Evaluating the transcriptional fidelity of cancer models

Da Peng[1*], Rachel Gleyzer[2*], Wen-Hsin Tai[2], Pavithra Kumar[2], Qin Bian[2], Bradley Issacs[2], Edroaldo Lumertz da Rocha[3], Stephanie Cai[1], Kathleen DiNapoli[4,5], Patrick Cahan[1,2,6]

[1]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore MD 21205 USA

[2]Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore MD 21205 USA

[3]Department of Biochemistry and Molecular Pharmacology, Boston Children's Hospital and Harvard Medical School, Boston MA 02115 USA

[4]Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, MD 21205 USA

[5]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore MD 21218 USA

[6]Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore MD 21205 USA

* These authors made equal contributions.

Correspondence to: patrick.cahan@jhmi.edu

Article type: Analysis

Website: http://www.cahanlab.org/resources/cancerCellNet_web

Code: https://github.com/pcahan1/cancerCellNet

ABSTRACT

Cancer researchers use cell lines, patient derived xenografts, and genetically engineered mice as models to investigate tumor biology and to identify therapies. The generalizability and power of a model derives from the fidelity with which it represents the tumor type of investigation, however, the extent to which this is true is often unclear. The preponderance of models and the ability to readily generate new ones has created a demand for tools that can measure the extent and ways in which cancer models resemble or diverge from native tumors. Here, we present a computational tool, CancerCellNet, that measures the similarity of cancer models to 22 naturally occurring tumor types and 36 subtypes, in a platform and species agnostic manner. We applied this tool to 657 cancer cell lines, 415 patient derived xenografts, and 26 distinct genetically engineered mouse models, documenting the most faithful models, identifying cancers underserved by adequate models, and finding models with annotations that do not match their classification. By comparing models across modalities, we find that genetically engineered mice have higher transcriptional fidelity than patient derived xenografts and cell lines in four out of five tumor types. We have made CancerCellNet available as freely downloadable software and as a web application that can be applied to new cancer models.

INTRODUCTION

Models are widely used to investigate cancer biology and to identify potential therapeutics. Popular modeling modalities are cancer cell lines (CCLs)[1], genetically engineered mouse models (GEMMs)[2], and patient derived xenografts (PDXs)[3].These classes of models differ in the types of questions that they are designed to address. CCLs are often used to address cell intrinsic mechanistic questions[4], GEMMs to chart progression of molecularly defined-disease[5], and PDXs to explore patient-specific response to therapy in a physiologically relevant context[6]. Models also differ in the extent to which the they represent specific aspects of a cancer type[7]. Even with this intra- and inter-class model variation, all models should represent the tumor type or subtype under investigation, and not another type of tumor, and not a non-cancerous tissue. Therefore, cancer-models should be selected not only based on the specific biological question but also based on the similarity of the model to the cancer type under investigation[8,9].

Various methods have been proposed to determine the similarity of cancer models to their intended subjects. Domcke et al devised a 'suitability score' as a metric of the molecular similarity of CCLs to high grade serous ovarian carcinoma based on a heuristic weighting of copy number alterations, mutation status of several genes that distinguish ovarian cancer subtypes, and hypermutation status[10]. Other studies have taken analogous approaches by either focusing on transcriptomic or ensemble molecular profiles (e.g. transcriptomic and copy number alterations) to quantify the similarity of cell lines to tumors[11–13]. These studies were tumor-type specific, focusing on CCLs that model, for example, hepatocellular carcinoma or breast cancer. More recently, Yu et al compared the transcriptomes of CCLs to the Cancer Genome Atlas (TCGA) by correlation analysis, resulting in a panel of CCLs recommended as most representative of 22 tumor types[14]. While all of these studies have provided valuable information, they leave two major challenges unmet. The first challenge is to determine the fidelity of GEMMs and PDXs and whether there are stark differences between these classes of models and CCLs. The other major unmet challenge is to enable the rapid assessment of new, emerging cancer models. This challenge is especially relevant now as technical barriers to generating models have been substantially lowered[15,16], and because each PDX can be considered a distinct entity requiring individual validation[17].

To address these challenges, we developed CancerCellNet (CCN), a computational tool that uses transcriptomic data to quantitatively assess the similarity between cancer models and 22 naturally occurring tumor types and 36 subtypes in a platform- and species-agnostic manner. Here, we describe CCN's performance, and the results of applying it to assess 657 cancer cell lines, 415 patient derived xenografts, and 26 distinct genetically engineered mouse models. This has allowed us to identify the most faithful models currently available, to document cancers underserved by adequate models, and to find models with inaccurate tumor type annotation. Moreover, because CCN is open-source and easy to use, it can be readily applied to newly generated cancer models as a means to assess their fidelity.

RESULTS

CancerCellNet classifies samples accurately across species and technologies

Previously, we had developed a computational tool using the Random Forest classification method to measure the similarity of engineered cell populations to their *in vivo* counterparts based on transcriptional profiles[18,19]. More recently, we elaborated on this approach to allow for classification of single cell RNA-Seq data in a manner that allows for cross-platform and cross-species analysis[20]. Here, we used an analogous approach to quantitatively compare cancer models to naturally occurring patient tumors (**Fig 1A**). In brief, we used TCGA RNA-seq expression data from 22 solid tumor types to train a top-pair multi-class Random forest classifier. We combined training data from Rectal Adenocarcinoma (READ) and Colon Adenocarcinoma (COAD) into one COAD_READ category because READ and COAD are considered to be virtually indistinguishable at a molecular level[21]. We included an 'Unknown' category trained using randomly shuffled gene-pair profiles generated from the training data of 22 tumor types to identify query samples that are not reflective of any of the training data.

We assessed the performance of this approach by computing the area under the precision recall (AUPR) curves derived by 50 iterations of cross validation (**Fig 1B, Supp Fig 1A**). In the cross validations, the mean AUPR exceeded 0.95 in most of the tumor types. In addition to achieving high mean AUPRs on held-out TCGA data, we found that CCN also achieved high AUPR (above 0.9) when we applied it to independent testing data from the International Cancer Genome Consortium (ICGC) consisting of RNA-Seq data from 886 tumors across 5 tumor types (**Supp Fig 1B**)[22]**.**

As one of the central aims of our study is to compare distinct cancer models, including GEMMs, our method needed to be able to classify samples from mouse and human samples equivalently. We used the Top-Pair transform[20] to achieve this and we tested the feasibility of this approach by assessing the performance of a normal (i.e non-tumor) cell and tissue classifier trained on human data as applied to mouse samples. Consistent with prior applications[23], we found that the cross-species classifier performed well, achieving mean AUPR of 0.96 when applied to mouse data (**Supp Fig 1C**)**.**

To evaluate cancer models at a finer resolution, we also developed an approach to perform tumor subtype classifications (**Supp Fig 1D**). We constructed 11 different cancer subtype classifiers based on the availability of expression or histological subtype information[21,24–34]. We also included non-cancerous, normal tissues as categories for several subtype classifiers when sufficient data was available: breast invasive carcinoma (BRCA), COAD_READ, head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC) and uterine corpus endometrial carcinoma (UCEC). The 11 subtype classifiers all achieved high overall average AUPRs ranging from 0.75 to 0.99 (**Supp Fig 1E**)**.**

Fidelity of cancer cell lines

Having validated the performance of CCN, we then used it to determine the fidelity of CCLs. We mined RNA-Seq expression data of 657 different cell lines across 20 cancer types from the Cancer Cell Line Encyclopedia (CCLE) and applied CCN to them, finding a wide classification range for cell lines of each tumor type (**Fig 2A, Supp Tab 1**). To verify the classification results, we applied CCN to CCLE expression profiles generated through microarray expression profiling[35]. To ensure that CCN would function on microarray data, we first tested it by applying a CCN classifier created to test microarray data to 720 expression profiles of 12 tumor types. The cross-platform CCN classifier performed well, based on the comparison to study-provided annotation, achieving a mean AUPR of 0.944 (**Supp Fig 2A**). Next, we applied this cross-platform classifier to microarray expression profiles of CCLE (**Supp Fig 2B**). From the classification results of 571 cell lines that have both RNA-seq and microarray expression profiles, we found a strong positive association between the classification scores from RNA-seq and those from microarray (**Supp Fig 2C**). This comparison supports the notion that the classification scores for each cell line are not artifacts of profiling methodology. Moreover, this comparison shows that the scores are consistent between the times that the cell lines were first assayed by microarray expression profiling in 2012 and by RNA-Seq in 2019. We also observed high level of correlation between our analysis and the analysis done by Yu et al[14](**Supp Fig 2D**), further validating the robustness of the CCN results.

Next, we assessed the extent to which CCN classifications agreed with their nominal tumor type of origin. We annotated cell lines based their CCN score profile as follows. 'Correct' Cell lines with CCN score > 0.3 for the tumor type of origin were annotated 'correct'. Those with CCN scores > 0.3 in the tumor type of origin and at least one other tumor type were annotated as 'mixed'. Cell lines with CCN scores > 0.3 for tumor types other than that of the cell lines origin were annotated as 'other', and those lines that did not received a CCN score > 0.3 for any tumor type were annotated as 'none' (**Fig 2B**). We selected a decision threshold of 0.3 based on the average of the threshold that produced the highest Macro F1 measure, harmonic mean of precision and recall, across 50 cross validations. We found that majority of cell lines originally annotated as Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Skin Cutaneous Melanoma (SKCM), Colorectal Cancer (COAD_READ) and Sarcoma (SARC) fell into the 'correct' category (**Fig 2B**). On the other hand, no Esophageal carcinoma (ESCA) or Brain Lower Grade Glioma (LGG) were classified as 'correct', demonstrating the need for more transcriptionally faithful cell lines that model those general cancer types.

There are several possible explanations for cell lines not receiving a 'correct' classification. One possibility is that the sample was incorrectly labeled in the study from which we harvest the expression data. Consistent with this explanation, we found that colorectal cancer line NCI-H684[36,37], a cell line labelled as liver hepatocellular carcinoma (LIHC) by CCLE, was classified strongly as COAD_READ (**Supp Tab 1**). Another possibility to explain low CCN score is that cell lines were derived from subtypes of tumors that are not well-represented in TCGA. To explore this hypothesis, we first performed tumor subtype classification on the CCLE lines from 11 tumor types for which we had trained subtype classifiers (**Supp Tab 2**). We reasoned that if a cell was a good model for a rarer subtype, then it would receive a poor general classification but a high classification for the subtype that it models well. Therefore, we counted the number of lines that fit this pattern. We found that of the 223 lines with no general classification, 54 (24%) were classified as a specific subtype, suggesting that derivation from rare subtypes is not the major contributor to the poor overall fidelity of CCLs.

Another potential contributor to low scoring cell lines is intra-tumor stromal and immune cell impurity in the training data. If impurity were a confounder of CCN scoring, then we would expect a strong positive correlation between mean purity and mean CCN classification of CCLs per general tumor type. However, the Pearson correlation coefficient between the mean purity of general tumor type and mean CCN classification scores of CCLs in the corresponding general tumor type was low (0.059), suggesting that tumor purity is not a major contributor to the low CCN scores across CCLE (**Supp Fig 2E**).

To more directly assess the impact of intra-tumor heterogeneity in the training data on evaluating cell lines, we constructed a classifier using cell types found in human melanoma and glioblastoma scRNA-seq data[38, 39]. Previously, we have demonstrated the feasibility of using our classification approach on scRNA-seq data[23]. Our scRNA-seq classifier achieved a high average AUPR (0.95) when applied to held-out data (**Supp Fig 3A-B**). Comparing the CCN score from bulk RNA-seq general classifier and scRNA-seq classifier, we observed a high level of correlation (Pearson correlation of 0.89) between the SKCM CCN classification scores and scRNA-seq SKCM malignant CCN classification scores for SKCM cell lines (**Fig 2C, Supp Fig 3C**). Among the 37 SKCM cell lines that were classified as SKCM in general classification, 36 SKCM cell lines were also classified as SKCM malignant cells in scRNA-seq classifier. Interestingly, we also observed a high correlation between the SARC CCN classification score and scRNA-seq cancer associated fibroblast (CAF) CCN classification scores (Pearson correlation of 0.89). Six of the 10 SKCM cell lines that had been classified as SARC by CCN were classified as CAF by the scRNAseq classifier (**Fig 2D, Supp Fig 3C**), which suggests the possibility that these cell lines were derived from CAF or other mesenchymal populations, or that they have acquired a mesenchymal character through their derivation. The high level of agreement between scRNA-seq and bulk RNA-seq classification results shows that heterogeneity in the training data of general CCN classifier has little impact in the classification of SKCM cell lines.

In contrast, we observed a weaker correlation between GBM CCN classification scores and scRNA-seq GBM neoplastic CCN classification scores (Pearson correlation of 0.58) for GBM cell lines (**Fig 2E, Supp Fig 3D**). Of the 32 GBM lines that were not classified as GBM with CCN, 26 were classified as GBM neoplastic cells with the scRNAseq classifier. Among the 27 GBM lines that were classified as SARC with CCN, 15 cell lines were classified as CAF (**Fig**

2F), and 10 of 15 lines were classified as both GBM neoplastic and CAF in the scRNA-seq classifier. Similar to the situation with SKCM lines that classify as CAF, this result is consistent with the possibility that some GBM lines classified as SARC by CCN could be derived from mesenchymal subtypes exhibiting both strong mesenchymal signatures and glioblastoma signatures or that they have acquired a mesenchymal character through their derivation[40]. The lower level of agreement between scRNA-seq and bulk RNA-seq classification results for GBM models suggests that the heterogeneity of glioblastomas[41] can impact the classification of GBM cell lines, and that the use of scRNA-seq classifier can resolve this deficiency.

Next, we explored the subtype classification of CCLs from three general tumor types in more depth, focusing first on UCEC. The histologically defined subtypes of UCEC, endometrioid and serous, differ in prevalence, molecular properties, prognosis, and treatment. For instance, the endometrioid subtype, which accounts for approximately 80% of uterine cancers, retains estrogen receptor and progesterone receptor status and is responsive towards progestin therapy[42,43]. Serous, a more aggressive subtype, is characterized by the loss of estrogen and progesterone receptor and is not responsive to progestin therapy[42,43]. CCN classified the majority of the UCEC cell lines as serous except for JHUEM-1 which is classified as endometrioid (**Fig 3A**). The preponderance CCLE lines of serous versus endometroid character may be due to properties of serous cancer cells that promote their *in vitro* propagation, such as upregulation in cell adhesion[44]. Some of our subtype classification results are consistent with prior observations. For example, HEC-1A, HEC-1B, and KLE were previously characterized as type II endometrial[45]. On the other hand, our subtype classification results contradict prior observations in at least one case. For instance, the Ishikawa cell line was derived from type I endometrial cancer (endometrioid histological subtype)[45,46], however CCN classified a derivative of this line, Ishikawa 02 ER-, as serous. The high serous CCN score could result from a shift in phenotype of the line concomitant with its loss of estrogen receptor (ER) as this is a distinguishing feature of type II endometrial cancer (serous histological subtype)[42]. Taken together, these results indicate a need for more endometroid-like CCLs.

Next, we examined the subtype classification of Lung Squamous Cell Carcinoma (LUSC) cell lines (**Fig 3C**). We found that of the 22 lines unclassified or misclassified in the general classifier, 6 (27%) were classified as a subtype. Among the LUSC cell lines that were classified as unknown in the general classifier and classified with a subtype, several cell lines had general classification scores modestly below the threshold. All the LUSC lines with at least one subtype classification had an underlying primitive subtype classification. This is consistent either with the ease of deriving lines from tumors with a primitive character, or with a process by which cell line derivation promotes similarity to more the primitive subtype, which is marked by increased cellular proliferation[26]. Some of our results are consistent with prior reports that have investigated the resemblance of some lines to LUSC subtypes. For example, HCC-95, classified as classical and primitive subtype, has previously been characterized as classical[26,47]. Further, LUDLU-1, classified as a primitive subtype, has classification signal in classical subtype which was previously characterized as resembling classical[47]. Likewise, although EPLC-272H was also classified as primitive subtype, it has relatively high CCN score in the basal subtype, which corresponds to its previous characterization as basal[47]. Lung Adenocarcinoma (LUAD) cell lines had classification results similar to LUSC: most lines did not classify as LUAD in the general classifier (66 of 76) (**Fig 3B**). The cell lines that were classified as a subtype were either classified as proximal inflammation, proximal proliferation or a mix of the two. RERF-LC-Ad1 had the highest general classification score and the highest proximal inflammation subtype classification score. Taken together, these subtype classification results have revealed an absence of cell lines models for basal, classical, and secretory LUSC, and for the Terminal respiratory unit (TRU) LUAD subtype.

Finally, we sought to measure the extent to which cell line transcriptional fidelity related to model use. We used the number of papers in which a model was mentioned, normalized by the number of years since the cell line was documented, as a rough approximation of model usage. To explore this metric, we plotted the normalized citation count versus general classification score, labeling the highest cited and highest classified cell lines from each general tumor type (**Fig 3D**). For most of the general tumor types, the highest cited cell line is not the highest classified cell line except for Hep G2 and ML-1, representing liver hepatocellular carcinoma (LIHC) and thyroid carcinoma (THCA), respectively. On the other hand, the general scores of the highest cited cell lines representing BLCA, SKCM, BRCA, PRAD and COAD_READ fall below the classification threshold of 0.3. Notably, each of these tumor types have other lines with scores exceeding 0.7, which should be considered as more faithful transcriptional models when selecting lines for a study (**Supp Table 1 and http://www.cahanlab.org/resources/cancerCellNet_results/**).

Evaluation of patient derived xenografts

Next, we sought to evaluate a more recent class of cancer models: PDX. To do so, we subjected the RNA-Seq expression profiles of 415 PDX models from 13 different types of cancer types generated previously[17] to CCN. Similar to the results of CCLE, the PDXs exhibited a wide range of classification scores (**Fig 4A, Supp Tab 3**). By categorizing the CCN scores of PDX based on the proportion of samples associated with each tumor type that were correctly classified**,** we found that SARC, SKCM, COAD_READ and BRCA have higher proportion of correctly classified PDX than those of other cancer categories (**Fig 4B**). In contrast to CCLE, we found a higher proportion of correctly classified PDX in Stomach adenocarcinoma (STAD) and KIRC (**Fig 4B**)**.** However, similar to CCLE, no ESCA PDXs were classified as such. This held true when we performed subtype classification on PDX samples: none of the PDX in ESCA were classified as any of the ESCA subtypes (**Supp Tab 4**). UCEC PDXs had both endometrioid subtypes, serous subtypes, and mixed subtypes, which provided a broader representation than in CCLE (**Fig 4C**). Many LUSC PDXs that were classified as a subtype were also classified as Head and Neck squamous cell carcinoma (HNSC) (**Fig 4D**). This could be due to the similarity in expression profiles of basal and classical subtypes of HNSC and LUSC[26,48], which is consistent with the observation that these PDXs were also subtyped as basal and classical. No LUSC PDXs were classified as the secretory subtype. While eight of the LUAD PDX samples were classified as the unknown subtype class classification, the remaining six classified as proximal proliferative or proximal inflammatory (**Fig 4E**). Finally, similar to the CCLE, there were no TRU subtypes in the PDX cohort. In summary, we found that while individual PDXs can reach extremely high transcriptional fidelity to both general tumor types and subtypes, many PDXs were not classified as the general tumor type from which they originated.

Evaluation of GEMMs

Next, we used CCN to evaluate GEMMs of six general tumor types from nine studies for which expression data was publicly available[49–57]. As was true for CCLs and PDXs, GEMMs also had a wide range of CCN scores (**Fig 5A, Supp Tab 5**). We next categorized the CCN scores based on the proportion of samples associated with each tumor type that were correctly classified (**Fig 5B**)**.** In contrast to CCLs and PDXs, the GEMM dataset included multiple replicates per model, which allowed us to examine intra-GEMM variability. Both at the level of CCN score and at the level of categorization, GEMMs were highly invariant. For example, replicate of UCEC GEMMs driven by Prg(cre/+)Pten(lox/lox) received almost identical general and subtype classification profiles (**Supp Fig 4, Supp Tab 6**). GEMMs sharing genotypes across studies such as LUAD GEMMs driven by Kras mutation and loss of p53[49,55,57] received similar general and subtype classification scores (**Fig. 5A,B,D**)**.** Even GEMMs with mixed classifications received consistent CCN scores. For example, LGG GEMMs, generated by Nf1 mutations expressed in different neural progenitors in combination with Pten deletion[56], consistently received mixed classification as both LGG and GBM (**Fig 5A**)**.**

To explore the extent to which driver genotype impacted subtype classification, we examined two general tumor types in which there were GEMMs with different tumor drivers: LUSC and LUAD. The LUSC GEMMs were generated using loss of Lkb1 and either overexpression of Sox2 (via two distinct mechanisms) or loss of Pten[55]. Although most of the lenti-Sox2-Cre-infected;Lkb1^{fl}/^{fl} and Rosa26LSL-Sox2-IRES-GFP;Lkb1^{fl}/^{fl} samples were classified as unknown, their general LUSC CCN scores were only modestly lower than the decision threshold and consistently throughout (**Fig 5C**). Those two models also classified mostly as secretory subtype of LUSC. The consistency is not surprising given both models overexpress Sox2 and lose Lkb1. Most of the Lkb1$^{fl/fl}$;Pten$^{fl/fl}$ GEMMs received unknown general classifications with general LUSC CCN scores substantially lower than those of lenti-Sox2-Cre-infected;Lkb1^{fl}/^{fl} samples and Rosa26LSL-Sox2-IRES-GFP;Lkb1^{fl}/^{fl} samples. Moreover, our subtype classification indicated that this GEMM was mostly classified as unknown, in contrast to prior reports suggesting that it is most similar to a basal subtype[58]. The lenti-Sox2-Cre-infected;Lkb1^{fl}/^{fl} samples received high secretory subtype scores, whereas the Rosa26LSL-Sox2-IRES-GFP;Lkb1^{fl}/^{fl} samples were classified as a more balanced mix of secretory and primitive subtypes. None of the three LUSC GEMMs have strong classical or basal sub-type CCN scores.

All of the LUAD GEMMs, which were generated using various combinations of activating Kras mutation, loss of Trp53, and loss of Smarca4L[49,55,57], were correctly classified (**Fig 5D**). There were no substantial differences in general, or subtype classification across driver genotypes. Notably, the subtypes tended to have CCN scores in mixture of proximal proliferation, proximal inflammation and TRU. Taken together, this analysis suggests that there is a degree of similarity, and perhaps plasticity between the primitive and secretory (but not basal or classical) subtypes of LUSC. On the other hand, while the LUAD GEMMs classify strongly as LUAD, do not have strong particular subtype classification -- a result that does not vary by genotype.

Comparison of CCLs, PDXs, and GEMMs

Finally, we sought to estimate the comparative transcriptional fidelity of the three cancer models modalities, limiting our comparison to those five general tumor types for which there were at least two examples per modality: UCEC, Pancreatic adenocarcinoma (PAAD), LUSC, LUAD, and LIHC. We compared the general CCN scores of each model on a per tumor type basis (**Fig 6A**). In the case of GEMMs, we used the mean classification score of all samples with shared genotypes. We found that GEMMs had the highest median general classification scores in four out of the five tumor types. However, some PDXs achieved the highest classification scores. In UCEC, LUAD and LIHC, the maximum classification score of PDXs exceeded 0.75 and were thus comparable to the majority of scores on held out TCGA data, highlighting the potential for PDXs to mirror the transcriptional state of natural tumors (**Fig 6A**). Because the CCN score is based on a moderate number of gene pairs (i.e. 1647) relative to the total number of protein-coding genes, it is possible that a cancer model with a high CCN score might not have a high global similarity to a naturally occurring tumor. Therefore, we also calculated the GRN status, a metric of the extent to which tumor-type specific gene regulatory network is established[18], for all models (**Supp Fig 5**). We observed high level of correlation between the two similarity metrics, which suggests that although CCN classifies on a selected set of genes, its scores are highly correlated with global assessment of transcriptional similarity.

We also sought to compare model modalities in terms of the diversity of subtypes that they represent (**Supp Fig 6**). As a reference, we also included in this analysis the overall subtype incidence, as approximated by incidence in TCGA. In models of UCEC, there is a notable difference in endometroid incidence, and the proportion of models classified as endometroid, with only PDX having any representatives (**Fig 6B**). The vast majority of CCLs and all of the GEMM models of PAAD have an unknown subtype classification. However, the majority of PDXs are subtyped as either a mixture of basal and classical, or basal and classical alone. LUSC have proximal inflammation and proximal proliferation subtypes modelled by CCLs and PDX, and TRU subtype modelled by GEMMs exclusively (**Fig 6B**). Likewise, LUAD have basal, classical and primitive subtypes modelled by CCLs and PDXs, and secretory subtype modelled by GEMMs exclusively (**Fig 6B**). Taken together, these results demonstrate the need to carefully select different model systems to more suitably model certain cancer subtypes.

DISCUSSION

A major goal in the field of cancer biology is to develop models that mimic naturally occurring tumors with enough fidelity to enable therapeutic discoveries. However, methods to measure the extent to which cancer models resemble or diverge from native tumors are lacking. This is especially problematic now because there are many existing models from which to choose, and it has become easier to generate new models. Here, we present CancerCellNet (CCN), a computational tool that measures the similarity of cancer models to 22 naturally occurring tumor types and 36 subtypes. Because CCN is platform- and species-agnostic, it can be applied across many model modalities, including CCLs, PDXs, and GEMMs, and thus it represents a consistent platform to compare models across modalities. Here, we applied CCN to 657 cancer cell lines, 415 patient derived xenografts, and 26 distinct genetically engineered mouse models. Several lessons emerged from our computational analyses that have implications for the field of cancer biology.

First, CancerCellNet indicates that GEMMs are transcriptionally the most faithful models of four out of five general tumor types for which data from all modalities was available. This is consistent with the fact that GEMMs are typically derived by recapitulating well-defined driver mutations of natural tumors, and thus this observation corroborates the importance of genetics in the etiology of cancer[59]. Moreover, in contrast to most PDXs, GEMMs are typically generated in immune replete hosts. Therefore, the higher fidelity of GEMMs may also be a result of the influence of a native immune system on GEMM tumors[60]. Second, PDXs and CCLs have lower scores that are comparable to each other. This is consistent with the observation that PDXs can undergo selective pressures in the host that distort the progression of genomic alterations away from what is observed in natural tumors[61]. Furthermore, the observation that a few PDXs have very high classification scores, approaching a level that is indistinguishable from held out TCGA data, suggests that under certain conditions, PDX can almost perfectly mimic natural tumors transcriptionally. It is unclear what are these conditions; it may be that these few PDXs were profiled prior to the acquisition of non-typical genomic alterations. Third, we have found that none of the samples that we evaluated here are transcriptionally adequate models of ESCA, and therefore this tumor type requires further attention to derive new models. Fourth, we found that in several tumor types, GEMMs tend to reflect mixtures of subtypes rather than conforming strongly to single subtypes. The reasons for this are not clear but it is possible that in the cases that we examined the histologically defined subtypes have a degree of plasticity that is exacerbated in the murine host environment. We have made the results of our analyses available online so that researchers can easily explore the performance of selected models or identify the best models for any of the 22 general tumor types and the 36 subtypes presented here.

Currently, there are several limitations to our CCN tool, and caveats to our analyses which indicate areas for future work and improvement. First, CCN is based on transcriptomic data but other molecular readouts of tumor state, such as profiles of the proteome[62], epigenome[63], non-coding RNA-ome[63], and genome[59] would be equally, if not more important, to mimic in a model system. Therefore, it is possible that some models reflect tumor behavior well, and because this behavior is not well predicted by transcriptome alone, these models have lower CCN scores. To both measure the extent that such situations exist, and to correct for them, we plan in the future to incorporate other omic data into CCN so as to make more accurate and integrated model evaluation possible. A second limitation is that in the cross-species analysis, CCN implicitly assumes that homologs are functionally equivalent. The extent to which they are not functionally equivalent determines how confounded the CCN results will be. This possibility seems to be of limited consequence based on the high performance of the normal tissue cross-species classifier and based on the fact that GEMMs have the highest median CCN scores. Finally, the TCGA training data is made up of RNA-Seq from bulk tumor samples, which necessarily includes non-tumor cells, whereas the CCLs are by definition cell lines of tumor origin. Therefore, CCLs theoretically could have artificially low CCN scores due to the presence of non-tumor cells in the training data. This problem appears to be limited as we found no correlation between tumor purity and CCN score in the CCLE samples. However, this problem is related to the question of intra-tumor heterogeneity. We demonstrated the feasibility of using CCN and single cell RNA-seq data to refine the evaluation of cancer cell lines contingent upon availability of scRNA-seq training data. As more sufficient training single cell RNA-Seq data accrues, CCN would be able to not only evaluate models on a per cell type basis, but also based on cellular composition.

To ensure that CCN is widely available we have developed a free web application, which performs CCN analysis on user-uploaded data and allows for direct comparison their to the cancer models evaluated here. We have also made the CCN code freely available under an Open Source license and as an easily installed R package, and we are actively supporting its further development. The documentation describes how to analyze model(s) and compare the results to the panel of models that we evaluated here, thereby allowing researchers to immediately compare their models to the broader field in a comprehensive and standard fashion.

Online Methods

Training General CancerCellNet Classifier

To generate training data sets, we downloaded 8991 patient tumor RNA-seq expression count matrix and their corresponding sample table across 22 different tumor types from TCGA using TCGAWorkflowData, TCGAbiolinks[64] and SummarizedExperiment[65] packages. We used all the patient tumor samples for training the general CCN classifier. Later, we found the intersecting genes between TCGA dataset and all the query samples (CCLs, PDXs, GEMMs), and used them as features for the feature engineering and selection process of building the classifier. To train the top pair Random Forest classifier, we used a method similar to our previous method[23]. CCN first normalized the training counts matrix by down-sampling the counts to 500,000 counts per cell. To significantly reduce the time and resource of generating gene pairs for all possible genes, CCN then selected 30 up-regulated genes, 30 down-regulated genes and 30 least differentially expressed genes for each of the 23 cancer categories using template matching[66] as the genes to generate top scoring gene pairs. In short, for each tumor type, CCN defined a template vector that labelled the training tumor samples in cancer type of interest as 1 and all other tumor samples as 0 CCN then calculated the Pearson correlation coefficient between template vector and gene expressions for all genes. The genes with strong match to template as either upregulated or downregulated had large absolute Pearson correlation coefficient. CCN chose the upregulated, downregulated and least differentially expressed genes based on the magnitude of Pearson correlation coefficient.

After CCN selected the genes for each cancer type, CCN generated gene pairs among those genes. Gene pair transformation was a method inspired by the top-scoring pair classifier[67] to allow compatibility of classifier with query expression profiles that were collected through different platforms (e.g. microarray query data applied to RNA-seq training data). In brief, the gene pair transformation compares 2 genes within an expression sample and encodes the "gene1_gene2" gene-pair as 1 if the first gene has higher expression than the second gene. Otherwise, gene pair transformation would encode the gene-pair as 0. Using all the gene pair combinations generated through the gene sets across all cancer types, CCN then selected top 75 discriminative gene pairs for each category using template matching (with large absolute Pearson correlation coefficient) described above.

After the top discriminative gene pairs were selected for each cancer categories, CCN concatenate all the gene pairs into a vector and gene pair transformed the training samples into a binary matrix with all the discriminative gene pairs as row names and all the training samples as column names. Using the binary gene pair matrix, CCN randomly shuffled the binary values across rows then across column generating random profiles that should not resemble training data from any of the cancer categories. CCN then sampled 70 random profiles, annotated them as “Unknown” and appended them to the training gene pair binary matrix as training data for the “Unknown” category.

Using gene pair binary training matrix, CCN constructed a multi-class Random Forest classifier of 2000 trees and used stratified sampling of 60 sample size to ensure balance of training data in constructing the decision trees. The specific parameters for the final CCN classifier using the function “broadClass_train” in the package cancerCellNet are in **Supp Tab 7**. The gene-pairs are in **Supp Tab 8.**

Validating General CancerCellNet Classifier

2/3 of patient tumor data from each cancer type were randomly sampled as training data to construct a CCN classifier. After the classifier was built, 35 held-out samples from each cancer categories were sampled and 40 “Unknown” profiles were generated for validation. CCN gene pair transformed the held-out data for assessment based on the top gene-pairs selected to construct the classifier. The process of randomly sample training set from 2/3 of all patient tumor data, train classifier and validate using validation set was repeated 50 times to have a more comprehensive assessment of the classifier. We used precision-recall curve and area under the precision-recall curve (AUPR) as our metric of assessing the classifiers.

Classifying Query Data into General Cancer Categories

We downloaded the RNA-seq cancer cell lines expression profiles and sample table from (https://portals.broadinstitute.org/ccle/data), and microarray cancer cell lines expression profiles and sample table from Barretina et al [35]. We received PDX expression estimates and sample annotations from the authors of Gao et al [17]. We gathered GEMM expression profiles from 9 different studies[49–57]. To use CCN classifier on GEMM data, the mouse genes from GEMM expression profiles were converted into their human homologs. The query samples were gene pair transformed using gene pairs selected from the training step, and then inputted into CCN classifier for classification. Each query classification profile was labelled as one of the four classification categories: “correct”, “mixed”, “none” and “other” based on classification profiles. If a sample has a CCN score higher than the decision threshold (0.3) in the labelled cancer category, we assign that as “correct”. If a sample has CCN score higher than the decision threshold in labelled cancer category and in other cancer categories, we assign that as “mixed”. If a sample has no CCN score higher than the decision threshold in any cancer category or has the highest CCN score in ‘Unknown’ category, then we assign it as “none”. If a sample has CCN score higher than the decision threshold in a cancer category or categories not including the labelled cancer category, we assign it as ”other”. We analyzed and visualized the results using R and R packages pheatmap[68] and ggplot2[69].

Cross-Species Assessment

To assess the performance of cross-species classification, we downloaded 1003 labelled human tissue/cell type and 1993 labelled mouse tissue/cell type RNA-seq expression profiles from Github (https://github.com/pcahan1/CellNet). We first converted the mouse genes into human homologous genes. Then we found the intersecting genes between mouse tissue/cell expression profiles and human tissue/cell expression profiles. Limiting the input of RNA-seq profiles to the intersecting genes, we trained a CCN classifier with all the human tissue/cell expression profiles. The parameters used for the function “broadClass_train” in the package cancerCellNet are in **Supp Tab 7.** After the classifier was trained, we randomly sampled 75 samples from each tissue category in mouse tissue/cell data and applied the classifier on those samples to assess performance.

Cross-Technology Assessment

To assess the performance of CCN in applications to microarray data, we gathered 6,219 patient tumor microarray profiles across 12 different cancer types from more than 100 different projects (**Supp Tab 9**). We found the intersecting genes between the microarray profiles and TCGA patient RNA-seq profiles. Limiting the input of RAN-seq profiles to the intersecting genes, we created a CCN classifier with all the TCGA patient profiles using parameters for the function "broadClass_train" listed in **Supp Tab 7**. After the microarray specific classifier was trained, we randomly sampled 60 microarray patient samples from each cancer category, and applied CCN classifier on them as assessment of the cross-technology performance in **Supp Fig 3A**. The same CCN classifier was used to assess microarray CCL samples **Supp Fig 3B**.

Training and validating scRNA-seq Classifier

We extracted labelled human melanoma and glioblastoma scRNA-seq expression profiles[38,39], and compiled the two datasets excluding 3 cell types T.CD4, T.CD8 and Meyloid due to low number of cells for training. 60 cells from each of the 11 cell types were sampled for training a scRNA-seq classifier. The parameters for training a general scRNA-seq classifier using the function "broadClass_train" are in **Supp Tab 7.** 25 cells from each of the 11 cell types from the held-out data were selected to assess the single cell classifier**.** Using the PR curve and maximizing Macro F1 measure, we selected the decision threshold of 0.255. We then applied the scRNA-seq classifier on SKCM CCLs and GBM CCLs.

Training Subtype CancerCellNet

We found 11 cancer types (BRCA, COAD, ESCA, HNSC, KIRC, LGG, PAAD, UCEC, STAD, LUAD, LUSC) which have meaningful subtypes based on either histology or molecular profile and have sufficient samples to train a subtype classifier with high AUPR. We also included normal tissues samples from BRCA, COAD, HNSC, KIRC, UCEC to create a normal tissue category in the construction of their subtype classifiers. Training samples were either labelled as a cancer subtype for the cancer of interest or as “Unknown” if they belong to other cancer types. Similar to general classifier training, CCN performed gene pair transformation and selected the most discriminate gene pairs for each cancer subtype. In addition to the gene pairs selected to discriminate cancer subtypes, CCN also performed general classification of all training data and appended the classification profiles of training data with gene pair binary matrix as additional features. The reason behind using general classification profile as additional features is that many general cancer types may share similar subtypes, and general classification profile could be important features to discriminate the general cancer type of interest from other cancer types before performing finer subtype classification. The specific parameters used to train individual subtype classifiers using “subClass_train” function of CancerCellNet package can be found in **Supp Tab 7** and the gene pairs are in **Supp Tab 8**.

Validating Subtype CancerCellNet

Similar to validating general class classifier, we randomly sampled 2/3 of all samples in each cancer subtype as training data and sampled an equal amount across subtypes in the held-out data for assessing subtype classifiers. We repeated the process 20 times for more comprehensive assessment of subtype classifiers.

Classifying Query Data into Subtypes

We assigned subtype to query sample if the query sample has CCN score higher than the decision threshold. If a query sample has CCN scores higher than decision threshold, which was chosen through maximizing Macro F1 measure. The table of decision threshold for subtype classifiers are in **Supp Tab 10**. If a query sample with no CCN score higher than decision threshold in any subtype or has the highest CCN score in 'Unknown' category, then we assigned that sample as 'Unknown'. Analysis and visualizations were done in R and ComplexHeatmap package[70].

Tumor Purity Analysis

We used the R package ESTIMATE[71] to calculate the ESTIMATE scores from TCGA tumor expression profiles that we used as training data for CCN classifier. To calculate tumor purity we used the equation described in YoshiHara et al., 2013[71]:

$$\text{Tumour purity} = \cos\left(0.6049872018 + 0.0001467884\ \times\ \text{ESTIMATE score}\right)$$

Extracting Citation Counts

We used the R package RISmed[72] to extract the number of citations for each cell line through query search of "*cell line name*[Text Word] AND cancer[Text Word]" on PubMed. The citation counts were normalized by dividing the citation counts with the number of years since first documented.

$$Normalizedcitationcounts = \frac{citation\ counts}{\#\ years\ since\ first\ documented}$$

GRN construction and GRN Status

GRN construction was extended from our previous method[18]. 80 samples per cancer type were randomly sampled and normalized through down sampling as training data for the CLR GRN construction algorithm. Cancer type specific GRNs were identified by determining the differentially expressed genes per each cancer type and extracting the subnetwork using those genes.

To extend the original GRN status algorithm[18] across different platforms and species, we devised a rank-based GRN status algorithm. Like the original GRN status, rank based GRN status is a metric of assessing the similarity of cancer type specific GRN between training data in the cancer type of interest and query samples. Hence, high GRN status represents high level of establishment or similarity of the cancer specific GRN in the query sample compared to those of the training data. The expression profiles of training data and query data were transformed into rank expression profiles by replacing the expression values with the rank of the expression values within a sample (highest expressed gene would have the highest rank and lowest expressed genes would have a rank of 1). Cancer type specific mean and standard deviation of every gene's rank expression were calculated using training data. The modified Z-score values for genes within cancer type specific GRN were calculated for query sample's rank expression profiles to quantify how dissimilar the expression values of genes in query sample's cancer type specific GRN compared to those of the reference training data:

$$Z - score(gene\ i)_{mod} =$$

$$\begin{cases} 0,\ if\ Z - score\ is\ positive\ and\ the\ gene\ is\ found\ to\ be\ upregulated \\ 0,\ if\ Z - score\ is\ negative\ and\ the\ gene\ is\ found\ to\ be\ downregulated \\ abs(Z - score),\ otherwise \end{cases}$$

If a gene in the cancer type specific GRN is found to be upregulated in the specific cancer type relative to other cancer types, then we would consider query sample's gene to be similar if the ranking of the query sample's gene is equal to or greater than the mean ranking of the gene in training sample. As a result of similarity, we assign that gene of a Z-score of 0. The same principle applies to cases where the gene is downregulated in cancer specific subnetwork.

GRN status for query sample is calculated as the weighted mean of the $(1000 - Zscore(gene\ i)_{mod})$ across genes in cancer type specific GRN. 1000 is an arbitrary large number, and larger dissimilarity between query's cancer type specific GRN indicate high Z-scores for the GRN genes and low GRN status.

$$RGS = \sum_{i=1}^{n} (1000 - Zscore(gene\ i)_{mod})weight_{gene\ i}$$

$$GRN\ Status = \frac{RGS}{\sum_{i=1}^{n} weight_{gene\ i}}$$

The weight of individual genes in the cancer specific network is determined by the importance of the gene in the Random Forest classifier. We later normalize the GRN status in respect to the GRN status of the cancer type of interest and the cancer type with the lowest mean GRN status.

$$Normalized\ GRN\ status = \frac{GRN\ status_{query} - avg(GRN\ status_{min\ cancer})}{avg(GRN\ status_{cancer\ type\ interest})}$$

Where "min cancer" represents the cancer type where its training data have the lowest mean GRN status in the cancer type of interest, and $avg(GRN\ status_{min\ cancer})$ represents the average GRN status of cancer type with the lowest average GRN status in the "min cancer". $avg(GRN\ status_{cancer\ type\ interest})$ represents average GRN status of the cancer type of interest in the training data.

Code availability

CancerCellNet code and documentation is available at GitHub: https://github.com/pcahan1/cancerCellNet

FIGURE LEGENDS

Fig. 1 CancerCellNet (CCN) workflow and performance. **(A)** Schematic of CCN training (top) and usage (bottom). CCN was designed to assess and compare the expression profiles of cancer models such as CCLs, PDXs, and GEMMs with native patient tumors. First, CCN takes patient tumor expression profiles of 23 different cancer types from TCGA to train a multi-class

Random Forest classifier and performs gene-pair transformation on tumor expression profiles. Then CCN selects the most discriminative gene pairs for each cancer type as features. Lastly, CCN trains a multi-class Random Forest classifier using gene-pair transformed training data and feature gene pairs. To use trained classifier, CCN inputs the query samples (e.g. expression profiles from CCLs, PDXs, GEMMs) and generates a classification profile for the query samples. The column names of the classification heatmap represent sample annotation and the row names of the classification heatmap represent different cancer types. Each grid is colored from black to yellow representing the lowest classification score (e.g. 0) to highest classification score (e.g. 1). **(B)** Mean and standard deviation of area under the precision recall curve (AUPR) of classifiers based on 50 iterations of cross-validation: random sampling of training data (2/3 of samples for each cancer category), training CCN classifiers using training data and testing the classifiers on held-out data (1/3 of samples for each cancer category).

Fig. 2 Evaluation of cancer cell lines. **(A)** General classification heatmap of CCLs extracted from CCLE. Column annotations of the heatmap represent the labelled cancer category of the CCLs given by CCLE and the row names of the heatmap represent different cancer categories. CCLs' general classification profiles are categorized into 4 categories: correct (red), correct mixed (pink), no classification (light green) and other classification (dark green) based on the decision threshold of 0.3. **(B)** Bar plot represents the proportion of each classification category in CCLs across cancer types ordered from the cancer types with the highest proportion of correct and correct mixed CCLs to lowest proportion. **(C)** Comparison between SKCM general CCN scores from bulk RNA-seq classifier and SKCM malignant CCN scores from scRNA-seq classifier for SKCM CCLs. **(D)** Comparison between SARC general CCN scores from bulk RNA-seq classifier and CAF CCN scores from scRNA-seq classifier for SKCM CCLs. **(E)** Comparison between GBM general CCN scores from bulk RNA-seq classifier and GBM neoplastic CCN scores from scRNA-seq classifier for GBM CCLs. **(F)** Comparison between SARC general CCN scores and CAF CCN scores from scRNA-seq classifier for GBM CCLs. The green lines indicate the decision threshold for scRNA-seq classifier and general classifier.

Fig. 3 Subtype classification of CCLs. The heatmap visualizations represent subtype classification of UCEC CCLs **(A),** LUAD CCLs **(B)** and LUSC CCLs **(C)**. The row names represent CCLs and column names represent cancer subtypes of UCEC, LUSC and LUAD. The bar plots to the right of the subclass heatmaps represent the general classification scores in cell lines' annotated cancer category, and the color strips to the right of the subclass heatmaps represent subclass classification (left) and general classification (right). **(D)** Comparison of normalized citation counts and general CCN classification scores of CCLs. Labelled cell lines either have the highest CCN classification score in their labelled cancer category or highest normalized citation count. Hep G2 and ML-1 have both the highest CCN classification score in their labelled cancer category and the highest normalized citation count. Each citation count was normalized by number of years since first documented.

Fig. 4 Evaluation of patient derived xenografts. **(A)** General classification heatmap of PDXs. Column annotations represent annotated cancer type of the PDXs, and row names represent cancer categories. **(B)** Proportion of classification categories in PDXs across cancer types is visualized in the bar plot and ordered from the cancer type with highest proportion of correct and mixed correct classified PDXs to the lowest. Subtype classification heatmaps of UCEC PDXs **(C),** LUSC PDXs **(D)** and LUAD PDXs **(E)**.

Fig. 5 Evaluation of genetically engineered mouse models. **(A)** General classification heatmap of GEMMs. Column annotations represent annotated cancer type of the GEMMs, and row names represent cancer categories. **(B)** Proportion of classification categories in GEMMs across cancer types is visualized in the bar plot and ordered from the cancer type with highest proportion of correct and mixed correct classified GEMMs to the lowest. Subtype classification heatmap of LUSC GEMMs **(C)** and LUAD GEMMs **(D)**.

Fig. 6 Comparison of CCLs, PDXs, and GEMMs. **(A)** Box-and-whiskers plot comparing general CCN scores across CCLs, GEMMs, PDXs of five general tumor types. **(B)** Proportion of UCEC (top-left), PAAD (top-right), LUAD (bottom-left) and LUSC (bottom-right) subtypes across cancer model modalities and TCGA patient data. For GEMMs, all classification profiles of replicates with the same genotype from the same study are averaged into one classification profile when calculating the proportion.

Supplementary Information

Supplementary Figure 1 Assessment of CCN general classifier and subtype classifier. **(A)** Mean and range of CCN classifier's PR curves from 50 cross validations. **(B)** AUPR of CCN classifier when applied to independent patient tumor data from ICGC. **(C)** AUPR of CCN human tissue classifier when applied to mouse tissue data. **(D)** The schematic of training a subtype classifier in CCN. CCN uses patient tumor expression profiles from cancer of interest as training data. CCN performs gene-pair transformation and selects the most discriminative gene pairs among the cancer subtypes from training data as features. CCN then applies the general classification on training data and uses the general classification profile as features in addition to gene pairs for training a Random Forest classifier. The weight of the general classification profiles as features can be tuned to maximize AUPR. **(E)** The mean and standard deviation of AUPR for 11 subtype classifiers based on 20 iterations of random sampling of training and held-out data, training subtype classifier using training data, classification of held-out data, and calculation of recall and precision.

Supplementary Figure 2 Further validation of CCN and classification results. To validate the cross-platform classification performance of CCN, a new classifier specifically trained to classify microarray data was trained using RNA-seq data from TCGA as training data and intersecting genes between RNA-seq data and microarray data. **(A)** AUPR of CCN classifier when applied to primary microarray testing tumor data. **(B)** Classification heatmap of CCLs using microarray expression data. **(C)** Pearson correlation between CCN scores of CCLE lines generated from RNA-seq data and microarray data. **(D)** Comparison between CCLs' CCN scores and the median correlation results from Yu et al. **(E)** Comparison of mean tumor purity of training data and mean CCN score of CCLs for each cancer category.

Supplementary Figure 3 Single-cell classification of SKCM and GBM cell lines. **(A)** Classification heatmap of held-out scRNA-seq data. **(B)** AUPR of the scRNA-seq classifier when applied to scRNA-seq held-out data. **(C)** Single-cell classification of SKCM CCLs. **(D)** Single-cell classification of GBM CCLs.

Supplementary Figure 4 Cancer subtype classification heatmap of UCEC GEMMs.

Supplementary Figure 5 Correlation between cancer type subnetwork GRN status and general CCN scores.

Supplementary Figure 6 Proportion of cancer subtypes in different cancer models and TCGA tumor data across 11 general cancer types.

Supplementary Table 1 General classification profiles of CCLs.

Supplementary Table 2 Subtype classification profiles of CCLs.

Supplementary Table 3 General classification profiles of PDXs.

Supplementary Table 4 Subtype classification profiles of PDXs.

Supplementary Table 5 General classification profiles of GEMMs

Supplementary Table 6 Subtype classification profiles of GEMMs.

Supplementary Table 7 Specific parameters used for training of all classifiers.

Supplementary Table 8 Gene-pairs selected for final training of CCN general and subtype classifiers.

Supplementary Table 9 Accessions of tumor microarray data used in validation.

Supplementary Table 10 Decision thresholds for subtype classifiers.

REFERENCES

1. Sharma, S. V., Haber, D. A. & Settleman, J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat. Rev. Cancer* **10,** 241–253 (2010).
2. Kersten, K., de Visser, K. E., van Miltenburg, M. H. & Jonkers, J. Genetically engineered mouse models in oncology research and cancer medicine. *EMBO Mol. Med.* **9,** 137–153 (2017).
3. Hidalgo, M. *et al.* Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.* **4,** 998–1013 (2014).
4. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33,** 306–312 (2015).
5. Koren, S. *et al.* PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. *Nature* **525,** 114–118 (2015).
6. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat. Med.* **17,** 1514–1520 (2011).
7. Sharpless, N. E. & Depinho, R. A. The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat. Rev. Drug Discov.* **5,** 741–754 (2006).
8. Mouradov, D. *et al.* Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res.* **74,** 3238–3247 (2014).
9. Stuckelberger, S. & Drapkin, R. Precious GEMMs: emergence of faithful models for ovarian cancer research. *J. Pathol.* **245,** 129–131 (2018).
10. Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4,** 2126 (2013).
11. Jiang, G. *et al.* Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. *BMC Genomics* **17 Suppl 7,** 525 (2016).
12. Chen, B., Sirota, M., Fan-Minogue, H., Hadley, D. & Butte, A. J. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. *BMC Med. Genomics* **8 Suppl 2,** S5 (2015).
13. Vincent, K. M., Findlay, S. D. & Postovit, L. M. Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. *Breast Cancer Res.* **17,** 114 (2015).

14. Yu, K. *et al.* Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10,** 3574 (2019).
15. Guernet, A. & Grumolato, L. CRISPR/Cas9 editing of the genome for cancer modeling. *Methods* **121-122,** 130–137 (2017).
16. Gargiulo, G. Next-Generation in vivo Modeling of Human Cancers. *Front. Oncol.* **8,** 429 (2018).
17. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21,** 1318–1325 (2015).
18. Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158,** 903–915 (2014).
19. Radley, A. H. *et al.* Assessment of engineered cells using CellNet and RNA-seq. *Nat. Protoc.* **12,** 1089–1102 (2017).
20. Tan, Y. & Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst.* **9,** 207–213.e2 (2019).
21. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
22. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford)* **2011,** bar026 (2011).
23. Tan, Y. & Cahan, P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *BioRxiv* (2018). doi:10.1101/508085
24. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
25. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
26. Wilkerson, M. D. *et al.* Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res.* **16,** 4864–4875 (2010).
27. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu & Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32,** 185–203.e13 (2017).
28. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
29. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541,** 169–175 (2017).
30. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517,** 576–582 (2015).
31. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
32. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17,** 98–110 (2010).
33. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).

34. Hu, B. *et al.* Gastric cancer: Classification, histology and application of molecular pathology. *J. Gastrointest. Oncol.* **3,** 251–261 (2012).
35. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
36. Medico, E. *et al.* The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat. Commun.* **6,** 7002 (2015).
37. Park, J.-G. *et al.* Characteristics of Cell Lines Established from Human Colorectal Carcinoma. *Cancer Res.* (1987).
38. Jerby-Arnon, L. *et al.* A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell* **175,** 984–997.e24 (2018).
39. Darmanis, S. *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **21,** 1399–1410 (2017).
40. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9,** 391–403 (2006).
41. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).
42. Black, J. D., English, D. P., Roque, D. M. & Santin, A. D. Targeted therapy in uterine serous carcinoma: an aggressive variant of endometrial cancer. *Womens Health (Lond. Engl.)* **10,** 45–57 (2014).
43. Yang, S., Thiel, K. W. & Leslie, K. K. Progesterone: the ultimate endometrial tumor suppressor. *Trends Endocrinol. Metab.* **22,** 145–152 (2011).
44. Huszar, M. *et al.* Up-regulation of L1CAM is linked to loss of hormone receptors and E-cadherin in aggressive subtypes of endometrial carcinomas. *J. Pathol.* **220,** 551–561 (2010).
45. Kozak, J., Wdowiak, P., Maciejewski, R. & Torres, A. A guide for endometrial cancer cell lines functional assays using the measurements of electronic impedance. *Cytotechnology* **70,** 339–350 (2018).
46. Korch, C. *et al.* DNA profiling analysis of endometrial and ovarian cell lines reveals misidentification, redundancy and contamination. *Gynecol. Oncol.* **127,** 241–248 (2012).
47. Wu, D. *et al.* Gene-expression data integration to squamous cell lung cancer subtypes reveals drug sensitivity. *Br. J. Cancer* **109,** 1599–1608 (2013).
48. Walter, V. *et al.* Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PLoS One* **8,** e56823 (2013).
49. Adeegbe, D. O. *et al.* BET Bromodomain Inhibition Cooperates with PD-1 Blockade to Facilitate Antitumor Response in Kras-Mutant Non-Small Cell Lung Cancer. *Cancer Immunol Res* **6,** 1234–1245 (2018).
50. Blaisdell, A. *et al.* Neutrophils oppose uterine epithelial carcinogenesis via debridement of hypoxic tumor cells. *Cancer Cell* **28,** 785–799 (2015).
51. Fitamant, J. *et al.* YAP inhibition restores hepatocyte differentiation in advanced HCC, leading to tumor regression. *Cell Rep.* **10,** 1692–1707 (2015).
52. Jia, D. *et al.* Crebbp loss drives small cell lung cancer and increases sensitivity to HDAC inhibition. *Cancer Discov.* **8,** 1422–1437 (2018).

53. Kress, T. R. *et al.* Identification of MYC-Dependent Transcriptional Programs in Oncogene-Addicted Liver Tumors. *Cancer Res.* **76,** 3463–3472 (2016).
54. Li, L. *et al.* GKAP acts as a genetic modulator of NMDAR signaling to govern invasive tumor growth. *Cancer Cell* **33,** 736–751.e5 (2018).
55. Mollaoglu, G. *et al.* The Lineage-Defining Transcription Factors SOX2 and NKX2-1 Determine Lung Cancer Cell Fate and Shape the Tumor Immune Microenvironment. *Immunity* **49,** 764–779.e9 (2018).
56. Pan, Y. *et al.* Whole tumor RNA-sequencing and deconvolution reveal a clinically-prognostic PTEN/PI3K-regulated glioma transcriptional signature. *Oncotarget* **8,** 52474–52487 (2017).
57. Lissanu Deribe, Y. *et al.* Mutations in the SWI/SNF complex induce a targetable dependence on oxidative phosphorylation in lung cancer. *Nat. Med.* **24,** 1047–1057 (2018).
58. Xu, C. *et al.* Loss of Lkb1 and Pten leads to lung squamous cell carcinoma with elevated PD-L1 expression. *Cancer Cell* **25,** 590–604 (2014).
59. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
60. Balkwill, F. R., Capasso, M. & Hagemann, T. The tumor microenvironment at a glance. *J. Cell Sci.* **125,** 5591–5596 (2012).
61. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.* **49,** 1567–1575 (2017).
62. Hristova, V. A. & Chan, D. W. Cancer biomarker discovery and translation: proteomics and beyond. *Expert Rev Proteomics* **16,** 93–103 (2019).
63. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150,** 12–27 (2012).
64. Silva, T. C. *et al.* TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. [version 2; peer review: 1 approved, 2 approved with reservations]. *F1000Res.* **5,** 1542 (2016).
65. Morgan, M., Obenchain, V., Hester, J. & Pag`es, H. *SummarizedExperiment: SummarizedExperiment container.* (2018).
66. Pavlidis, P. & Noble, W. S. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol.* **2,** RESEARCH0042 (2001).
67. Geman, D., d Avignon, C., Naiman, D. Q. & Winslow, R. L. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol* **3,** Article19 (2004).
68. Kolde, R. *pheatmap: Pretty Heatmaps*. (CRAN, 2019).
69. Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis* . (Springer-Verlag New York, 2016). doi:10.1007/978-0-387-98141-3
70. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).
71. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
72. Kovalchik, S. *RISmed: Download Content from NCBI Databases*. (CRAN.R-project, 2017).

Figure 1

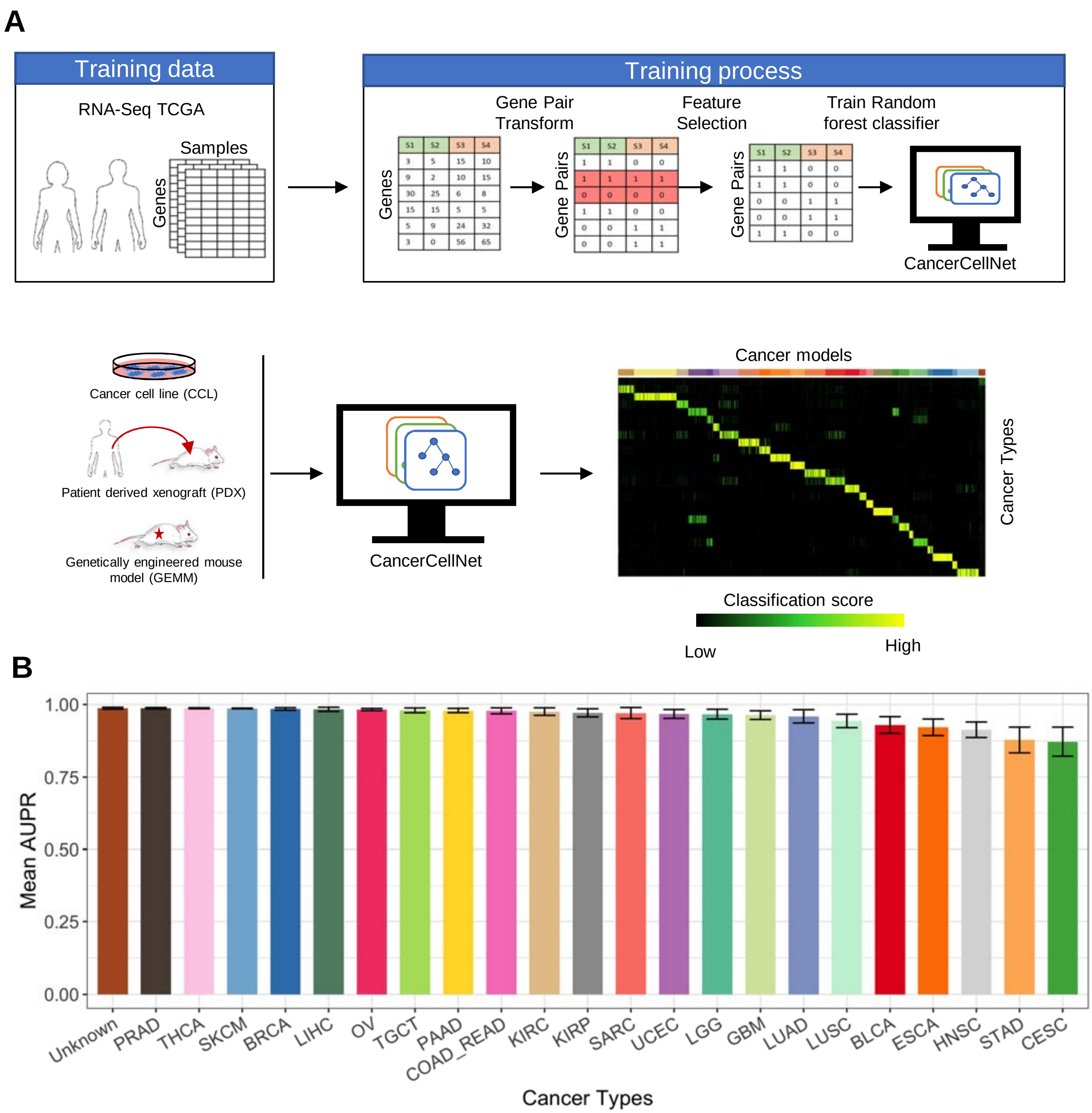

A
Training data
RNA-Seq TCGA
Samples
Genes
Training process
Gene Pair Transform
Feature Selection
Train Random forest classifier
Genes
Gene Pairs
Gene Pairs
CancerCellNet
Cancer cell line (CCL)
Patient derived xenograft (PDX)
Genetically engineered mouse model (GEMM)
CancerCellNet
Cancer models
Cancer Types
Classification score
Low
High
B
Mean AUPR
1.00
0.75
0.50
0.25
0.00
Unknown
PRAD
THCA
SKCM
BRCA
LIHC
OV
TGCT
PAAD
COAD_READ
KIRC
KIRP
SARC
UCEC
LGG
GBM
LUAD
LUSC
BLCA
ESCA
HNSC
STAD
CESC
Cancer Types

Figure 2

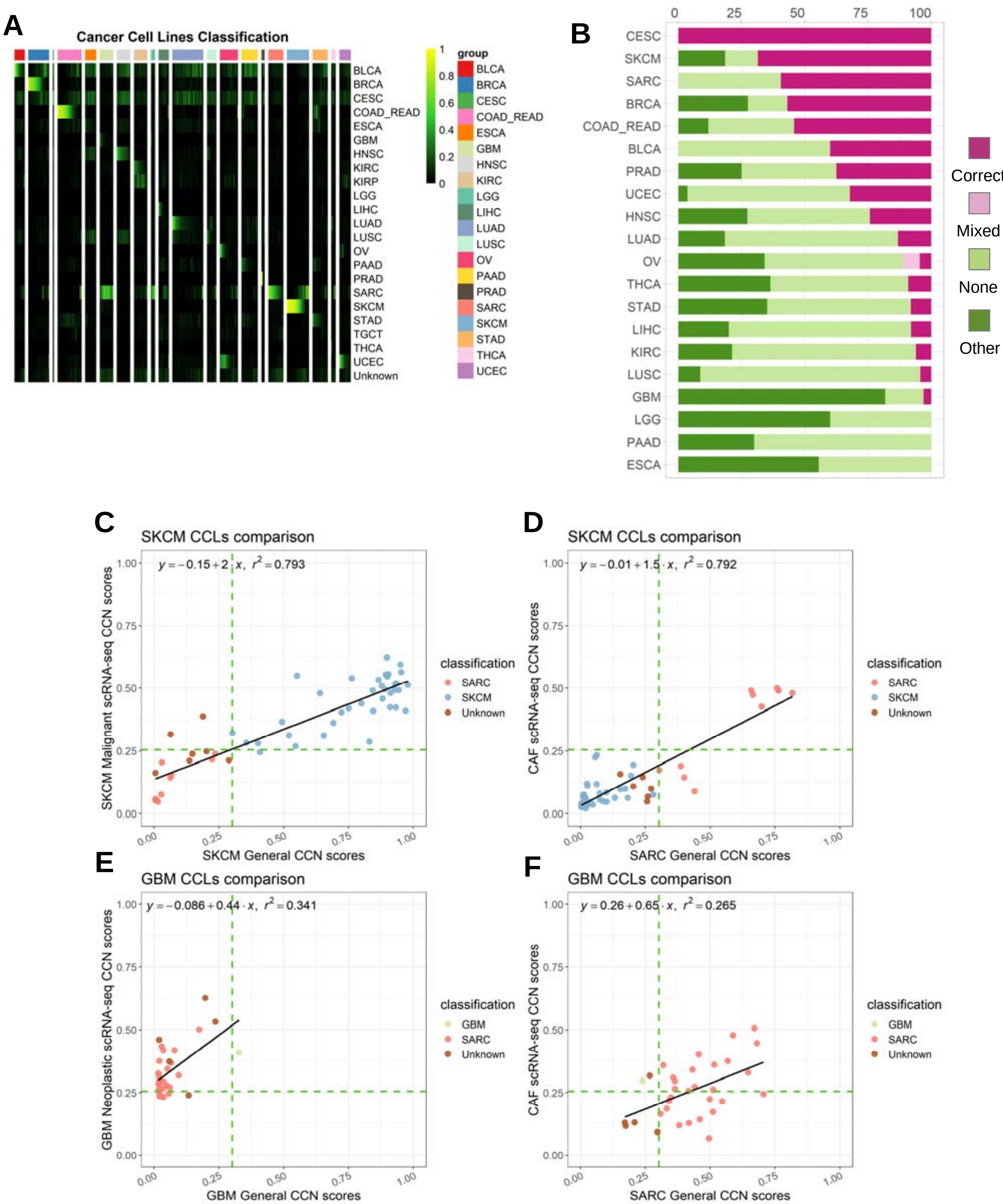

A
Cancer Cell Lines Classification
B
C
SKCM CCLs comparison
SKCM Malignant scRNA-seq CCN scores
SKCM General CCN scores
D
SKCM CCLs comparison
CAF scRNA-seq CCN scores
SARC General CCN scores
E
GBM CCLs comparison
GBM Neoplastic scRNA-seq CCN scores
GBM General CCN scores
F
GBM CCLs comparison
CAF scRNA-seq CCN scores
SARC General CCN scores
classification
Correct
Mixed
None
Other

Figure 3

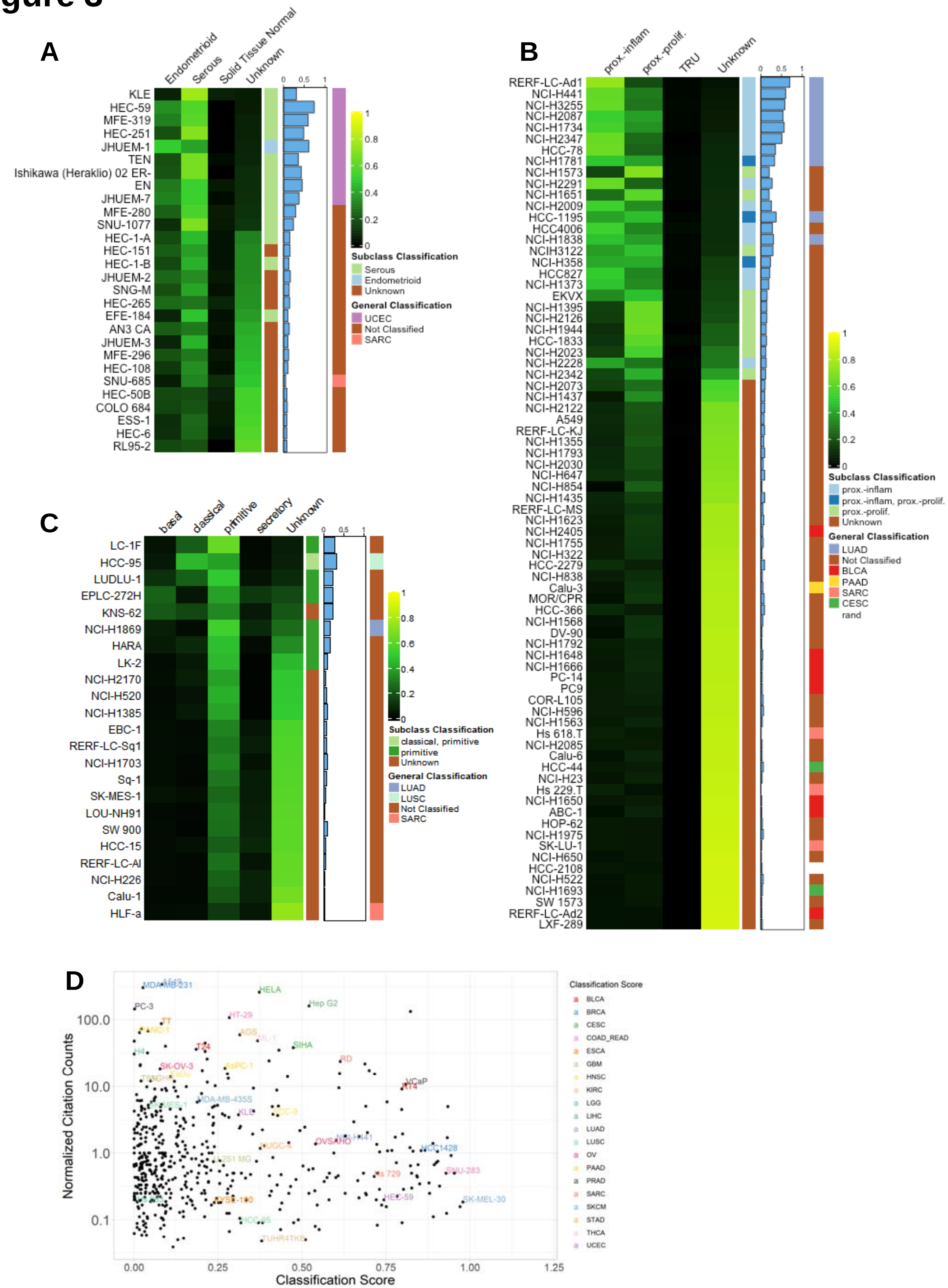

A
Endometrioid
Serous
Solid Tissue Normal
Unknown
KLE
HEC-59
MFE-319
HEC-251
JHUEM-1
TEN
Ishikawa (Heraklio) 02 ER-
EN
JHUEM-7
MFE-280
SNU-1077
HEC-1-A
HEC-151
HEC-1-B
JHUEM-2
SNG-M
HEC-265
EFE-184
AN3 CA
JHUEM-3
MFE-296
HEC-108
SNU-685
HEC-50B
COLO 684
ESS-1
HEC-6
RL95-2
Subclass Classification
Serous
Endometrioid
Unknown
General Classification
UCEC
Not Classified
SARC
B
prox.-inflam
prox.-prolif.
TRU
Unknown
RERF-LC-Ad1
NCI-H441
NCI-H3255
NCI-H2087
NCI-H1734
NCI-H2347
HCC-78
NCI-H1781
NCI-H1573
NCI-H2291
NCI-H1651
NCI-H2009
HCC-1195
HCC4006
NCI-H1838
NCIH3122
NCI-H358
HCC827
NCI-H1373
EKVX
NCI-H1395
NCI-H2126
NCI-H1944
HCC-1833
NCI-H2023
NCI-H2228
NCI-H2342
NCI-H2073
NCI-H1437
NCI-H2122
A549
RERF-LC-KJ
NCI-H1355
NCI-H1793
NCI-H2030
NCI-H647
NCI-H854
NCI-H1435
RERF-LC-MS
NCI-H1623
NCI-H2405
NCI-H1755
NCI-H322
HCC-2279
NCI-H838
Calu-3
MOR/CPR
HCC-366
NCI-H1568
DV-90
NCI-H1792
NCI-H1648
NCI-H1666
PC-14
PC9
COR-L105
NCI-H596
NCI-H1563
Hs 618.T
NCI-H2085
Calu-6
HCC-44
NCI-H23
Hs 229.T
NCI-H1650
ABC-1
HOP-62
NCI-H1975
SK-LU-1
NCI-H650
HCC-2108
NCI-H522
NCI-H1693
SW 1573
RERF-LC-Ad2
LXF-289
Subclass Classification
prox.-inflam
prox.-inflam, prox.-prolif.
prox.-prolif.
Unknown
General Classification
LUAD
Not Classified
BLCA
PAAD
SARC
CESC
rand
C
basal
classical
primitive
secretory
Unknown
LC-1F
HCC-95
LUDLU-1
EPLC-272H
KNS-62
NCI-H1869
HARA
LK-2
NCI-H2170
NCI-H520
NCI-H1385
EBC-1
RERF-LC-Sq1
NCI-H1703
Sq-1
SK-MES-1
LOU-NH91
SW 900
HCC-15
RERF-LC-AI
NCI-H226
Calu-1
HLF-a
Subclass Classification
classical, primitive
primitive
Unknown
General Classification
LUAD
LUSC
Not Classified
SARC
D
Normalized Citation Counts
Classification Score
Classification Score
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
THCA
UCEC
MDA-MB-231
PC-3
HELA
Hep G2
HT-29
SIHA
RD
VCaP
MDA-MB-435S
HCC1428
SNU-283
SK-MEL-30
HEC-59
SK-OV-3
OVSAHO

Figure 4

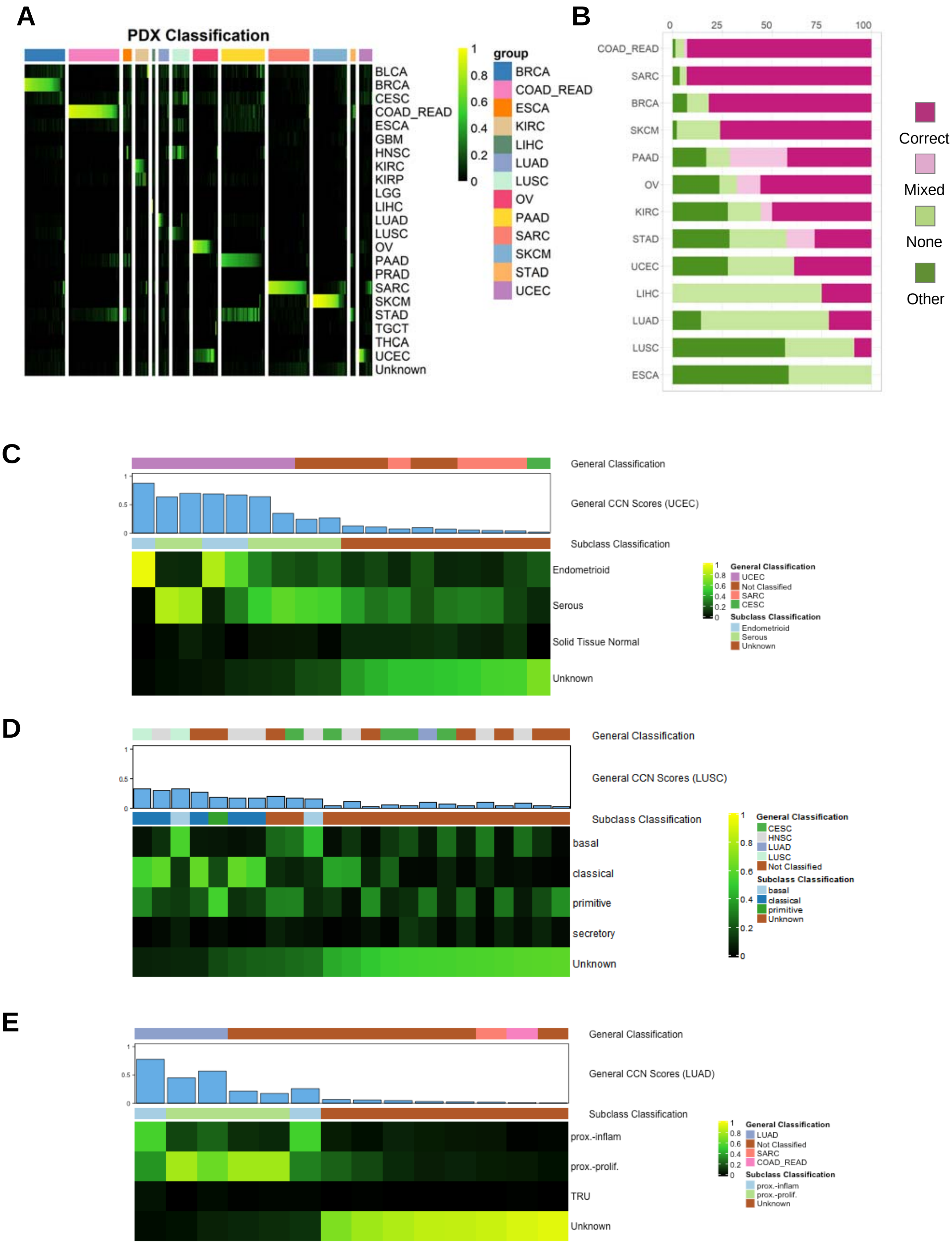

A
PDX Classification
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
group
BRCA
COAD_READ
ESCA
KIRC
LIHC
LUAD
LUSC
OV
PAAD
SARC
SKCM
STAD
UCEC
B
COAD_READ
SARC
BRCA
SKCM
PAAD
OV
KIRC
STAD
UCEC
LIHC
LUAD
LUSC
ESCA
Correct
Mixed
None
Other
C
General Classification
General CCN Scores (UCEC)
Subclass Classification
Endometrioid
Serous
Solid Tissue Normal
Unknown
General Classification
UCEC
Not Classified
SARC
CESC
Subclass Classification
Endometrioid
Serous
Unknown
D
General Classification
General CCN Scores (LUSC)
Subclass Classification
basal
classical
primitive
secretory
Unknown
General Classification
CESC
HNSC
LUAD
LUSC
Not Classified
Subclass Classification
basal
classical
primitive
Unknown
E
General Classification
General CCN Scores (LUAD)
Subclass Classification
prox.-inflam
prox.-prolif.
TRU
Unknown
General Classification
LUAD
Not Classified
SARC
COAD_READ
Subclass Classification
prox.-inflam
prox.-prolif.
Unknown

Figure 5

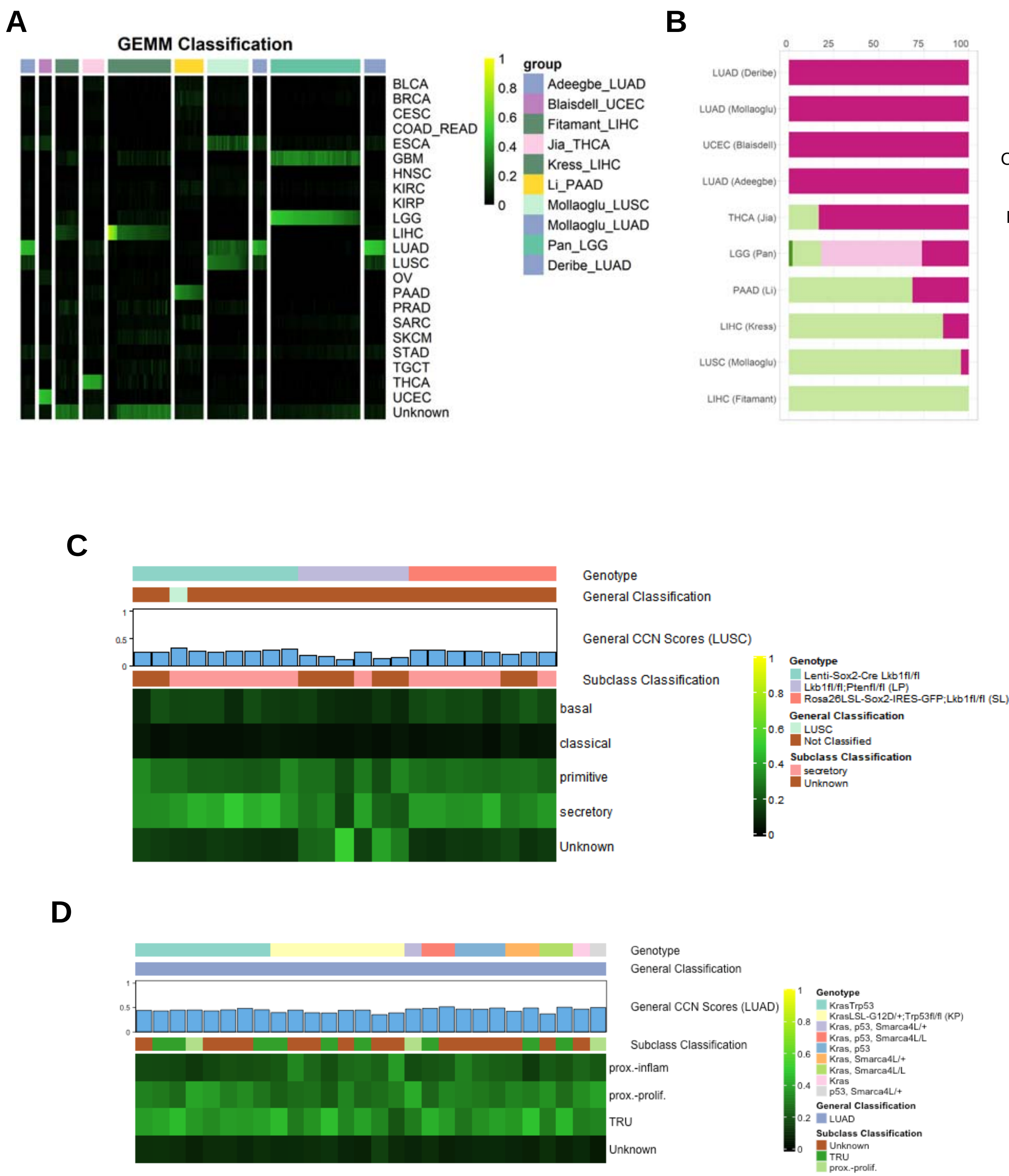
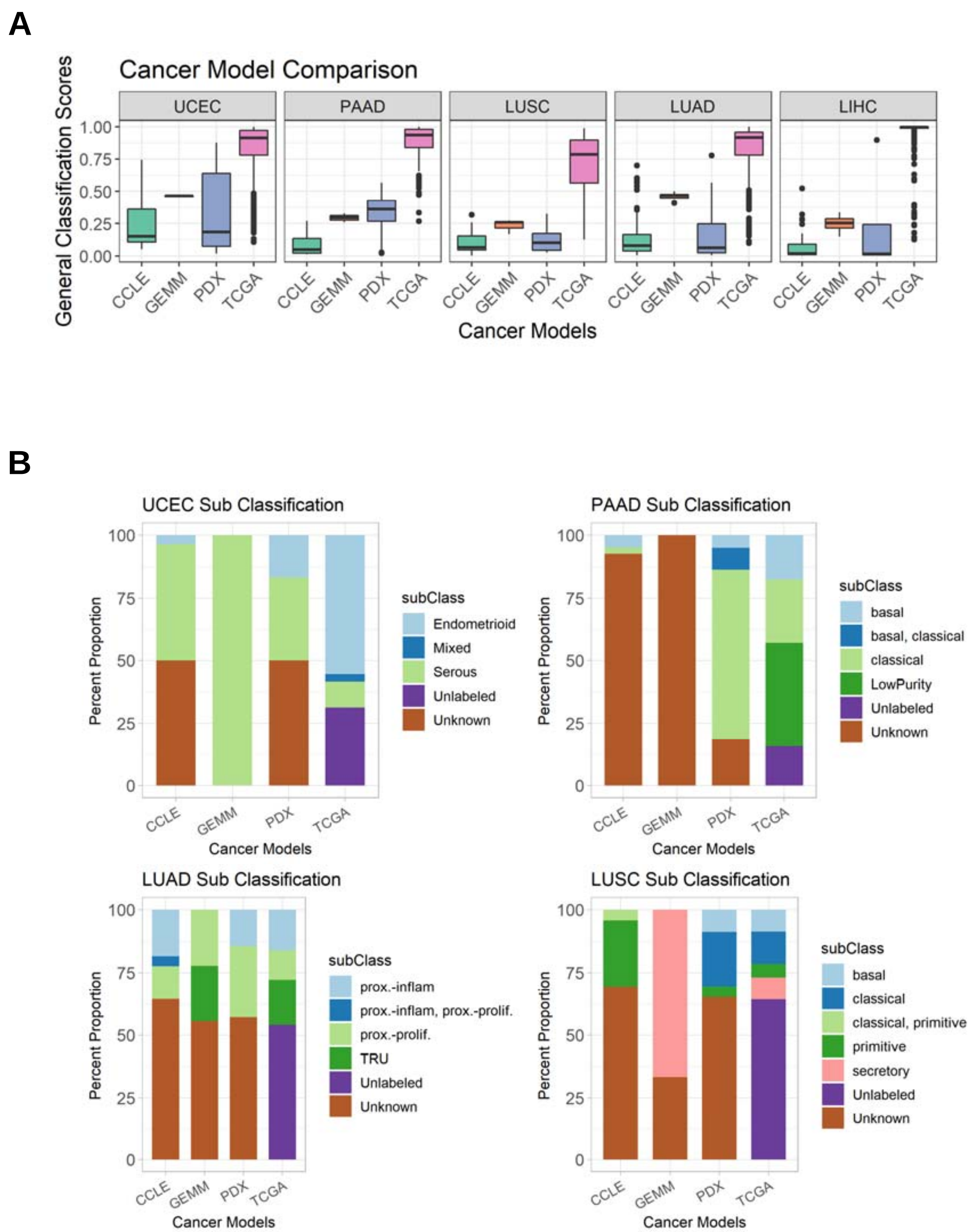
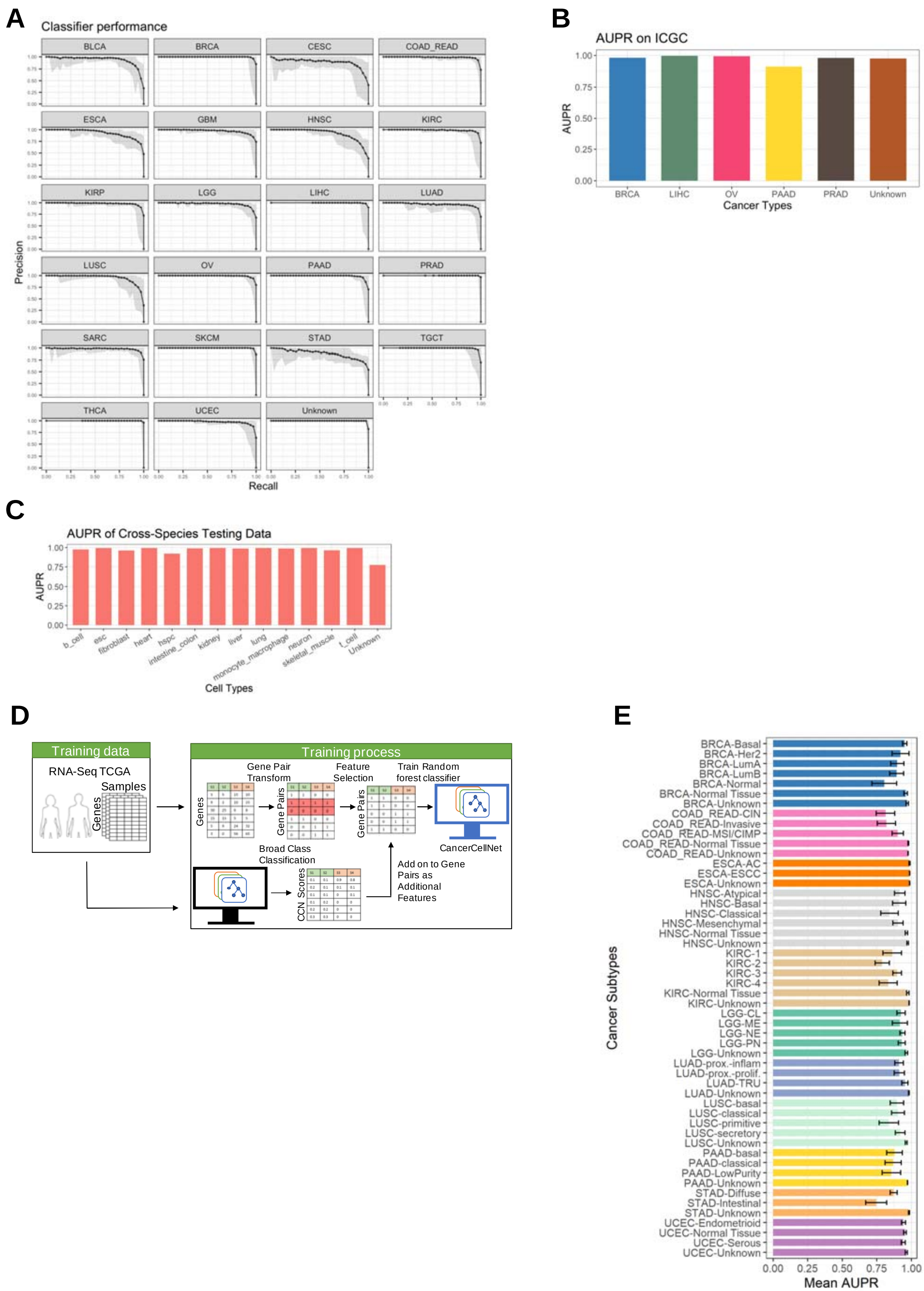

A
GEMM Classification
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
1 0.8 0.6 0.4 0.2 0
group
Adeegbe_LUAD
Blaisdell_UCEC
Fitamant_LIHC
Jia_THCA
Kress_LIHC
Li_PAAD
Mollaoglu_LUSC
Mollaoglu_LUAD
Pan_LGG
Deribe_LUAD
B
0 25 50 75 100
LUAD (Deribe)
LUAD (Mollaoglu)
UCEC (Blaisdell)
LUAD (Adeegbe)
THCA (Jia)
LGG (Pan)
PAAD (Li)
LIHC (Kress)
LUSC (Mollaoglu)
LIHC (Fitamant)
Correct
Mixed
None
Other
C
Genotype
General Classification
General CCN Scores (LUSC)
Subclass Classification
basal
classical
primitive
secretory
Unknown
Genotype
Lenti-Sox2-Cre Lkb1fl/fl
Lkb1fl/fl;Ptenfl/fl (LP)
Rosa26LSL-Sox2-IRES-GFP;Lkb1fl/fl (SL)
General Classification
LUSC
Not Classified
Subclass Classification
secretory
Unknown
D
Genotype
General Classification
General CCN Scores (LUAD)
Subclass Classification
prox.-inflam
prox.-prolif.
TRU
Unknown
Genotype
KrasTrp53
KrasLSL-G12D/+;Trp53fl/fl (KP)
Kras, p53, Smarca4L/+
Kras, p53, Smarca4L/L
Kras, p53
Kras, Smarca4L/+
Kras, Smarca4L/L
Kras
p53, Smarca4L/+
General Classification
LUAD
Subclass Classification
Unknown
TRU
prox.-prolif.

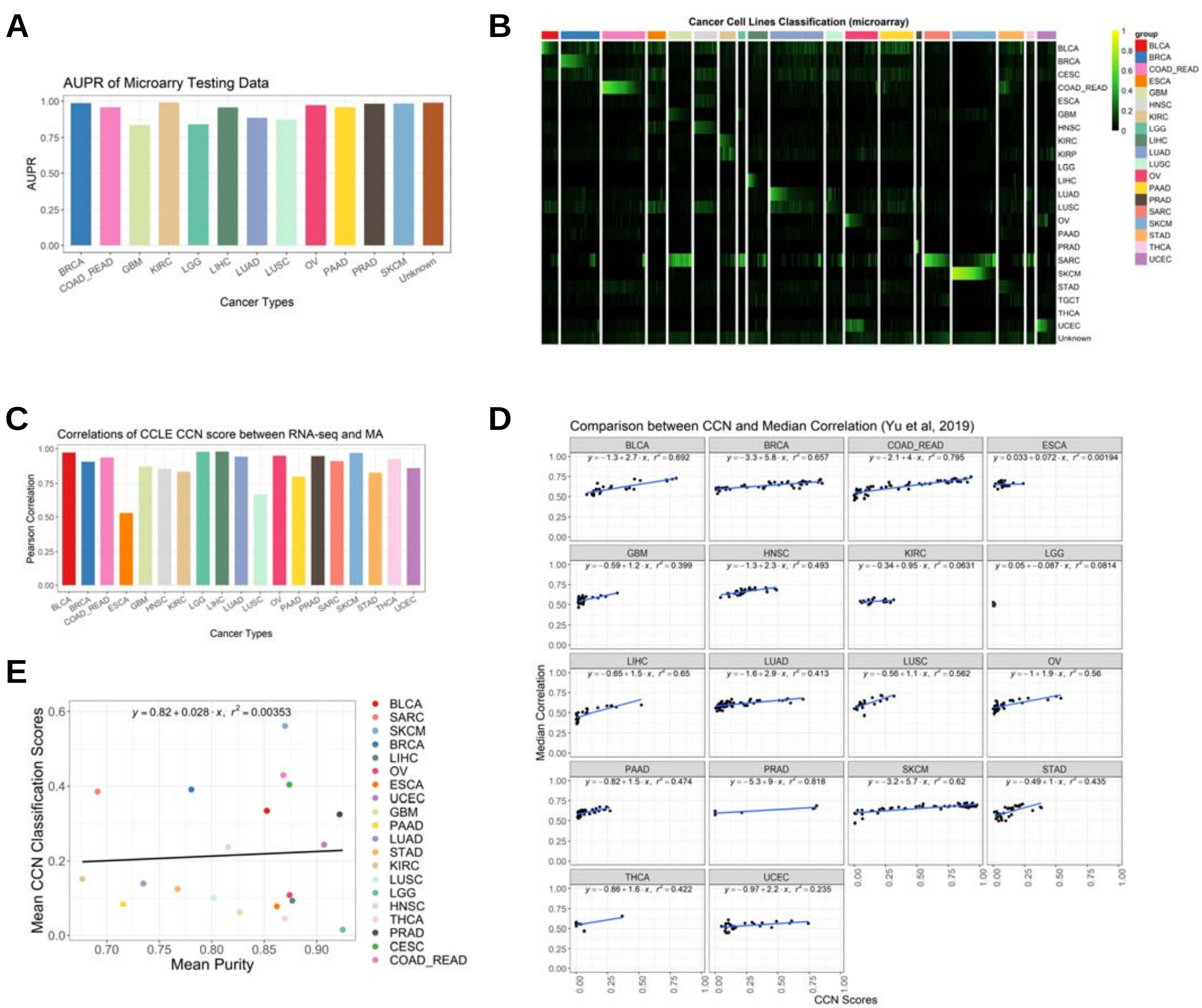
Figure 6


A
Cancer Model Comparison
UCEC
PAAD
LUSC
LUAD
LIHC
General Classification Scores
1.00
0.75
0.50
0.25
0.00
CCLE
GEMM
PDX
TCGA
Cancer Models
B
UCEC Sub Classification
PAAD Sub Classification
LUAD Sub Classification
LUSC Sub Classification
Percent Proportion
100
75
50
25
0
subClass
Endometrioid
Mixed
Serous
Unlabeled
Unknown
basal
basal, classical
classical
LowPurity
prox.-inflam
prox.-inflam, prox.-prolif.
prox.-prolif.
TRU
classical, primitive
primitive
secretory

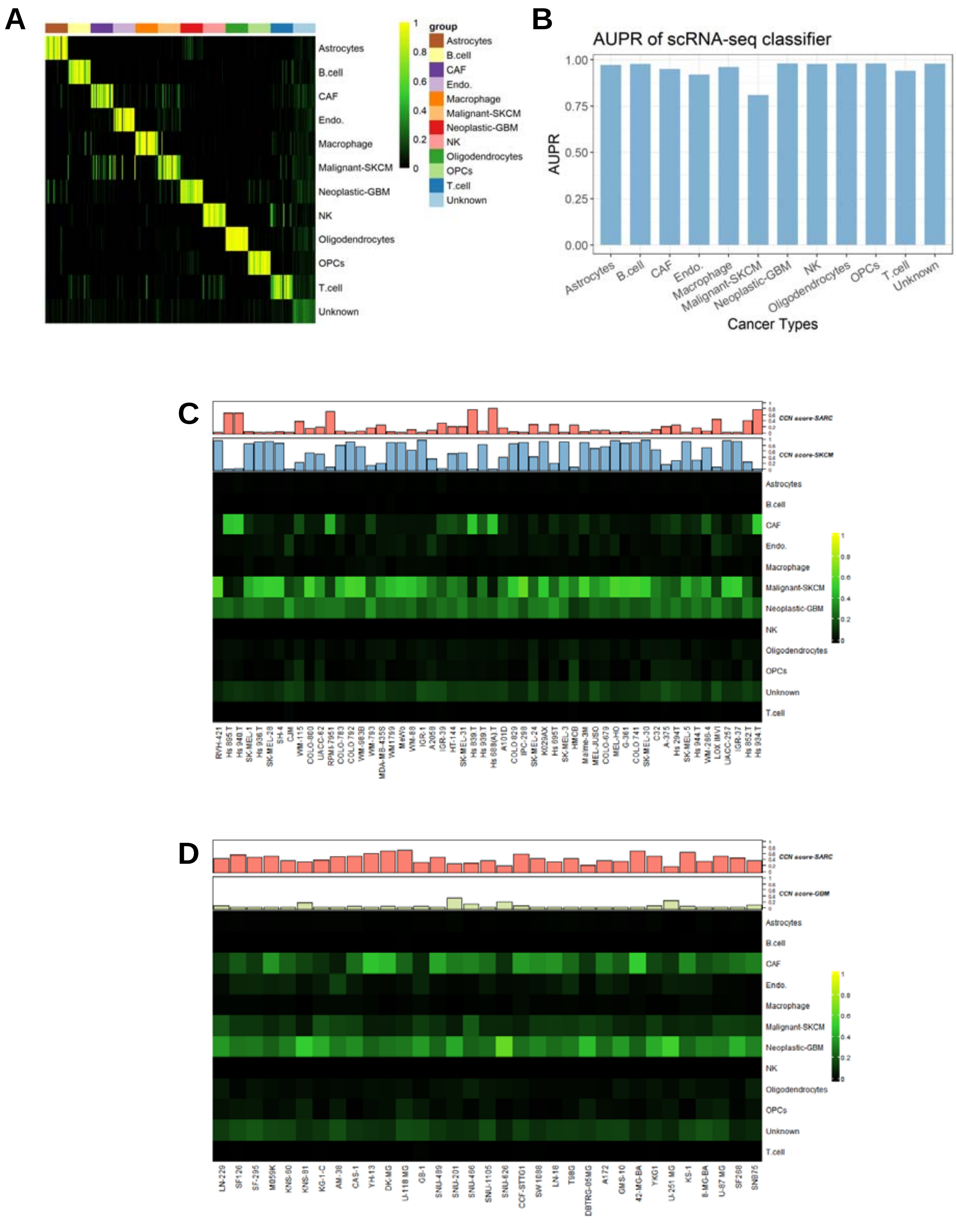
Supplemental Figure 1



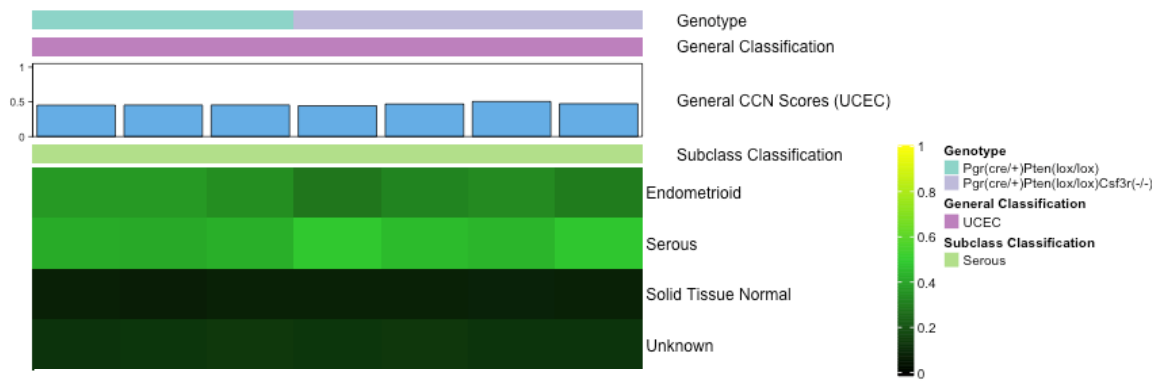
Supplemental Figure 2



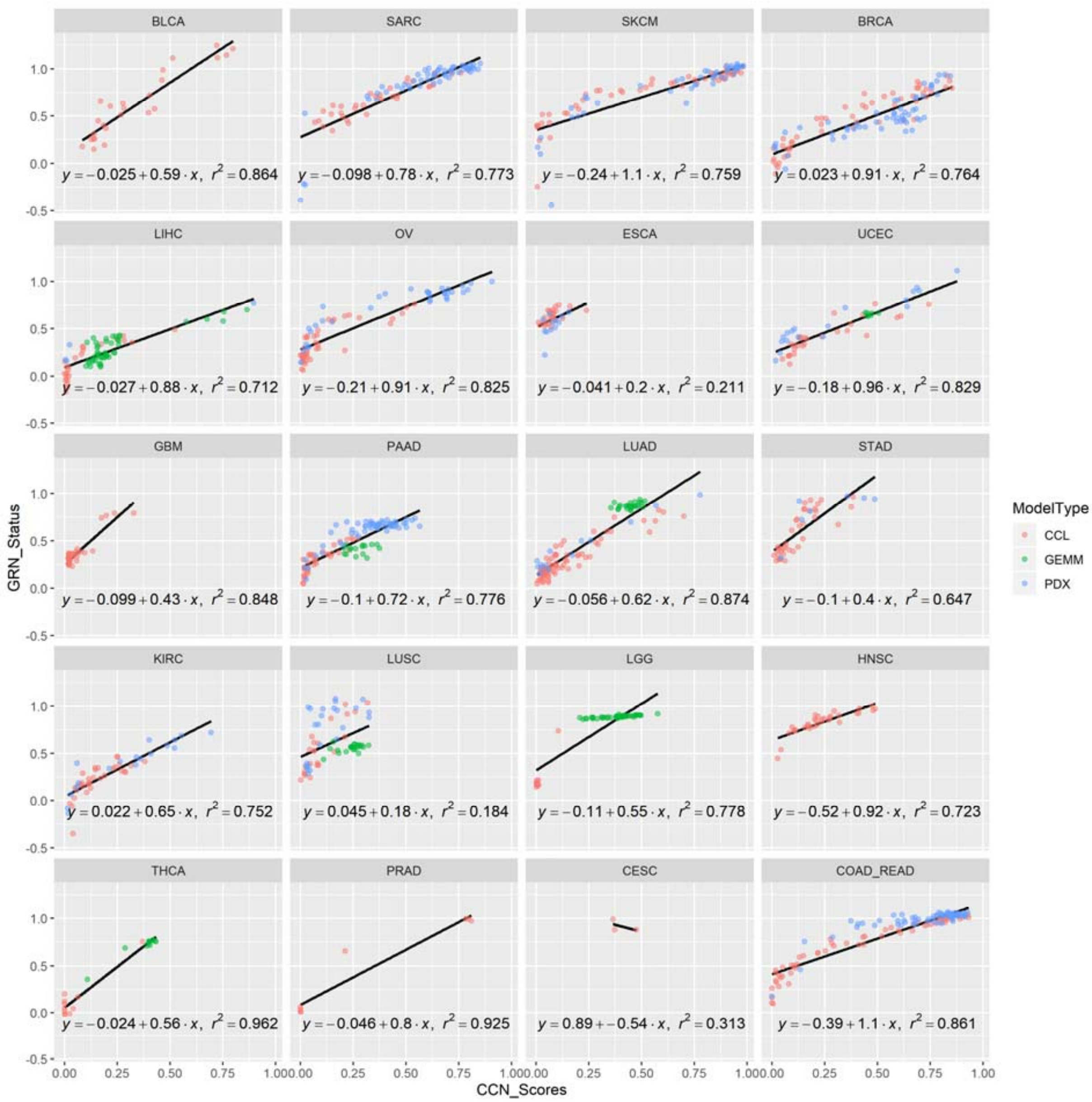
Supplemental Figure 3



Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6