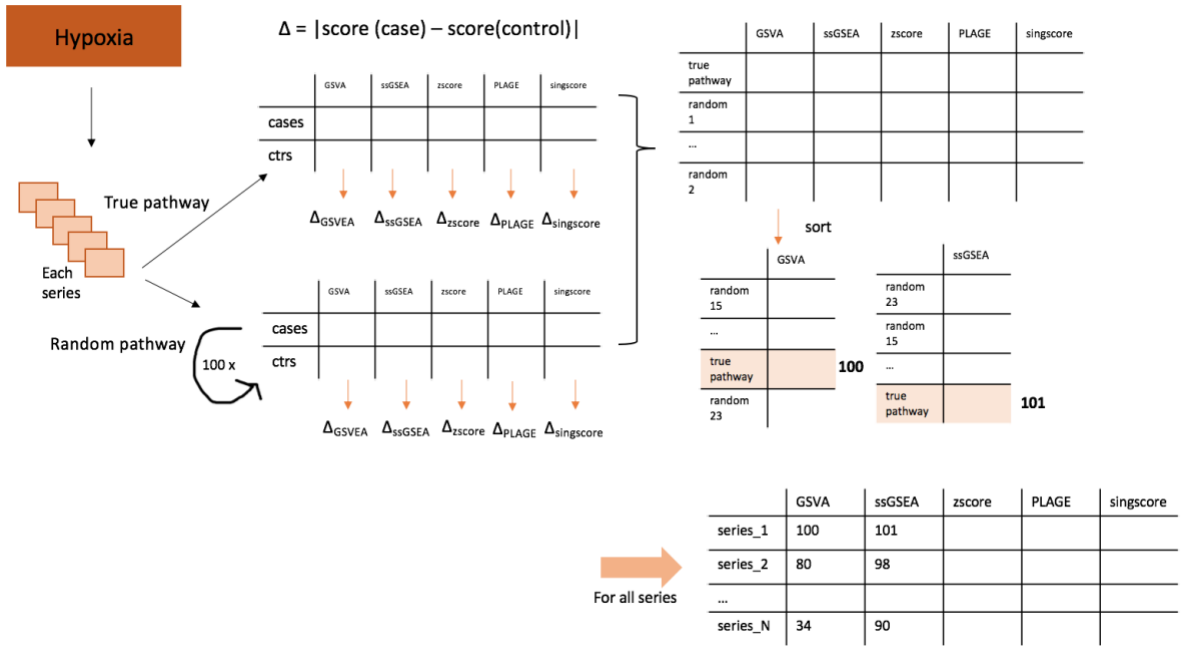


Figure S 1: Workflow pathway scoring evaluation for identification of true pathway out of 100 random gene sets. For each series within a specific pathway, the pathway scores for the true pathway was calculated using all pathway scoring methods. Then, the score difference between cases and controls were calculated. Similarly, this was repeated for 100 random pathways and the resulting differences inserted to a dataframe over all methods and pathways. The data was then sorted for each pathway separately in an ascending order, leaving the pathway with the largest pathway score difference at the bottom. The rank of the true pathway was then noted for each series and method and the distribution of the ranks plotted.

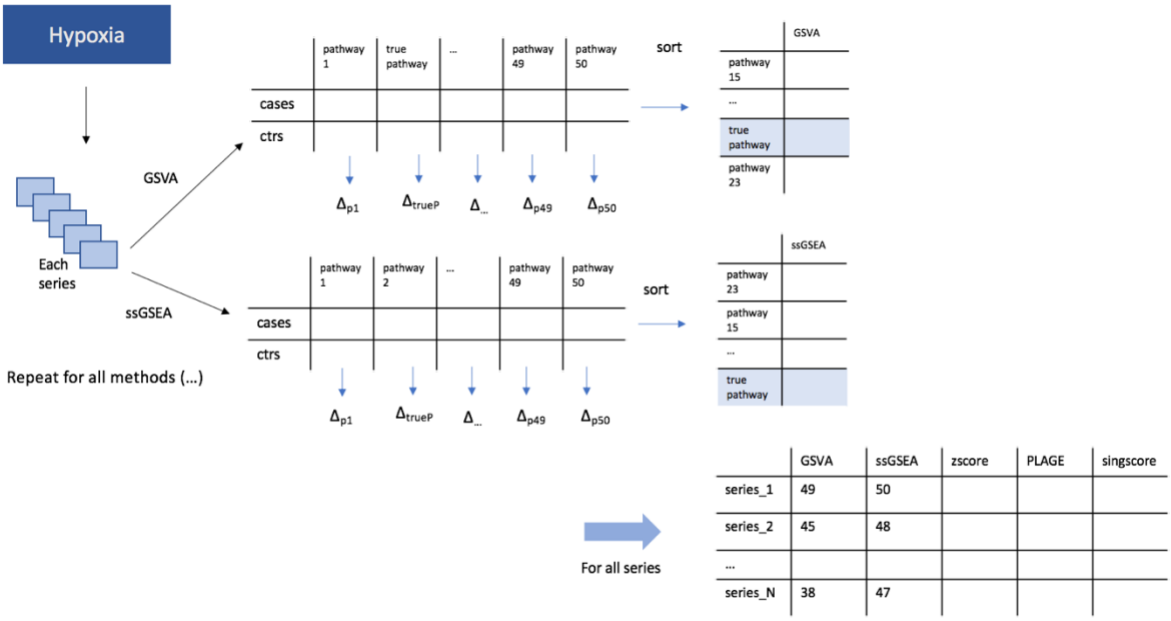


Figure S 2: Workflow pathway scoring evaluation for identification of true pathway out of 50 hallmark gene sets. For each series, the pathway scores for all 50 pathways were computed using one method and the score difference between case and control calculated. These differences were then sorted ascendingly leaving the pathway with the largest score difference at the bottom. The rank of the true pathway was then noted for each series. This was repeated for all methods and the distribution of true ranks plotted.

Table S 4: Overview of formulas and functions used for the different studies

Study	Function in R	Regression formula
Gender based pathway differences: KIRP	lm	pi3k ~ age_at_index + days_to_birth + gender + race + treatment_type + hypoxia + notch + p53 + cell_cycle + metastasis p53 ~ age_at_index + gender + race + hypoxia + notch + pi3k + cell_cycle + metastasis + ethnicity + treatment_or_therapy metastasis ~ age_at_index + gender + race + hypoxia + notch + p53 + pi3k + cell_cycle + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy notch ~ age_at_index + days_to_birth + gender + race + treatment_type + hypoxia + p53 + cell_cycle + metastasis hypoxia ~ age_at_index + gender + race + notch + p53 + pi3k + cell_cycle + metastasis + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy cell cycle ~ age_at_index + days_to_birth + gender + race + treatment_type + hypoxia + notch + p53 + metastasis
Gender based pathway differences: Glioblastoma	lm	Pathway score ~ age_at_index + gender + race + hypoxia + notch + p53 + pi3k + cell_cycle + metastasis + treatment_or_therapy
Hepatocellular carcinoma analysis: early stage	lm	pathway score ~ age_at_index + gender + race + hypoxia + notch + p53 + pi3k + cell_cycle + metastasis + ethnicity + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy
Hepatocellular carcinoma analysis: late stage	lm	pi3k ~ age_at_index + gender + race + hypoxia + notch + p53 + cell_cycle + metastasis + ethnicity + ajcc_pathologic_n+ prior_malignancy

		p53 ~ age_at_index + gender + race + hypoxia + notch + pi3k + cell_cycle + metastasis + ethnicity + ajcc_pathologic_n+ ajcc_pathologic_m metastasis/notch/hypoxia ~ age_at_indexgender + race + hypoxia + notch + p53 + pi3k + cell_cycle + metastasis + ethnicity + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy cell cycle ~ age_at_index + gender + race + hypoxia + notch + p53 + cell_cycle + metastasis + ethnicity + ajcc_pathologic_n+ prior_malignancy
Breast cancer analysis: early stages	lm	pi3k ~ age_at_diagnosis + gender + race + hypoxia + notch + p53 + cell_cycle + metastasis p53 ~ age_at_index + gender + race + hypoxia + notch + pi3k + cell_cycle + metastasis + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy metastasis ~ age_at_index + gender + race + hypoxia + notch + pi3k + cell_cycle + p53 + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy notch ~ age_at_index + gender + race + hypoxia + p53 + pi3k + cell_cycle + metastasis + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy hypoxia ~ age_at_index + gender + race + p53 + notch + pi3k + cell_cycle + metastasis + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy cell cycle ~ age_at_index + gender + race + hypoxia + notch + pi3k + p53 + metastasis + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy

Breast cancer analysis: late stages		pi3k ~ age_at_index + days_to_birth + race + treatment_type + hypoxia + notch + p53 + cell_cycle + metastasis p53 ~ age_at_index + days_to_birth + race + treatment_type + hypoxia + notch + pi3k + cell_cycle + metastasis metastasis ~ age_at_index + race + hypoxia + notch + p53 + pi3k + cell_cycle + ethnicity + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy notch ~ age_at_index + race + hypoxia + p53 + pi3k + cell_cycle + metastasis + ethnicity + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy hypoxia ~ age_at_index + race + notch + p53 + pi3k + cell_cycle + metastasis + ethnicity + treatment_or_therapy + ajcc_pathologic_m + ajcc_pathologic_n + ajcc_pathologic_stages + prior_malignancy cell cycle ~ age_at_index + days_to_birth + race + treatment_type + hypoxia + notch + p53 + pi3k + metastasis
Thymoma analysis	lm	pathway score ~ age_at_index + gender + race + hypoxia + notch + p53 + pi3k + cell_cycle + metastasis + ethnicity + treatment_or_therapy

7.3 RESULTS

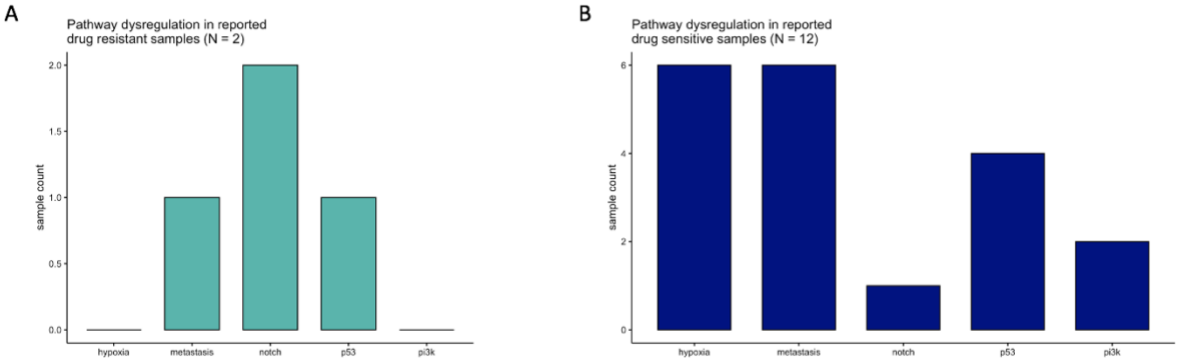


Figure S 3: Pathway analysis of palbociclib resistant and sensitive cells. A: Pathway dysregulation in drug resistant samples (N=2) was summarized in a barplot. All resistant samples showed differential regulation in notch, one sample in metastasis and p53, respectively. No differential regulation between resistant treated and untreated samples was observed for hypoxia and pi3k. B: Pathway dysregulation in drug sensitive samples (N=12) was summarized in a barplot. Half of the sensitive samples showed differential regulation in hypoxia and metastasis, four samples in p53 and two samples in pi3k. Only one sensitive sample showed differential regulation between treated and untreated samples for notch.