nature portfolio

Peer Review File

Unbiased discovery of cancer pathways and therapeutics using Pathway Ensemble Tool and Benchmark

Editorial Note: Parts on page 21 and 29 of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

Reviewers' comments:

Reviewer #1 (Remarks to the Author): Expert in computational cancer genomics, drug response prediction, and benchmarking of bioinformatics methods

Wang et al have generated a benchmark pipeline to evaluate pathway enrichment methods and introduce their own method, called PET, based on the combination of ranks of other methods. They characterized different cancer types leveraging pathway enrichment analysis. Then they identify combinations of genes inside differential enriched pathways as potential biomarkers. Then, by using drug-target information, they rank drugs based on their effect on the found biomarkers. Finally, they experimentally validate one of these drugs in vitro and in vivo.

While this is an active and meaningful topic of research, the manuscripts falls short in novelty both in terms of the approach and the findings, as outlined below:

Major points:
1. While the benchmark data and pipeline called “Benchmark” will be useful to evaluate methods, the method PET is not novel. Other works have already explored the aggregation of multiple methods such as PIANO (Väremo L. et al. 2013), Enrichment Browser (Geistlinger L. et al. 2016), EGSEA (Alhamdoosh M. et al. 2017) available in R, and decoupleR (Badia-i-Mompel P. et al. 2022) available in R and Python. Additionally, these frameworks offer multiple methods while PET currently only contains 3. In all these works, it was already shown that aggregation strategies work better than single methods.
2. Regarding the benchmark dataset the authors should show what is the overlap between gene sets across cell lines because this will affect the assumption that a gene set from the same perturbation across cell lines should be similar. Additionally, since the extraction of gene sets from perturbation experiments, eCLIP and ChIP-seq data is different, they should show the mentioned overlap and method performances in each of these three data types independently.
3. The prediction that CCR068127 has effect on cancer growth in bladder cancer is of limited novelty, as it already has reported effectiveness against human colon cancer and melanoma cell lines (ref 56).
4. Why does PET use GSEA as one of its methods for the aggregation if it has been consistently shown in the benchmark that it is not a good performing method? Moreover, for the same reason, why is it used as a standalone method for some of the validations?
5. Methods section is generally sparse and disorganized. For example, when calculating the false positive rate, which DEA test was used between cancer and control samples? Which thresholds were used to consider significant changes? How many permutations were done when shuffling samples?
6. Authors mention that pathways have not been previously used as prognostic markers, which is not true. There are many works showing the importance of characterizing cancers at the pathway level for better prognosis. Too many here to cite; just a google search offers many such studies. Additionally, authors also mention that collagen genes have not been used as biomarkers for renal cancer which is alsos not true, there are many examples available (Best, S.L., et al 2019).
7. When comparing combinations of top pathway genes against DEGs, it's not clear if the authors used the top 100 DEGs or the 10 most upregulated and 10 downregulated genes. In any case, they should use the correct signed top hits to not mix signals and make a fair comparison. Additionally, it's true that it does not reach significance, but the compound 0175029 is still the top ranked drug for BLCA when using

DEGs, maybe this also changes and should be discussed anyways.

Minor points:
1. The manuscript could use better wording, some sentences contain the necessary content but are hard to understand and some concepts are not defined, for example, what does it mean to normalize predicted pathways in line 90?
2. What is the difference between ora' and ora? I guess ora' is enrichr but it is not mentioned. Why are some times in the benchmark both included and others only one? This is also the case for GSEA.
3. In the prognosis prediction section, LAUD does actually have significant association to biological gender. Also, BLCA and COAD are almost significantly associated with age. This is something that should be discussed or pointed out.
4. It would be helpful to add a title mentioning if the shown pathways are enriched in the favorable or unfavorable groups for Fig2c and its supplemental. The same thing for Fig2d-e and their supplementals.
5. I would rename "PET-derived biomarkers" to "enrichment-derived biomarkers" as a more suitable and understandable term.
6. When talking about candidate compounds it is always good to show their chemical structure.

Reviewer #2 (Remarks to the Author): Expert in bladder cancer genomics and bioinformatics

Wang et. al.'s manuscript describing the establishing "Benchmark" and an ensemble method for performing pathway enrichment address an important problem, how reliable and relevant are the current methods of pathway enrichment. Through the establishment of a gold standard dataset, the authors evaluated commonly used pathway enrichments tools. Their results clearly show a needed for improved performance. With optimized setting, they used a combinatorial approach to best perform this task, which was packaged and made available with "Pathway Ensemble Tool (PET)". This first half of the paper is well thought-out and clearly executed, although I have reservations regarding the Benchmark data. If PET performs as well going forward as it appears to do in the manuscript, I feel it should supplant the other pathway enrichment tools.
The authors then move on demonstrating its utility through and application through generating prognostic gene signature based on supervised set of genes/pathways in bladder cancer. A CDK2/9 inhibitor is then identified as a top hit and the author's move forward performing in vitro and in vivo validation of this hit. This second half of the manuscript is far weaker than the first. Although, the manuscript does a very nice job taking the reader through how PET could be best utilized to impact drug selection and patient care, the manuscript lacks critical detail on patient/cohort selection which may introduce bias as well as making some underlying assumptions regarding prognosis, subtypes and their approach for validation.
Overall, I feel this is a quality manuscript that will be a benefit to the research community, but that message is muddled through an attempt to demonstrate its utility thought wet-bench validation. However, this portion of the manuscript may be strengthened by addressing the below points.

Concerns:

1) I would guess that most of signatures within the signature databases are based signaling from solid tumor, yet the benchmark dataset only uses 2/7 cell lines (A549 and HepG2) that were derived from solid tumors. Does the cell line choice and/or underlying genetic transformative event influence the authors performance metrics? Is there equal performance using data from the CCLE or DepMap?

2) The authors state that they used PET look for signal associated with prognosis during “early stages in 12 cancer types” (line 82). However, the stage listed for the tumors in Figure 2a are not necessarily “early stage” (e.g. BLCA Stage II is already muscle invasive and PAAD Stage II has spread to the peripancreatic tissue). Additionally, in the supplementally file listing samples, all stages are present, which seems to contradict the text.

3) Was the binning of prognosis restricted to DSS? If OS was used, are the same pathways relevant in a patient that survived 5 years post diagnosis vs 1 year? Additionally, are the same markers independently discovered in other dataset with more mature follow up data (TCGA survival data is rather poor with a short follow-up).

4) The AUC seem rather low for a robust biomarker. How do the leading combination of gene compare against well established biomarkers? For LIHC and BRCA can you please clarify your statement that there were no significant predictors, in 2b there are sig. pathway and in 2d AUC are equal to that of HNSC and LUAD.

5) The supplement lists the Hedegaard data set as validation for BLCA, however it is not cited or mentioned in the main text. Where and how was it used? It is especially important to note when using Hedegaard et. al. that it is a combination of NMIBC and MIBC, which are genetically different. Likewise, the use of the terms low/high risk BLCA should be carefully used as they have a specific meaning, low-risk usually denotes NMIBC which not likely to recur or progress and high-risk which are likely to recur/progress and are clinically treated more aggressively.

6) Robertson et. al 2017 and Kamoun et. al. 2020, represent large datasets with more complex subtype structure, how does the prognosis fit with these subtypes.

7) Line 332 states that CDKi was identified, but then the authors restate they screen compounds in line 342, what was different? Figure 5b provides no value, could the author show expression of CDK2/9 related gene vs CDK4/6 relate genes. Does expression of CDK2/9 correlate to any other clinical or molecular features more strongly they prognosis?

8) Why were T24 cells used in vitro but not in vivo? If UMUC-3 cells used in vivo their corresponding in vitro data should be placed alongside in the main figure. Are these cells basal/luminal or predicated to be favorable or un-favorable?

Minor points:

Fig. S3a legend does not line up with the title nor does the text seem accurate.

In Figure S3b, the authors list n=27 and n=29 samples, however the text states that 52 samples were used, please align these numbers.

Please label S4b/c x-axi

Reviewer #3 (Remarks to the Author): Expert in computational genomics, bioinformatics, and pathway and gene set enrichment analysis

This manuscript describes two approaches to, respectively, benchmark and run gene set enrichment analysis (GSEA) algorithms on high-throughput genomics data. These two types of bioinformatic analyses have been intensively investigated and the manuscript justifies in the introduction the need for the presented work by saying that the results from tools such as GSEA or Enrichr, that is the pathways reported as most relevant by those tools, "are often not the very top ranked pathway among all pathways returned by the analysis". This is a quite vague and speculative statement that neglects the fact that analytical methods are not the only responsible for poor or incomplete results in this type of analysis, but also the quality and breadth of the available pathway databases, which may contain partial, noisy or even incorrect information, due to the fact that only a small fraction of that information has been experimentally verified, and this is even worse outside human and mainstream model organisms.

The authors make the point that most methods have been evaluated either using computer simulated pathways, verified at a small scale biological datasets, or using low-resolution microarray technology. This manuscript attempts to address these shortcomings by introducing a new approach to benchmark gene set enrichment algorithms based on RNA-seq datasets from the ENCODE project on two cell lines (HepG2 and K562) that profile the transcriptome under the systematic knockout (KO) of 197 genes. For each such datasets, the top-200 differentially expressed (DE) genes is defined as a gene set associated with the KO-gene from which the dataset was derived. Then, a GSEA method is evaluated by deriving enrichment scores on the datasets of one of the cell lines using those previously derived gene sets, and tracking the rank of the gene set associated with the KO-gene, which should be high under the hypothesis that such a gene set should be highly ranked whenever the associated KO-gene has been perturbed (knocked-out), independently of the cell line. This approach makes therefore the reasonable assumption that the knockout of a gene in two different cell lines perturbs sets of genes that more alike than with gene sets derived from other KO-genes. While the approach makes sense, the very limited sample size of the ENCODE gene sets, with only two replicates per condition, may easily generate inaccurate lists of differentially expressed (DE) genes, and it is currently way below the number of samples typically assayed nowadays in transcriptomics studies.

The approach in fact is very similar to the one described in the review by Nguyen et al. (2019) and while Nguyen et al. (2019) used microarray data, the fact that the benchmarking involves only well characterised protein-coding genes does not make such a big difference, and the larger sample sizes

employed by Nguyen et al. (2019), cast doubt on whether the benchmarking approach presented here is less unbiased and more exhaustive than the one described in Nguyen et al. (2019).

With respect to the GSEA ensemble approach, called Pathway Ensemble Tool (PET), it integrates three GSEA methods (one-tailed Fisher's exact test, Enrichr and GSEA) by calculating their average ranking and the combined P-value by Stouffer's method, produced by those three methods for each pathway. This kind of ensemble approach is already implemented in the R/Bioconductor package EGSEA (Alhamdoosh et al., 2016), which provides the integration of 11 GSEA methods by a handful of different ensembl scoring functions, including the combination of p-values and averaging the ranking of pathways by each method. From this perspective, the functionality provided by EGSEA includes the one provided by PET. Since the benchmarking described in the manuscript shows that PET performs much better than EGSEA, the only explanation for that result is that the input for EGSEA, e.g., the combined methods, is different than for PET, unless the authors clarify how is it possible that two different implementations of the same methodology give different results.

The manuscript nicely shows meaningful applications of PET in cancer datasets from TCGA, including the prediction of prognostic pathways and drugs that inhibit cell growth in cancer cell lines, demonstrated through in vitro and in vivo bladder cancer data generated for this manuscript.

One important shortcoming of the benchmark and the ensembl approach presented in the manuscript as a contribution to the available open-source software for bioinformatics, is in the way in which the software is provided and documented. At the moment the presented software is only available through a GitHub repo, which could disappear from one day to the next. The instructions for its installation are limited to asking the users to clone the GitHub repo, while as any ensemble approach, it has an important number of upstream software and system dependencies. The GitHub repo provides some further instructions in scripts and in Jupyter notebooks, but these require quite a high level of expertise with the use of the OS command line, and the Python and R packaging installing systems. I could not find the code that allows one to reproduce the benchmarking reported in the manuscript, which is very important for this type of contribution. For instance, it is unclear how the sample specific enrichment scores from ssGSEA are integrated into the approach presented here.

Reviewer #4 (Remarks to the Author): Expert in bladder cancer molecular biology, preclinical models, biomarkers, and CDK inhibitor therapy

This manuscript describes the development of a new platform called Benchmark and a method named Pathway Ensemble Tool (PET) to identify pathway alterations, develop prognostic biomarkers and select therapeutic leads. It is novel and has potential translational and clinical applications. However, there are several issues.

1. Line 111-117: “Benchmark was designed to simulate a common scenario where the investigated genesets (i.e; IGSs) are derived from different sources (e.g. cell lines, species, conditions)”. This should be considered as a strength of Benchmark as it develops a platform applicable across different conditions,

i.e., cell lines, species, conditions. However, this may also miss many important pathways. For example, some biological pathways are important in some cancers, but not in other cancers (for example, the HER2 pathway is important in breast, but not so in bladder cancer).

2. Line 242-249: The comparison of the prognostic biomarkers developed through PET with PAI-1 and uPA. These two genes are rarely used, if ever, in clinic for their prognostic applications. It may be more appropriate to compare with other more commonly used prognostic biomarkers, such as BRAF mutation in colon cancer.

3. The prognostic biomarkers/risk groups described at Figure 2f, Figure S2e, Figure 3b and S3a can efficiently distinguish good versus poor prognosis. However, they are the same databases that were used to develop those prognostic biomarkers. These biomarkers should be validated in independent/different databases and maybe at different stages of cancers of the same database. For example, of the 413 TCGA bladder cancer cases, only 129 Stage II bladder cancer cases were studied here. Will the same prognostic biomarkers be applicable to Stage IV bladder cancer?

4. Line 337-338: “CDK4/6 inhibitors including Palbociclib did not meet the primary endpoint during phase II clinical trials and have performed poorly in additional clinical trials.” It seems that this sentence refers to CDK4/6 inhibitors in any cancer types, not just in bladder cancer. In fact, CDK4/6 inhibitors have been approved in breast cancer.

5. Figure 5a compares the efficacy of a CDK4/6 inhibitor Palbociclib, CDK2 inhibitor CCT068127 and a CDK2/7/9 inhibitor seliciclib in Hela and T24 cells. The targets, CDK2 and CDK4/6, were identified from large groups of cancers. It is not known whether the findings identified from large groups of cancers are applied to individual cell lines. These two cell lines should be analyzed with PET to determine whether they share the same pathway alterations as shown at Figure 4. This is especially concerning since several other studies showed that the CDK4/6 pathway alterations are potential drivers in bladder cancer (Rubio et al. Clin Cancer Res. 2019; 25:390; Long et al. Cancer Immunol Immunother. 2020; 69:2305; Tong et al. J Exp Clin Cancer Res. 2019;38:322) while this study shows that Palbociclib is possibly the least active drug (Fig S4c). One of the previous studies (Long et al. Cancer Immunol Immunother. 2020; 69:2305) was performed in two different patient-derived bladder cancer xenografts (PDXs). PDXs are directly derived from patient cancers, retain the morphology and genomic alterations of parental cancer, at least in bladder cancer (Pan et al PLoS One. 2015;10:e0134346) and are considered better reflecting the behavior of patient cancers. It should be discussed why the prediction of this platform and the findings in these two cell lines are different from multiple previously published studies.

6. Figure S4c” “It is worth noting that our approach also predicted that cisplatin desirably normalizes prognostic genes, however, its capacity was inferior to that of 0175029 (Fig. S4c)”: since cisplatin is one of the most effective drugs in bladder cancer, this claim should be validated in experiments. More specifically, it should be tested in cell lines or PDXs that have inferior response to cisplatin than 0175029 as predicted by PET.

7. Even though the efficacy study described in Figure 5 was done based on the prognostic prediction of pathway alterations and supported the importance of this pathways in cancers, it should be pointed out

and mentioned that prognostic biomarkers of patient outcomes and predictive biomarkers of the efficacy of individual drugs are frequent different.

Many thanks for the insightful comments and constructive feedback. We are pleased that the reviewers found our manuscript to be of interest. In particular, reviewer #2 considered the work to be “a quality manuscript that will be a benefit to the research community”, reviewer #3 considered that “the manuscript nicely shows meaningful applications of PET in cancer datasets from TCGA, including the prediction of prognostic pathways and drugs that inhibit cell growth in cancer cell lines, demonstrated through in vitro and in vivo bladder cancer data generated for this manuscript ” and reviewer #4 believed that the work “is novel and has potential translational and clinical applications.” The strength of our study is the development of an appropriate benchmarking tool for the systematic exploration of the accuracy of multiple biological pathway analysis tools, identification of the optimal parameters for conducting pathway analysis using the most commonly used methods, development of a bespoke ensemble method termed PET, which significantly outperforms all existing methods, including those that employ ensemble approaches and extensive validation experiments as proof-of-principle. By using PET, we have uncovered multiple novel cancer biomarkers that can be used in combination for predicting favorable or unfavorable outcomes and for therapeutic drug prediction. By using this approach, we uncovered a novel class of drugs as top predictions in bladder cancer and confirmed this both in vitro and in vivo.

Our conclusion from reading the point-by-point comments of the reviewers is that they were positive about the data in the manuscript as they have asked principally for more in-depth information, additional analyses, and further discussions. Multiple reviewers pointed out that there are ensemble methods for biological pathway analysis already in existence (reviewers #1 and #3) and that the sample size and choice of cell line for construction of our Benchmark may potentially be a limiting factor (Reviewers #2 and #3). Accordingly, we would like to address these points up-front, followed by point-by-point responses to reviewer comments below. We have, therefore, structured our response letter into two parts, addressing these specific points in the first part, followed by point-by-point responses to additional concerns in the second part.

Together with this response letter, we have uploaded a marked copy of the revised manuscript files, with all changes marked in red. In summary, the revised manuscript now contains **>10 new and revised panels** that confirm and extend our initial conclusions. All these data are now incorporated in the manuscript.

Part 1.1. Rationale behind PET with respect to existing ensemble approaches

Reviewer #1 stated that other works have already explored the aggregation of multiple methods such as PIANO, Enrichment Browser, EGSEA and decoupleR, that these frameworks offer multiple methods while PET currently only contains 3 and that in these works it was already shown that aggregation strategies work better than single methods. These comments aligned with those from reviewer #3, who also cited EGSEA but queried whether the superior performance of PET might be due to different input methods aggregated together. We concur with both reviewers that ensemble approaches have been attempted and have generally yielded enhanced outcomes than single method approaches. Indeed, for this reason, we had initially evaluated EGSEA, one of the ensemble methods specifically mentioned by the reviewers in our benchmark. However, we found that EGSEA only performed better than its underlying methods by a small margin when employed with default settings, and when the underlying methods and input parameters were chosen “blindly" without careful consideration. The difference between EGSEA and our study is that we have first systematically demonstrated that optimizing the input parameters of individual methods (e.g. for EnrichR, GSEA, and ORA) enhances their performance, making them considerably superior to EGSEA per se, before aggregating them in the PET ensemble package. Thus, the ensemble method by PET generates superior improvements that are indispensable for discovery and surpasses all other currently available methods by some distance (e.g. the median rank of

the correct pathway for EGSEA and PET are 8 and 1, respectively, and the average precision at 10 (AP@10) are 44% and 73% for the two methods, respectively). It is also worth noting that EGSEA uses traditional methods that were originally intended for microarray analysis and forces the use of Limma for differential gene expression analysis, which is also designed for microarray data, and therefore may not be optimal for pathway discovery from RNA-seq data.

To provide a more comprehensive evaluation, we have now also assessed the performance of PIANO (Väremo L. et al., 2013) and Enrichment Browser (Geistlinger L. et al., 2016), two additional methods mentioned by reviewer #1. We evaluated performance using four different geneset rankings, including the DESeq2-based ranking that we previously determined to be an optimal parameter for pathway discovery. The median rank and AP@10 for PIANO were recorded as 28-45 and 6-13%, respectively. The median rank and AP@10 for Enrichment browser were recorded as 15-48 and 17-33%, respectively. Notably, these values were significantly lower than those achieved by PET (i.e. median rank of 1 and AP@10 of 73%). Collectively, these results suggest that “blind” ensemble approaches are not optimal for pathway discovery.

With respect to the reviewer #1’s comment regarding decoupleR (Badia-i-Mompel P. *et al*., 2022), it is worth mentioning that decoupleR primarily focuses on calculating pathway activity scores (PAS) for a single condition, which differs from pathway analysis designed to compare pathways between two conditions. Although our Benchmark dataset could potentially be used to evaluate PAS methods, our main emphasis in this paper was on pathway analysis. Consequently, we did not include decoupleR in our evaluations.

We have now included the analyses above into revised **Figs. 1b**, **1e**, and **S1a**. For convenience, we have included some of these results below in **Fig. R1**. Furthermore, we have elaborated on the significance of these findings in our discussion, as per the statements above (lines 439-447, pages 16-17).

Collectively, the revised results indicate significant variability in the performance of existing pathway analysis methods when it comes to identifying the expected pathways. Importantly, as the results show, arbitrarily aggregating all available methods without careful consideration only marginally improves the overall performance. Therefore, it is crucial to identify the optimal setting for these methods prior to ensemble, as we have done with PET, in order to achieve optimal results for unbiased pathway discovery. Most importantly, none of the currently available ensemble methods achieve the accuracy of PET for blinded pathway discovery.

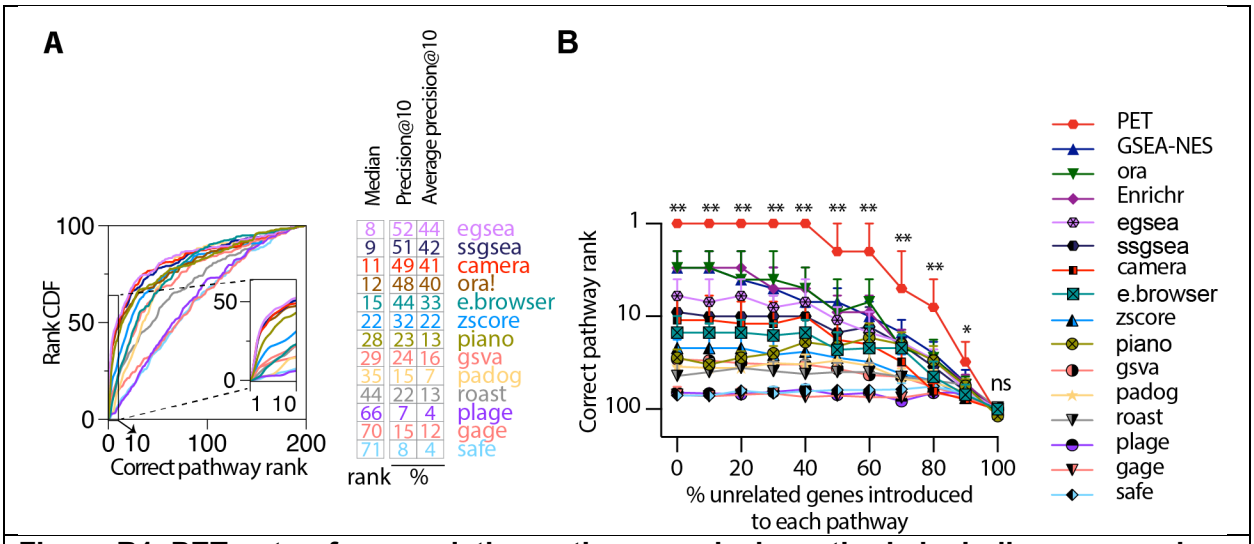

Figure R1. PET outperforms existing pathway analysis methods including egsea, piano and enrichment browser (e.browser). A) cumulative distribution function (CDF) plots of correct pathway ranks comparing indicated pathway analysis methods. Median ranks for correct pathways, the fraction of correct pathways among top 10 reported pathways (precision@10/P@10, and average precision@10/AP@10) are shown. For comparison the median rank, P@10 and AP@10 for PET are 1, 82% and 72%, respectively. **B)** line plots showing the performance of various methods as increasing percentage of unrelated random genes ("noise") are introduced to target genesets. This evaluates the robustness of methods to inaccuracies (or noise) in target genesets. Lines and error bars show median + 95% confidence interval from n=100 iterations. *p<0.05; **p < 0.01; p< 0.0001 by two-sided Wilcoxon rank sum test corrected for multiple hypothesis. ora! and ora are essentially identical, with the only difference being that they are implemented using different programming languages. Specifically, ora! is implemented in R by EGSEA, while ora is implemented in Python by PET.

Part 1.2. Clarification of cell line choice and sample size for construction of Benchmark

Reviewers #2 and #3 specifically asked for clarification about the choice of cell lines and sample size for construction of Benchmark. Reviewer #2 also queried whether equal performance could be achieved by using data from the Cancer Cell Line Encyclopedia (CCLE) or DepMap. Reviewer #3 specifically asked whether the quality of differentially expressed genes, given small sample sizes, could led to inaccuracies. These are very insightful points. As correctly assumed by the reviewers, the availability of appropriate data for benchmarking purposes in which we could clearly define related pathways was, in fact, a limiting factor for creating a reliable Benchmark. However, we believe that the choice of cell lines does not significantly impact the performance evaluation. Specifically, this is supported by our findings, where the performance of most methods remained stable even with the introduction of up to 40% of "unrelated genes" into each geneset (**Fig. 1e** of the original submission), suggesting that the evaluation results cover broad ranges, are resilient to the choice of cell type and that the quality of differential gene identification does not appear to be a major concern. Additionally, we have observed that utilizing 2-3 samples in each group yields an expected false positive rate of 5% (**Fig. S1g**), indicating that the number of samples may not be a limiting factor when constructing a benchmark.

With respect to the reviewer's suggested sources for additional data, we concluded that the CCLE dataset was not suitable for the purpose of constructing an evaluation benchmark for pathway analysis methods. This is because the high similarity between cell lines within each lineage will

result in large overlaps of their differentially expressed genes in comparison to other lineages. Thus, distinguishing ISGs and TSGs under this scenario would be the equivalent of asking analysis methods to distinguish apples from oranges. This would not be challenging and most methods should be near-perfect at achieving this. Such a level of similarity is uncommon in typical unbiased pathway analyses, which involve querying pathways across various species and diverse conditions. For this reason, we constructed our benchmark using more related cell types, in essence creating a benchmark in which pathway analysis methods would be asked to distinguish different types of apples from each other. This is far more challenging and a more rigorous way to assess the performance of pathway analysis tools compared to each other.

To illustrate this point, we have now followed the reviewer's suggestion and constructed another benchmark using cell line expression data from the CCLE. To this end, we first categorized all cell lines based on Depmap lineage annotation (e.g. COAD, STAD) and retained all n=65 lineages with a minimum of 6 distinct cell lines. For each lineage, we randomly split the associated cell lines into two groups, input and target groups. Additionally, we randomly selected n=20 cell lines as controls. We next generated input genesets by identifying the differentially expressed genes between input and control samples using ratio of classes, signal2noise, edgeR, and DESeq2 similar to our previous Benchmark. Subsequently, we employed ora, Enrichr, GSEA and PET pathway analysis methods comparing target and control groups against the input genesets. Under this setting, an ideal pathway analysis method would rank the corresponding lineage as the top result among all the input lineages. All four methods demonstrated excellent performance, with a median rank of correct pathway equal to one and precision@10 exceeding 95% (**Fig. R2A**). This corroborates our suspicion that CCLE-based benchmark is unchallenging for all these methods, making this form of benchmark not useful for comparing the reliability of methods at correct pathway identification. Consistently, even if we introduce up to 90% noise to genesets, most methods exhibited robust performance. Nevertheless, despite these observations, PET still performed better or comparable to all other methods, reaffirming its status as the top-performing approach (**Fig. R2B**). We are willing to include it in the manuscript if the reviewer deems it necessary or beneficial.

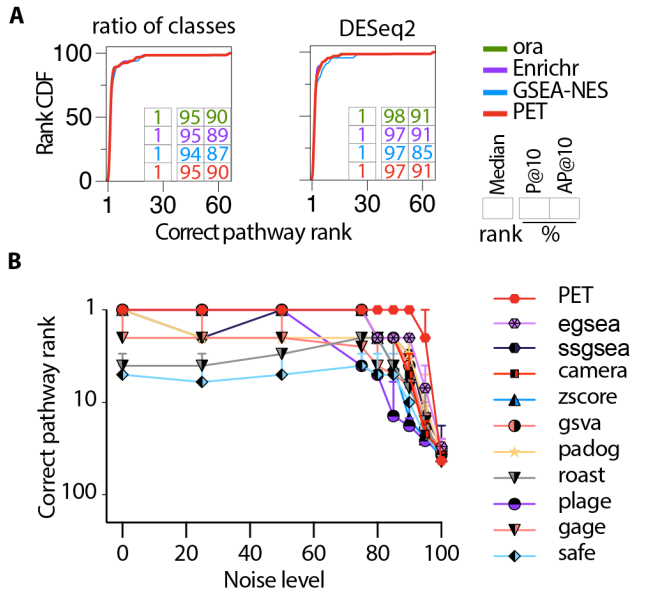

Figure R2. CCLE-based benchmark. A, cumulative distribution function (CDF) plots of correct pathway ranks comparing indicated pathway analysis methods on CCLE genesets. Median rank for correct pathways, precision@10(P@10), and average precision@10(AP@10) are shown. Input and target genesets were defined as the top 200 genes by indicated method (i.e. ratio of classes, DESeq2). **B,** line plots showing the performance of various methods as increasing percentage of unrelated random genes ("noise") are introduced to target genesets. This evaluates the robustness of methods to inaccuracies (or noise) in target genesets. Lines and error bars show median + 95% confidence interval at each noise level.

Part 2. Point-by-point responses to remaining reviewer comments

Reviewers' comments:

Reviewer #1 (Remarks to the Author): Expert in computational cancer genomics, drug response prediction, and benchmarking of bioinformatics methods

Wang et al have generated a benchmark pipeline to evaluate pathway enrichment methods and introduce their own method, called PET, based on the combination of ranks of other methods. They characterized different cancer types leveraging pathway enrichment analysis. Then they identify combinations of genes inside differential enriched pathways as potential biomarkers. Then, by using drug-target information, they rank drugs based on their effect on the found biomarkers. Finally, they experimentally validate one of these drugs in vitro and in vivo.

While this is an active and meaningful topic of research, the manuscripts falls short in novelty both in terms of the approach and the findings, as outlined below:

Major points:

1. While the benchmark data and pipeline called “Benchmark” will be useful to evaluate methods, the method PET is not novel. Other works have already explored the aggregation of multiple methods such as PIANO (Väremo L. et al. 2013), Enrichment Browser (Geistlinger L. et al. 2016), EGSEA (Alhamdoosh M. et al. 2017) available in R, and decoupleR (Badia-i-Mompel P. et al. 2022) available in R and Python. Additionally, these frameworks offer multiple methods while PET currently only contains 3. In all these works, it was already shown that aggregation strategies work better than single methods.

We thank the reviewer for this suggestion and have addressed this as detailed in **Part 1.1** of this response letter above. Briefly, we have now also included evaluations of PIANO and Enrichment browser and discussed decoupleR. We show that by systematically optimizing the underlying methods, outputs from PET are superior to all other methods that operate without optimization of the underlying methods aggregated in their outputs. As such, none of these methods perform as well as PET due to the lack of optimization.

2. Regarding the benchmark dataset the authors should show what is the overlap between genesets across cell lines because this will affect the assumption that a geneset from the same perturbation across cell lines should be similar. Additionally, since the extraction of genesets from perturbation experiments, eCLIP and ChIP-seq data is different, they should show the mentioned overlap and method performances in each of these three data types independently.

We thank the reviewer for this suggestion, which we found insightful. In our initial submission we refrained from performing this analysis because some of the methods we evaluated, such as GSEA, do not primarily depend on the overlap size of the genesets but on their cumulative enrichment instead. Nevertheless, as suggested by the reviewer, we have now conducted an analysis of the geneset overlap in our shRNA-seq benchmark across HepG2 and K562 cell lines. As an additional control, we have generated sized-matched (n=200 genes) random genesets and measured the extent of overlap. As expected, the overlap between random genesets was minimal, with a median of 1 gene out of 200), and non-significant, whereas all four ranking metrics (i.e. Signal2noise, Ratio of classes, EdgeR and DESeq2) resulted in larger, and statistically significant, overlap sizes than that of random genesets. Notably, however, we observed that

genesets generated using the signal2noise metric exhibited a smaller overlap size (median=5 out of 200) compared to genesets generated using other three metrics. This finding provides an explanation for the poor performance of methods when the signal2noise metric is employed (**Fig. R3A**). For matched ChIP-seq genesets, the median overlap sizes were 42 for both extraction methods (**Fig. R3B**). However, in the case of matched eChLP-seq genesets, the median overlap sizes were significantly higher for q-value based extraction (median =49) compared to signal value-based extraction (median = 31) (**Fig. R3C**). The increased overlap sizes observed among genesets in ChIP-seq or eCLIP-seq based Benchmarks, compared to the shRNA-seq based Benchmark, could explain the overall superior performance trend of pathway analysis methods on these benchmarks. These results are now provided in **new Figs. S1B, S1f** and replicated below in **Fig. R3** for convenience.

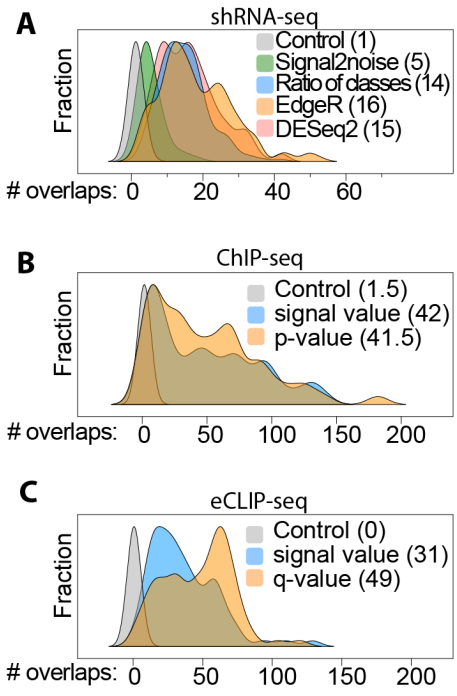

Figure R3. A-C) histograms showing the number of overlapping genes (out of 200) between matched IGSs and TGSs from shRNA-seq (**A**), ChIP-seq (**B**), and eCLIP-seq (**C**), as well as randomly generated control IGSs, for each indicated method employed to generate these genesets. The median of overlap sizes is shown in parentheses. All comparisons against random are statistically significant (adjusted p<0.0001; one-way ANOVA)

3. The prediction that CCR068127 has effect on cancer growth in bladder cancer is of limited novelty, as it already has reported effectiveness against human colon cancer and melanoma cell lines (ref 56).

Cancer types differ significantly from one another in terms of their response to drugs. Clinically, it is common for a drug that demonstrates efficacy in one cancer type to not demonstrate similar efficacy in other cancer types. This is especially true in aggressive forms of bladder cancer, which were the focus of our validation experiments. In fact, one of the major hurdles in oncology is the difficulty in predicting drug efficacy for individual cancers. Thus, we consider our unbiased identification of a CDK9 inhibitor that performs considerably well both *in vitro* and *in vivo* in an aggressive form of bladder cancer as novel, important and of special interest. Furthermore, this is proof-of-principle of the value of using PET for biomarker discovery and therapeutic drug repurposing.

4. Why does PET use GSEA as one of its methods for the aggregation if it has been consistently

shown in the benchmark that it is not a good performing method? Moreover, for the same reason, why is it used as a standalone method for some of the validations?

We apologize for not having made this clearer. In our study, we did not simply use GSEA as one of the methods for aggregation. We first identified the optimal settings under which GSEA, ora and EnrichR need to operate for unbiased pathway discovery. Under these settings they perform significantly better than under their native settings and better than EGSEA (**Figs. 1c-d**). However, because each of these individual methods still fall short of the desired level of accuracy, we developed PET, which combines the outputs of all three operating under their optimized settings, achieving a very high level of accuracy. All results in the subsequent sections of the manuscript are either based on optimized GSEA for visualization purposes or based on PET results.

5. Methods section is generally sparse and disorganized. For example, when calculating the false positive rate, which DEA test was used between cancer and control samples? Which thresholds were used to consider significant changes? How many permutations were done when shuffling samples?

We apologize for the lack of details in our previous version. We have now updated this section as follows: “To calculate false positive rates, we first collected raw read counts and normalized gene expression data from TCGA breast invasive carcinoma (BRCA), TCGA lung adenocarcinoma (LUAD), TCGA prostate adenocarcinoma (PRAD), and GTEx breast, lung and prostate tissues. We randomly selected 2N samples (i.e. N=2-10, 15, 20, 25, 30) from each tissue/cancer and split them into two groups, and ran pathway analysis methods on the expression matrix of the two groups against all canonical pathways (c2.cp.v7.1.symbols.gmt), where the permutations are repeated 100 times for each N. For methods requiring a pre-ranked gene list or set of genes, we utilized DESeq2 to perform the differential expression analysis and used –log10(p-value) or top 200 DEGs as input for those methods. Since the samples were randomly selected biological replicates, the expectation is that any reported pathway significantly different between the groups represents false positive, where significance is defined as the adjusted p-value<=0.05. Thus, the proportion of significant pathways among all the pathways is defined as the false-positive rate.” (lines 585-595, page23). We would welcome any other suggestions that the reviewer has for further clarifying any other areas of the methods sections.

6. Authors mention that pathways have not been previously used as prognostic markers, which is not true. There are many works showing the importance of characterizing cancers at the pathway level for better prognosis. Too many here to cite; just a google search offers many such studies. Additionally, authors also mention that collagen genes have not been used as biomarkers for renal cancer which is also not true, there are many examples available (Best, S.L., et al 2019).

We did not intend to suggest that pathways and gene signatures have not been utilized as prognostic markers and apologize if we gave this impression. We have revised our paper to reflect this accordingly by removing the sentence referred to by the reviewer. Our intention was not to suggest that collagen has not been implicated in renal cancer. With regards to the connection between collagen and renal cancer, we apologize for not conducting a more comprehensive literature search. While our initial acknowledgment in the paper recognized that "collagen expression plays a key role in promoting tumor cell invasion in renal cancers, and vitronectin is linked to inducing the differentiation of cancer stem cells and the formation of tumors", we acknowledge that it was insufficient in covering all relevant studies on the topic. Nevertheless, our main contribution in this part of the study is identifying a novel combination of biomarkers that confers an unprecedented survival advantage using an unbiased discovery approach that limits the search space. We have now revised our manuscript to reflect this more accurately by stating

that “Previous studies have linked the expression of individual collagen genes such as COL5A1 and COL23A1 to renal cancer[31,32]. Collagen expression has also been associated with tumor grade[33] and plays a key role in promoting tumor cell invasion[34] in renal cancers. Furthermore, vitronectin is linked to inducing the differentiation of cancer stem cells and the formation of tumors[35]. Thus, it is biologically conceivable that high expression of this 3-gene combination biomarker provides the observed dramatically unfavorable survival outcome, surpassing the prognostic power of individual collagen genes alone.” (lines 254-255, page9).

7. When comparing combinations of top pathway genes against DEGs, it's not clear if the authors used the top 100 DEGs or the 10 most upregulated and 10 downregulated genes. In any case, they should use the correct signed top hits to not mix signals and make a fair comparison. Additionally, it's true that it does not reach significance, but the compound 0175029 is still the top ranked drug for BLCA when using DEGs, maybe this also changes and should be discussed anyways.

We apologize for any confusion caused by the unclear description in our manuscript. We have now thoroughly revised and rewritten the relevant methods section (i.e. “Prognostic power of individual genes, gene combinations and independent validation” and “Computational drug prediction”) to ensure clarity and avoid any potential confusion. Importantly, to ensure a fair comparison, as highlighted by the reviewer, we had conducted separate comparisons of the DEGs associated with unfavorable survival and those associated with favorable survival. We specifically double-checked and confirmed that we had not mixed any signals during our comparisons. We have also added a statement indicating that “However, it is worth noting that although (CDK)2/9 inhibitor emerges as the top prediction for bladder cancer using DEG-based approach, it does not reach the statistical significance threshold (**Fig. S5b**).” (lines 370-372, page 13).

Minor points:

1. The manuscript could use better wording, some sentences contain the necessary content but are hard to understand and some concepts are not defined, for example, what does it mean to normalize predicted pathways in line 90?

We apologize for the lack of clarity. What we meant by normalizing predicted/perturbed pathways was to up-regulate the genes associated with good prognosis and down-regulate the genes associated with poor prognosis. We have now clarified this in the text by stating that “We demonstrated that the predicted drug indeed restricts cancer cells by normalizing the expression of genes belonging to predicted prognostic pathways. Specifically, it can down-regulate the leading genes of pathways associated with unfavorable prognosis or up-regulate the leading genes of pathways associated with favorable prognosis.” (lines 98-100, page 3).

2. What is the difference between ora’ and ora? I guess ora’ is enrichr but it is not mentioned. Why are some times in the benchmark both included and others only one? This is also the case for GSEA.

We apologize for any confusion caused. We would like to clarify that ora! refers to the EGSEA implementation in R, which was not readily available to us. Therefore, we re-implemented it as ora in Python for the sake of convenience and ease of use. It is important to note that both ora and the EGSEA implementation yield the same results. We have now provided this clarification in the figure legend to avoid any misunderstanding, stating that “ora! and ora are essentially

identical, with the only difference being that they are implemented using different programming languages. Specifically, ora! is implemented in R by egsea, while ora is implemented in Python by PET." (lines 660-662, page 26)

3. In the prognosis prediction section, LAUD does actually have significant association to biological gender. Also, BLCA and COAD are almost significantly associated with age. This is something that should be discussed or pointed out.

This is a crucial observation to consider. We concur that there appears to be a "statistically significant" difference in the gender distribution between the alive and deceased samples in LAUD, as determined by Fisher's exact test. However, when we conducted a log-rank test to compare the survival rates between male and female samples, we did not observe a statistically significant difference. This suggests that although there may be a slight gender bias in LAUD sample distribution, it does not significantly impact the survival outcome. Therefore, we can expect that pathway analyses aimed at distinguishing high and low survival patients are not influenced by sex. However, to address the reviewer's concern, we performed a log-rank test to compare the survival rates between age groups in BLCA and COAD. Specifically, we divided the samples equally into two groups: samples with an age lower than the median (younger group) and samples with an age greater than the median (older group). We compared their survival using the log-rank test. Our analysis revealed no statistically significant differences in survival (p=0.18 and 0.35, respectively) among the age groups. Therefore, we conclude that the pathways distinguishing high and low survival patients in these two cancer types are not influenced by age.

4. It would be helpful to add a title mentioning if the shown pathways are enriched in the favorable or unfavorable groups for Fig2c and its supplemental. The same thing for Fig2d-e and their supplementals.

Thank you for the suggestion. We have now incorporated this suggestion into the title as suggested by the reviewer.

5. I would rename "PET-derived biomarkers" to "enrichment-derived biomarkers" as a more suitable and understandable term.

Thank you for the suggestion. We have now incorporated this suggestion throughout the manuscript.

6. When talking about candidate compounds it is always good to show their chemical structure.

Thank you for the suggestion. We have now added the chemical structure of the CDK inhibitors used in the manuscript in new **Fig. S4d**, which is replicated below as **Fig. R4** for convenience.

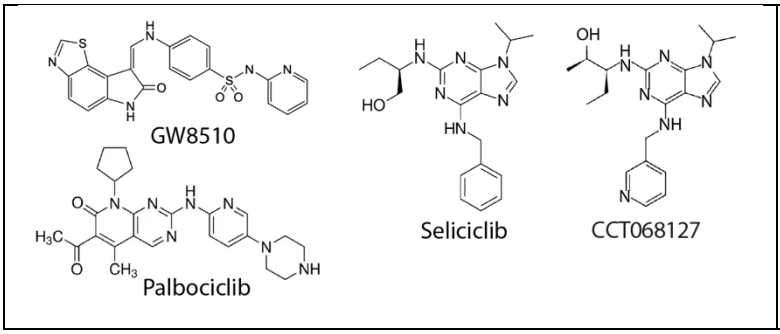

Figure R4. Chemical structures of CDK inhibitors used in this study. CW8510 (CDK2i), Palbociclib (CDK4/6i), Seliciclib (CDK2/7/9i), and CCT068127 (CDK2/7i) are shown.

Reviewer #2 (Remarks to the Author): Expert in bladder cancer genomics and bioinformatics

Wang et. al.'s manuscript describing the establishing "Benchmark" and an ensemble method for performing pathway enrichment address an important problem, how reliable and relevant are the current methods of pathway enrichment. Through the establishment of a gold standard dataset, the authors evaluated commonly used pathway enrichments tools. Their results clearly show a needed for improved performance. With optimized setting, they used a combinatorial approach to best perform this task, which was packaged and made available with "Pathway Ensemble Tool (PET)". This first half of the paper is well thought-out and clearly executed, although I have reservations regarding the Benchmark data. If PET performs as well going forward as it appears to do in the manuscript, I feel it should supplant the other pathway enrichment tools.

The authors then move on demonstrating its utility through and application through generating prognostic gene signature based on supervised set of genes/pathways in bladder cancer. A CDK2/9 inhibitor is then identified as a top hit and the author's move forward performing in vitro and in vivo validation of this hit. This second half of the manuscript is far weaker than the first. Although, the manuscript does a very nice job taking the reader through how PET could be best utilized to impact drug selection and patient care, the manuscript lacks critical detail on patient/cohort selection which may introduce bias as well as making some underlying assumptions regarding prognosis, subtypes and their approach for validation.

Overall, I feel this is a quality manuscript that will be a benefit to the research community, but that message is muddled through an attempt to demonstrate its utility thought wet-bench validation. However, this portion of the manuscript may be strengthened by addressing the below points.

We thank the reviewer for their positive comments and constructive feedbacks on our manuscript. In the following sections, we have addressed all the specific concerns raised by the reviewer.

Concerns:

1) I would guess that most of signatures within the signature databases are based signaling from solid tumor, yet the benchmark dataset only uses 2/7 cell lines (A549 and HepG2) that were derived from solid tumors. Does the cell line choice and/or underlying genetic transformative event influence the authors performance metrics? Is there equal performance using data from the CCLE or DepMap?

We thank the reviewer for this insightful comment and for their suggestion regarding alternative sources of data for construction of Benchmark. We have addressed this as detailed in **Part 1.2** of this response letter above. Briefly, the performance of most methods appeared to deteriorate significantly even with up to 40% simulated noise, indicating that choice of cell lines and sample size is not a major problem. We have followed the reviewer suggestion and created CCLE based benchmark. However, we found that this alternative Benchmark is suboptimal for comparing the performance of pathway analysis methods as illustrated and discussed in **part 1.2** of this response letter above.

2) The authors state that they used PET to look for signal associated with prognosis during "early stages in 12 cancer types" (line 82). However, the stage listed for the tumors in Figure 2a are not necessarily "early stage" (e.g. BLCA Stage II is already muscle invasive and PAAD Stage II has

spread to the peripancreatic tissue). Additionally, in the supplementally file listing samples, all stages are present, which seems to contradict the text.

We apologize for not being clear. Our intention was to state that we looked at the earliest stages of cancer for which >20 samples are available in the dataset. We have removed "early" qualifier in the revised manuscript. We apologize for not being clear regarding the samples used in our study. We had used Pathologic stage (NCI Thesaurus Code: C28257) and/or Pathologic T (TNM) (NCI Thesaurus Code: C48881 and C48739) indicators, depending on their availability. In cases where both indicators were present but contradictory, we prioritized the indicator that corresponds to an earlier stage of tumor staging. We have now clarified this in the revised methods section (lines 536-538, page 21) and rectified staging information in **Fig. 2a**.

3) Was the binning of prognosis restricted to DSS? If OS was used, are the same pathways relevant in a patient that survived 5 years post diagnosis vs 1 year? Additionally, are the same markers independently discovered in other dataset with more mature follow up data (TCGA survival data is rather poor with a short follow-up).

This is an excellent question. In our study, we compared the overall survival, not the disease free survival, of patients who were at the same tumor stage and age-matched. Specifically, we compared patients who had died due to the disease with those who were still alive, to minimize the impact of potential confounding factors in our analyses. Importantly, the length of follow-up of alive and deceased patients was quite similar (680 and 563 days, respectively). This information is now amended to modified **Table S2b**.

To address independent validations of biomarkers, we had included analyses using six independent datasets from four different cancer types in our original submission (please see **Fig. S2f** of the original manuscript). Moreover, we had also shown an association between enrichment-derived biomarkers and established independent prognostic markers, such as CDKN2A alteration, in either renal or papillary kidney cancer types (**Fig. 3c** of the original submission). We acknowledge that this was not clearly described in the methods. We have now updated this section in our revised manuscript.

To directly address the reviewer's question regarding whether "the same pathways relevant in patients that survived longer post diagnosis vs those died early", we turned our focus to bladder cancer samples, as most of our experimental validations were conducted in this particular cancer type. We performed pathway analysis using PET comparing deceased patients who survived less than 1 year versus those that survived more than 3 years and identified significant pathways (adjusted p-value<=0.05). We opted for a 3-year timeframe instead of the recommended 5-year survival post-diagnosis because the number of patients meeting the 5-year cutoff was extremely limited. As a result, we identified 21 pathways associated with favorable (>3 years survival) and 56 pathways associated with unfavorable prognosis (<1year survival). A significant number of those pathways overlapped with previously identified prognostic pathways in bladder cancer, (fisher exact test p-values of 0.0085 and 0.0346 for favorable and unfavorable prognostic pathways, respectively). These data are not currently in the revised manuscript but could be added if the reviewer deems it to be essential.

4) The AUC seem rather low for a robust biomarker. How do the leading combination of gene compare against well-established biomarkers? For LIHC and BRCA can you please clarify your statement that there were no significant predictors, in 2b there are sig. pathway and in 2d AUC are equal to that of HNSC and LUAD.

We would like to express our gratitude to the reviewer for bringing up this question. Although we acknowledge that some AUC values seem relatively low, we have demonstrated strong survival differences using these biomarkers, as illustrated in our results (e.g. **Fig. 2f**). Additionally, we would like to emphasize that none of the individual gene biomarkers displayed a higher AUC than the combinations identified. This is one of the unique aspects of our approach, highlighting that our combinations outperform any individual biomarker. We have clarified this in the revised manuscript by stating, “Notably, none of the individual gene biomarkers displayed a higher AUC than the combinations identified.” (line 245, page 9)

Regarding LIHC and BRCA, even though the AUC values were comparable to those of HNSC and LUAD, the FDR-adjusted logrank p-values for distinguishing alive versus deceased patients were not significant (FDR<0.05). Therefore, the corresponding figures (**Figs. 2e, S2d**), which only display the significant biomarkers, do not include these. We believe that our misplacement of the Legends for cancers that did not yield any significant predictors may have contributed to the confusion. We have now rectified this issue by moving this to the appropriate section of the legend.

5) The supplement lists the Hedegaard data set as validation for BLCA, however it is not cited or mentioned in the main text. Where and how was it used? It is especially important to note when using Hedegaard et. al. that it is a combination of NMIBC and MIBC, which are genetically different. Likewise, the use of the terms low/high risk BLCA should be carefully used as they have a specific meaning, low-risk usually denotes NMIBC which not likely to recur or progress and high-risk which are likely to recur/progress and are clinically treated more aggressively.

We would like to apologize for the oversight of missing the citation in the main manuscript. We have now added citations for all the independent datasets used in our validations in the main text and in the methods sections. In relation to the Hedegaard et al datasets, we had also noted a mixture of NMIBC (non-muscle invasive bladder cancer) and MIBC (muscle invasive bladder cancer) patients. Upon further analysis, we had discovered that among patients with an EORTC risk score of NMIBC=0, only 4 patients progressed to stage T2, while 282 patients remained as non-progressors. In contrast, among patients with an EORTC risk score of NMIBC=1, there were 27 progressors compared to 147 non-progressors. Considering the limited number of progressors in the NMIBC=0 category, we decided to utilize the entire dataset for validating our biomarkers, regardless of the NMIBC status.

We also thank the reviewer for drawing our attention to the clinical interpretation of low and high-risk cancer patients. In order to eliminate any potential misinterpretation, we have replaced the terms "low-risk" with "resilient" and "high-risk" with "vulnerable" throughout the manuscript to accurately reflect the intended categories. If the reviewer feels that different terms would be more appropriate, we welcome their suggestions.

6) Robertson et. al 2017 and Kamoun et. al. 2020, represent large datasets with more complex subtype structure, how does the prognosis fit with these subtypes.

We are grateful to the reviewer for bringing additional datasets to our attention. The dataset from Robertson *et al.*, 2017 corresponds to the same TCGA BLCA cohort that we have already

examined in our manuscript. Additionally, Kamoun *et al.*, 2020 curated multiple cohorts, including TCGA and GSE87304, which are the two largest cohorts with available annotations. In response to the reviewer's comment, we examined basal and luminal subtypes in these datasets.

We observed higher expression levels of the enrichment-derived unfavorable prognostic biomarkers in the basal subtype compared to the luminal subtypes of bladder cancer in both the TCGA and GSE87304 cohorts. These observations are consistent with the data (GSE160693) in our initial submission showing that the unfavorable prognostic markers are significantly elevated in basal subtype compared to luminal subtype bladder cancer (**Fig. S3b**). These data are now included in updated **Fig. S3b** and replicated below for convenience (**Fig. R5**).

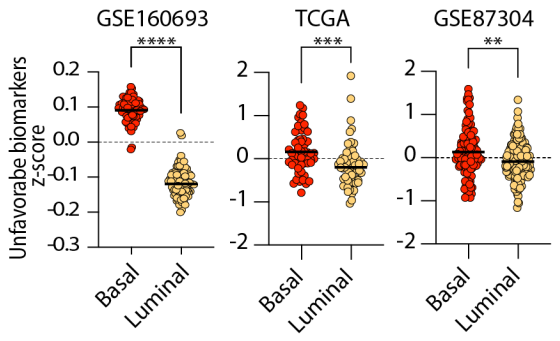

Figure R5. plots showing the expression (z-score normalized) of top 100 significant unfavorable prognostic biomarkers in RNA-seq of luminal versus basal sub-types of bladder cancer in three indicated independent cohorts. Each dot represents the mean z-score of an unfavorable prognostic biomarker across all luminal or basal samples. ** p<0.01; *** p<0.001; **** p< 0.0001 by two-sided Wilcoxon rank sum test.

7) Line 332 states that CDKi was identified, but then the authors restate they screen compounds in line 342, what was different? Figure 5b provides no value, could the author show expression of CDK2/9 related gene vs CDK4/6 relate genes. Does expression of CDK2/9 correlate to any other clinical or molecular features more strongly they prognosis?

We apologize for the confusion. To clarify, line 332 pertained to computational screening of a small dataset of approximately 1600 compounds, mostly comprising FDA-approved drugs which did not include CDK4/6 inhibitor or cisplatin, while line 342 refers to a larger dataset that includes these and additional ~5000 compounds. We have now clarified this in the text stating that "We scrutinized whether PET-based drug prediction specifically advocates CDK2/9 or CDK4/6 inhibition for bladder cancer. Since the drug database we used for screening did not include any CDK4/6 inhibitors, we computationally screened an additional 5425 compounds from LINC1000 perturbation database that included the CDK4/6 inhibitor Palbociclib." (lines 373-375, page 13)

We acknowledge the reviewer's point regarding the original **Fig. 5b**, and we sincerely appreciate their valuable suggestion regarding the inclusion of CDK9 and CDK4/6 target genes to enhance this figure. To this end we sourced an independent list of target genes for Palbociclib (CDK4/6 specific inhibitor) [Redacted] and a list of target genes for VIP152 (CDK9 specific inhibitor curren tly in clinical trials) (Sher et al, Leukemia, 2023, PMID: 36376377). We compared transcriptomes of T24 cells treated with CCT068127 or DMSO control using optimized GSEA. We observed that the CDK2/9 inhibitor, CCT068127, was able to significantly reduce the expression of genes ass ociated with CDK9 inhibition, while showing no significant impact on genes associated with CDK4/6 inhibition. We have now updated **Fig. 5b** to reflect this information. For convenience, we have replicated the same results below in **Fig. R6**.

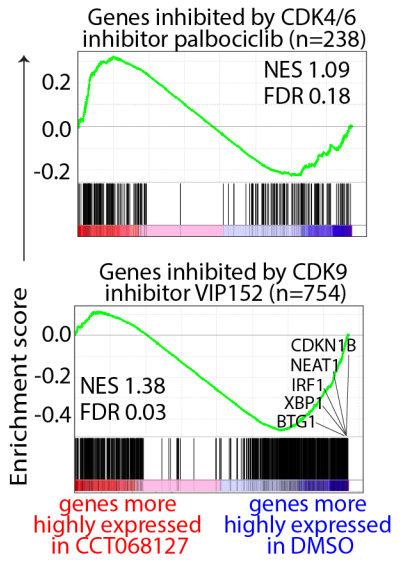

Figure R6. GSEA plots showing enrichment of genes repressed by the CDK4/6 inhibitor Palbociclib (top) or genes repressed by the CDK9 inhibitor VIP152 (bottom) in T24 cells treated 48hr with the CDK2/9 inhibitor CCT068127 (1uM) or DMSO control. Palbociclib (SRP404373) or VIP152 inhibited genes are publicly sourced.

To directly answer the reviewer's query regarding whether the expression of CDK2/9 correlate to molecular features, we evaluated the mRNA expression of CDK2 and CDK9 in luminal and basal subtypes of bladder cancer by analyzing RNAseq from multiple datasets. Indeed, we found that CDK9, but not CDK2, expression is more highly expressed in basal subtype of bladder cancer in all three independent cohorts. This observation has now been included in **new Fig. S5e** in the revised manuscript and is reproduced below as **Fig. R7** for convenience.

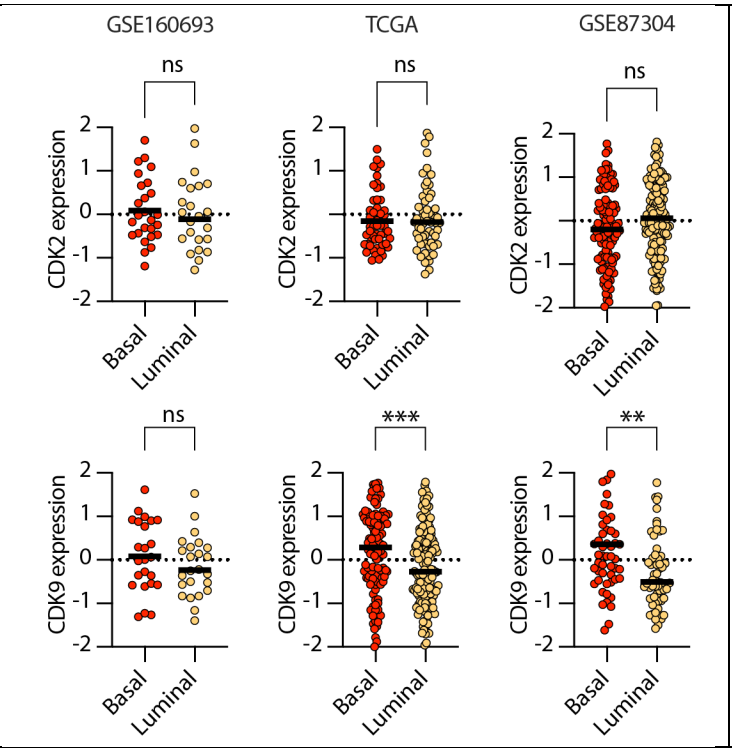

Figure R7. mRNA expression of CDK2 and CDK9 in RNA-seq of luminal versus basal sub-types of bladder cancer in three indicated independent cohorts. Each dot represents a sample in the cohort. ** p<0.01; *** p<0.001 by two-sided Wilcoxon rank sum test. y-axes denote z-scores of expression values.

8) Why were T24 cells used in vitro but not in vivo? If UMUC-3 cells used in vivo their

corresponding in vitro data should be placed alongside in the main figure. Are these cells basal/luminal or predicated to be favorable or un-favorable?

UMUC-3 was chosen for the *in vivo* experiment due to its highly aggressive nature as a form of bladder cancer (PMC6055076), with the intention of assessing the effectiveness of predicted drugs even against such aggressive tumors. As suggested by the reviewer, we have moved the *in vitro* UMUC3 data from supplementary figures to the main figures (now revised **Fig 5a**).

Minor points:

Fig. S3a legend does not line up with the title nor does the text seem accurate.
We apologize for this oversight. We have now split **Fig. S3** into two distinct figures, ensuring that they are accompanied by appropriate titles.

In Figure S3b, the authors list n=27 and n=29 samples, however the text states that 52 samples were used, please align these numbers.
We have now rectified this.

Please label S4b/c x-axis
We have now rectified this.

Reviewer #3 (Remarks to the Author): Expert in computational genomics, bioinformatics, and pathway and geneset enrichment analysis

This manuscript describes two approaches to, respectively, benchmark and run geneset enrichment analysis (GSEA) algorithms on high-throughput genomics data. These two types of bioinformatic analyses have been intensively investigated and the manuscript justifies in the introduction the need for the presented work by saying that the results from tools such as GSEA or Enrichr, that is the pathways reported as most relevant by those tools, "are often not the very top ranked pathway among all pathways returned by the analysis". This is a quite vague and speculative statement that neglects the fact that analytical methods are not the only responsible for poor or incomplete results in this type of analysis, but also the quality and breadth of the available pathway databases, which may contain partial, noisy or even incorrect information, due to the fact that only a small fraction of that information has been experimentally verified, and this is even worse outside human and mainstream model organisms.

We apologize for the vague statement in our introduction. We have now rectified the sentence to state “One common application of these methods has been to test whether specific pathways hypothesized by researchers are, in fact, enriched in a dataset. In this approach, pathway enrichment analyses have generally proven successful, but pre-suppose *a priori* knowledge or hypothesis. An alternative, highly desired application, especially in the context of exponentially growing omics datasets, is the adoption of unbiased discovery-based approaches. Here, the objective is to rank the "correct pathways" above all other pathways in the reported results without imparting bias from *a priori* knowledge. However, the performance and suitability of existing pathway analysis methods for this form of blinded or unbiased discovery remains largely unknown.” (lines 64-70, page 2)

We also fully agree with the reviewer that analytical methods are not solely responsible for poor results and acknowledge that there are several factors that can affect the outcome of a study, such as sample size, quality of data, study design, and biological variability. Our focus here has been choosing appropriate and optimized analytical methods as one of the critical steps towards improving the quality and reliability of results despite the presence of other biological and non-biological variables.

The authors make the point that most methods have been evaluated either using computer simulated pathways, verified at a small-scale biological datasets, or using low-resolution microarray technology. This manuscript attempts to address these shortcomings by introducing a new approach to benchmark geneset enrichment algorithms based on RNA-seq datasets from the ENCODE project on two cell lines (HepG2 and K562) that profile the transcriptome under the systematic knockout (KO) of 197 genes. For each such datasets, the top-200 differentially expressed (DE) genes is defined as a geneset associated with the KO-gene from which the dataset was derived. Then, a GSEA method is evaluated by deriving enrichment scores on the datasets of one of the cell lines using those previously derived genesets, and tracking the rank of the geneset associated with the KO-gene, which should be high under the hypothesis that such a geneset should be highly ranked whenever the associated KO-gene has been perturbed (knocked-out), independently of the cell line. This approach makes therefore the reasonable assumption that the knockout of a gene in two different cell lines perturbs sets of genes that more alike than with genesets derived from other KO-genes. While the approach makes sense, the very limited sample size of the ENCODE genesets, with only two replicates per condition, may easily generate inaccurate lists of differentially expressed (DE) genes, and it is currently way below the number of samples typically assayed nowadays in transcriptomics studies.

We acknowledge the limitations of constructing a benchmark resulting from the scarcity of available datasets, as pointed out by the reviewer. This was also a point raised by reviewer #2 and has been addressed in **part 1.2** of this response letter above. Briefly, utilizing 2-3 samples in each group yields an expected false positive rate of 5% (**Fig. S1g**), indicating that the number of samples may not be a limiting factor when constructing a benchmark. Likewise, robustness to noise (**Fig. 1e** of the initial submission) indicates that the quality of differential gene identification does not appear to be a major concern. To further address this issue, in accordance with the recommendation provided by reviewer #2, we also established an additional benchmark using data from the cancer cell line expression database (CCLE). This new benchmark was constructed by utilizing 65 DepMap lineages that consist of a minimum of 6 cell lines. We found that our inferences from our original benchmark remain consistent when using this benchmark.

The approach in fact is very similar to the one described in the review by Nguyen et al. (2019) and while Nguyen et al. (2019) used microarray data, the fact that the benchmarking involves only well characterised protein-coding genes does not make such a big difference, and the larger sample sizes employed by Nguyen et al. (2019), cast doubt on whether the benchmarking approach presented here is less unbiased and more exhaustive than the one described in Nguyen et al. (2019).

We thank the reviewer for this comment. It is likely that our failure to clarify the distinction between our benchmark and the one introduced by Nguyen *et al*., 2019 may have given a false impression and undermined the novelty of our work. In our revised manuscript, we have made a clear distinction between our benchmark and the one introduced by Nguyen *et al*., 2019, which we summarize below.

There are clear points of distinction that indicate the advantages of our approach over that of Nguyen *et al.,* 2019. Firstly, our benchmark utilizes a large number (~1000) of well-defined pathways as ground truth, whereas the benchmark created by Nguyen *et al*. does not. Instead, the authors manually annotated 75 datasets from 15 diseases and associated them with "related" target pathways from the KEGG database. This is a comparatively small set and the manual association to KEGG pathways is subjective. As a consequence, performance measurements were indirect and all methods tested by Nguyen *et al.,* 2019 indicated poor performance (i.e. the median rank of target pathways), with many target pathways not reaching statistical significance. This makes comparison among methods unreliable. In addition, we evaluated two novel measures (P@10 and AP@10), which are important considerations for the application of these tools by biologists. We have made these distinctions clear in the revised manuscript so that the novelty of our approach and its inherent advantages are well-defined.

With respect to the GSEA ensemble approach, called Pathway Ensemble Tool (PET), it integrates three GSEA methods (one-tailed Fisher's exact test, Enrichr and GSEA) by calculating their average ranking and the combined P-value by Stouffer's method, produced by those three methods for each pathway. This kind of ensemble approach is already implemented in the R/Bioconductor package EGSEA (Alhamdoosh et al., 2016), which provides the integration of 11 GSEA methods by a handful of different ensembl scoring functions, including the combination of p-values and averaging the ranking of pathways by each method. From this perspective, the functionality provided by EGSEA includes the one provided by PET. Since the benchmarking described in the manuscript shows that PET performs much better than EGSEA, the only explanation for that result is that the input for EGSEA, e.g., the combined methods, is different than for PET, unless the authors clarify how is it possible that two different implementations of the same methodology give different results.

We thank the reviewer for this very insightful comment, which was raised by other reviewers and has been addressed in **part 1.1** of this response letter above. Briefly, the reviewer's intuition is essentially correct – we systematically determined the optimal functioning parameters of underlying methods aggregated in PET, which is not an option for other ensemble methods, resulting in far superior performance. Technical limitations may also limit the capabilities of ensemble approaches, as discussed above. We have also further evaluated and discussed additional ensemble methods in our revised manuscript.

The manuscript nicely shows meaningful applications of PET in cancer datasets from TCGA, including the prediction of prognostic pathways and drugs that inhibit cell growth in cancer cell lines, demonstrated through in vitro and in vivo bladder cancer data generated for this manuscript.

We thank the reviewer for their positive comments in recognizing the significance and application of our study.

One important shortcoming of the benchmark and the ensembl approach presented in the manuscript as a contribution to the available open-source software for bioinformatics, is in the way in which the software is provided and documented. At the moment the presented software is only available through a GitHub repo, which could disappear from one day to the next. The instructions for its installation are limited to asking the users to clone the GitHub repo, while as any ensemble approach, it has an important number of upstream software and system dependencies. The GitHub repo provides some further instructions in scripts and in Jupyter notebooks, but these require quite a high level of expertise with the use of the OS command line, and the Python and R packaging installing systems. I could not find the code that allows one to

reproduce the benchmarking reported in the manuscript, which is very important for this type of contribution. For instance, it is unclear how the sample specific enrichment scores from ssGSEA are integrated into the approach presented here.

We apologize for the lack of details in our previous version. We have now amended details in the methods section of our revised manuscript and have updated the GitHub repository, as suggested by the reviewer. Briefly, we have now provided environment files (in YAML format) for both Linux and OSX systems to install all Python and R dependency packages using a single command. We have also now provided template scripts for running all methods, such as ssGSEA, that were evaluated by Benchmark. We acknowledge the reviewer's concern regarding the requirement of expertise and basic programming skills for running PET. In response, we have taken steps to simplify this process and identified the sections that needed improvement to enhance accessibility. We believe that the process is now improved, and installation has been simplified. We are open to any additional suggestions from the reviewer for further clarification of these steps and to make the process even more user-friendly.

Reviewer #4 (Remarks to the Author): Expert in bladder cancer molecular biology, preclinical models, biomarkers, and CDK inhibitor therapy

This manuscript describes the development of a new platform called Benchmark and a method named Pathway Ensemble Tool (PET) to identify pathway alterations, develop prognostic biomarkers and select therapeutic leads. It is novel and has potential translational and clinical applications. However, there are several issues.

We extend our appreciation to the reviewer for their positive feedback and their recognition of the novelty of our work. In the following sections, we address all the specific concerns raised by the reviewer.

1. Line 111-117: “Benchmark was designed to simulate a common scenario where the investigated genesets (i.e; IGSs) are derived from different sources (e.g. cell lines, species, conditions)”. This should be considered as a strength of Benchmark as it develops a platform applicable across different conditions, i.e., cell lines, species, conditions. However, this may also miss many important pathways. For example, some biological pathways are important in some cancers, but not in other cancers (for example, the HER2 pathway is important in breast, but not so in bladder cancer).

We fully agree with the reviewer that it is crucial to ensure that evaluations are unbiased and capture differences among cancer types. In our study, we addressed this concern by evaluating specificity and sensitivity. Our results demonstrate that PET has controlled specificity (false positive rate, **Figs. S2e-f** of the initial submission) and sensitivity (**Fig. 1e** of the original submission). Furthermore, our evaluations demonstrate that PET identifies cancer-specific pathways and biomarkers (**Figs. 2 and 3**), indicating its ability to capture differences among cancer types. We would like to also refer the reviewer to **part 1.2** above, which highlights that most pathway analysis methods excel in distinguishing highly specific pathways, especially when these pathways are well-documented in existing databases.

2. Line 242-249: The comparison of the prognostic biomarkers developed through PET with PAI-

1 and uPA. These two genes are rarely used, if ever, in clinic for their prognostic applications. It may be more appropriate to compare with other more commonly used prognostic biomarkers, such as BRAF mutation in colon cancer.

We thank the reviewer for bringing this to our attention. Our aim in this study was to compare available biomarkers based on gene expression for a fair comparison. However, we recognize the value that mutational information holds as biomarkers in clinical settings. Therefore, as suggested by the reviewer, we focused on the three most frequently mutated genes in cancer samples, BRAF, TP53 and KRAS. For each TCGA cancer type with a sufficient number of samples exhibiting mutations (n>15), we split the samples into two groups based on their mutational status: wild-type samples without the mutation and samples containing mutant oncogenes. To assess whether these mutations could serve as indicators of survival, we conducted log-rank tests between the wild-type (WT) and mutant (mut) samples for each cancer type. Among all the cancer types examined, pancreatic cancer (PAAD) was the only one in which the presence of KRAS mutations served as a significant indicator of poorer overall survival. (Hazard ratio = 1.8; logrank p-value=0.02) (**new Fig. S2g**). Nevertheless, the KRAS mutation exhibited less discriminatory power compared to PET-based biomarkers in distinguishing between favorable (**Fig. S2e,** HR=0.41(or 2.4)) and unfavorable (**Fig. 2f,** HR=2.6) survival outcomes. Overall, these data further underscore the added value provided by our predicted gene combination biomarkers across various types of cancer. We have now included these in our revised manuscript as **new Fig. S2g** and replicated below in **Fig. R8** for convenience.

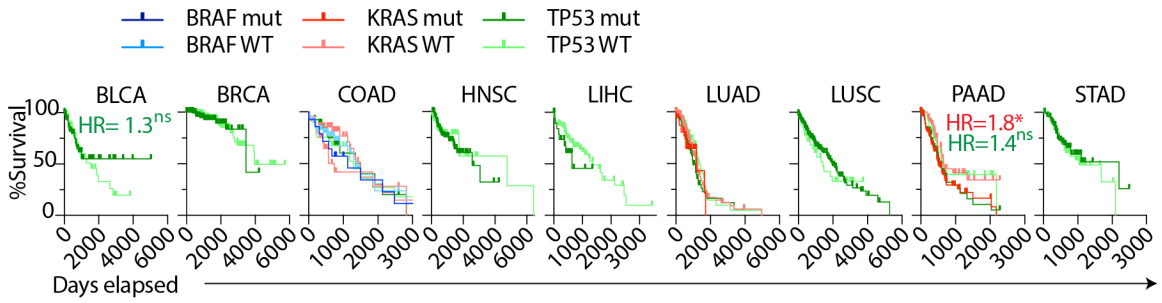

Figure R8. Kaplan–Meier (KM) survival curves showing the power of the indicated gene mutation to stratify survival in each cancer type. Samples were split into two groups, with one group having the indicated gene mutation and the other group with no mutation. Hazard ratio (HR) and logrank test p-values are reported. * p<0.01.

3. The prognostic biomarkers/risk groups described at Figure 2f, Figure S2e, Figure 3b and S3a can efficiently distinguish good versus poor prognosis. However, they are the same databases that were used to develop those prognostic biomarkers. These biomarkers should be validated in independent/different databases and maybe at different stages of cancers of the same database. For example, of the 413 TCGA bladder cancer cases, only 129 Stage II bladder cancer cases were studied here. Will the same prognostic biomarkers be applicable to Stage IV bladder cancer?

We are in agreement with the reviewer that validation datasets should be independent from test datasets. For this reason, we had indeed used independent datasets for validation. For example,

in **Fig. S2f** we showed that ~70-95% of favorable/unfavorable biomarkers in TCGA cohorts agreed with favorable/unfavorable survival outcome in independent cohorts of pancreatic, breast, bladder and stomach cancers. We have now highlighted this aspect further in our revised submission. Furthermore, we have obtained molecular subtype information from an additional bladder cancer database (GSE87304) by Kamoun *et. al*., 2020, which was brought to our attention by reviewer #2. Using this, we show that the predicted unfavorable prognostic biomarkers are indeed more elevated in patients with the basal subtype of bladder cancer. For a more in-depth description, we refer the reviewer to our response to point **#6 of reviewer #2** above.

As suggested by the reviewer, we have also tested whether the identified prognostic biomarkers from earlier stages of cancer could predict outcomes in later stages of cancers. Specifically, we investigated the percentage of combination gene biomarkers that consistently exhibited predictive power in later-stage samples. The comprehensive results of this analysis are presented in **Table S2h-i**, and a concise summary is provided below (**Table R1**). The findings indicate that in certain cancer types (e.g. HNSC, PAAD), the early-stage biomarkers consistently predict outcomes even in later stages of cancer. However, in other cancer types (e.g. BLCA, KIRP), only a small fraction of the early-stage biomarkers demonstrate consistency in their predictive ability. We have now added these analyses to our revised manuscript (lines 270-274, page 10).

Cancer type (TCGA)	%Consistency in later stage
BLCA	10.0
BRCA	26.4
CESC	54.0
COAD	34.5
HNSC	76.5
KIRC	76.8
KIRP	5.0
LIHC	40.0
LUAD	11.1
LUSC	42.0
PAAD	85.5
STAD	70.3

Table R1. For each cancer type shown are the percentage of combination gene markers associated with early-stage prognosis that consistently exhibited predictive power in previously unseen later-stage samples.

4. Line 337-338: "CDK4/6 inhibitors including Palbociclib did not meet the primary endpoint during phase II clinical trials and have performed poorly in additional clinical trials." It seems that this sentence refers to CDK4/6 inhibitors in any cancer types, not just in bladder cancer. In fact, CDK4/6 inhibitors have been approved in breast cancer.

We apologize for any confusion caused. To clarify, the statement was intended to refer solely to bladder cancer. We have revised the sentence to make it specific to bladder cancer. The sentence now reads, "CDK4/6 inhibitors including Palbociclib did not meet the primary endpoint during phase II clinical trials for bladder cancer[62-65]." (lines 367-368, page 13)

5. Figure 5a compares the efficacy of a CDK4/6 inhibitor Palbociclib, CDK2 inhibitor CCT068127 and a CDK2/7/9 inhibitor seliciclib in Hela and T24 cells. The targets, CDK2 and CDK4/6, were identified from large groups of cancers. It is not known whether the findings identified from large

groups of cancers are applied to individual cell lines. These two cell lines should be analyzed with PET to determine whether they share the same pathway alterations as shown at Figure 4. This is especially concerning since several other studies showed that the CDK4/6 pathway alterations are potential drivers in bladder cancer (Rubio et al. Clin Cancer Res. 2019; 25:390; Long et al. Cancer Immunol Immunother. 2020; 69:2305; Tong et al. J Exp Clin Cancer Res. 2019;38:322) while this study shows that Palbociclib is possibly the least active drug (Fig S4c). One of the previous studies (Long et al. Cancer Immunol Immunother. 2020; 69:2305) was performed in two different patient-derived bladder cancer xenografts (PDXs). PDXs are directly derived from patient cancers, retain the morphology and genomic alterations of parental cancer, at least in bladder cancer (Pan et al PLoS One. 2015;10:e0134346) and are considered better reflecting the behavior of patient cancers. It should be discussed why the prediction of this platform and the findings in these two cell lines are different from multiple previously published studies.

This is an important set of questions. We agree with the reviewer that relating the drug predictions from enrichment-derived biomarkers to the cell lines tested is a very important point. A similar query was raised by reviewer #2 (point #7) above. To specifically address this, we have sourced an independent list of target genes repressed by Palbociclib (a CDK4/6-specific inhibitor) [Redacted] and a list of genes repressed by VIP152 (a specific inhibitor of CDK9 currently under going clinical trials) (Sher et al, Leukemia, 2023, PMID: 36376377). We then analyzed the transcriptomes of T24 cells treated with either a CDK2/9 inhibitor (CCT068127) or DMSO control for enrichment of these genesets using optimized GSEA. We observed that the CDK2/9 inhibitor significantly reduced expression of CDK9-, but not CDK4/6-repressed genes, in T24 cells. We have now updated **Fig. 5b** to reflect this information. For convenience, we have included the results in **Fig. R6** above.

We do not dispute that altered expression of CDK4/6 has been observed and proposed as a driver in bladder cancer. In fact, this has been nicely shown in the paper by Rubio *et al.* cited by the reviewer. In our hands, we also found that T24 cells were susceptible to Palbociclib in culture as a significant reduction in survival was observed in dose-response to this drug (**Fig. 5a** of our original submission). In our hands, the magnitude of this effect was less than that seen in some of the publications cited by the reviewer, but the same pattern was observed. The different magnitude could potentially be attributable to differences in seeding density, culture medium used or readout method between the papers cited by the reviewer and our own data (all these factors differ between the different reports). Nevertheless, in direct comparison, we found that the inhibitory effects of Seliciclib (CDK2,7,9 inhibitor) and CC068127 (CDK2,9 inhibitor) were of a far greater magnitude. These data illustrate the power of therapeutic drug prediction based on sets of dysregulated genes or biomarkers (**Figs. 5a** in our original submission) rather than single gene alterations (e.g. Fig. 1A of Rubio *et al*).

We have now discussed these points in our revised manuscript, by specifically stating, “The prediction of CDK9 but not CDK4/6 by our approach was somewhat unexpected but aligned with the observation that Palbociclib (a CDK4/6 inhibitor) did not meet the primary endpoint during phase II clinical trials in bladder cancer[62-65]. Previous studies have indicated that alterations in the CDK4/6 pathway could potentially drive bladder cancer[71-73] but the cell lines and assays used in these studies are slightly different than ours. Consistently, our cell viability measurements, conducted using the CCK-8 assay, also demonstrated a modest yet significant effect of Palbociclib on T24 cells. Further *in vivo* investigation using patient-derived bladder cancer xenografts (PDX) models are warranted[74] to compare the effects of CDK9 inhibitors and CDK4/6 inhibitors on primary bladder tumors in future studies”.

6. Figure S4c" "It is worth noting that our approach also predicted that cisplatin desirably normalizes prognostic genes, however, its capacity was inferior to that of 0175029 (Fig. S4c)": since cisplatin is one of the most effective drugs in bladder cancer, this claim should be validated in experiments. More specifically, it should be tested in cell lines or PDXs that have inferior response to cisplatin than 0175029 as predicted by PET.

We greatly appreciate the valuable suggestion provided by the reviewer. In response, we have now conducted the suggested experiments in both T24 and UMUC cell lines. We found that, as expected, Cisplatin is a highly effective drug in bladder cancer cell lines. We found that the tested CDK2/9 inhibitor inhibited proliferation to a similar degree as cisplatin (updated **Fig. 5a** and replicated in **Fig. R9** below for convenience). We have now adjusted text in the manuscript to reflect this point. (lines 387-389, page 14)

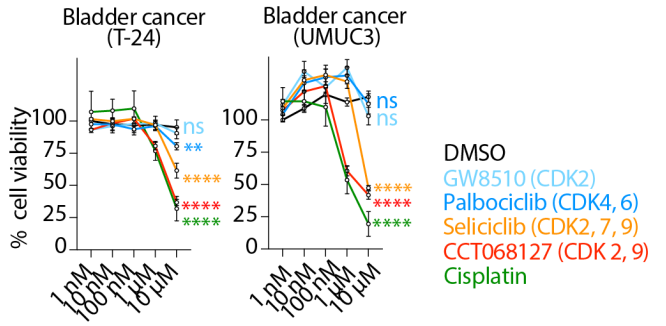

Figure R9. Viability, measured by CCK-8 assay, of T24 and UMUC3 bladder cancer cell lines treated with increasing concentrations of vehicle or the indicated drugs for 72 hrs. The selective targets of CDK inhibitors are shown in parenthesis. Mean +/- sem from n>3 biological replicates. **p<0.01, ****p<0.0001 by by two-way ANOVA with multiple comparisons adjustment.

7. Even though the efficacy study described in Figure 5 was done based on the prognostic prediction of pathway alterations and supported the importance of this pathways in cancers, it should be pointed out and mentioned that prognostic biomarkers of patient outcomes and predictive biomarkers of the efficacy of individual drugs are frequent different.

We acknowledge the reviewer's excellent point. In our original discussion, we mentioned that "determining whether predicted prognostic pathways contribute directly to the pathophysiology (rather than merely representing biomarkers) can only be established through experimentation." We have now also incorporated the distinction pointed out by the reviewer by stating that, "It is also important to note that prognostic biomarkers, which predict patient outcomes, and predictive biomarkers, which determine the efficacy of individual drugs, are often distinct from each other. With increasing omics data from responders and non-responders, future research will be needed to apply pathway enrichment analysis to identify not only prognostic, but also predictive, pathways." (lines 478-482, page 18)

Conclusion

In summary, in response to the insightful comments of the reviewers, we have performed a number of additional experiments and further analyses that support and extend our original conclusions. We thank the reviewers for the time they have taken to carefully peer-review our work and their help in improving our manuscript. We believe that the manuscript has been strengthened by inclusion of the suggested additions. We feel that, given the interest in the subject, this work will be of great value to both the general readership of Nature Communications as well as researchers working with high-throughput assays from which pathway level information needs to be extracted. We have uploaded both marked (all changes are in red) and clean copies of the revised manuscript and thank the editorial team and the reviewers for the opportunity to improve our study. We look forward to your response.

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

We are glad to see that Wang et al have taken the time to address most of our comments and concerns. We certainly value their efforts for improving the quality of the manuscript.

However, two major points still remain that need to be addressed:

1) It is expected that if the better performing DEG selection is used for PET's methods, PET is going to outperform the other ensemble methods that ran without this optimization. The authors should show the results of the other ensemble methods using the enrichment scores obtained from the optimized selection of DEGs. This can be generated using the functions consensusScores and combResults for piano and Enrichment browser, respectively.

2) In the prognosis prediction section, the authors have run a log-rank test to test for possible confounding covariates in the LAUD, BLCA and COAD cell lines. However, they only mention the p-value in the response letter and do not plot the survival curves. The authors should plot them and add a brief comment, similar to what they already wrote in the response letter, to the manuscript.

As a minor comment:
3) In the comparison of ensemble methods, authors include several methods but not decoupled, claiming that it can not be used for contrast level statistics, but it actually can and that is even done in the original decoupler paper.

Reviewer #2 (Remarks to the Author):

First, I would like to thank the authors for their thoughtful response to this reviewer's comments, however many of their responses fall short of addressing the comment or only speak to part of the comment. Below, I have again presented my concerns, hopefully in a clearer fashion, so that they can be addressed.

Rebuttal point 1.2:

Yes, the authors are correct that if they used the CCLE to validate the data it would not be challenging as lineage specific differences would be the primary source of variation, just as if they TCGA PanCan data as a whole. As mentioned within the original comment, and I apologize if it was not initially clear, but the intent was to compare cell lines within lineage not between lineages. Specifically, within bladder cell lines, there are groups of tumors which have specific alteration that drive differences in transcriptional programs. When FGFR3 mutant vs wild type cell lines are compared does PET identify the relevant signaling pathways; similarly comparisons could be made for ER+ vs ER- breast cells lines or EGFR mutant

lung cancer cell lines.

Response point 3:

The authors in their response state that they compared “overall survival not the disease free survival”, however my comment was regarding disease specific survival. They then go on to state we compared patients who died of disease. This latter statement implying they used disease specific survival (DSS) and not overall survival (OS), although this wording may seem trivial, it could make differences as OS treats a person who dies of a heart attack the same as one who died because of the cancer.
I then asked if they could validate the data in another cohort, as TCGA as poor follow up. They bring that point to the forefront in that the alive vs dead follow (assuming they are referring to the median) was 680 vs 563 days or only 1.8 years vs 1.5 years. To address the point they mention other tumors type, however again, this misconstrues the point, can these pathways be validated in other bladder datasets with survival, such as UC-GENOME, IMvigor, ABACUS, or PURE-01.

Response Point 4:
We understand that the AUC was greatest for the multi gene combination but was it significantly different that the single predictors.

Response Point 5:
Whereas the BLCA analysis was performed OS previously and biomarker gene were selected from this group, the validation was performed in E-MTAB-4321, which is a majority NMIBC and only has progression data, which is not comparable to survival data. Additionally, their results may simply be separating out HG vs LG samples, which would not be clinically relevant. If true validation is to be performed, it should be in done in a MIBC, as mentioned above.

Response Point 6:
In this comment, I was intended to ask how the unfavorability score was distributed among the TCGA or Consensus Subtypes. Also, does the score do a better job at delineating survival then subtype alone. Their use of a comparison from another paper to illustrate that the basal/luminal subtypes are not accurate for prognosis, is not appropriate and should be the analysis should be performed by the authors on their same data using a published and validated method such as the packaged from Kamoun et. al. or signatures from Damrauer el. al or Choi et. al.

Response Point 8:

The Authors did not address my question as to whether the T24 and UMUC3 cells were favorable or unfavorable. Does the CDK9 effect favorable and unfavorable cell line equally? Likewise, based on depmap J82 and 5637 cells have high CDK9 expression, whereas UBLC1 has low expression, are they predicted to be unfavorable and favorable, respectively.

Reviewer #3 (Remarks to the Author):

The authors should be commended for their very well structured efforts to address the reviewers' concerns. I still have, though, a few points of disagreement and suggestions, which I enumerate below.

1. In your response letter you cleared up that the reason why integrating the results of GSEA, EnrichR and ORA with PET is superior as any other ensemble method that also integrates them, is because PET optimizes the input parameters of GSEA, EnrichR and ORA. Looking at the code in the Jupyter notebook 'run_PET_tutorial' and the GSEA documentation in https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html I would say that PET is using the default parameters for GSEA. The other two methods, EnrichR and ORA, use the exact Fisher's test, and I don't see much to optimize there either. The only place I found in the manuscript where an optimization seems to be described is in page 5, where it says that a geneset size of 200 genes was near optimal for both GSEA and Enrichr, but I do not see how a fixed number of called DE genes can be optimum irrespective of the dataset and the experimental design that generated it, usually genes are called DE on the basis of cutoff values for FDR and/or magnitude of the fold change. As a side note, the rationale of using an ensembl method is normally to integrate disparate methods that implement "orthogonal" approaches, while ORA and EnrichR are based on the same statistical device, the Fisher's exact test. It would be therefore interesting to know why you integrate two methods based in the Fisher's exact test.

2. In the light of the previous comment, the main difference between PET and EGSEA must be in how the DE genes (what it's called input genesets -IGS in the manuscript) and their statistics are derived. PET is using DESeq2, while EGSEA uses limma. This is in fact highlighted in part 1.1 of the response letter, where it says "EGSEA uses traditional methods that were originally intended for microarray analysis and forces the use of Limma for differential gene expression analysis, which is also designed for microarray data, and therefore may not be optimal for pathway discovery from RNA-seq data". This statement is wrong, limma provides procedures to analyze both, microarray and RNA-seq data, the latter concretely through the limma-voom pipeline, which is also widely used with RNA-seq data (the corresponding paper Law et al. (2014), https://doi.org/10.1186/gb-2014-15-2-r29 accumulates nearly 5,000 citations) and is the one used by EGSEA. So, my current guess is that the main difference between PET and EGSEA is in the way in which the DE statistics are derived, which means that the superior performance of PET cannot be attributed to an innovative way of integrating pathway analysis methods, but to a more careful way of performing a DE analysis with two replicates per condition (GLMs on negative binomial distributions are probably better suited for limited replication than limma-voom).

3. Embedding DE methods within pathway analysis may seem appealing, but it has the challenge that is not entirely trivial how to accomodate the flexibility of the embedded tool. Note that PET is interfacing DESeq2 for a two-group comparison only, so more complex experimental designs (i.e., adjusting for an extra covariate) are not feasible. In this sense, EGSEA overcome that limitation by allowing one to pass a design matrix to limma, you may want to consider this possibility too with DESeq2.

4. Regarding the limited sample size used to find DE genes, you argue that utilizing 2-3 samples in each group yields an expected false positive rate (FPR) of 5%, illustrated in Figure S1g. The FPR=FP/(FP+TN), so it has on its denominator the number of true negative (TN) instances, which in our context (truly non-DE

genes?) is much larger than the number of false positive (FP) instances, and therefore it is relatively easy to get a low FPR. A much relevant parameter in this context is the false discovery rate (FDR). In fact, the performance of DE methods has been thoroughly investigated in the past, for instance in the work of Schurch et al. RNA, 2016 (https://10.1261/rna.053959.115), they used a clean yeast data set generated for benchmarking purposes and they show that while most of the DE methods successfully control their FDR < 5% with as low as 3 replicates (note that this study is using two, so we cannot infer from that work whether 2 replicates also allow one to control the FDR), only 20%-40% of the expected DE genes are found. So, it is hard to believe that the target gene sets (TGS) used for benchmarking constitute a reliable gold standard to rank methods.

5. I appreciate the efforts of the authors to improve the documentation. I believe the installation and use of the software could be made somewhat simpler if the authors would attempt to implement PET in a single platform, either Python or R. For Python, the authors could use the recent implementation of DESeq2 in Python by Muzellec et al. (2023, https://doi.org/10.1093/bioinformatics/btad547), and GSEApy (https://github.com/zqfang/GSEApy) for GSEA and even EnrichR. For R, the authors could use the Bioconductor package fgsea for GSEA and the CRAN package enrichR for EnrichR.

Reviewer #4 (Remarks to the Author):

Most of my previous review comments were properly addressed. However, two issues still exist and need clarification.

1. Line 119-124: “Importantly, Benchmark was designed to simulate a common scenario where the investigated genesets (i.e. IGSs) are derived from different sources (e.g. cell lines, species, conditions) than established biological pathways (i.e. TGSs). To achieve this, Benchmark was created such that genesets from the same TF, RBP or gKD but from distinct cell lines are better matched to each other than any other genesets. As an example, the IGS that contains STAT1-bound genes in K562 cells is more closely related to the TGS that contains STAT1-bound genes in HepG2 cells than any other TGS.”

I raised this issue at the previous review that this approach will identify common genesets shared by those cell lines/species/conditions used to generate Benchmark, but miss those genesets/pathways unique to a specific cancer/cell line/species/condition. Similar to the comments raised by Reviewer #3 during the previous round of review as well, this approach makes the assumption that the alteration of one specific gene in two different cell lines perturb sets of genes that are more alike than with genesets derived from alterations of other genes. Alteration of one specific gene, but at different genetic background/in different cancers, frequently has different biological effects. This is frequently seen at clinic that the same targeted therapy may be effective in some cancer types, but not in others even when they contain the same target gene alterations. Even within the same cancer type, targeted therapy works in some patients, but not in others even when their cancers harbor the same alterations. For example, the FGFR2/3 inhibitor erdafitinib has been approved for bladder cancer harboring FGFR2/3 activation mutation or translocation. However, the response rate is only 40%. This suggests that the

same FGFR2/3 alterations can induce different geneset alterations in different bladder cancers. In this specific case, Benchmark may identify genesets/pathways shared by K562 and HepG2. But will likely miss many unique to each ccell. For example, perturbation of the Bcr-Abl will affect K562, but is unlikely to affect HepG2.

2. I assume that the elbow plots at Figure S5c was obtained from combined data of many bladder cancer cell lines/clinical tissues. Based on this elbow plots, his article then determined the drug efficacy in three cell lines, cervical cancer Hela, bladder cancer T24 and UMUC3. Can PET be used to generate the elbow plots specifically for each of these three cell lines? Are the elbow plots of these three cell lines the same as the one shown at Figure S5c? If PET can be used to generate specific elbow plots for those cell lines, drug sensitivity assays should be done based on the drugs shown on the specific elbow plots and determine whether PET can predict drug sensitivity and guide drug selection.

Reviewers' comments:

We deeply appreciate the insightful comments and constructive feedback provided by all reviewers during this second round of review. We are pleased that the reception of our initial point-by-point response was positive overall. Additionally, we value the constructive suggestions that has further strengthened the rigor of our manuscript and are grateful for the opportunity to refine our work.

In this response letter, we have systematically addressed the additional concerns and suggestions provided by the reviewers. To ensure clarity and ease of review, we have detailed all changes made to the manuscript directly within this letter and have also provided a marked-up version of the manuscript, highlighting all revisions in red. In summary, the revised manuscript now over 10 new and revised panels and supplementary tables. We believe that the changes enhance our initial findings and extend our initial conclusions, providing a more comprehensive understanding of the work.

Reviewer #1 (Remarks to the Author):

We are glad to see that Wang et al have taken the time to address most of our comments and concerns. We certainly value their efforts for improving the quality of the manuscript.

We thank the reviewer for their thoughtful review and constructive feedback. We are delighted to hear that our initial revisions have addressed most of their comments and concerns which undoubtedly enhanced the quality of the manuscript. We are grateful for the opportunity to improve and refine our work based on the remaining issues, which we have addressed below.

However, two major points still remain that need to be addressed:
1) It is expected that if the better performing DEG selection is used for PET's methods, PET is going to outperform the other ensemble methods that ran without this optimization. The authors should show the results of the other ensemble methods using the enrichment scores obtained from the optimized selection of DEGs. This can be generated using the functions consensusScores and combResults for piano and Enrichment browser, respectively.

We appreciate the insightful feedback from the reviewer. In response, we have evaluated PIANO and Enrichment browser using the consensusScores and combResults functions, as suggested. For both ensemble methods, we utilized DESeq2 for differential gene expression analyses, consistent with the approach used in PET. Further details are provided in a **new methods section** entitled, "Running ensemble methods". The results show PIANO and Enrichment browser achieving precision at 10 (P@10) scores of 44% and 59%, respectively (**Figs. R1A-B**; all ensemble methods are shown for convenience and comparison purposes). Notably, these methods showed improved performance with the use of DESeq2 (**Fig. R1A**) versus the use of other ranking metrics (**Fig. R1C**, showing signal2noise as an example), underscoring the importance of DEG selection in ensemble performance, aligning with the reviewer's observation. However, these methods

exhibit lower performance than PET, which demonstrated a P@10 of 82% (**Figs. R1A-B**). This disparity reinforces our finding that the choice of underlying pathway analysis tools and their settings is equally crucial for the performance of ensemble methods, a key focus of our study. This dual emphasis on DEG selection and pathway tool configuration is pivotal to understanding the dynamics of ensemble methods. The revised **Figs. 1b, 1e** and **S1a** now include these analyses, providing a more comprehensive comparison under optimized conditions. Further technical information is provided in the new Methods section entitled, "Running Ensemble Tools", detailing how the analysis was performed (page 23 lines 606-620). These enhancements not only address the reviewer's concerns but also add depth to our analysis.

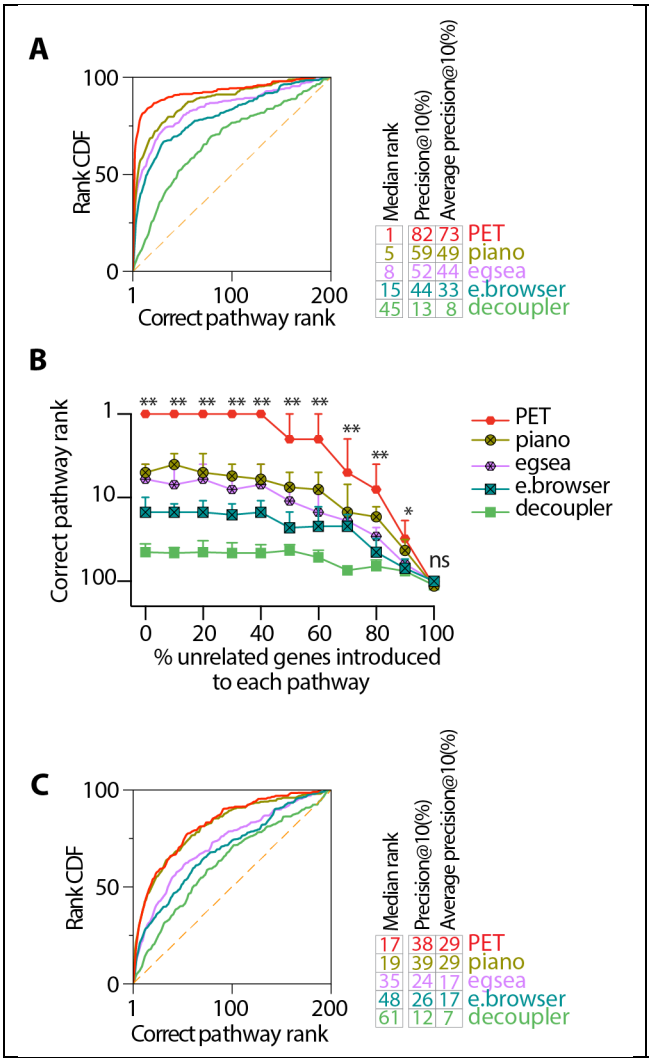

Figure R1. PET outperforms existing pathway ensemble methods A) cumulative distribution function (CDF) plots of correct pathway ranks comparing indicated pathway analysis methods using DESeq2 as ranking statistics. Median ranks for correct pathways, the fraction of correct pathways among the top 10 reported pathways (precision@10, P@10, and average precision@10, AP@10) are shown. **B)** line plots showing the performance of various methods as increasing percentage of unrelated random genes ("noise") are introduced to target genesets. All methods were run with the same optimal settings as in (**A**). This evaluates the robustness of methods to inaccuracies (or "noise") in target genesets. Lines and error bars show median + 95% confidence interval from n=100 iterations. *p<0.05; **p < 0.01 by two-sided Wilcoxon rank sum test corrected for multiple hypothesis testing, comparing PET against other methods. **C)** CDF plots of correct pathway ranks comparing indicated pathway analysis methods using signal2noise as ranking statistics. Median ranks for correct pathways, the fraction of correct pathways among top 10 reported pathways (precision@10, P@10, and average precision@10, AP@10) are shown.

2) In the prognosis prediction section, the authors have run a log-rank test to test for possible confounding covariates in the LAUD, BLCA and COAD cell lines. However, they only mention the p-value in the response letter and do not plot the survival curves. The authors should plot them and add a brief comment, similar to what they already wrote in the response letter, to the manuscript.

We apologize for the oversight in our previous submissions and thank the reviewer for pointing out the need to show the survival curves. In response, we have included these as new **Fig. S3b**, as suggested by the reviewer, which is replicated below (**Fig. R2**) for convenience. These figures provide a visual representation of the survival analysis for LUAD, BLCA and COAD cancers and are added to the revised manuscript, accompanied by a brief commentary, as requested, in the figure legend indicating that we are showing, "Kaplan-Meier overall survival curves comparing samples below versus above median age **(b)** and female versus male **(c)** in the indicated cancer types. Hazard ratio (HR) and logrank test p-values are reported." (page 31, lines 805-814). Our original text has been retained in the main description, "We confirmed that the groups were matched in age and sex and that age and sex were not determinants of outcome (**Fig. S3**)" (page 8, lines 215-216).

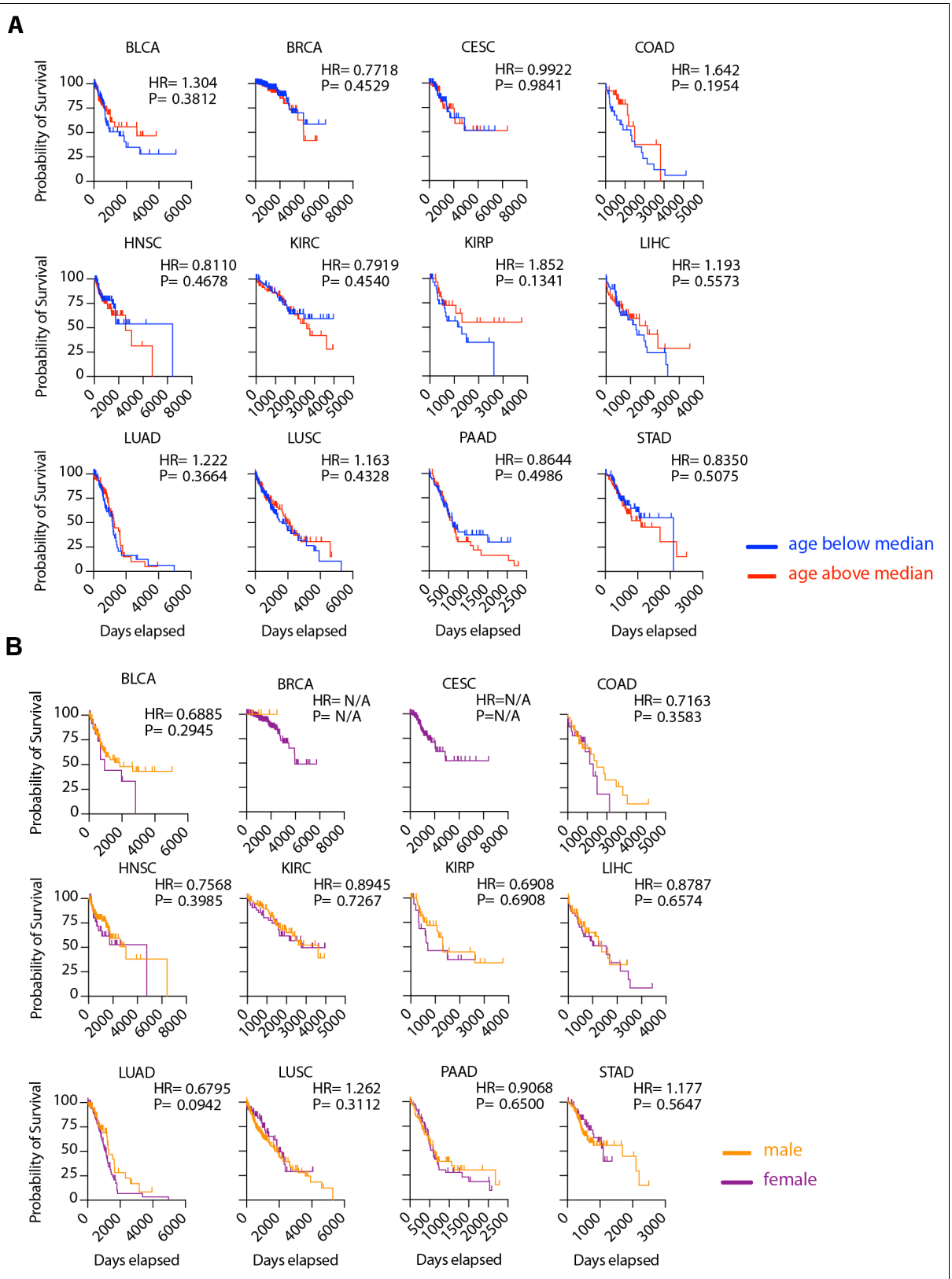

Figure R2. A-B, Kaplan-Meier survival curves showing samples below versus above median age **(A)** and male versus female **(B)** in the indicated cancer types. Hazard ratio (HR) and log-rank test p-values are

reported. Samples from breast invasive carcinoma and cervical squamous cell carcinoma are biased towards female donors due to the nature of these cancer types. BLCA, bladder carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma; COAD, colon adenocarcinoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LAUD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PAAD, pancreatic adenocarcinoma; STAD, stomach adenocarcinoma.

As a minor comment:

3) In the comparison of ensemble methods, authors include several methods but not decoupler, claiming that it can not be used for contrast level statistics, but it actually can and that is even done in the original decoupler paper.

We sincerely apologize for our initial oversight in not including Decoupler in our comparison of ensemble methods. We have now evaluated and included Decoupler using our benchmark. The addition of Decoupler further supports our initial finding that the choice of underlying pathway analysis tools and their settings is crucial for the performance of ensemble methods, as detailed in our response to the reviewer's first major point above. Further information is provided in the new Methods section entitled, "Running Ensemble Tools", detailing how the analysis was performed (page 23 lines 606-620). The results of this inclusion are reflected in updated **Figures 1b, 1e** and **S1a**, and, for convenience, **Figure R1** in this response. These updated figures now present a more comprehensive comparison of ensemble methods, including the performance and applicability of Decoupler in our study context.

Reviewer #2 (Remarks to the Author):

First, I would like to thank the authors for their thoughtful response to this reviewer's comments, however many of their responses fall short of addressing the comment or only speak to part of the comment. Below, I have again presented my concerns, hopefully in a clearer fashion, so that they can be addressed.

We thank the reviewer for their positive perspective on our manuscript, apologize for not fully addressing their concerns during the first round of revisions, and appreciate their diligence in providing clarifications. We apologize for any shortcomings in our initial responses and are grateful for the opportunity to address each point more comprehensively, and to improve and refine our work based on the reviewer's concerns, as detailed below.

Rebuttal point 1.2:

Yes, the authors are correct that if they used the CCLE to validate the data it would not be challenging as lineage specific differences would be the primary source of variation, just as if they TCGA PanCan data as a whole. As mentioned within the original comment, and I apologize if it was not initially clear, but the intent was to compare cell lines within lineage not between lineages. Specifically, within bladder cell lines, there are groups of tumors which have specific alteration that drive differences in transcriptional programs. When FGFR3 mutant vs wild type cell lines are compared does PET identify the relevant signaling pathways; similarly, comparisons could be made for ER+ vs ER- breast cells lines or EGFR mutant lung cancer cell lines.

We really appreciate the reviewer's clarification and have undertaken a more targeted pathway analysis directly in line with their suggestion. This involved applying PET to compare specific cell line groups within the same lineage in the CCLE dataset, focusing on FGFR3 mutant vs. WT in bladder cancer (BLCA), ER^+ vs. ER^- in breast cancer (BRCA), and EGFR mutant vs. WT in lung cancer (LAUD) cell lines (all listed in **new Table S1c**). These comparisons are designed to assess PET's ability to identify relevant signaling pathways within lineages rather than between lineages, i.e. in scenarios with more subtle sources of variation. All 1988 canonical pathways annotated in MSigDB release 7.2 (Liberzon, Subramanian et al.; Bioinformatics; 2011) were assessed.

In the FGFR3 mutant vs. WT BLCA, PET notably identified the "MYC activity pathway" and the related "ribosomal RNA processing" among the top three enriched pathways in FGFR3 mutant cells. This finding aligns with known interactions between FGFR3 and MYC in BLCA, as supported by existing literature (Mahe, Dufour et al.; EMBO Mol Med; 2018) and MYC as regulator of ribosomal RNA synthesis and processing (van Riggelen, Yetil et al.; Nat Rev Cancer; 2010) (Grandori, Gomez-Roman et al.; Nat Cell Biol; 2005) (Arabi, Wu et al.; Nat Cell Biol; 2005) (Grewal, Li et al.; Nat Cell Biol; 2005).

For ER^+ vs. ER^- breast cancer cell lines, the top two enriched pathways identified by PET in ER^+ cell lines were "Estrogen dependent gene expression" and "Estrogen receptors (ESR) mediated signaling", which corresponds with the established understanding of estrogen receptor signaling in ER^+ breast cancer (Saha Roy and Vadlamudi; Int J Breast Cancer; 2012, Clusan, Ferriere et al.;

Int J Mol Sci; 2023), reinforcing PET's efficacy in detecting expected biological pathways. The third top enriched pathway is the Mta3 pathway, recognized as a crucial component of an estrogen-dependent module that regulates growth and differentiation in breast cancer (Fujita, Jaye et al.; Cell; 2003).

Likewise, in the comparison of EGFR mutant vs. WT LAUD cell lines, PET highlighted the following as the top 3 enriched pathways: “Surfactant metabolism”, “Integrin5 pathway”, and “Pentose Phosphate Pathway”. All three pathways have been previously associated with EGFR signaling in lung cancers. For example, there is cross-talk between surfactant proteins and EGF signaling. For example, while EGF stimulates surfactant apoprotein synthesis (Whitsett, Weaver et al.; J Biol Chem; 1987, Inoue, Xin et al.; Cancer Sci; 2008), surfactant protein D suppresses EGF signaling (Hasegawa, Takahashi et al.; Oncogene; 2015). Likewise, interactions between EGFR and integrin signaling can lead to EGFR activation (Ricono, Huang et al.; Cancer Res; 2009, Rubio, Romero-Olmedo et al.; Theranostics; 2023) and EGFR signaling regulates metabolic pathways that include the pentose phosphate pathway in EGFR-mutated lung adenocarcinoma cells (Makinoshima, Takita et al.; J Biol Chem; 2014).

These analyses demonstrate that PET is able to discern subtle yet significant variations in pathway activities across different genetic backgrounds. The detailed results of these pathway analyses have now been added to the manuscript as new **Figure S2**, which is replicated below as **Fig. R3** for convenience. A summarized description of the data is now provided on page 7 (lines 183-196). We believe the suggestion by the reviewer significantly strengthens our study, showcasing PET's versatility and accuracy in a more nuanced analytical setting.

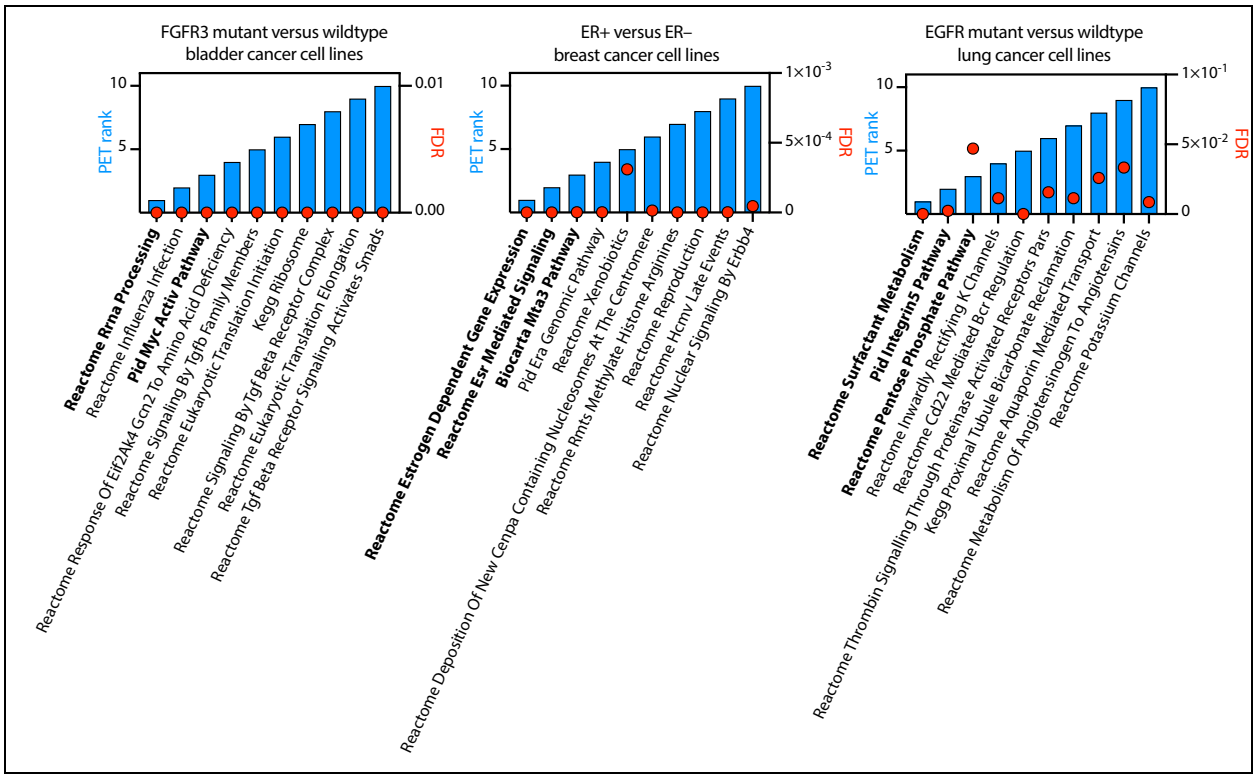

Figure R3. PET identifies relevant signaling pathways distinguishing genetic subtypes of cancers. Bar plots show the top 10 pathways reported by PET and their corresponding FDR values when comparing the indicated samples. Bold font highlights those pathways among the top three reported that are supported by literature.

Response point 3:

The authors in their response state that they compared "overall survival not the disease free survival", however my comment was regarding disease specific survival. They then go on to state we compared patients who died of disease. This latter statement implying they used disease specific survival (DSS) and not overall survival (OS), although this wording may seem trivial, it could make differences as OS treats a person who dies of a heart attack the same as one who died because of the cancer.

We apologize for the confusion in our initial response and appreciate the opportunity to clarify the survival metrics used in our analysis. In our study, we utilized "overall survival" (OS) rather than "disease specific survival" (DSS). OS considers all causes of mortality, whereas DSS focuses solely on deaths attributed to the cancer being studied. Our choice of OS was guided primarily by its comprehensive availability across all 12 TCGA cancer datasets and its relevance to our broad analysis objectives. We acknowledge that using OS instead of DSS may present certain limitations, as OS does not differentiate between cancer-related deaths and those from other causes. This distinction could potentially impact the interpretation of survival-related findings. However, we believe that OS still provides valuable and relevant insights into patient outcomes in the context of our study. To ensure clarity and consistency, we have updated the terminology throughout the manuscript to specifically refer to "overall survival". This change is intended to eliminate any ambiguity and accurately reflect the survival metric used in our analyses. Despite the broader scope of OS, we are confident that our findings offer significant contributions to understanding survival outcomes in the cancer types studied.

I then asked if they could validate the data in another cohort, as TCGA has poor follow up. They bring that point to the forefront in that the alive vs dead follow (assuming they are referring to the median) was 680 vs 563 days or only 1.8 years vs 1.5 years. To address the point, they mention other tumors type, however again, this misconstrues the point, can these pathways be validated in other bladder datasets with survival, such as UC-GENOME, IMvigor, ABACUS, or PURE-01.

We apologize for unintentional misunderstanding of the question and thank the reviewer for their clarification. We agree the reviewer's emphasis on the need for validation in independent datasets with robust survival data, beyond the TCGA cohort. As suggested, we have now further evaluated the biomarkers (please refer to response point #5 below) and pathways identified by PET from the TCGA cohort in muscle invasive bladder cancer (MIBC) samples from three other independent datasets, namely GSE160693 (Kardos, Rose et al.; PLoS One; 2020), GSE13507 (Lee, Leem et al.; J Clin Oncol; 2010) and GSE48276 (Choi, Porten et al.; Cancer Cell; 2014).

In this validation, we tested whether pathways identified by PET as associated with unfavorable outcomes in TCGA-BLCA overlap with those identified in these independent datasets. For

comparative rigor, we also included analyses using the piano ensemble method (with DESeq2 statistics), which was the best performing ensemble method (other than PET), optimized GSEA-NES, ora, and Enrichr. PET results showed significant overlap with all three datasets (**Fig. R4**), further validating PET-derived pathways.

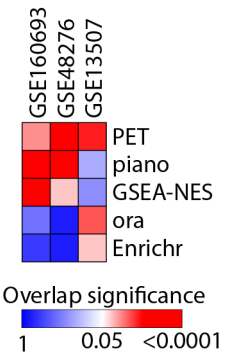

Figure R4. The significance of the overlap between the top reported pathways associated with unfavorable outcomes in TCGA-BLCA versus three independent datasets using the indicated methods. Fisher exact test p-values are displayed.

Unfortunately, we were unable to access ABACUS and PURE-01 due to controlled access restrictions, and the UC-GENOME dataset deposited in Cbioportal was not suitable due to its z-score normalization across genes, which is incompatible with our sample-wise analysis requirements. Nevertheless, the validation performed with the three accessible independent datasets, including one suggested by the reviewer, provides strong proof-of-principle support for our findings. These are now included as new **Fig. S4c** in the revised manuscript, and accompanied by the descriptor in the text, "As proof-of-principle, we validated the reported pathways associated with unfavorable outcome in bladder cancer in three independent cohorts (**Fig. S4c**)" (page 8 lines 223-224).

Response Point 4:

We understand that the AUC was greatest for the multi gene combination but was it significantly different that the single predictors.

We apologize for the omission of a direct statistical comparison in our initial response letter. To address this, we have conducted Mann-Whitney U rank test to compare the AUCs between single gene predictors ("1-core gene") and multi-gene combinations ("5-core genes") and between these and all genes. The results consistently demonstrate that the AUCs for multi-gene combinations are significantly higher than those for single-gene predictors across all analyzed cancer types. For instance, in TCGA-BLCA, the p-value for the comparison between 1-core gene vs. 5-core genes is 5.19E-39, indicating a highly significant improvement in AUC with multi-gene combinations. This pattern of significant improvement is consistent across the board, as detailed in the table below, and now included as **Table S2d** in the manuscript. We have also updated the **Fig 2d** legend to indicate that the comparisons are statistically significant when compared against all genes or single core genes (page 27 lines 716-718). These findings highlight the superior predictive power of multi-gene combinations over single-gene predictors in our analysis, reinforcing the robustness and potential of our approach for identifying effective biomarkers.

Cancer type	1-core gene *vs.* all genes	5-core genes *vs.* 1-core gene	5-core genes *vs.* all genes
TCGA-BLCA	1.89E-34	5.19E-39	<1E-300
TCGA-CESC	9.67E-30	3.34E-53	<1E-300
TCGA-COAD	7.71E-43	2.99E-40	<1E-300
TCGA-HNSC	4.30E-30	1.41E-14	<1E-300
TCGA-KIRC	3.87E-33	2.69E-44	<1E-300
TCGA-KIRP	1.28E-53	6.01E-36	<1E-300
TCGA-LUAD	7.60E-40	3.75E-47	<1E-300
TCGA-LUSC	8.99E-11	7.32E-10	<1E-300
TCGA-PAAD	3.17E-44	3.45E-55	<1E-300
TCGA-STAD	3.08E-11	1.12E-12	<1E-300
TCGA-BRCA	2.03E-30	1.38E-39	<1E-300
TCGA-LIHC	3.22E-20	5.25E-47	<1E-300

Response Point 5:

Whereas the BLCA analysis was performed OS previously and biomarker gene were selected from this group, the validation was performed in E-MTAB-4321, which is a majority NMIBC and only has progression data, which is not comparable to survival data. Additionally, their results may simply be separating out HG vs LG samples, which would not be clinically relevant. If true validation is to be performed, it should be in done in a MIBC, as mentioned above.

We appreciate the reviewer's critical assessment of our validation dataset choices. Upon review, we acknowledge that the E-MTAB-4321 dataset, predominantly consisting of non-muscle invasive bladder cancer (NMIBC) samples with progression data, is not suitable for direct comparison to overall survival data from TCGA's muscle invasive bladder cancer (MIBC) samples. Consequently, we have excluded E-MTAB-4321 from our validation analysis.

To ensure a more robust analysis, we have focused on datasets with survival data on MIBC samples, specifically GSE13507 (Lee, Leem et al.; J Clin Oncol; 2010) and GSE48276 (Choi, Porten et al.; Cancer Cell; 2014), along with GSE160693 (Kardos, Rose et al.; PLoS One; 2020), which we had previously included (**Fig S2f in original submission**) (updated **Table S2g**). In these datasets, we found that 71%, 83% and 73% of the TCGA-derived biomarkers, respectively remained consistent in predicting survival outcomes, underscoring the robustness of our biomarker identification approach using PET.

Notably, when comparing these results to combination biomarkers derived solely from conventional DEGs, the consistencies dropped to 18%, 42%, and 53%, for the respective datasets.

This contrast highlights the potential of our PET-driven approach in identifying more reliable and consistent prognostic biomarkers in MIBC samples from independent datasets. These findings are now incorporated in **Fig. S4g,** which is reproduced below for convenience as **Fig. R5**.

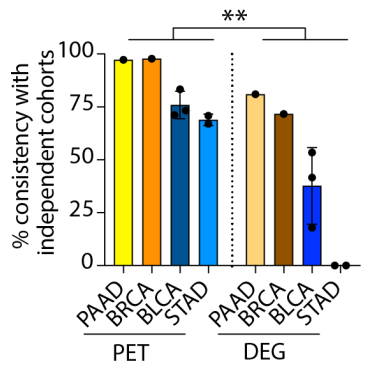

Figure R5. The level of consistency between significant prognostic biomarkers in TCGA and independent cohorts (see **Methods**). The prognostic biomarkers derived from PET pathways or differentially expressed genes (DEGs) are separated. Independent cohorts are listed in **Table S2g**. Error bars show mean ± s.d. **p<0.01 derived from two-tailed paired t-test. PAAD, pancreatic adenocarcinoma; BRCA, breast invasive carcinoma; BLCA, bladder carcinoma; STAD, stomach adenocarcinoma.

Response Point 6:

In this comment, I was intended to ask how the unfavorability score was distributed among the TCGA or Consensus Subtypes. Also, does the score do a better job at delineating survival then subtype alone. Their use of a comparison from another paper to illustrate that the basal/luminal subtypes are not accurate for prognosis, is not appropriate and the analysis should be performed by the authors on their same data using a published and validated method such as the packaged from Kamoun et. al. or signatures from Damrauer el. al or Choi et. al.

We apologize for any confusion caused by our previous response and appreciate the opportunity to clarify. We recognize the significance of molecular subtypes in bladder cancer prognosis and treatment. While we did not intend to imply that the basal/luminal subtypes are inaccurate for prognosis, we recognize that our choice of wording may have been inapt. We assume that the reviewer is referring to the section where we had mentioned, “our top combination biomarker for both favorable and unfavorable prognosis in this setting were not simply the markers distinguishing basal from luminal molecular sub-types but provided much more accurate prognosis (hazard ratio ~ 4.6 and logrank p-value <0.0001; Fig. S5a) compared to the current method of basal-luminal molecular subtyping”. As such, to prevent any misinterpretation, we have now removed this statement entirely.

Nevertheless, to address the reviewer’s question, we have conducted an analysis to evaluate the distribution of the unfavorability score across original TCGA (Cancer Genome Atlas Research; Nature; 2014) and Consensus Subtypes (Kamoun, de Reynies et al.; Eur Urol; 2020) within our dataset. This analysis included plotting the unfavorability scores among TCGA samples by their molecular subtypes and assessing the impact of these scores on survival outcomes compared to subtype classification alone. Our findings, presented in the new **Fig. R6**, reveal how the unfavorability score correlates with the established subtypes and its potential to provide insights into survival prognosis.

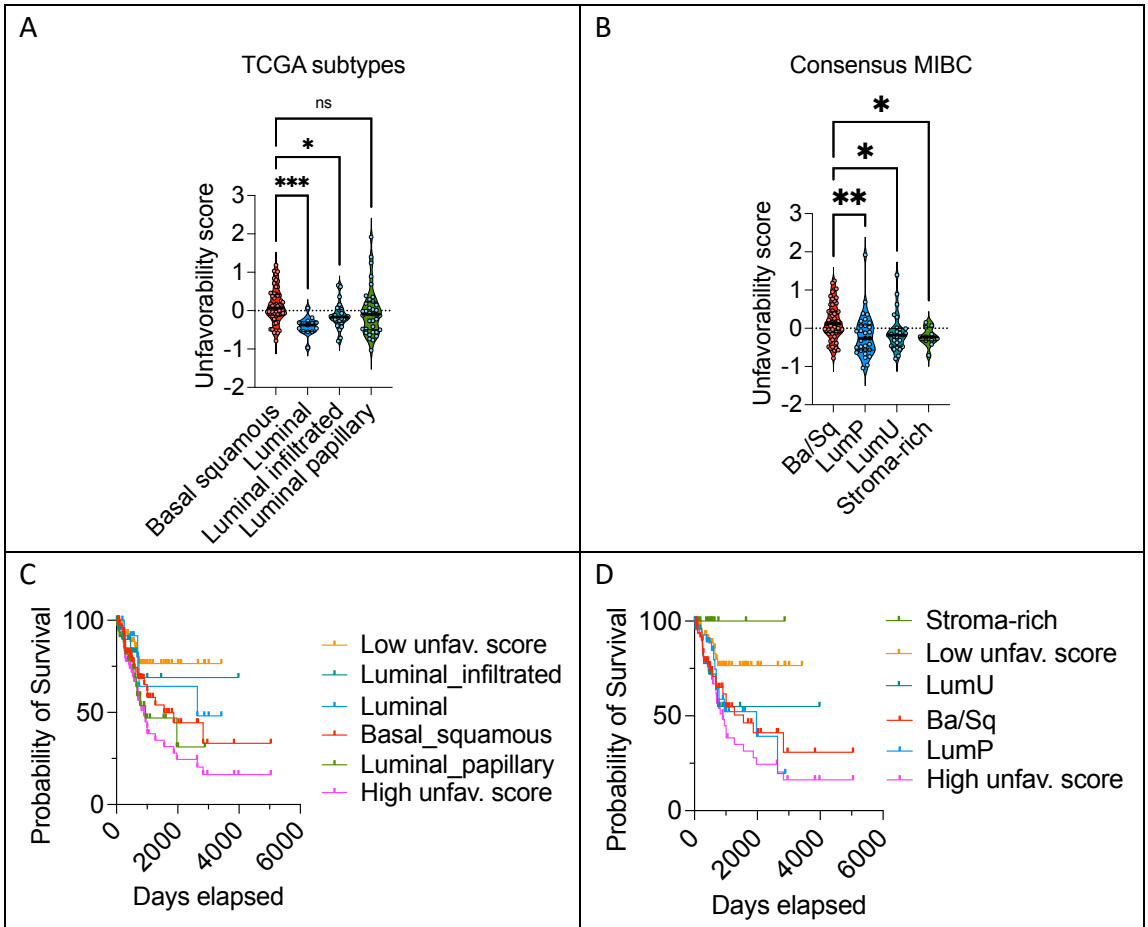

Figure R6. Unfavorability score distribution among bladder cancer subtypes. A-B) The distribution of the unfavorability (unfav.) score among the TCGA samples separated by TCGA or Consensus Subtypes. The unfavorability scores are calculated as the mean z-score expression of leading genes identified to be associated with an unfavorable outcome in each sample (genes are shown in **Table S2e**). The higher the unfavorability score the more unfavorable is the cancer prognosis (overall survival). P-values are by one-way ANOVA test. ns non-significant; * p<0.01; ** p<0.001; *** p< 0.0001. Ba/Sq = basal/squamous; LumP = luminal papillary; LumU = luminal unstable. **C-D**) Overall survival stratification based on subtypes (TCGA or Consensus) or unfavorability scores.

Response Point 8:

The Authors did not address my question as to whether the T24 and UMUC3 cells were favorable or unfavorable. Does the CDK9 effect favorable and unfavorable cell line equally? Likewise, based on depmap J82 and 5637 cells have high CDK9 expression, whereas UBLC1 has low expression, are they predicted to be unfavorable and favorable, respectively.

We apologize for not fully addressing the reviewer's previous question regarding the favorability or unfavorability of the T24 and UMUC3 cell lines. As suggested by the reviewer, we have now calculated “unfavorability scores” for all bladder cancer cell lines available in CCLE, including T24

and UMUC3, and compared against CDK9 expression. The unfavorability score for each sample is calculated as the mean z-score expression of leading genes identified to be associated with an unfavorable outcome in BLCA (genes are shown in **Table S2e**). The higher the unfavorability score the more unfavorable is the cancer prognosis. This analysis indicates an overall association between CDK9 expression and unfavorability score (Pearson correlation 0.71; p-value = 0.006; **Figure R7**). Within this framework, the T24 cell line has both a higher expression of CDK9 and a higher unfavorability score compared to the UMUC3 cell line, which in turn has higher expression of CDK9 and a higher unfavorability score compared to the UBLC1 cell line. Thus, T24 cells are highly unfavorable and UMUC3 cells medium unfavorable cell lines, as shown in **Fig. 5a** of the original submission.

To expand on this, we referenced publicly available data from the GDSC database where a range of BLCA cell lines were treated with a different CDK9 specific inhibitor (i.e. CDK9_5038) and IC50 values were calculated (Yang, Soares et al.; Nucleic Acids Res; 2013). Despite the difference in inhibitors, we observed a trend where cell lines with higher unfavorability scores required higher concentrations of the inhibitor to achieve similar inhibition of proliferation (Pearson r=0.54, p<0.06). Addressing the reviewer's point, UMUC3 and UBLC1 cells were not available in this study, but the analysis did include cell lines with comparable unfavorability scores, including HT1197 (unfavorability score = 1.05, which is similar to UMUC3; CDK9i IC50 = 0.56uM), SW1710 (unfavorability score = 0.96, which is similar to UMUC3; CDK9i IC50 = 0.16uM) and 647V (unfavorability score = 0.78, which is similar to UBLC1; CDK9i IC50 = 0.16uM). The trend observed suggests a broad efficacy of CDK9 inhibitors across BLCA cell lines, with efficacy correlating with unfavorability scores. These findings are shown below as **Fig. R7** and are now presented in the revised manuscript as new **Figure 5b**, together with the above description (pages 14-15, lines 396-404), and in **Table S5a**.

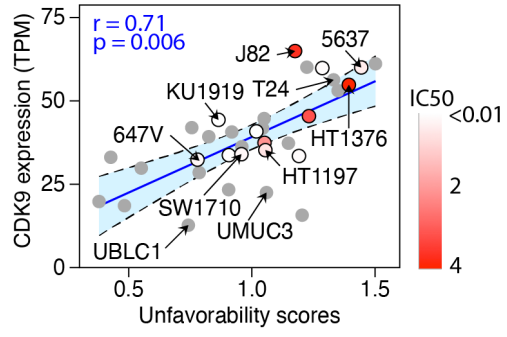

Figure R7. CDK9 expression is associated with unfavorability scores among bladder cancer cell lines. The scatter plot displays expression of the CDK9 gene versus the unfavorability score for each bladder cancer cell line in the CCLE database. The color scale represents IC50 values reported by the GDSC database for the CDK9-specific inhibitor CDK9_5038. Cell lines for which IC50 values are not available are highlighted in gray. The Pearson correlation value and corresponding p-value are shown in blue. The trend line and 95% confidence interval are highlighted and select cell lines are labeled.

Reviewer #3 (Remarks to the Author):

The authors should be commended for their very well-structured efforts to address the reviewers' concerns. I still have, though, a few points of disagreement and suggestions, which I enumerate below.

We thank the reviewer for their positive and constructive feedback on our efforts to address their concerns. We are grateful for the opportunity to improve and refine our work based on the remaining suggestions, which we have addressed below.

1. In your response letter you cleared up that the reason why integrating the results of GSEA, EnrichR and ORA with PET is superior as any other ensemble method that also integrates them, is because PET optimizes the input parameters of GSEA, EnrichR and ORA. Looking at the code in the Jupyter notebook 'run_PET_tutorial' and the GSEA documentation in https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html I would say that PET is using the default parameters for GSEA. The other two methods, EnrichR and ORA, use the exact Fisher's test, and I don't see much to optimize there either. The only place I found in the manuscript where an optimization seems to be described is in page 5, where it says that a geneset size of 200 genes was near optimal for both GSEA and Enrichr, but I do not see how a fixed number of called DE genes can be optimum irrespective of the dataset and the experimental design that generated it, usually genes are called DE on the basis of cutoff values for FDR and/or magnitude of the fold change. As a side note, the rationale of using an ensembl method is normally to integrate disparate methods that implement "orthogonal" approaches, while ORA and EnrichR are based on the same statistical device, the Fisher's exact test. It would be therefore interesting to know why you integrate two methods based in the Fisher's exact test.

We appreciate the reviewer's insightful observations and would like to provide further clarifications on our methodological choices in PET. Concerning GSEA, while a pre-ranked statistics option (e.g. from DESeq2) is available, most researchers commonly use the default signal2noise metric in the GSEA GUI. Through benchmarking, we discovered that DESeq2-based pre-ranking yielded more optimal results. This discovery influenced our choice within PET, ensuring that users benefit from this optimal setting.

For ORA and EnrichR, we acknowledge their shared use of the Fisher exact test. However, in the case of EnrichR, we utilized the “combined score” output, which differs from Fisher exact statistics by incorporating additional background correction through z-score permutation. This methodological distinction, coupled with the diverse ranking of the top pathways by ORA and EnrichR observed during benchmarking, influenced our decision to integrate both into PET. This integration allows us to leverage the unique strengths of each approach, enhancing the performance of PET.

Regarding the fixed number of DEGs, our analysis established that utilizing the top 200 statistically significant DEGs was near-optimal for gene set size (**Fig. S1d**). This decision was based on empirical evidence indicating that including more DEGs often led to diminished performance,

possibly due to the incorporation of “less reliable” DEGs that introduced noise into the genesets. We acknowledge that this may not be universally optimal across all datasets and experimental designs. However, in the context of our study (n>200 genesets), this approach provided a balance between inclusivity and accuracy. Nevertheless, we have also included an option in the PET pipeline for customizing DEG selection. Users can choose to set either a p-value/adjusted value threshold or specify the number of top differential genes to define a DEG list for investigation.

In summary, our methodological choices in PET were guided by extensive benchmarking and aimed at optimizing the pathway analysis performance. We believe that these choices, while sometimes unconventional, are justified by the improved outcomes demonstrated in our study.

2. In the light of the previous comment, the main difference between PET and EGSEA must be in how the DE genes (what it's called input genesets -IGS in the manuscript) and their statistics are derived. PET is using DESeq2, while EGSEA uses limma. This is in fact highlighted in part 1.1 of the response letter, where it says "EGSEA uses traditional methods that were originally intended for microarray analysis and forces the use of Limma for differential gene expression analysis, which is also designed for microarray data, and therefore may not be optimal for pathway discovery from RNA-seq data". This statement is wrong, limma provides procedures to analyze both, microarray and RNA-seq data, the latter concretely through the limma-voom pipeline, which is also widely used with RNA-seq data (the corresponding paper Law et al. (2014), https://doi.org/10.1186/gb-2014-15-2-r29 accumulates nearly 5,000 citations) and is the one used by EGSEA. So, my current guess is that the main difference between PET and EGSEA is in the way in which the DE statistics are derived, which means that the superior performance of PET cannot be attributed to an innovative way of integrating pathway analysis methods, but to a more careful way of performing a DE analysis with two replicates per condition (GLMs on negative binomial distributions are probably better suited for limited replication than limma-voom).

We appreciate the reviewer’s accurate observation regarding the limma pipeline and its capabilities in RNA-seq analysis. We apologize for our oversight in our previous statement. The limma-voom pipeline is indeed adept at analyzing RNA-seq data, as evidenced by its widespread use and citations. Indeed, all the differential expression analyses of EGSEA in the manuscript were conducted via the default option, limma-voom.

The reviewer is correct in pointing out that the differential analysis methodology is a crucial component, with DESeq2 playing a significant role in ensuring the robust performance of PET in contexts of limited replication. However, notably, it is not the sole factor contributing to the enhanced performance of PET. PET’s distinct advantage also lies in its selection of underlying pathway analysis tools and the optimization of their settings. For instance, our comparative analyses demonstrate that even other ensemble methods (e.g., PIANO, Enrichment Browser), which utilize the same DE statistics as PET, do not achieve the same level of performance (please see our response to reviewer #1 point #1). This indicates that PET's method selection, combined with its optimizations of settings, also plays a significant role in its superior efficacy. These aspects of PET, along with the choice of DESeq2 for differential expression analysis, contribute to a more effective pathway ensemble strategy.

3. Embedding DE methods within pathway analysis may seem appealing, but it has the challenge that is not entirely trivial how to accommodate the flexibility of the embedded tool. Note that PET is interfacing DESeq2 for a two-group comparison only, so more complex experimental designs (i.e., adjusting for an extra covariate) are not feasible. In this sense, EGSEA overcome that limitation by allowing one to pass a design matrix to limma, you may want to consider this possibility too with DESeq2.

We appreciate the reviewer for this excellent and important suggestion. We have now incorporated the option to pass the design matrix and contrasting samples to DESeq2.

4. Regarding the limited sample size used to find DE genes, you argue that utilizing 2-3 samples in each group yields an expected false positive rate (FPR) of 5%, illustrated in Figure S1g. The FPR=FP/(FP+TN), so it has on its denominator the number of true negative (TN) instances, which in our context (truly non-DE genes?) is much larger than the number of false positive (FP) instances, and therefore it is relatively easy to get a low FPR. A much relevant parameter in this context is the false discovery rate (FDR). In fact, the performance of DE methods has been thoroughly investigated in the past, for instance in the work of Schurch et al. RNA, 2016 (https://10.1261/rna.053959.115), they used a clean yeast data set generated for benchmarking purposes and they show that while most of the DE methods successfully control their FDR < 5% with as low as 3 replicates (note that this study is using two, so we cannot infer from that work whether 2 replicates also allow one to control the FDR), only 20%-40% of the expected DE genes are found. So, it is hard to believe that the target gene sets (TGS) used for benchmarking constitute a reliable gold standard to rank methods.

We thank the reviewer for this feedback. In **Figures S1g-h**, there was a mislabeling of the false discovery rate, which represents the proportion of falsely reported significant pathways. We concur with the reviewer that having more than two replicates offers an advantage in better controlling the FDR, as evidenced in **Fig. S1g** by a trend in FDR being lower when using 2 samples (FDR ~ 7%) versus 5 samples (FDR ~ 5%).

In our analyses, the average number of overlapping genes among related gene sets (n=16) compared to unrelated gene sets (n=1.3) across cell lines, obtained from 2 replicates (**Fig. S1b** and **S1f**), suggests that this sample size might be sufficient for accurately distinguishing related from unrelated gene sets. Nevertheless, the reviewer's point is an important one. To acknowledge this, we have added the following statements in our discussion: "The number of replicates in each group is an important consideration in controlling for false discovery rates. Our Benchmark is generally constructed from 2 replicates per group available in ENCODE. However, the average number of overlapping genes among related gene sets (n=16) compared to unrelated gene sets (n=1.3) across cell lines, obtained from 2 replicates (**Figure S1b** and **S1f**), suggests that this sample size is sufficient for accurately distinguishing related from unrelated gene sets. Future studies will extend the Benchmark to include additional genesets and replicates" (page 17 lines 466-471).

5. I appreciate the efforts of the authors to improve the documentation. I believe the installation and use of the software could be made somewhat simpler if the authors would attempt to implement PET in a single platform, either Python or R. For Python, the authors could use the recent implementation of DESeq2 in Python by Muzellec et al. (2023, https://doi.org/10.1093/bioinformatics/btad547), and GSEApy (https://github.com/zqfang/GSEApy) for GSEA and even EnrichR. For R, the authors could use the Bioconductor package fgsea for GSEA and the CRAN package enrichR for EnrichR.

We thank the reviewer for their suggestion. We have now incorporated the option to use PyDESeq2 by PET, allowing users to choose this option and have the entire platform in Python. Some users have observed discrepancies between results from PyDESeq2 and DESeq2 in R. Therefore, we have retained both options and kept the R version as the default. As suggested by the reviewer, we have also included GSEApy in PET as another option for running GSEA method. To acknowledge this, we have added the following statements in our method section: “PET implementation in python incorporates the option to use DESeq2 (Love, Huber et al.; Genome Biol; 2014) or PyDESeq2 (Muzellec, Telenczuk et al.; Bioinformatics; 2023), as well as GSEA (Subramanian, Tamayo et al.; Proc Natl Acad Sci U S A; 2005) or GSEApy (Fang, Liu et al.; Bioinformatics; 2023), providing flexibility in selecting the platform for their analysis” (page 23 lines 603-604).

Reviewer #4 (Remarks to the Author):

Most of my previous review comments were properly addressed. However, two issues still exist and need clarification.

We thank the reviewer for their positive and constructive feedback. We are grateful for the opportunity to improve and refine our work based on the remaining issues, which we have addressed below.

1. Line 119-124: "Importantly, Benchmark was designed to simulate a common scenario where the investigated genesets (i.e. IGSs) are derived from different sources (e.g. cell lines, species, conditions) than established biological pathways (i.e. TGSs). To achieve this, Benchmark was created such that genesets from the same TF, RBP or gKD but from distinct cell lines are better matched to each other than any other genesets. As an example, the IGS that contains STAT1-bound genes in K562 cells is more closely related to the TGS that contains STAT1-bound genes in HepG2 cells than any other TGS."

I raised this issue at the previous review that this approach will identify common genesets shared by those cell lines/species/conditions used to generate Benchmark, but miss those genesets/pathways unique to a specific cancer/cell line/species/condition. Similar to the comments raised by Reviewer #3 during the previous round of review as well, this approach makes the assumption that the alteration of one specific gene in two different cell lines perturb sets of genes that are more alike than with genesets derived from alterations of other genes. Alteration of one specific gene, but at different genetic background/in different cancers, frequently has different biological effects. This is frequently seen at clinic that the same targeted therapy may be effective in some cancer types, but not in others even when they contain the same target gene alterations. Even within the same cancer type, targeted therapy works in some patients, but not in others even when their cancers harbor the same alterations. For example, the FGFR2/3 inhibitor erdafitinib has been approved for bladder cancer harboring FGFR2/3 activation mutation or translocation. However, the response rate is only 40%. This suggests that the same FGFR2/3 alterations can induce different geneset alterations in different bladder cancers. In this specific case, Benchmark may identify genesets/pathways shared by K562 and HepG2. But will likely miss many unique to each cell. For example, perturbation of the Bcr-Abl will affect K562, but is unlikely to affect HepG2.

We acknowledge the reviewer's concerns about Benchmark potentially overlooking unique, condition-specific pathways and its assumption regarding the uniform effect of gene alterations across different cell lines or conditions. Benchmark was designed primarily to assess the precision of pathway analysis methods in identifying common gene sets across varied conditions, leveraging over 1,000 real-world biological datasets. Benchmark excels in this aspect, allowing us to optimize the running parameters of existing methods and to construct our own ensemble method, PET, for conducting biological pathway analysis. Benchmarking of PET revealed that it possesses much higher precision and accuracy than existing single method and ensemble tools.

Deployment of PET provided proof-of-concept insights on the first submission that pointed to a novel therapeutic for bladder cancers, which we validated experimentally.

We do agree with the reviewer and recognize the importance of being able to detect unique pathways specific to particular conditions or cancers. To address this, we have validated the ability of PET to capture both common and unique pathways. This was partly in response to concerns raised by reviewer #2 (please refer to reviewer #2: Rebuttal point 1.2 for the full response) and further exemplified by Reviewer #4's example of FGFR2/3 in bladder cancers. As proof-of-principle, we conducted pathway analysis using PET on FGFR3 mutant vs. WT bladder cancer (BLCA) cell lines, ER^+ vs. ER^- breast cancer (BRCA) cell lines, and EGFR mutant vs. WT lung cancer (LAUD) cell lines available in CCLE, assessing all canonical pathways annotated in MSigDB release 7.2 (Liberzon, Subramanian et al.; Bioinformatics; 2011). These analyses demonstrated that, in all three cases, PET effectively identifies top-reported pathways aligning with known differences in these cancers, as illustrated in the new **Figure S2**. This suggests that PET, benefiting from the optimization process using Benchmark, is adept at uncovering unique gene sets or pathways specific to certain cancer types or subtypes. We agree with the reviewer that continued, extensive validation in real-world settings is essential. This ongoing research, driven by the broader scientific community, will be essential for further substantiating the capabilities of PET in multiple experimental contexts over the coming years.

2. I assume that the elbow plots at Figure S5c was obtained from combined data of many bladder cancer cell lines/clinical tissues. Based on this elbow plots, his article then determined the drug efficacy in three cell lines, cervical cancer Hela, bladder cancer T24 and UMUC3. Can PET be used to generate the elbow plots specifically for each of these three cell lines? Are the elbow plots of these three cell lines the same as the one shown at Figure S5c? If PET can be used to generate specific elbow plots for those cell lines, drug sensitivity assays should be done based on the drugs shown on the specific elbow plots and determine whether PET can predict drug sensitivity and guide drug selection.

We thank the reviewer for the question regarding the generation and application of elbow plots in PET. **Figure S5c** was indeed generated by ordering drugs with known transcriptional targets according to their capacity to downregulate genes associated with an unfavorable outcome and upregulate genes linked to a favorable outcome. This ranking is derived from statistical overlap between drug targets and genes that are desired to be modulated. Regarding the generation of elbow plots for individual cell lines, such as T24 and UMUC3, it is important to note that the approach used for clinical samples may not directly apply. In clinical samples, we compared cancer patients who survived versus those who did not, enabling us to identify genes for targeted modulation. For individual cell lines, such a comparison is not feasible, which poses a challenge in creating a meaningful ordering of drugs based on their transcriptional impact.

Nevertheless, based on the advice from reviewer #2 (point #8), we have explored an alternative approach. We calculated unfavorability scores (indicating a poorer outcome) for all bladder cancer cell lines available in CCLE and examined their association with CDK9 expression. Our findings showed a significant correlation between high unfavorability scores and CDK9 expression.

Subsequently, we tested the efficacy of a CDK9 inhibitor on T24 and UMUC3 cells, which showed effectiveness for both of these cell lines, which had high and moderate unfavorability scores, as well as CDK9 expression, respectively. While this approach does not directly involve specific elbow plots for each cell line, it demonstrates the utility of PET in associating drug efficacy with gene expression profiles and unfavorability scores. This finding suggests the potential of PET in predicting drug sensitivity and guiding drug selection, a capability we aim to further develop and validate in future studies.

Conclusion

In summary, in response to the insightful and constructive comments of the reviewers, we have performed several additional analyses and provided clarifications that support and add rigor to our original conclusions. We appreciate the collaborative peer-review process, which has undoubtedly elevated the quality of our manuscript. We believe our work, enhanced by thus thorough review, holds significant value for the readership of *Nature Communications* and the broader scientific community engaged in high-throughput assays. We have submitted revised versions of our manuscript, including marked and clean copies, and extend our gratitude to the editorial team and reviewers for their guidance and support in refining our research. We look forward to your response.

List of References:

Arabi, A., et al. (2005). "c-Myc associates with ribosomal DNA and activates RNA polymerase I transcription." Nat Cell Biol **7**(3): 303-310.

Cancer Genome Atlas Research, N. (2014). "Comprehensive molecular characterization of urothelial bladder carcinoma." Nature **507**(7492): 315-322.

Choi, W., et al. (2014). "Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy." Cancer Cell **25**(2): 152-165.

Clusan, L., et al. (2023). "A Basic Review on Estrogen Receptor Signaling Pathways in Breast Cancer." Int J Mol Sci **24**(7).

Fang, Z., et al. (2023). "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python." Bioinformatics **39**(1).

Fujita, N., et al. (2003). "MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer." Cell **113**(2): 207-219.

Grandori, C., et al. (2005). "c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I." Nat Cell Biol **7**(3): 311-318.

Grewal, S. S., et al. (2005). "Myc-dependent regulation of ribosomal RNA synthesis during Drosophila development." Nat Cell Biol **7**(3): 295-302.

Hasegawa, Y., et al. (2015). "Surfactant protein D suppresses lung cancer progression by downregulation of epidermal growth factor signaling." Oncogene **34**(32): 4285-4286.

Inoue, A., et al. (2008). "Suppression of surfactant protein A by an epidermal growth factor receptor tyrosine kinase inhibitor exacerbates lung inflammation." Cancer Sci **99**(8): 1679-1684.

Kamoun, A., et al. (2020). "A Consensus Molecular Classification of Muscle-invasive Bladder Cancer." Eur Urol **77**(4): 420-433.

Kardos, J., et al. (2020). "Development and validation of a NanoString BASE47 bladder cancer gene classifier." PLoS One **15**(12): e0243935.

Lee, J. S., et al. (2010). "Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors." J Clin Oncol **28**(16): 2660-2667.

Liberzon, A., et al. (2011). "Molecular signatures database (MSigDB) 3.0." Bioinformatics **27**(12): 1739-1740.

Love, M. I., et al. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Mahe, M., et al. (2018). "An FGFR3/MYC positive feedback loop provides new opportunities for targeted therapies in bladder cancers." EMBO Mol Med **10**(4).

Makinoshima, H., et al. (2014). "Epidermal growth factor receptor (EGFR) signaling regulates global metabolic pathways in EGFR-mutated lung adenocarcinoma." J Biol Chem **289**(30): 20813-20823.

Muzellec, B., et al. (2023). "PyDESeq2: a python package for bulk RNA-seq differential expression analysis." Bioinformatics **39**(9).

Ricono, J. M., et al. (2009). "Specific cross-talk between epidermal growth factor receptor and integrin alphavbeta5 promotes carcinoma cell invasion and metastasis." Cancer Res **69**(4): 1383-1391.

Rubio, K., et al. (2023). "Non-canonical integrin signaling activates EGFR and RAS-MAPK-ERK signaling in small cell lung cancer." Theranostics **13**(8): 2384-2407.

Saha Roy, S. and R. K. Vadlamudi (2012). "Role of estrogen receptor signaling in breast cancer metastasis." Int J Breast Cancer **2012**: 654698.
Subramanian, A., et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.
van Riggelen, J., et al. (2010). "MYC as a regulator of ribosome biogenesis and protein synthesis." Nat Rev Cancer **10**(4): 301-309.
Whitsett, J. A., et al. (1987). "Differential effects of epidermal growth factor and transforming growth factor-beta on synthesis of Mr = 35,000 surfactant-associated protein in fetal lung." J Biol Chem **262**(16): 7908-7913.
Yang, W., et al. (2013). "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells." Nucleic Acids Res **41**(Database issue): D955-961.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

We are glad to see that Wang et al are trying to improve their manuscript, however, critical points remain unaddressed, and there are flaws in their revised analysis:

1) The authors claim that the observed good performance of PET is due to the optimization of DESeq2 results. The better performance of DESeq2 for DEA might be interesting to mention in a benchmark setting but it is not that relevant from a tool perspective: any of these ensemble methods can accept contrast statistics results from any DEA framework. Additionally, it then becomes crucial to use the same DESeq2 results across ensemble approaches to make them comparable, which is unclear if this is the case. For example, the authors mention that they rerun DESeq2 inside Ebrowser which depending on how it is done results might be different. Furthermore, the authors have run decoupler using as input the raw gene expression values and have come up with a custom contrast score. This is wrong and probably explains its poor performance compared to the rest of ensemble methods; decoupler works with contrast level statistics from DESeq2 such as the ones they use in PET: -log10(p-value)*sign(log(fold change)).

2) All this would be clearer if the code used in this analysis was in the GitHub repository but we cannot find it there. Can authors please make their code available?

Reviewer #1 (Remarks on code availability):

the code of the benchmark is not available.

Reviewer #2 (Remarks to the Author):

I would like to thank Wang et. al. for thoroughly addressing a vast majority of my concerns. Overall, I think the manuscript is well thought out, well written and will be a valuable tool for the research community moving forward. While I am convinced of the performance of PET, I still have some reservations regarding their proof of principle examples concerning the prognostic application of PET, mostly surrounding TCGA sample selection and how that may bias the analysis. Of the below concerns, I feel that a response to #1 (and to a lesser extent 2) are the most critical, while points 3 and 4 would be helpful to the reader but would not invalidate any of the conclusions.

1) Unfortunately, I still have strong concerns regarding the method in which TCGA samples were filtered and chosen for the analysis. As an example, the table in Figure 2a states for bladder that stage II/T2 was the earliest stage and that n=84/45 (alive/dead). Based on the text, the comparison should be alive vs dead (restricted to II/T2), however in the supplementary data, all “dead” were Stage II/T2, and all alive

were Stage III/T3 or greater. Therefore, based on my reading of the manuscript, TCGA patient “TCGA-XF-AAME” had an unfavorable prognosis even though they were T2 with an OS time was 2828 days, whereas “TCGA-GC-A4ZW” had a good prognosis because they were alive with T3 disease, but only had 15 days of follow-up.
Are the same results generated if the authors created 2 cohorts, the 1st with deceased patients which had OS below the median and equal representation of all stages and a 2nd with alive patients with follow up of greater than the median and equally split between the stages?
Based on my quick sorting, the “unfavorable” group would have 45 tumors with a median survival of 158 days and a ”favorable” group with 45 tumors and a median follow up of 1156 days. This would then allow validate to be performed on the other ~300 tumors in the cohort.

2) The concern regarding tumor filtering extends to R6c/d. The survival plotted is not representative of published relationship between survival and subtype.

3) The authors clearly show the overlap between validation datasets and TCGA for the top unfavorable outcomes from PET (Figure S4c), however it would be meaningful to show not only that the pathways overlap, but that your optimal biomarker or expression score (Figure 2f) was also predictive in the validation datasets.

4) What is the authors rational for choosing the most frequently mutated gene to stratify patients? If they were trying to show the benefit of PET over mutation alone, wouldn’t the more adapt comparison be to stratify patients by the gene or genes that had a significantly different alteration frequency between good and poor prognosis?

Reviewer #3 (Remarks to the Author):

I thank the authors for the continued efforts to address the comments of this and the other reviewers. I'm satisfied with their answers, but I'm concerned with two issues I raised in my first review, which have been only partly addressed, but which I think should be considered at the editor's discretion whether they are substantial enough to preclude publication:

1. At the moment the presented software is only available through a GitHub repo, which could disappear from one day to the next. Ideally, open-source software should be distributed through a trusted repository that provides some minimum standards of versioning, archiving, licensing and testing. In the case of Python, this could be The Python Package Index (PyPI at https://pypi.org), while in the case of R, this could be CRAN at https://cran.r-project.org or Bioconductor at https://bioconductor.org.

2. Reproducing the benchmarking reported in the manuscript. The authors have provided template scripts for running all methods, but I could not find the way to run the benchmarking and obtain at least one of the figures that show the benchmarking results, such as Figures 1b or 1e. Providing templates to run any given method is certainly useful, but in my opinion, a benchmarking manuscript should provide

the full code that allows one to reproduce the benchmarking figures as easy as possible. Here, I'm not referring to scripts that read values from CSV file and generate the figures, but scripts that run the analysis and obtain the results shown in those figures. This probably should reside in a different repository as the software, and ideally in some repository from the same journal (e.g., as supplementary material) or a third-party platform (e.g., https://figshare.com) that links those scripts with the main manuscript somehow (e.g., through a DOI).

Reviewer #3 (Remarks on code availability):

The authors now provide sufficient information to successfully install and run the code. I was able to do it. However, the two concerns regarding software distribution and reproducibility that I expressed in my first review, have been only partly addressed, as I explain in this review. The editor should decide whether these concerns are substantial as to preclude publication. In my opinion, open-source software upon which scientific conclusions may be drawn, should be distributed through a trusted repository, and benchmarking manuscripts should provide the full code that produces the benchmarking results and corresponding figures. Regarding the former, the software in this manuscript is distributed through a GitHub repo that can disappear from one day to the next. Regarding benchmarking, as far as I could check, the authors provide template scripts to allow one to run a method, but not the full code that they used to produce the results and the corresponding benchmarking figures.

Reviewer #4 (Remarks to the Author):

This is Reviewer #4. In my previous round of review, I had two comments: first, the significance of genomic and epigenetic alterations is context-dependent and the same alterations have different physiological impacts in different cancers. Benchmark was designed to identify common gene sets across varied conditions. Therefore, this comment cannot be addressed with Benchmark. However, the authors did show that some common gene set alterations shared by different cancers have similar physiological significance. So I think Benchmark has the merit to be published.

Second, the elbow plots were obtained from combined data of many cancer cell lines and clinical tissues. The data from the combined analysis may not be applicable to individual cell lines/patients. This comment has not been addressed. However, the information obtained from combined data analysis can still be used for drug discovery and preclinical studies even though it may not be applicable to individual patients at clinic. Hence, it also has some merit for publication.

Reviewer #5 (Remarks to the Author): Expert in bioinformatics, computational genomics, systems biology, and drug response prediction

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewers' comments:

We appreciate the insightful comments and constructive feedback provided by all reviewers during this third round of review. We are pleased that the reception of our point-by-point response was positive. Additionally, we value the constructive suggestions specifically regarding the enhanced availability of the source code that will improve the future utility of our study.

In this response letter, we have addressed the additional concerns and suggestions provided by the reviewers. To ensure clarity and ease of review, we have detailed all changes made to the manuscript directly within this letter and have also provided a marked-up version of the manuscript, highlighting all revisions in red. We believe that the changes enhance our initial findings and extend the initial conclusions, providing a more comprehensive understanding of the work.

Reviewer #1 (Remarks to the Author):

We are glad to see that Wang et al are trying to improve their manuscript, however, critical points remain unaddressed, and there are flaws in their revised analysis:

We thank the reviewer for their positive perspective on our manuscript, apologize for not fully addressing their concerns during our previous revisions, and appreciate their diligence. We apologize for any shortcomings in our responses and are grateful for the opportunity to address the remaining points more comprehensively, and to improve and refine our work based on the reviewer's concerns, as detailed below.

1) The authors claim that the observed good performance of PET is due to the optimization of DESeq2 results. The better performance of DESeq2 for DEA might be interesting to mention in a benchmark setting but it is not that relevant from a tool perspective: any of these ensemble methods can accept contrast statistics results from any DEA framework. Additionally, it then becomes crucial to use the same DESeq2 results across ensemble approaches to make them comparable, which is unclear if this is the case. For example, the authors mention that they rerun DESeq2 inside Ebrowser which depending on how it is done results might be different. Furthermore, the authors have run decoupler using as input the raw gene expression values and have come up with a custom contrast score. This is wrong and probably explains its poor performance compared to the rest of ensemble methods; decoupler works with contrast level statistics from DESeq2 such as the ones they use in PET: -log10(p-value)*sign(log(fold change)).

We thank the reviewer for the insightful questions. The observed enhanced performance can be credited to both the ranking metrics selected for genes, such as utilizing DESeq2 results, and the underlying methods selected, as the inclusion of poor-performing approaches in ensemble methods will clearly negatively impact the overall performance. We would also like to emphasize the significance of our study beyond the development of another ensemble pathway analysis method. Our study included a benchmark to determine optimal settings such as contrast statistics for pathway analyses, a reanalysis of TCGA data focusing on identifying pathways and combination biomarkers associated with enhanced survival, and the repurposing of drugs for modulating these genes, resulting in a novel use for a CDK9 inhibitor in bladder cancer, which was experimentally validated.

The reviewer is correct in noting that we did not run decoupler properly, as suggested. We sincerely apologize for this oversight but thank the reviewer for enabling us to correct the error. Initially, we struggled to identify the correct setting and improvised with a "custom contrast", which resulted in poor performance of decoupler. After seeking advice from colleagues with more experience in decoupler usage and considering the reviewer's input, we reran decoupler using contrast-level statistics generated by the 'run_consensus' function. These statistics involved using -log10(p-value)*sign(log(fold change)) from DESeq2, which aligns with the input required by other ensemble methods, including PET. Indeed, with the proper contrast statistics, decoupler performs as the second best method after PET, with its performance metrics reaching P@10=76% and AP@10=69%, comparable to but slightly inferior to PET's P@10=82% and AP@10=73%, respectively (**Fig. R1A**). In comparison to PET, decoupler was also more susceptible to noise (**Fig. R1B**). Decoupler results are now updated throughout in **Figs. 1b, 1e**, and **S1a** and the manuscript text, with **Figs. 1b** and **1e** replicated below as **Fig. R1** for convenience.

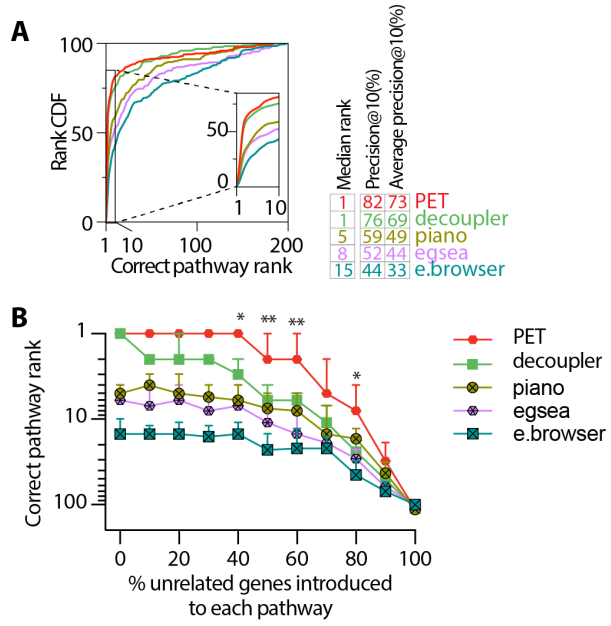

Figure R1. PET comparison with existing pathway ensemble methods A) cumulative distribution function (CDF) plots of correct pathway ranks comparing indicated pathway analysis methods using DESeq2 as ranking statistics. Median ranks for correct pathways, the fraction of correct pathways among the top 10 reported pathways (precision@10, P@10, and average precision@10, AP@10) are shown. **B)** line plots showing the performance of various methods as increasing percentage of unrelated random genes ("noise") are introduced to target genesets. All methods were run with the same optimal settings as in (**A**). This evaluates the robustness of methods to inaccuracies (or "noise") in target genesets. Lines and error bars show median + 95% confidence interval from n=100 iterations. *p<0.05; **p < 0.01 by two-sided Wilcoxon rank sum test corrected for multiple hypothesis testing, comparing PET against decoupler.

Collectively, we evaluated five ensemble-based methods using Benchmark, namely PET, Decoupler, PIANO, EGSEA, and Enrichment Browser (e.browser), alongside 10 individual methods. Topping the list were PET, Decoupler, PIANO, and EGSEA respectively, with Decoupler achieving the highest P@10 and AP@10 of existing methods, utilizing the optimal contrast-level statistics from DESeq2 results. Of note, the exact same statistics were provided as input for PIANO, but for EGSEA, a customized contrast level was not available without modifying the source code. Interestingly, the e.browser also employs DESeq2 statistics, and upon examination of its source code, we confirmed that it invokes DESeq2 identically to what we used for PET or Decoupler.

However, to the best of our knowledge, the reason for e.browser's suboptimal performance is its negligence of the directionality of pathway enrichment when ranking the relevant pathways.

2) All this would be clearer if the code used in this analysis was in the GitHub repository but we cannot find it there. Can authors please make their code available?

We apologize for missing this earlier. The scripts for executing and evaluating various methods are now deposited on GitHub (https://github.com/hedgehug/Benchmark). We have also improved the instructions for running PET (https://github.com/hedgehug/PET) and additionally deposited the entire project to Figshare (https://doi.org/10.6084/m9.figshare.c.7252324) to ensure long-term accessibility via DOI.

Reviewer #2 (Remarks to the Author):

I would like to thank Wang et. al. for thoroughly addressing a vast majority of my concerns. Overall, I think the manuscript is well thought out, well written and will be a valuable tool for the research community moving forward. While I am convinced of the performance of PET, I still have some reservations regarding their proof of principle examples concerning the prognostic application of PET, mostly surrounding TCGA sample selection and how that may bias the analysis. Of the below concerns, I feel that a response to #1 (and to a lesser extent 2) are the most critical, while points 3 and 4 would be helpful to the reader but would not invalidate any of the conclusions.

We thank the reviewer for their thoughtful review and constructive feedback. We are delighted to hear that our second revisions have addressed most of their comments and concerns which undoubtedly enhanced the quality of the manuscript. We are grateful for the opportunity to improve and refine our work based on the remaining issues, which we have addressed below.

1) Unfortunately, I still have strong concerns regarding the method in which TCGA samples were filtered and chosen for the analysis. As an example, the table in Figure 2a states for bladder that stage II/T2 was the earliest stage and that n=84/45 (alive/dead). Based on the text, the comparison should be alive vs dead (restricted to II/T2), however in the supplementary data, all "dead" were Stage II/T2, and all alive were Stage III/T3 or greater. Therefore, based on my reading of the manuscript, TCGA patient "TCGA-XF-AAME" had an unfavorable prognosis even though they were T2 with an OS time was 2828 days, whereas "TCGA-GC-A4ZW" had a good prognosis because they were alive with T3 disease, but only had 15 days of follow-up. Are the same results generated if the authors created 2 cohorts, the 1st with deceased patients which had OS below the median and equal representation of all stages and a 2nd with alive patients with follow up of greater than the median and equally split between the stages? Based on my quick sorting, the "unfavorable" group would have 45 tumors with a median survival of 158 days and a "favorable" group with 45 tumors and a median follow up of 1156 days. This would then allow validate to be performed on the other ~300 tumors in the cohort.

We thank the reviewer for the questions and great suggestions. We sincerely wish we had thought of conducting our live/dead sample selections according to the reviewer's suggestion, as

it is more accurate and intuitive. To test the impact of the suggested sample selection on identifying pathways, biomarkers, and drugs for bladder cancer samples, we followed the reviewer's advice. We classified deceased samples as those with survival times less than the median, and alive samples as those with days to last follow-up more than the median of all samples. We then used PET to identify the prognostic pathways and pathway-guided biomarkers, and importantly ranked drugs with the potential to normalize these new markers. We found a significant overlap between the original prognostic pathways and biomarkers with those from new sample selections. Specifically, approximately 50% of them were shared between both selection methods (overlap p-value <0.0001). With respect to drug prediction, where most of our validation experiments focused, 4 out of the previous top 5 remained within the top 5 with the new selection method. Importantly, the CDK9 inhibitor 0175029 remained the top predicted drug for normalizing the new set of prognostic biomarkers. This aligns with our findings that PET is robust to the presence of "noise" in input genesets. The only different drug in the top 5 in the new sample selection method was vinblastine. Our collaborators and others have previously shown that vinblastine has moderate efficacy in dogs with bladder cancer as stand-alone or in combination with other drugs (PMID: 22092632; PMID: 37624316). Taken together, these analyses indicate that sample selection is an important consideration, as noted by the reviewer, but that the conclusions from our original analyses remain robust. The results with the suggested selection are provided in **Fig. R2** below.

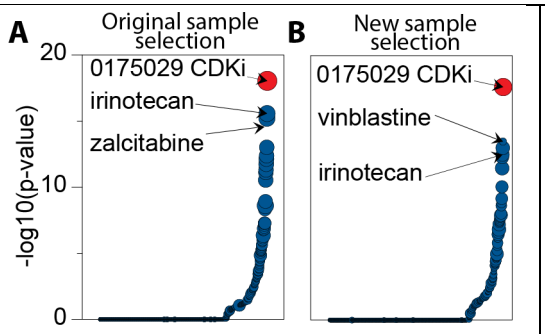

Figure R2. Drug prediction based on original versus suggested sample selection A-B) elbow plots showing the significance of overlap between drug up (or down) regulated genes with genes from prognostic pathways from original (A) or reviewer suggested (B) sample selection. Drugs are sorted based on the significance. Top four drugs are highlighted.

We have now included a statement about the importance of sample selection stating that "It is important to recognize that the selection of samples can influence the identification of prognostic pathways and biomarkers. In our analyses, we classified patients with unfavorable outcomes as those in early cancer stages who passed away, and age and sex matched patients with favorable outcomes as those in later stages who were marked as alive in the TCGA cohort. A more precise selection criterion would have been to classify deceased patients with survival times less than the median, and age and sex matched alive individuals with days to last follow-up more than the median of all samples as unfavorable and favorable selections, respectively."

2) The concern regarding tumor filtering extends to R6c/d. The survival plotted is not representative of published relationship between survival and subtype.

We thank the reviewer for the question. We have now also investigated the plots in the previous rebuttal panels **Figs. R6c-d** using the new tumor filtering criteria suggested by the reviewer and

have observed similar survival patterns as we had shown in those panels previously. We believe that the differences between our analyses and the published relationships between survival and subtypes stem from the selection in our study of early-stage samples versus all cancer samples in the published studies. Nevertheless, this section has already been removed from the paper to prevent any confusion and so should no longer be a major concern.

3) The authors clearly show the overlap between validation datasets and TCGA for the top unfavorable outcomes from PET (Figure S4c), however it would be meaningful to show not only that the pathways overlap, but that your optimal biomarker or expression score (Figure 2f) was also predictive in the validation datasets.

We thank the reviewer for the question and agree. In fact, **Fig. S4g** in the previous submission was meant to do just that. This panel shows that the majority of the significant prognostic combination biomarkers identified in TCGA for four different cancer types remain consistent in independent cohorts and are also predictive in the validation datasets.

4) What is the authors rational for choosing the most frequently mutated gene to stratify patients? If they were trying to show the benefit of PET over mutation alone, wouldn't the more adapt comparison be to stratify patients by the gene or genes that had a significantly different alteration frequency between good and poor prognosis?

We thank the reviewer for the question. This was requested by reviewer #4 during the first round of revisions, as these genes often confer poor prognosis. We have now added references for the statement.

Reviewer #3 (Remarks to the Author):

I thank the authors for the continued efforts to address the comments of this and the other reviewers. I'm satisfied with their answers, but I'm concerned with two issues I raised in my first review, which have been only partly addressed, but which I think should be considered at the editor's discretion whether they are substantial enough to preclude publication:

We thank the reviewer for their positive and constructive feedback on our efforts to address their concerns. We are grateful for the opportunity to improve and refine our work based on the feedback about the source code availability, which we have addressed below.

1. At the moment the presented software is only available through a GitHub repo, which could disappear from one day to the next. Ideally, open-source software should be distributed through a trusted repository that provides some minimum standards of versioning, archiving, licensing and testing. In the case of Python, this could be The Python Package Index (PyPI at https://pypi.org), while in the case of R, this could be CRAN at https://cran.r-project.org or Bioconductor at https://bioconductor.org.

We apologize for missing this earlier. The scripts for executing and evaluating various methods are now deposited on GitHub (https://github.com/hedgehug/Benchmark). We have also improved the instructions for running PET (https://github.com/hedgehug/PET) and additionally

deposited the entire project to Figshare (https://doi.org/10.6084/m9.figshare.c.7252324) to ensure long-term accessibility via DOI.

2. Reproducing the benchmarking reported in the manuscript. The authors have provided template scripts for running all methods, but I could not find the way to run the benchmarking and obtain at least one of the figures that show the benchmarking results, such as Figures 1b or 1e. Providing templates to run any given method is certainly useful, but in my opinion, a benchmarking manuscript should provide the full code that allows one to reproduce the benchmarking figures as easy as possible. Here, I'm not referring to scripts that read values from CSV file and generate the figures, but scripts that run the analysis and obtain the results shown in those figures. This probably should reside in a different repository as the software, and ideally in some repository from the same journal (e.g., as supplementary material) or a third-party platform (e.g., https://figshare.com) that links those scripts with the main manuscript somehow (e.g., through a DOI).

The scripts for executing and evaluating various methods to obtain the reported results are now deposited on GitHub (https://github.com/hedgehug/Benchmark). We have also improved the instructions for running PET (https://github.com/hedgehug/PET) and additionally deposited the entire project to Figshare (https://doi.org/10.6084/m9.figshare.c.7252324) to ensure long-term accessibility via DOI.

Reviewer #4 (Remarks to the Author):

This is Reviewer #4. In my previous round of review, I had two comments: first, the significance of genomic and epigenetic alterations is context-dependent and the same alterations have different physiological impacts in different cancers. Benchmark was designed to identify common gene sets across varied conditions. Therefore, this comment cannot be addressed with Benchmark. However, the authors did show that some common gene set alterations shared by different cancers have similar physiological significance. So I think Benchmark has the merit to be published. Second, the elbow plots were obtained from combined data of many cancer cell lines and clinical tissues. The data from the combined analysis may not be applicable to individual cell lines/patients. This comment has not been addressed. However, the information obtained from combined data analysis can still be used for drug discovery and preclinical studies even though it may not be applicable to individual patients at clinic. Hence, it also has some merit for publication.

We thank the reviewer for their effort and consideration in helping to improve our manuscript, and we are pleased to see that we have fully addressed their concerns and received their support for publication.

Reviewer #5 (Remarks to the Author): Expert in bioinformatics, computational genomics, systems biology, and drug response prediction

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the reviewer for their additional consideration and have now addressed the issues above.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have addressed our comments and in particular corrected how they run the different methods and provided detailed code and data. In that regard, we are content with this revision.

Reviewer #2 (Remarks to the Author):

I would like to thank the reviewers for their responses to my concerns. All my questions have been answered and I have no outstanding comments or concerns.

Reviewer #3 (Remarks to the Author):

I thank the authors for making available all the data and scripts through figshare. They have not followed my suggestion of making the PET software available through a trusted repository, besides the GitHub repo. I think this will not help reaching a wide user community. If they want to use GitHub only as a software distribution platform, I'd suggest them to exploit the features of GitHub for CI/CD (GitHub Actions) in different platforms, and for having stable release versions of the software (see https://docs.github.com/en/repositories/releasing-projects-on-github/managing-releases-in-a-repository).

Testing the software across different operating systems, allows developers to anticipate and reproduce installation problems, and making different release versions of the software facilitates the reproducibility of the analyses obtained through it.

Note, for instance, that the 'environment.yml' file provided in the GitHub repo to install the software requires specific versions of R (4.2) and Python (3.10). While this is a fine choice for an environment where you want to reproduce the results of the paper, this might be suboptimal when distributing the software for general use because, for instance, R 4.2 is as of June 2024, already one year old, and consequently the DESeq2 version used in PET is also one year old. This precludes PET from being up to date with bug fixes and feature improvements in DESEq2, while this does not happen with EGSEA or decoupleR, because they form part of the Bioconductor project, which forces them to be continuously tested and up to date, resulting in a more stable experience for users, and a long-term compromise to keep the software being useful to the community.

Reviewer #3 (Remarks on code availability):

I was able to install and run it, but the fact that the authors chose not to distribute the software meeting some minimum standards of archiving and testing, compromises its usability; see my comments to the authors. Distributing the software under those standards implies a compromise to distribute and maintain the software throughout a longer period of time than, maybe, what the authors are willing to invest. The journal should choose where to put the bar for scientific open source software distribution.

Reviewer #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #1 (Remarks to the Author):

The authors have addressed our comments and in particular corrected how they run the different methods and provided detailed code and data. In that regard, we are content with this revision.
We thank the reviewer for the insightful feedback throughout the review process.

Reviewer #2 (Remarks to the Author):

I would like to thank the reviewers for their responses to my concerns. All my questions have been answered and I have no outstanding comments or concerns.
We thank the reviewer for the insightful feedback throughout the review process.

Reviewer #4 (Remarks to the Author):

Previously addressed.

Reviewer #5 (Remarks to the Author):
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.
We thank the reviewer for the insightful feedback throughout the review process.

Reviewer #3 (Remarks to the Author):

I thank the authors for making available all the data and scripts through figshare. They have not followed my suggestion of making the PET software available through a trusted repository, besides the GitHub repo. I think this will not help reaching a wide user community. If they want to use GitHub only as a software distribution platform, I'd suggest them to exploit the features of GitHub for CI/CD (GitHub Actions) in different platforms, and for having stable release versions of the software (see https://docs.github.com/en/repositories/releasing-projects-on-github/managing-releases-in-a-repository).

Testing the software across different operating systems, allows developers to anticipate and reproduce installation problems, and making different release versions of the software facilitates the reproducibility of the analyses obtained through it.

Note, for instance, that the 'environment.yml' file provided in the GitHub repo to install the software requires specific versions of R (4.2) and Python (3.10). While this is a fine choice for an environment where you want to reproduce the results of the paper, this might be suboptimal

when distributing the software for general use because, for instance, R 4.2 is as of June 2024, already one year old, and consequently the DESeq2 version used in PET is also one year old. This precludes PET from being up to date with bug fixes and feature improvements in DESEq2, while this does not happen with EGSEA or decoupleR, because they form part of the Bioconductor project, which forces them to be continuously tested and up to date, resulting in a more stable experience for users, and a long-term compromise to keep the software being useful to the community.

We thank the reviewer for their feedback. We respectfully argue that Figshare and GitHub are trusted repositories that are well-adopted by the community and have been recommended by Nature's software policy. We have now also released a stable version, as suggested by the reviewer and will explore other options in future releases.

Reviewer #3 (Remarks on code availability):

I was able to install and run it, but the fact that the authors chose not to distribute the software meeting some minimum standards of archiving and testing, compromises its usability; see my comments to the authors. Distributing the software under those standards implies a compromise to distribute and maintain the software throughout a longer period of time than, maybe, what the authors are willing to invest. The journal should choose where to put the bar for scientific open source software distribution.

Thank you for the feedback and for taking the time to install and run our software. We appreciate the comments. Our software has been distributed on both GitHub and Figshare, where we have made efforts to meet the standards of testing to ensure usability. We have included comprehensive documentation, installation instructions (as tested by the reviewer), and examples to facilitate use and reproducibility.