Meta-analysis of 1,200 transcriptomic profiles identifies a prognostic model for pancreatic ductal adenocarcinoma

Authors

Vandana Sandhu[1,2], Knut Jorgen Labori[3], Ayelet Borgida[4], Ilinca Lungu[5], John Bartlett[5], Sara Hafezi-Bakhtiari[6], Rob Denroche[8], Gun Ho Jang[8], Danielle Pasternack[7], Faridah Mbaabali[7], Matthew Watson[7], Julie Wilson[8], Elin H. Kure[2,9], Steven Gallinger[8,10.11], Benjamin Haibe-Kains[1,8,12,13]

Affiliations

1. Princess Margaret Cancer Centre, University Health Network, Toronto, M5G 1L7, Canada
2. Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital - The Norwegian Radium Hospital, Oslo, Norway
3. Department of Hepato-Pancreato- Biliary Surgery, Oslo University Hospital, Oslo, Norway
4. Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, Toronto
5. Diagnostic Development, Ontario Institute of Cancer Research
6. Department of Pathology, University Health Network, Toronto, M5G 1L7, Canada
7. Department of Genomics, Ontario Institute of Cancer Research
8. PanCuRx Translational Research Initiative, Ontario Institute of Cancer Research
9. Department of Natural Sciences and Environmental Health, Faculty of Technology, Natural Sciences and Maritime Sciences, University of Southeastern Norway, Bø in Telemark, Norway
10. Wallace McCain Centre for Pancreatic Cancer, University Health Network, Toronto, M5G 1L7, Canada
11. Hepatobiliary/pancreatic Surgical Oncology Program, University Health Network, Toronto, M5G 1L7, Canada
12. Department of Medical Biophysics, University of Toronto, Toronto, M5G 1L7, Canada
13. Department of Computer Science, University of Toronto, Toronto, M5S 2E4, Canada

***Corresponding author:** Benjamin Haibe-Kains

101 College Street, PMCRT 11-310, M5G1L7, Toronto, Canada

benjamin.haibe.kains@utoronto.ca, Phone: +1 (416) 581-7628

Abbreviations: AUROC: Area under the receiver operating curve, GO: Gene annotation, OS: Overall survival, PCOSP: Pancreatic cancer overall survival predictor, PDAC: Pancreatic ductal adenocarcinoma, TSP: Top scoring pairs.

Keywords:

1. Pancreatic cancer
2. Overall survival predictive model
3. Meta-analysis
4. One sample classifier
5. Transcriptome analysis
6. Top Scoring pairs

ABSTRACT

Background

With a dismal 8% median 5-year overall survival (OS), pancreatic ductal adenocarcinoma (PDAC) is highly lethal. Only 10-20% of patients are eligible for surgery, and over 50% of these will die within a year of surgery. Identify molecular predictors of early death would enable the selection of PDAC patients at high risk.

Methods

We developed the Pancreatic Cancer Overall Survival Predictor (PCOSP), a prognostic model built from a unique set of 89 PDAC tumors where gene expression was profiled using both microarray and sequencing platforms. We used a meta-analysis framework based on the binary gene pair method to create gene expression barcodes robust to biases arising from heterogeneous profiling platforms and batch effects. Leveraging the largest compendium of PDAC transcriptomic datasets to date, we show that PCOSP is a robust single-sample predictor of early death (≤1 yr) after surgery in a subset of 823 samples with available transcriptomics and survival data.

Results

The PCOSP model was strongly and significantly prognostic with a meta-estimate of the area under the receiver operating curve (AUROC) of 0.70 (P=1.9e-18) and hazard ratio (HR) of 1.95(1.6-2.3) (P=2.6e-16) for binary and survival predictions, respectively. The prognostic value of PCOSP was independent of clinicopathological parameters and molecular subtypes. Over-representation analysis of the PCOSP 2619 gene-pairs (1070 unique genes) unveiled pathways associated with Hedgehog signalling, epithelial mesenchymal transition (EMT) and extracellular matrix (ECM) signalling.

Conclusion

PCOSP could improve treatment decision by identifying patients who will not benefit from standard surgery/chemotherapy and may benefit from alternate approaches.

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal malignancy with 5-year overall survival rate less than 8%[1]. The majority of patients (> 80%) are inoperable due to locally advanced or metastatic disease at time of diagnosis. While surgical resection is the key to curative treatment, it rarely results in long-term survival[2]. Hence, completion of multimodality treatment - surgery combined with adjuvant or neoadjuvant chemotherapy - is the standard of care for treatment of PDAC. However, even after surgical resection with curative intent, median survival does not exceed 28 months and half of those who undergo surgery develop recurrent disease, and die within a year after surgery[3,4]. Therefore, there is a need for a robust prognostic model to identify patients with high risk of early death based on molecular profiles of their tumors. Such a prognostic model would assist clinicians in identifying patients who might not benefit from surgery and standard adjuvant chemotherapy and may benefit from alternative treatment strategies.

Various clinical factors are prognostic following PDAC surgery such as lymph node metastasis status[5], tumor grade[6], margins[7], degree of differentiation[8] and protein biomarker *CA-19-9*[9]. However, the prognostic value of these clinical variables are insufficient to accurately stratify patients based on risk of disease recurrence[10,11]. With the advent of high-throughput next-generation molecular profiling technologies, multiple studies have released transcriptomic profiles of PDAC to the public domain. These gene expression profiles have been leveraged to identify molecular subtypes of PDACs[12–16]. While overlap between these subtypes[15] supports the biological relevance of these published classification schemes[15], they have not been designed to optimize prognostic value.

Previously published prognostic models were developed from small number of samples lacking proper validation in multiple datasets[17–21]. Attempts have been made recently to build a prognostic gene signature using pooled samples from multiple cohorts to identify patients at high risk of short-term survival post surgery[22–24]. However, they used samples profiled using either array or sequencing based method as the learning cohort,

therefore the classifiers may perform better for subjects whose samples were profiled using only one of the two platforms.

To address these issues, we took advantage of a unique set of 89 PDAC profiled using both microarray and sequencing technologies to develop the Pancreatic Cancer Overall Survival Predictor (PCOSP) model. Using an independent set of PDAC transcriptomic profiles from 823 primary resected patients, we show that PCOSP is a robust single-sample predictor of early death (≤1 yr) after surgery with a meta-estimate of the area under the receiver operating characteristics curve of 0.70 (p=1.9e-18). We also show that PCOSP is significantly prognostic (meta-estimate of hazard ratio of 1.95; p=2.6e-16). Furthermore, we show that PCOSP performs significantly better than published prognostic models across microarray and sequencing datasets (Superiority test, $P < 0.01$). Our results support PCOSP as a potential tool to assist clinicians in decision making.

MATERIALS AND METHODS

The meta-analysis pipeline used to develop the PCOSP model and evaluate its prognostic value is provided in Figure 1.

Datasets

We surveyed the literature and curated 17 datasets including 1,236 PDAC patients from public domain for which transcriptome data of PDAC are available (Supplementary Table S1). We further filtered samples based on the availability of overall survival (OS) and sample size (>10) after dichotomization into high and low survival groups based on an OS cut-off of 1-year (Figure 2). This resulted in a total of four sequencing studies and seven array-based studies providing transcriptomic and clinical data for 1,001 PDAC patients. A total of 12,430 protein-coding genes commonly assessed across all the cohorts were used for further analysis. Clinicopathological features of all the cohorts are presented in Supplementary Table S2. The different cohorts had similar clinical presentation, and were treated with

curative surgery followed by adjuvant chemotherapy. Two third of the patients completed multimodal treatment (i.e., surgery and adjuvant chemotherapy; Supplementary Table S2).

Prognostic model

To develop a robust predictor for early death, we used the gene expression profiles of 89 PDAC patient samples whose tumors have been profiled using both microarray and sequencing platforms within the International Cancer Genome Consortium (ICGC) cohort. Approximately half of the patients of the training cohort which where eligible for surgery relapsed within 1 year, we used this threshold to predict PDAC patients with high risk of early death (≤1 yr) post surgery. We excluded 7 samples from the training cohort as these patients were censored before one year of follow-up.

To make gene expression profiles comparable between the training and validation sets, we transformed the original gene expression profiles into binary gene pair barcodes. The advantages of considering pairs of genes with a binary value ("1" if expression of gene *i* > gene *j*, "0" otherwise) are; (*i*) it transforms the feature space in a way that mitigates platform biases and potential batch effects; (*ii*) it makes the model robust to any data processing that preserves the gene order[25,26]. We implemented k-Top Scoring disjoint Pairs (k-TSP) classifier predictor[27] using the Wilcoxon rank sum method as filtering function in the *SwitchBox* package (version 1.12.0)[28].

The decision rules are based on the relative ordering of gene expression values within the same sample, where the *k* top scoring gene pairs are used to build the classifier. The samples were resampled 1000 times, where 40 samples from each group were selected in each run to build a *k*-TSP model and the model was further tested on the 49 out-of-bag samples. The models were selected if the balanced accuracy was above 0.6 else the model was rejected. We then froze the parameters of the predictive model and validated it in the remaining compendium of independent datasets. The class probability of the sample was

calculated as the frequency of sample predicted as one class divided by the total number of models.

Random classifiers

To test whether the prognostic value of the PCOSP model could be achieved by random chance alone, we implemented two permutation tests. To test whether the gene expression profiles were associated with survival, we shuffled the actual class labels while maintaining the expression values. To test whether the gene pairs selected in the PCOSP model were robustly associated with survival, we randomly assigned genes to the k-TSP model and assessed its prognostic value. Both procedures were performed 1000 times. As a pre-validation set we compared the balanced accuracy of all the 1000 random models generated using both the approaches to PCOSP using the Wilcoxon rank sum test. Further, we trained the k-TSP classifier models from both approaches in the same way as we built our consensus PCOSP model. We then froze the parameters of the prognostic model and validated it in the compendium of independent datasets, and compared the meta-estimates for both the models against the PCOSP model.

Early death prediction

The meta analysis was performed for the PDAC sequencing cohorts, PDAC array-based cohorts and the overall combined cohorts to assess and statistically compare the performance of the PCOSP. The patient samples were dichotomized into two groups based on the outcome variable (time from surgery to death ≤ 1 year). Samples censored before 1 year of follow-up were excluded from the analysis of meta-estimate of the area under the receiver operating characteristics curve (AUROC). The AUROC plots the sensitivity vs. 1-Specificity and is used as a criterion to measure the discriminatory ability of the model[29]. The AUROC was computed using *pROC* package (version 1.10.0), and the p-value was

estimated using the Mann-Whitney test statistics estimating whether the AUROC curve estimate is significantly different from 0.5 (random classifier). The meta-estimate of AUROC was estimated using the random effect model[30] implemented in *survcomp* package (version 1.26.0)[31,32].

Survival prediction

Prognostic value and statistical significance of survival difference between the predicted classes were assessed using the D-Index, which is a robust estimate of the hazard ratio comparing two equal-sized prognostic groups[33]. In addition, we used the concordance index (C-index) which estimates the probability that, for a random pair of patients, the PCOSP score for the patient with shorter survival is higher than the patient with longer survival[34]. Both the robust hazard ratio (HR) and the C-index were calculated using the *survcomp* package. The meta estimate of HR and C-index were calculated for the PDAC sequencing cohorts, the PDAC array-based cohorts and the combined PDAC sequencing and array-based cohorts using the random effect model[30] implemented in *survcomp* package.

Subtyping of PDAC cohorts

The PDAC cohorts were classified into basal and classical transcriptomic subtype using the Moffitt classification method[13].

Clinicopathological features based model to predict early death

The clinical model was built by fitting the logistic regression model using common clinicopathological features available from all the cohorts, i.e., age, gender, TNM status and tumor grade.

Gene set enrichment analysis

To categorize genes in the PCOSP, we performed gene set enrichment analysis using RunGSAhyper function implemented in piano package (version 1.16.4)[35]. The genes

selected in the PCOSP model were compared against Gene Ontology (GO) gene sets, canonical pathways and hallmark gene sets using as background the protein-coding genes commonly assessed across the gene expression profiling platforms in our data compendium. Enrichment p-values were corrected for multiple testing using the false discovery rate approach (FDR < 5%)[36].

Comparison to existing classifiers

We calculated the Birnbaum signature scores[22] and Chen signature scores[23] using the published coefficients of the 25 and 15 classifier genes, respectively, as weight parameter in the *sig.score* function implemented in the *genefu* R package (version 2.10.0)[37]. The Haider signature scores were used as courtesy of the author[24]. The C-index and HR were computed for the three classifiers using eight validation cohorts excluding the cohorts used for training by PCOSP and other classifiers in comparison. Further, we compared the meta-estimates of C-index of each classifier with PCOSP at P<0.05 (one-sided t-test) as implemented in *survcomp* package.

Research reproducibility

Our code and documentation are open-source and publicly available through the PDACSurv GitHub repository (github.com/bhklab/PDACsurv). A detailed tutorial describing how to run our pipeline and reproduce our analysis results is available in the GitHub repository. A virtual machine reproducing the full software environment is available on Code Ocean. Our study complies with the guidelines outlined in[38–40]. All the data are available in the form of R package MetaGxPancreas (http://bioconductor.org/packages/MetaGxPancreas/).

RESULTS

Overall survival predictive model

To predict the patients with early death (≤ 1 year after surgery), the PCOSP model was trained on the 89 ICGC cohort samples profiled using both microarray and sequencing transcriptomic profiles (Supplementary Table S1). To develop a predictor that can be applied to multiple profiling platforms, we transformed the gene expression profiles into binary gene pairs (x=1 if expression of gene i > gene j, x=0 otherwise) and used these transcriptomic barcodes in an ensemble of 1000 *k-TSP* predictive models. The PCOSP score is subsequently calculated using the majority voting rule. We tested the prognostic value of PCOSP score in three independent sequencing cohorts, including the Pancreatic Cancer Sequencing Initiative (PCSI)[41], TCGA-PAAD[15] and Kirby[42] cohorts, and seven independent array-based cohorts composed of ICGC-array (excluding the 89 samples used for training)[43], UNC[13], OUH[44], Chen[23], Zhang[45], Winter[46] and Collisson cohorts[12](Supplementary Table S1). We first tested the predictive value of early death by calculating the AUROC for each dataset separately (Figure 3A). PCOSP was highly significant overall (AUROC=0.70; P<1.9E-18; Figure 3A), although higher in the datasets generated using sequencing platforms compared to microarrays (AUROC 0.72 vs 0.68 for sequencing and array datasets, respectively). PCOSP was significantly predictive of early death in all cohorts (AUROC∈[0.67,0.76]; P<0.05) except the Winter and OUH cohorts (P>0.48) and was almost significant for the Collisson cohort (AUROC=0.69; P=0.051). To determine whether the early death predictive value of the PCOSP model can be achieved by random chance alone, we first computed meta-estimates of AUROC by randomly shuffling the class labels (early deaths) 1000 times and applying the same training procedure used for the PCOSP model. We observed that the gene expression profiles were significantly associated with survival as none of the random models could yield a predictive value greater or equal to PCOSP (p<0.001; Supplementary Figure S1A). We further tested whether the gene pairs selected in the PCOSP model were robustly associated with early death events, by randomly assigning genes to the PCOSP model. Again, we observed that the genes selected in PCOSP yielded significantly more

predictive information than the models comprised of random genes ($p<0.001$; Supplementary Figure S1B), supporting the biological relevance of the PCOSP gene set.

Prognostic relevance of the PCOSP model

To assess the prognostic value of the PCOSP model, we calculated the C-indices and HR using the overall survival data for all the cohorts. The C-index is significant overall (C-index=0.63, P=3.5E-21; Figure 3B). In agreement with the results of early death prediction, the PCOSP prognostic value was higher for the sequencing datasets when compared to the arrays (C-index=0.65 vs 0.61 for sequencing and array datasets, respectively; P<2.1E-11; Figure 3B). Similar to the C-index, the PCOSP HR was strong and significant overall (HR =1.95, P=2.7E-16; Figure 3C), and stronger for the sequencing datasets (HR = 2.24 vs 1.83; Figure 3C). To assess whether the prognostic value of PCOSP depends on PDAC molecular subtypes, we stratified PDAC samples into the basal and classical subtypes and calculated meta-estimates of C-index and HR (Supplementary Figures S2A and S2B). We found that PCOSP was prognostic in validation cohorts independently of molecular subtypes. We further tested whether PCOSP prognostic value was complementary to clinicopathological parameters and molecular subtypes by fitting both a multivariate Cox proportional hazard model to predict survival and a logistic regression model to predict binary outcome (death >1yr or not) (Supplementary Table S3).

Clinicopathological model to predict OS

Patient-specific clinicopathologic features were available for the PCSI, ICGC-sequencing, ICGC-array, TCGA and OUH cohorts. The common variables available were age, gender, TNM status and tumor grade. We fitted the logistic regression model using these clinicopathological features to predict early death of PDAC patients. The clinicopathological model was significant overall with however a weak predictive value (C-index=0.55; P=0.04; Figure 4A). Contrary to PCOSP, the clinicopathological model was not predictive in the

sequencing cohort (C-index=0.53 vs 0.58 with P=0.25 vs 0.03 for the sequencing and the array datasets, respectively; Figure 4A). Only nodal status, tumor grade and molecular classes were significant in the univariate analysis (Supplementary Table S3). We compared the prognostic value of the clinicopathological model against PCOSP (Figure 4B,C). PCOSP was significantly more prognostic than the clinicopathological model (one-sided t-test P < 0.01; Figure 4D).

Comparison with published prognostic models

We compared the prognostic value of PCOSP to three published PDAC prognostic models, referred to as Birnbaum[22], Chen[23]and Haider[24]. The overall prognostic value of the three published models was significant (P<8.5E-7; Figure 5A,C). PCOSP significantly outperformed published prognostic models in all cases (P<0.05, Figure 5C,D); except for the HR of the Chen classifier where the superiority of the PCOSP prognostic value showed a trend to significance (one sided t-test P=0.10) .

Pathway analysis of prognostic genes

Gene enrichment analysis for PCOSP signature genes was performed using hypergeometric test using the hallmarks gene sets, GO molecular function, GO cellular component terms and canonical pathways in MSigDb[47] with the 1,070 unique genes from the PCOSP model. The Extracellular matrix (ECM), Epithelial Mesenchymal transition (EMT) and hedgehog signalling pathway genes were enriched in the PCOSP model at false discovery rate (FDR) <5%. The complete list of GO terms and pathways significantly enriched in the PCOSP model are listed in Supplementary Table S4A- 4D.

DISCUSSION

We performed a meta-analysis of the transcriptomic profiles of 1,236 PDAC patients and developed PCOSP, a new prognostic model to identify patients with high risk of early death

after surgery. The model is built from a unique set of 89 patients profiled using both array-based and sequencing platforms, and validated on a compendium of ten independent datasets, including 823 patients. The prognostic value of the PCOSP model was highly significant for both early death (≤1 year) and overall survival (P<0.001; Figure 3).

Contrary to published prognostic signatures fitted on small number of samples and lacking validation in large independent datasets[17–21], PCOSP has been trained and validated on a large compendium of datasets. Comparison of PCOSP with existing classifiers[22–24] showed that the Birnbaum, Chen and Haider models yielded significant (P<0.8.5E-7) but significantly weaker prognostic value than PCOSP (Figure 5C,D). Importantly, PCOSP performs significantly better than existing classifiers for both microarray and sequencing platforms, likely due to simplifying the continuous expression space into binary pair barcodes. This enables PCOSP to be used as a single sample predictor robust to profiling platforms, potential batch effects and normalization methods compared to other classifiers.

Comparison of PCOSP against known prognostic clinicopathological variables showed that PCOSP outperformed the clinicopathological model in predicting early death (Figure 4). PCOSP prognostic value was significant, even after adjusting for molecular subtyping (classical vs basal) and clinicopathological parameters (age, sex, TNM status, differentiation grade of tumor and molecular classes); Supplementary Figure S2A,B and Supplementary Table S3).

The PCOSP model incorporates 2,619 unique gene pairs, totalling 1,070 unique genes. Functional analysis of 1,070 genes showed enrichment of Hedgehog signalling, ECM and EMT pathway. Numerous studies have suggested the involvement of EMT in invasion and metastasis of PDAC[48]. EMT enhances cell motility through loss of cell-cell adhesion, escaping from extracellular matrix and overcoming the apoptosis process[48]. The ECM and EMT pathways are not only associated with the metastatic spread of tumor but also with chemoresistance that leads to worse survival[49].

PDAC is a heterogeneous and genetically highly complex disease, supporting the molecular[13,14] and morphological[50] characterization of a given tumor as an important cornerstone for the development of future therapies. We provide the largest compendium of 17 PDAC datasets as a gold standard for future PDAC analyses. The new meta-analysis framework implemented in PCOSP maximizes robustness and performance across the cohorts. In order to implement PCOSP as a clinical assay, we tested different feature set sizes for the k-TSP models and compared the performance of the reduced models. We achieved accuracy comparable to the 1,070 gene-PCOSP model by including only 256 unique genes, supporting the potential of a smaller PCOSP based useful in the clinic (Supplementary Figure S3). Endoscopic ultrasound (EUS) biopsies could be utilized prior to curative surgery to estimate the prognosis of PDAC patients using PCOSP. This may assist clinicians in making postoperative treatment decision, i.e., palliative therapy, standard adjuvant chemotherapy or alternative chemotherapy regimens, which holds the potential to improve the overall survival of the patients.

The current study has potential limitations. First, there are inherent tumor sample collection biases as the different datasets were collected and sampled at different centers. The levels of tumor cellularity varied highly across cohorts as PCSI and Collisson datasets were generated using laser microdissection prior to sequencing, Kirby and Chen datasets were macrodissected, while TCGA, ICGC, OUH, Zhang and Winter datasets used bulk tumors for profiling. Second, the transcriptomic profiles in our data compendium were generated using different gene expression profiling technologies for sequencing (Illumina HiSeq 2000/2500) and microarray platforms (Agilent, Affymetrix, and Illumina). Third, all samples were normalized using the published processing methods, which depend on the profiling platforms (Supplementary Table S2). Despite these limitations, PCOSP yielded robust prognostic value across the heterogeneous datasets, indicating that the gene expression barcode transformation is robust to the inevitable biases present in large meta-analyses.

The lack of available clinical and treatment information across the cohorts is also a limiting factor in our meta-analysis. However, comparison of cohort specific clinical information for the cohort were not significantly different across the cohorts (Supplementary table S2). During the time period of sample collection, standard of care treatment for PDAC was curative-intent surgery followed by adjuvant chemotherapy with gemcitabine or 5-FU. New approaches using doublet and triplet chemotherapy regimens are now standard of care in the palliative setting and randomised trials using these agents in the adjuvant setting will be reported shortly. Neoadjuvant therapy is also being evaluated in many centres. Thus, heterogeneity in treatment is expected within and between different cohorts, we will need to test our PCOSP model using new clinical datasets, or preferably within the context of randomized trials.

CONCLUSION

We leveraged the largest compendium of PDAC transcriptomes to develop PCOSP, a prognostic model identifying PDAC patients at high risk of early death independently of, and superior to, clinicopathological features and molecular subtypes. PCOSP may be useful in the clinical setting as a single sample classifier to identify patients who could be at higher risk of early death following surgery and adjuvant chemotherapy, potentially facilitating treatment decisions, including the use of neoadjuvant chemotherapy or alternate treatment strategies.

Acknowledgements

This study was conducted with the support of the Ontario Institute for Cancer Research (OICR, PanCuRx Translational Research Initiative) through funding provided by the Government of Ontario (Ministry of Research, Innovation, and Science), and a charitable donation from the Canadian Friends of the Hebrew University (Alex U. Soyka). V.S was supported by grants from The Radium Hospital Foundation, Oslo University Hospital, and

the PanCuRx Translational Research Initiative at the OICR. B.H.K was supported by the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre, the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Ministry of Economic Development and Innovation/Ministry of Research & Innovation of Ontario (Canada). We thank all the patients who participated in the study.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin*. 2017;67(1):7-30.

2. Winter JM, Brennan MF, Tang LH, et al. Survival after resection of pancreatic adenocarcinoma: results from a single institution over three decades. *Ann Surg Oncol*. 2012;19(1):169-175.

3. Labori KJ, Katz MH, Tzeng CW, et al. Impact of early disease progression and surgical complications on adjuvant chemotherapy completion rates and survival in patients undergoing the surgery first approach for resectable pancreatic ductal adenocarcinoma – A population-based cohort study. *Acta Oncol* . 2016;55(3):265-277.

4. Neoptolemos JP, Palmer DH, Ghaneh P, et al. Comparison of adjuvant gemcitabine and capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-4): a multicentre, open-label, randomised, phase 3 trial. *Lancet*. 2017;389(10073):1011-1024.

5. Slidell MB, Chang DC, Cameron JL, et al. Impact of total lymph node count and lymph node ratio on staging and survival after pancreatectomy for pancreatic adenocarcinoma: a large, population-based analysis. *Ann Surg Oncol*. 2008;15(1):165-174.

6. Lüttges J, Schemm S, Vogel I, Hedderich J, Kremer B, Klöppel G. The grade of pancreatic ductal carcinoma is an independent prognostic factor and is superior to the immunohistochemical assessment of proliferation. *J Pathol*. 2000;191(2):154-161.

7. Richter A, Niedergethmann M, Sturm JW, Lorenz D, Post S, Trede M. Long-term results of partial pancreaticoduodenectomy for ductal adenocarcinoma of the pancreatic head: 25-year experience. *World J Surg*. 2003;27(3):324-329.

8. Imaoka H, Shimizu Y, Mizuno N, et al. Clinical characteristics of adenosquamous carcinoma of the pancreas: a matched case-control study. *Pancreas*. 2014;43(2):287-290.

9. Tas F, Karabulut S, Ciftci R, et al. Serum levels of LDH, CEA, and CA19-9 have prognostic roles on survival in patients with metastatic pancreatic cancer receiving gemcitabine-based chemotherapy. *Cancer Chemother Pharmacol*. 2014;73(6):1163-1171.

10. Le N, Sund M, Vinci A, GEMS collaborating group of Pancreas 2000. Prognostic and predictive markers in pancreatic adenocarcinoma. *Dig Liver Dis*. 2016;48(3):223-230.

11. Martinez-Useros J, Garcia-Foncillas J. Can Molecular Biomarkers Change the Paradigm of Pancreatic Cancer Prognosis? *Biomed Res Int*. 2016;2016:4873089.

12. Collisson EA, Sadanandam A, Olson P, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med*. 2011;17(4):500-503.

13. Moffitt RA, Marayati R, Flate EL, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet*. 2015;47(10):1168-1178.

14. Bailey P, Chang DK, Nones K, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*. 2016;531(7592):47-52.

15. Raphael BJ, Hruban RH, Aguirre AJ, et al. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell*. 2017;32(2):185-203.e13.

16. Sandhu V, Wedge DC, Bowitz Lothe IM, et al. The Genomic Landscape of Pancreatic and Periampullary Adenocarcinoma. *Cancer Res*. 2016;76(17):5092-5102.

17. Van den Broeck A, Vankelecom H, Van Delm W, et al. Human pancreatic cancer contains a side population expressing cancer stem cell-associated and prognostic genes. *PLoS One*. 2013;8(9):e73968.

18. Donahue TR, Tran LM, Hill R, et al. Integrative survival-based molecular profiling of human pancreatic cancer. *Clin Cancer Res*. 2012;18(5):1352-1363.

19. Sergeant G, van Eijsden R, Roskams T, Van Duppen V, Topal B. Pancreatic cancer circulating tumour cells express a cell motility gene signature that predicts survival after surgery. *BMC Cancer*. 2012;12:527.

20. Newhook TE, Blais EM, Lindberg JM, et al. A thirteen-gene expression signature predicts survival of patients with pancreatic cancer and identifies new genes of interest. *PLoS One*. 2014;9(9):e105631.

21. Stratford JK, Bentrem DJ, Anderson JM, et al. A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS Med*. 2010;7(7):e1000307.

22. Birnbaum DJ, Finetti P, Lopresti A, et al. A 25-gene classifier predicts overall survival in resectable pancreatic cancer. *BMC Med*. 2017;15(1):170.

23. Chen D-T, Davis-Yadley AH, Huang P-Y, et al. Prognostic Fifteen-Gene Signature for Early Stage Pancreatic Ductal Adenocarcinoma. *PLoS One*. 2015;10(8):e0133562.

24. Haider S, Wang J, Nagano A, et al. A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. *Genome Med*. 2014;6(12):105.

25. Patil P, Bachant-Winner P-O, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. *Bioinformatics*. 2015;31(14):2318-2323.

26. Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol Cancer Res Treat*. 2010;9(2):149-159.

27. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*. 2005;21(20):3896-3904.

28. Afsari B, Fertig EJ, Geman D, Marchionni L. switchBox: an R package for k-Top Scoring Pairs classifier development. *Bioinformatics*. 2015;31(2):273-274.

29. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543-2546.

30. Cochran WG. The Combination of Estimates from Different Experiments. *Biometrics*. 1954;10(1):101-129.

31. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*. 2011;27(22):3206-3208.

32. Haibe-Kains B, Desmedt C, Sotiriou C, Bontempi G. A comparative study of survival

models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*. 2008;24(19):2200-2208.

33. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*. 2004;23(5):723-748.

34. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.

35. Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res*. 2013;41(8):4378-4391.

36. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289-300.

37. Gendoo DMA, Ratanasirigulchai N, Schröder MS, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2016;32(7):1097-1099.

38. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. *PLoS Comput Biol*. 2013;9(10):e1003285.

39. Gentleman R. Reproducible research: a bioinformatics case study. *Stat Appl Genet Mol Biol*. 2005;4:Article2.

40. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting. *JAMA*. 2000;283(15):2008-2012.

41. Notta F, Chan-Seng-Yue M, Lemire M, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*. 2016;538(7625):378-382.

42. Kirby MK, Ramaker RC, Gertz J, et al. RNA sequencing of pancreatic adenocarcinoma tumors yields novel expression patterns associated with long-term survival and reveals a role for ANGPTL4. *Mol Oncol*. 2016;10(8):1169-1182.

43. Nones K, Waddell N, Song S, et al. Genome-wide DNA methylation patterns in pancreatic ductal adenocarcinoma reveal epigenetic deregulation of SLIT-ROBO, ITGA2 and MET signaling. *Int J Cancer*. 2014;135(5):1110-1118.

44. Sandhu V, Bowitz Lothe IM, Labori KJ, et al. Molecular signatures of mRNAs and miRNAs as prognostic biomarkers in pancreatobiliary and intestinal types of periampullary adenocarcinomas. *Mol Oncol*. 2015;9(4):758-771.

45. Zhang G, Schetter A, He P, et al. DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. *PLoS One*. 2012;7(2):e31507.

46. Winter C, Kristiansen G, Kersting S, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*. 2012;8(5):e1002511.

47. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740.

48. Zheng X, Carstens JL, Kim J, et al. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature*. 2015;527(7579):525-530.

49. Gaianigo N, Melisi D, Carbone C. EMT and Treatment Resistance in Pancreatic Cancer. *Cancers* . 2017;9(9). doi:10.3390/cancers9090122

50. FUKUSHIMA, N. Ductal adenocarcinoma variants and mixed neoplasm of the pancreas. *WHO Classification of Tumours of the Digestive System*. 2010:292-295.

FIGURES

Figure 1. Pipeline showing the approach used for building the Pancreatic Cancer Overall Survival Predictor (PCOSP).

Figure 2: Flowchart showing the inclusion criteria for the pancreatic adenocarcinoma samples.

Figure 3. Predictive value of PCOSP for early death and overall survival. (A) Area under the ROC curves for all the cohorts and the meta estimates for sequencing cohorts, array-based cohorts and for both the platforms combined. **(**Forestplot reporting (**B**) the concordance indices (C-index) and (**C**) the hazard ratio (HR) for all the cohorts and the meta estimates for sequencing cohorts (orange), array-based cohorts (blue) and for both the platforms combined (grey).

Figure 4. Comparison of the prognostic value of the clinicopathological model and PCOSP. (A) Barplot reporting the AUROCs for the clinical model and the PCOSP model. **(**Forestplot reporting the the (**B**) concordance index (C-index) and (**C**) Hazard ratio (HR) of validation cohorts computed using PCOSP, and clinicopathological model.

Figure 5. Comparison of existing classifiers with PCOSP. The forestplot reports the meta-estimate of (**A**) concordance indices (C-index) and (**B**) hazard ratio (HR) for PCOSP and existing classifiers.

SUPPLEMENTARY FIGURES

Supplementary Figure S1: Density plot showing the distribution of balanced accuracy for random models. Distribution of meta-estimates of 1000 models generated using (**A**) random reshuffling of labels and (**B**) random assignment of genes to TSP models. The meta-estimates were independently calculated for all the cohorts combined, sequencing

cohorts and array-based cohort. The pink, green, blue dashed lines represent meta-estimate of AUROC from PCOSP model for overall, sequencing and array-based cohorts respectively.

Supplementary Figure S2: Forestplot of (**A**) concordance index (C-index) and (**B**) hazard ratio (HR) for all the cohorts divided based on the molecular subtypes. The grey, green and pink color in the forestplot depicts meta-estimate of C-index for overall cohort, the basal subtype and the classical subtype of the cohorts, respectively.

Supplementary Figure S3: The scatterplot shows the meta-estimate of AUROC (orange) and total number of unique genes (blue) in the PCOSP model at different balanced accuracy thresholds. The threshold used in the PCOSP is marked as dashed line at 0.6.

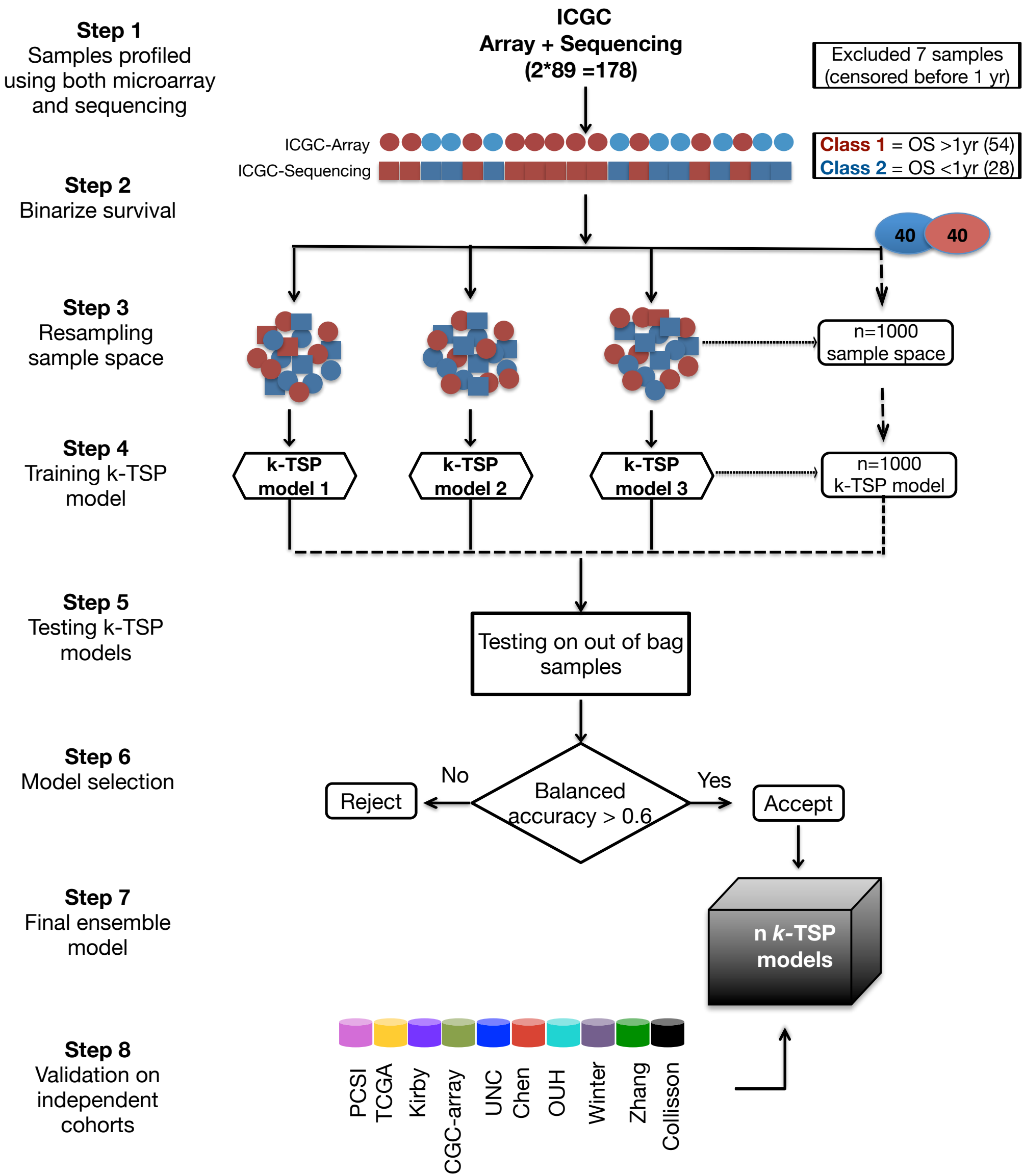
SUPPLEMENTARY TABLES

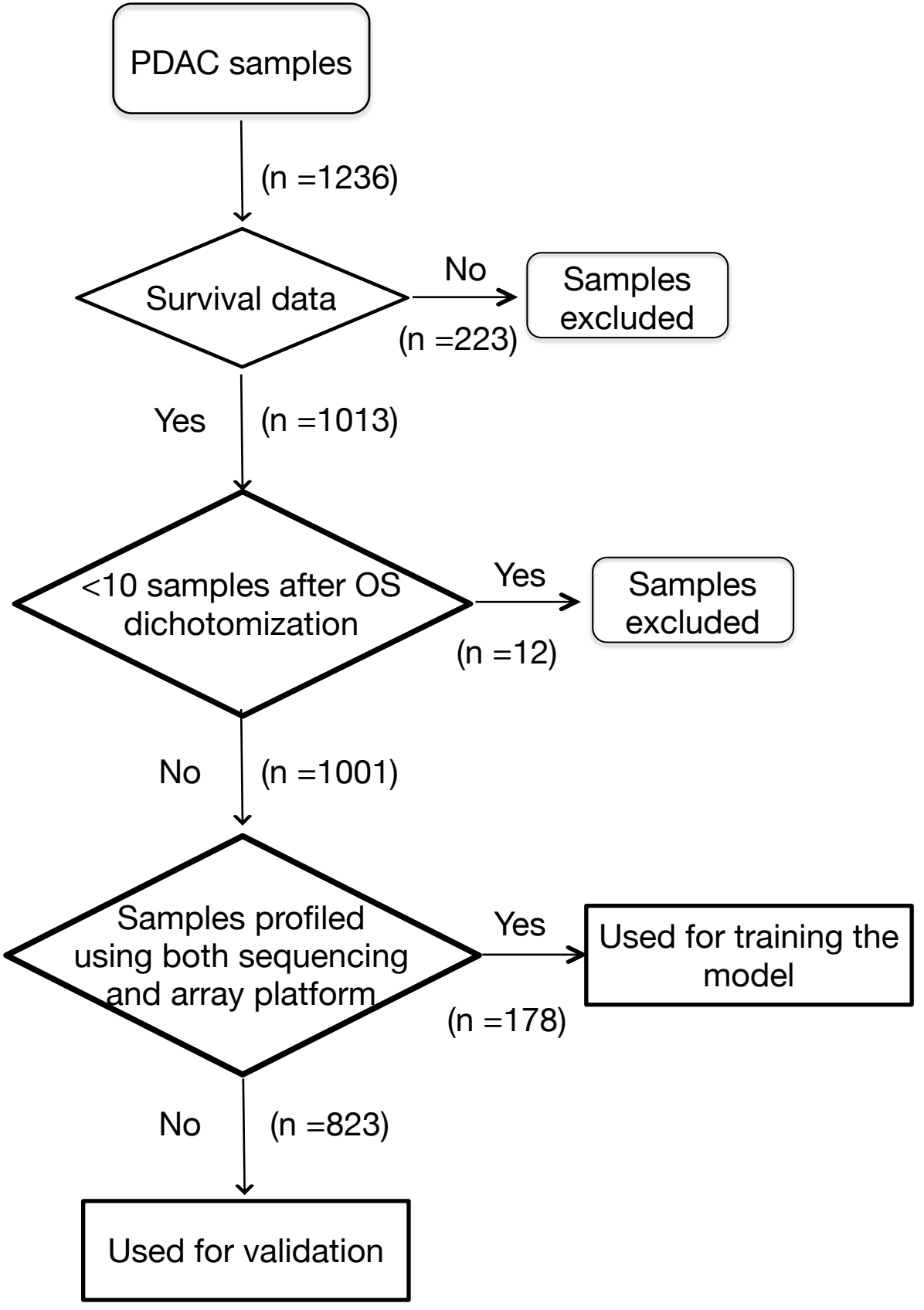
Supplementary Table S1: The table shows the datasets used in the project for meta-analysis.

Supplementary Table S2: The table shows the clinicopathological information of the validation cohorts used in the analysis.

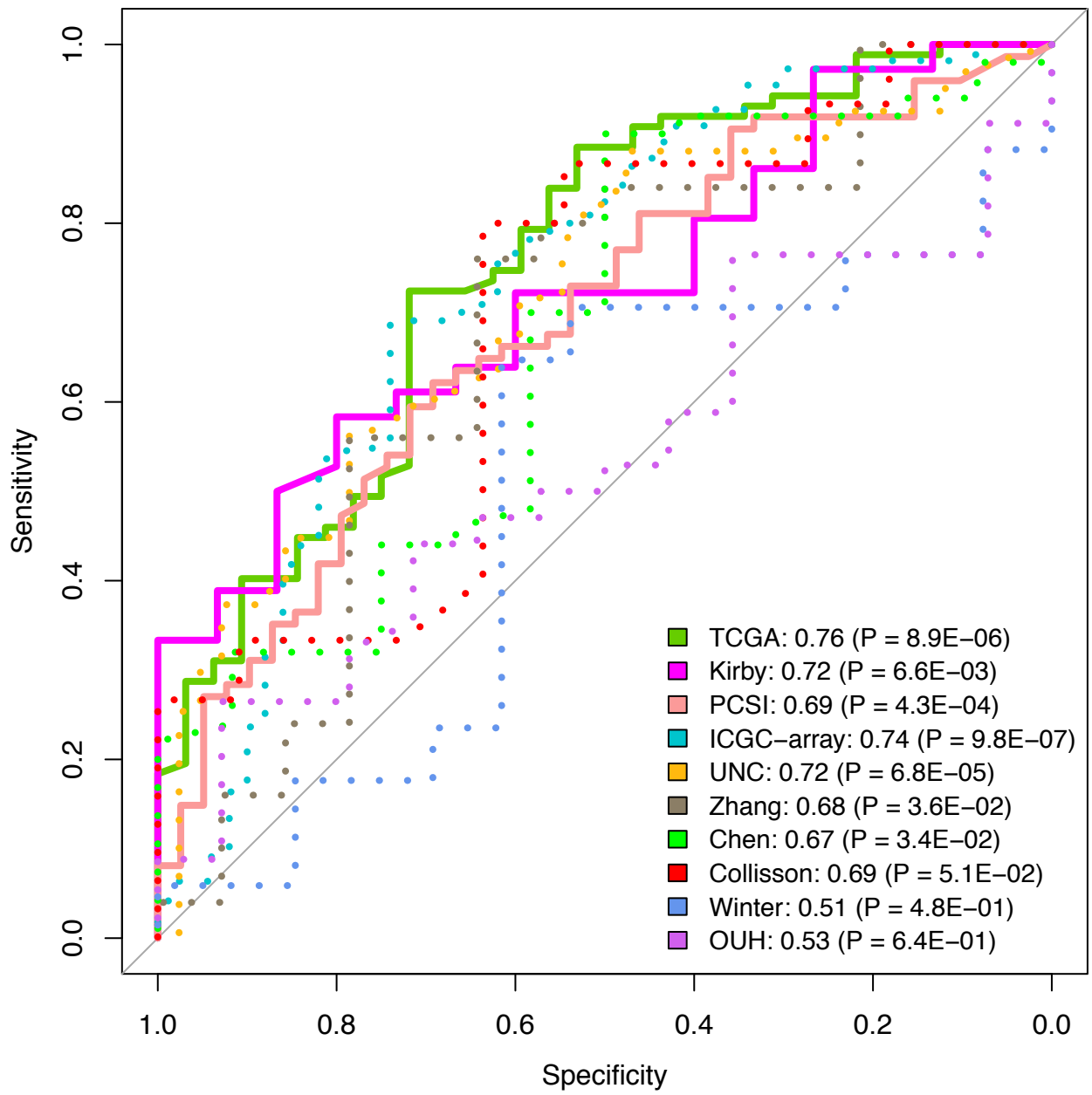
Supplementary Table S3: Univariate and multivariate regression analysis.from (**A**) logistic regression model to predict early death (death >1 yr or not), and (**B**). the Cox regression model using clinicopathological features, molecular subtypes and PCOSP model probabilities for validation cohorts.

Supplementary table S4: The table shows the pathways overrepresented in the PCOSP model genes using (**A**) hallmark gene sets, (**B**) canonical pathways, (**C**) GO-molecular function term (and (**D**) GO cellular component terms from MSigDB.


Step 1
Samples profiled using both microarray and sequencing
ICGC
Array + Sequencing
(2*89 =178)
Excluded 7 samples (censored before 1 yr)
ICGC-Array
ICGC-Sequencing
Class 1 = OS >1yr (54)
Class 2 = OS <1yr (28)
Step 2
Binarize survival
40
40
Step 3
Resampling sample space
n=1000 sample space
Step 4
Training k-TSP model
k-TSP model 1
k-TSP model 2
k-TSP model 3
n=1000 k-TSP model
Step 5
Testing k-TSP models
Testing on out of bag samples
Step 6
Model selection
Balanced accuracy > 0.6
No
Reject
Yes
Accept
Step 7
Final ensemble model
n k-TSP models
Step 8
Validation on independent cohorts
PCSI
TCGA
Kirby
ICGC-array
UNC
Chen
OUH
Winter
Zhang
Collisson


PDAC samples
(n =1236)
Survival data
No
(n =223)
Samples excluded
Yes
(n =1013)
<10 samples after OS dichotomization
Yes
(n =12)
Samples excluded
No
(n =1001)
Samples profiled using both sequencing and array platform
Yes
(n =178)
Used for training the model
No
(n =823)
Used for validation

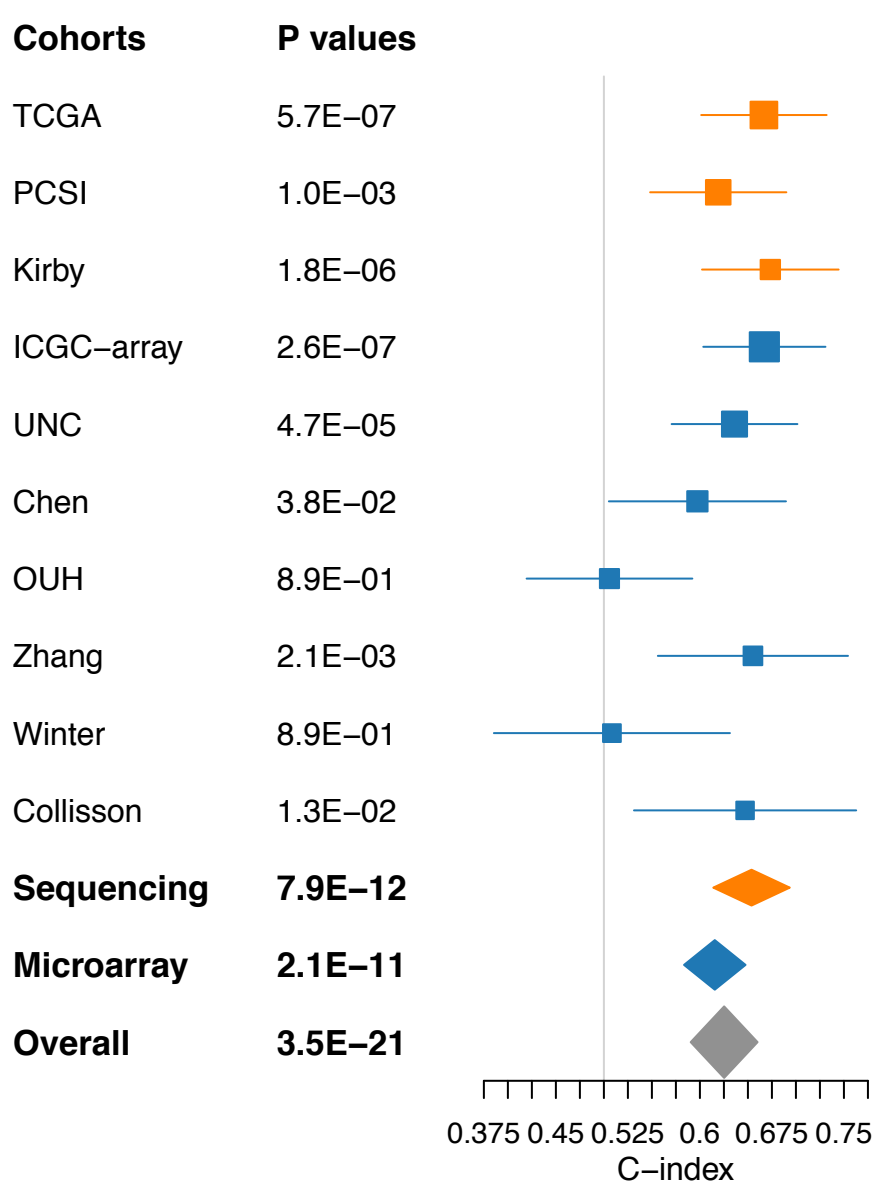
3A



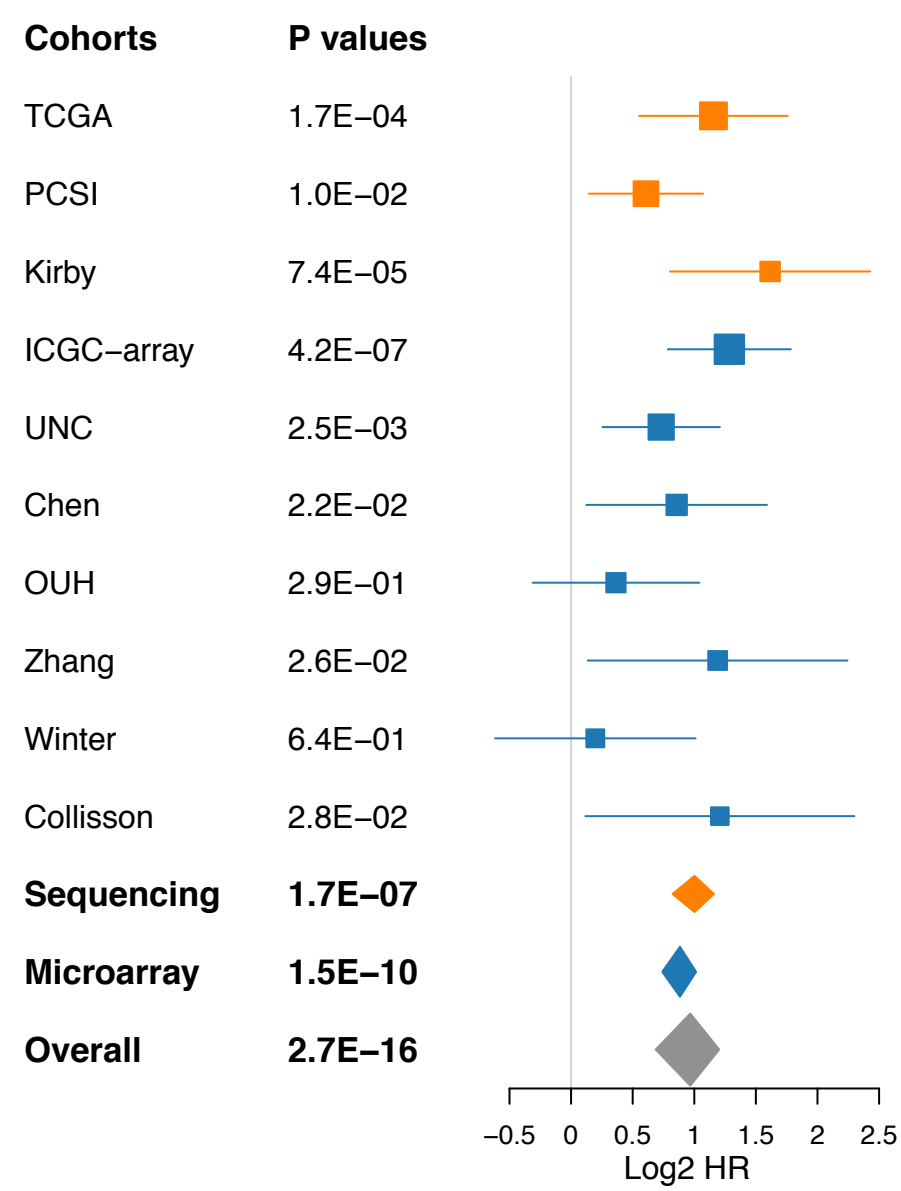
AUC Meta-estimates

Sequencing cohorts: 0.72 (P = 8.0E-09)
Microarray cohorts: 0.68 (P = 7.0E-11)
Overall: 0.70 (P = 1.9E-18)

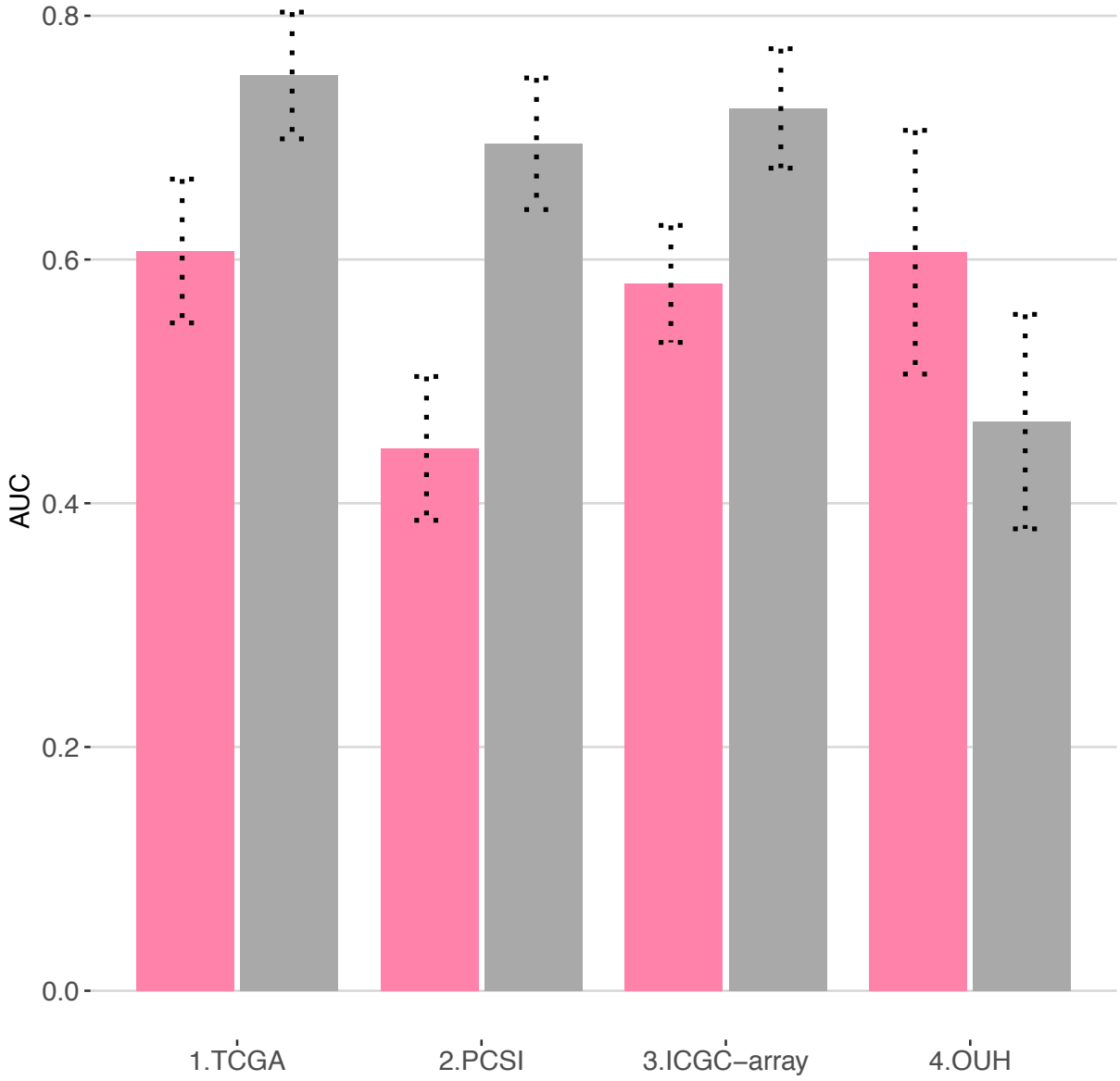
3B



3C



4A



Clinicopathological model

AUC Meta-estimates

Sequencing cohorts = 0.53 (P= 0.25)
Microarray cohorts = 0.58 (P= 0.03)
Overall = 0.55 (P= 0.04)

PCOSP model

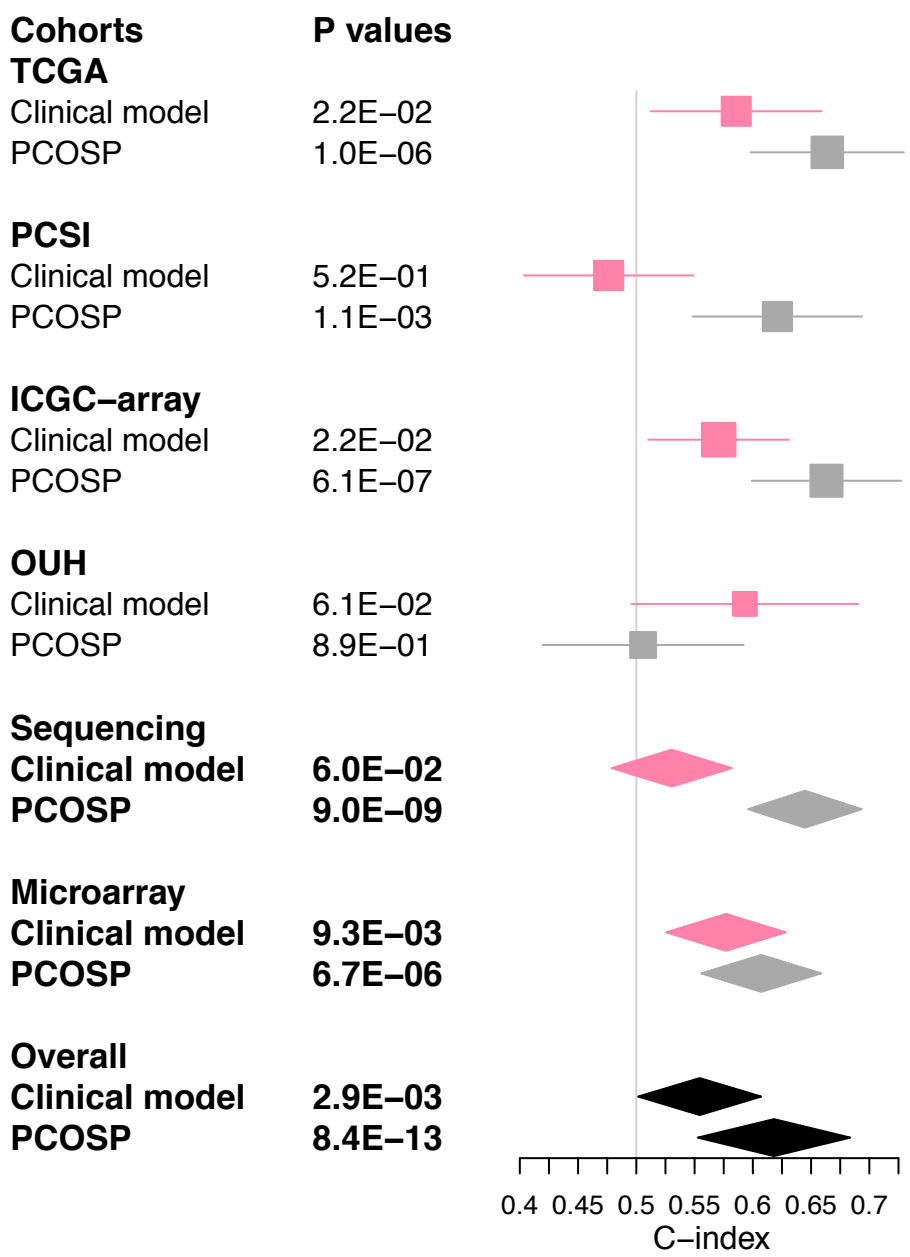
AUC Meta-estimates

Sequencing cohorts = 0.72 (P= 6.7E-08)
Microarray cohorts = 0.61 (P= 2.5e-05)
Overall = 0.68 (P= 2.2E-11)

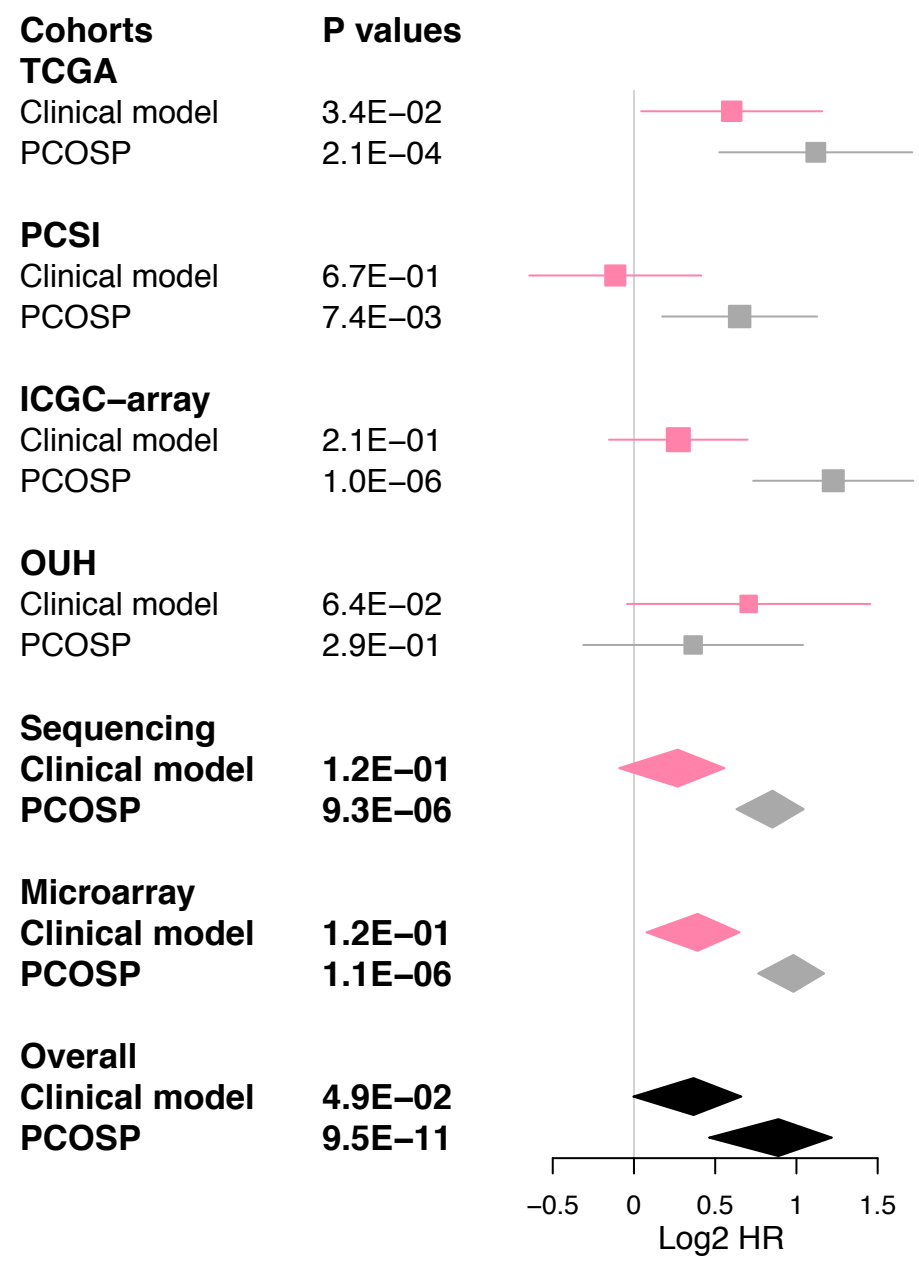
4D

Classifiers	PCOSP vs. Clinical model	Test of Superiority (P value)
C-index	0.63 vs. 0.55	2.1E-02
HR	1.84 vs. 1.23	1.4E-02

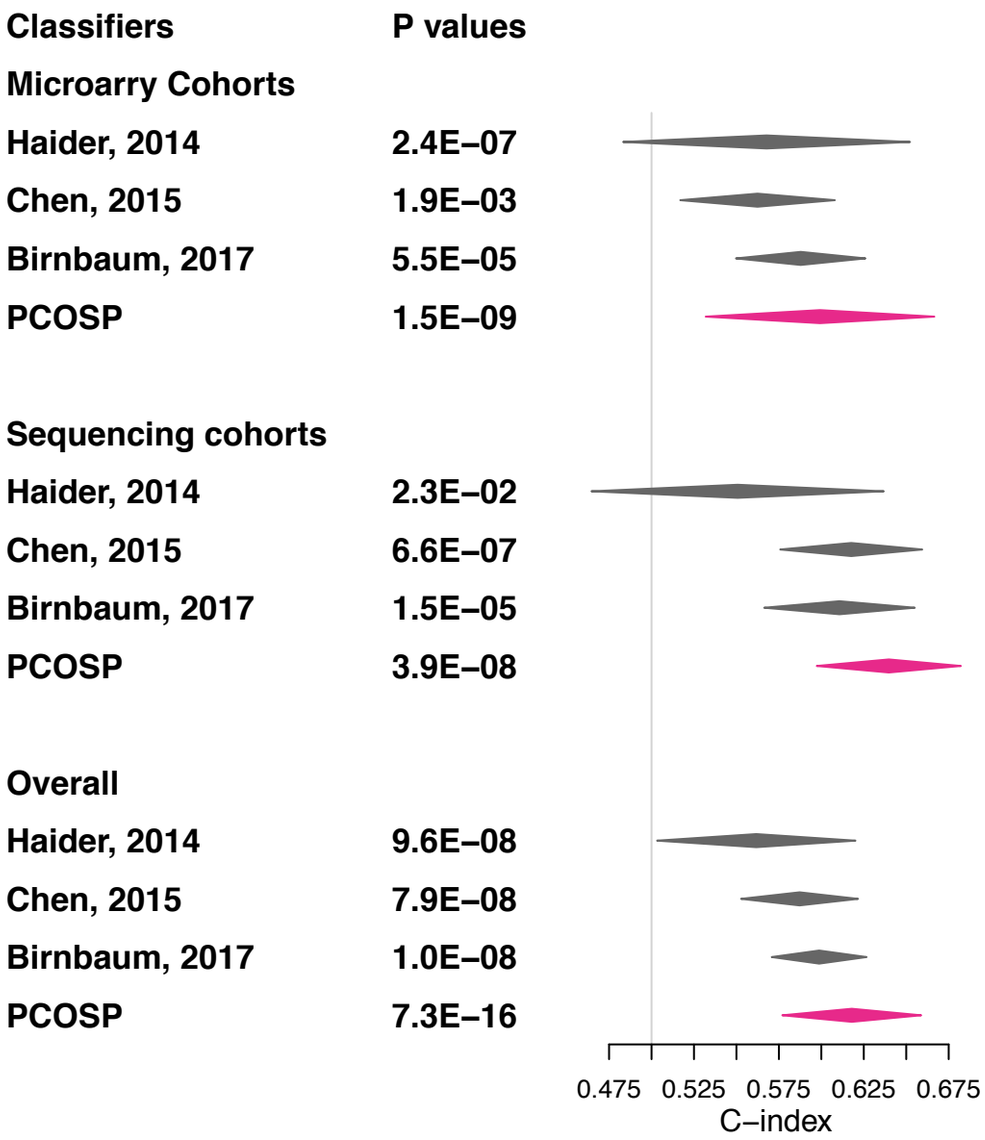
4B



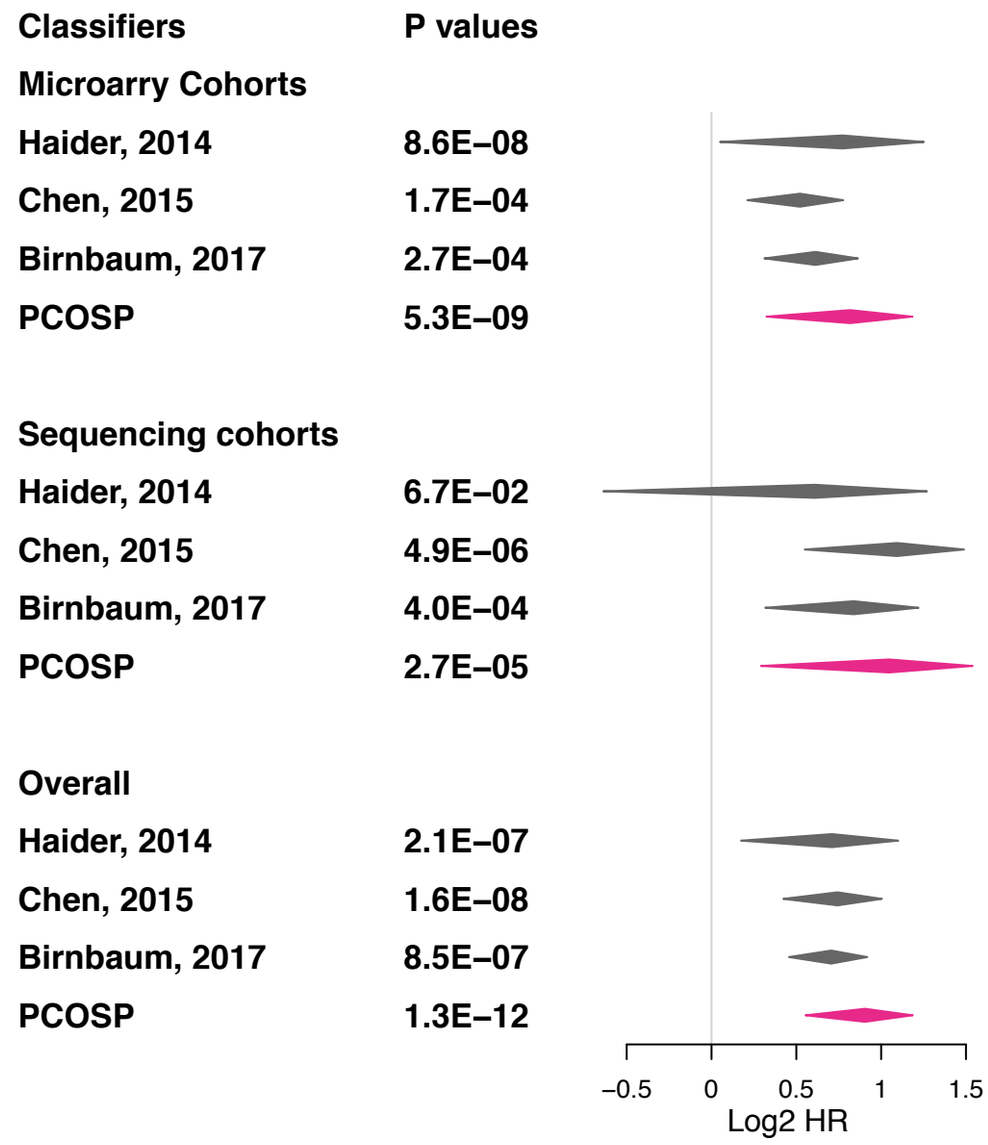
4C



5A



5B



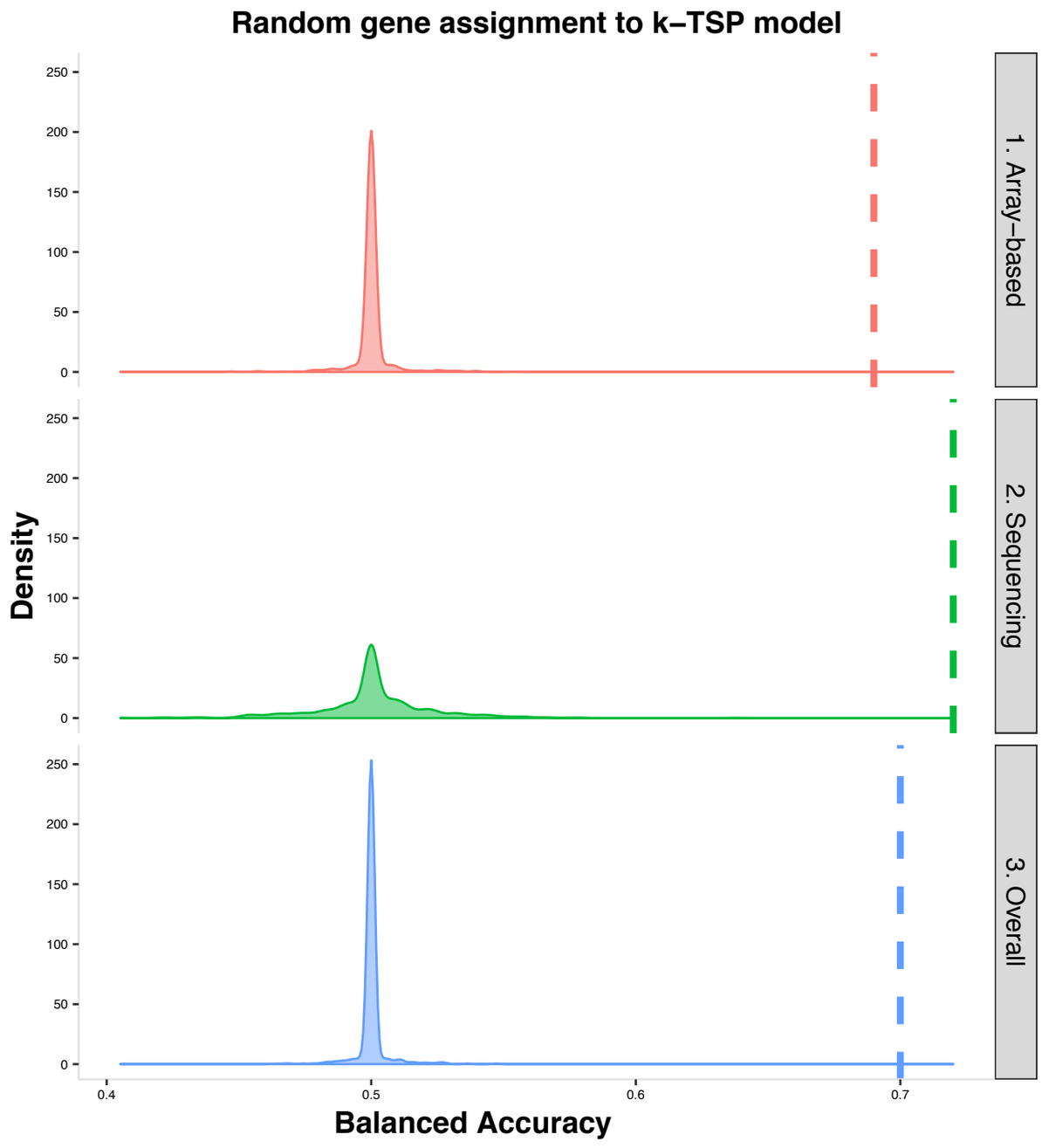
5C Meta concordance index comparison

Classifiers	PCOSP vs. other classifier	Test of Superiority (P value)
PCOSP vs. Birnbaum, 2017	0.62 vs. 0.59	3.4E-02
PCOSP vs. Chen, 2015	0.62 vs. 0.58	1.0E-02
PCOSP vs. Haider, 2014	0.62 vs. 0.56	2.6E-04

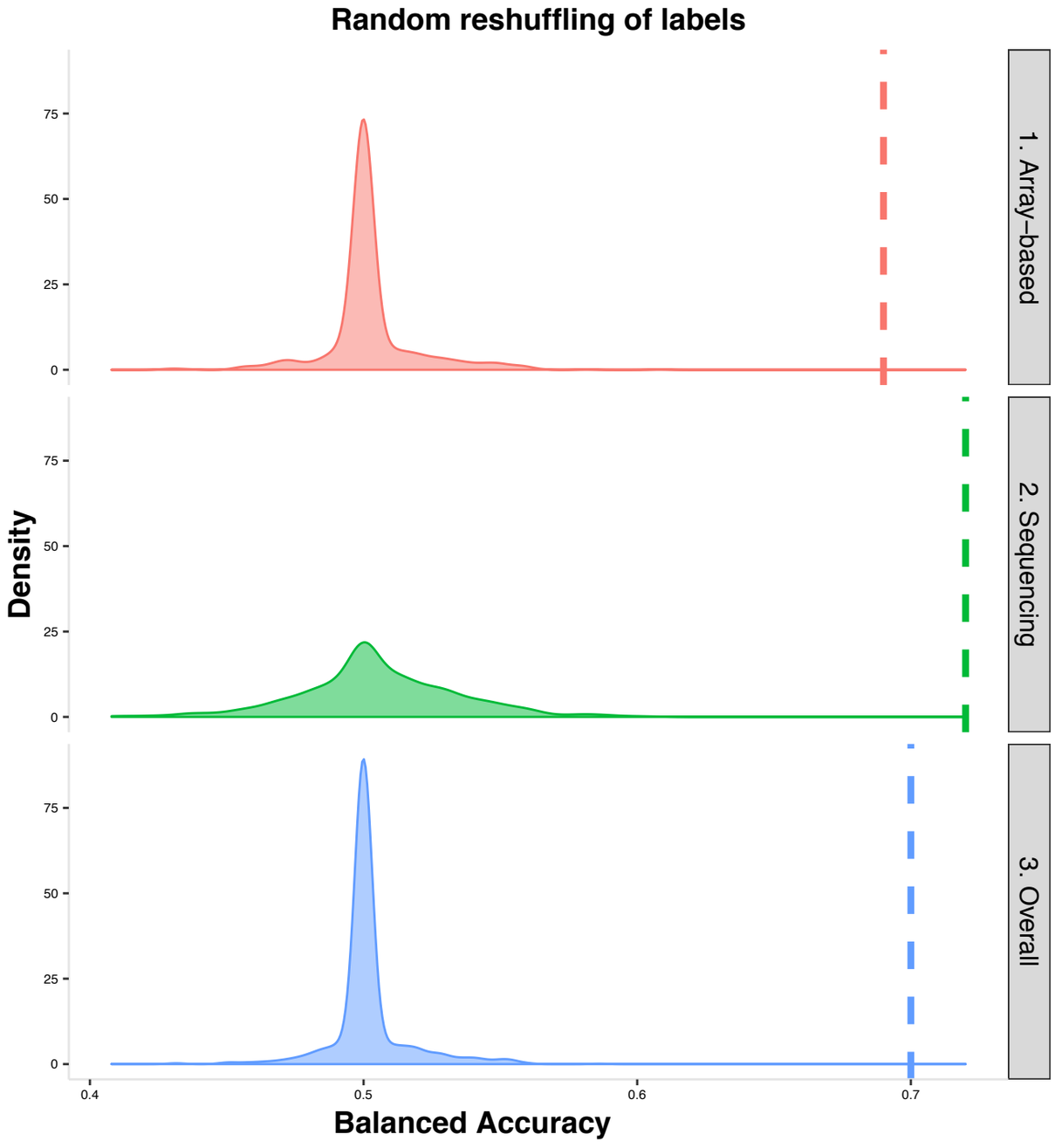
5D Meta HR comparison

Classifiers	PCOSP vs. other classifier	Test of Superiority (P value)
PCOSP vs. Birnbaum, 2017	1.76 vs. 1.55	4.5E-02
PCOSP vs. Chen, 2015	1.76 vs. 1.63	1.0E-01
PCOSP vs. Haider, 2014	1.76 vs. 1.43	7.4E-03

S1A

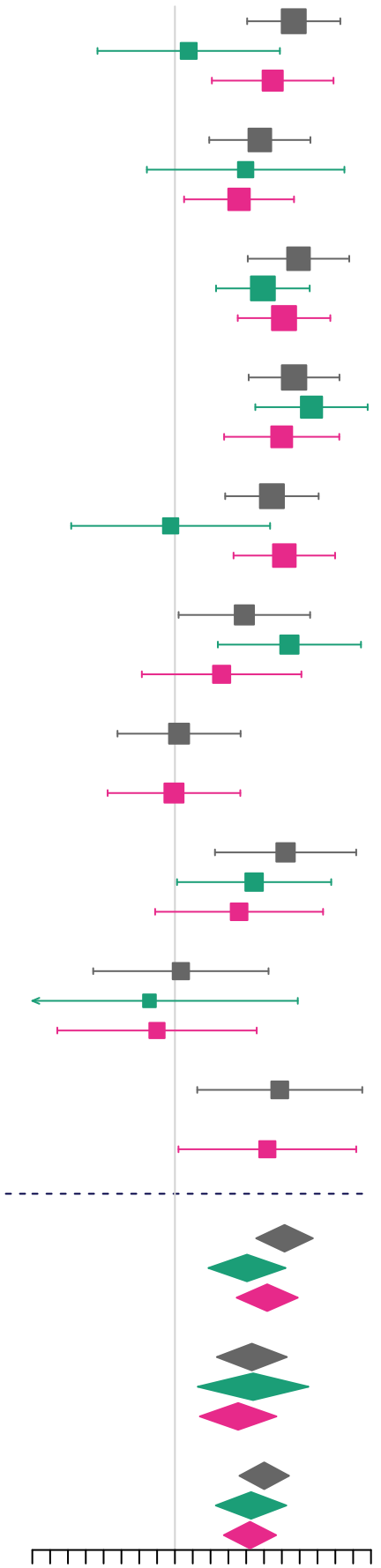

Random gene assignment to k−TSP model
Density
Balanced Accuracy
1. Array−based
2. Sequencing
3. Overall
0
50
100
150
200
250
0.4
0.5
0.6
0.7

S1B


Random reshuffling of labels
Density
Balanced Accuracy
1. Array−based
2. Sequencing
3. Overall
0
25
50
75
0.4
0.5
0.6
0.7

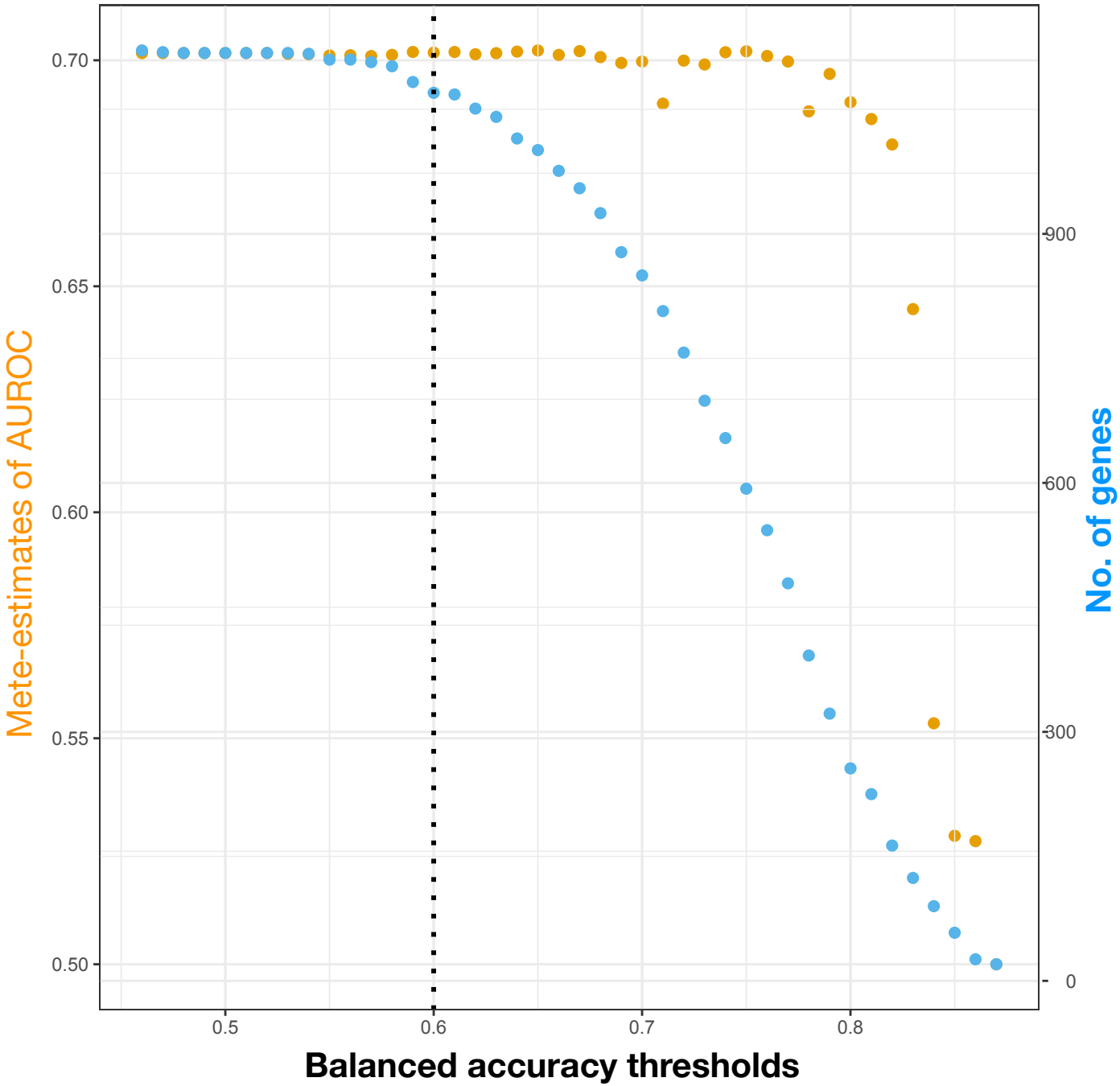
S2A

Cohorts	P values
TCGA (n=146)	5.7E−07
Basal−subtype (n=29)	7.7E−01
Classical−subtype (n=117)	1.6E−03
PCSI (n=118)	1.0E−03
Basal−subtype (n=25)	1.6E−01
Classical−subtype (n=93)	2.2E−02
Kirby (n=51)	1.8E−06
Basal−subtype (n=13)	2.3E−04
Classical−subtype (n=38)	3.8E−06
ICGC−array (n=178)	2.6E−07
Basal−subtype (n=156)	1.9E−06
Classical−subtype (n=22)	2.8E−04
UNC (n=125)	4.7E−05
Basal−subtype (n=27)	9.3E−01
Classical−subtype (n=98)	2.4E−05
Chen (n=63)	3.8E−02
Basal−subtype (n=20)	1.7E−03
Classical−subtype (n=43)	2.5E−01
OUH (n=48)	8.9E−01
Basal−subtype (n=08)	NA
Classical−subtype (n=40)	9.8E−01
Zhang (n=42)	2.1E−03
Basal−subtype (n=13)	4.4E−02
Classical−subtype (n=29)	1.3E−01
Winter (n=30)	8.9E−01
Basal−subtype (n=8)	7.4E−01
Classical−subtype (n=22)	7.2E−01
Collisson (n=27)	1.3E−02
Basal−subtype (n=1)	
Classical−subtype (n=26)	4.2E−02
Sequencing	**7.9E−12**
Basal−subtype	**1.1E−01**
Classical−subtype	**2.7E−06**
Microarray	**2.1E−11**
Basal−subtype	**7.1E−06**
Classical−subtype	**7.7E−07**
Overall	**3.5E−21**
Basal−subtype	**7.7E−06**
Classical−subtype	**1.9E−11**



S2B

Cohorts	P values
TCGA (n=146)	1.7E−04
Basal−subtype (n=29)	4.3E−01
Classical−subtype (n=117)	1.7E−02
PCSI (n=118)	1.0E−02
Basal−subtype (n=25)	5.2E−01
Classical−subtype (n=93)	1.6E−02
Kirby (n=51)	7.4E−05
Basal−subtype (n=13)	1.1E−01
Classical−subtype (n=38)	6.2E−03
ICGC−array (n=178)	4.2E−07
Basal−subtype (n=156)	2.4E−02
Classical−subtype (n=22)	1.2E−04
UNC (n=125)	2.5E−03
Basal−subtype (n=27)	9.7E−01
Classical−subtype (n=98)	4.6E−03
Chen (n=63)	2.2E−02
Basal−subtype (n=20)	9.1E−02
Classical−subtype (n=43)	4.4E−02
OUH (n=48)	2.9E−01
Basal−subtype (n=08)	
Classical−subtype (n=40)	3.5E−01
Zhang (n=42)	2.6E−02
Basal−subtype (n=13)	2.8E−01
Classical−subtype (n=29)	4.9E−01
Winter (n=30)	6.4E−01
Basal−subtype (n=8)	6.0E−01
Classical−subtype (n=22)	7.5E−01
Collisson (n=27)	2.8E−02
Basal−subtype (n=1)	
Classical−subtype (n=26)	6.3E−02
Sequencing	**1.7E−07**
Basal−subtype	**3.1E−01**
Classical−subtype	**2.1E−04**
Microarray	**1.5E−10**
Basal−subtype	**8.3E−02**
Classical−subtype	**1.1E−06**
Overall	**2.7E−16**
Basal−subtype	**7.5E−02**
Classical−subtype	**2.1E−09**


Mete-estimates of AUROC
0.70
0.65
0.60
0.55
0.50
No. of genes
900
600
300
0
0.5
0.6
0.7
0.8
Balanced accuracy thresholds