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 Lummertz da Rocha[3], Stephanie Cai[1], Kathleen DiNapoli[4,5], Franklin W Huang[6], Patrick Cahan[1,2,7]

[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 Microbiology, Immunology and Parasitology, Federal University of Santa Catarina, Florianópolis SC, Brazil

[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]Division of Hematology/Oncology, Department of Medicine; Helen Diller Family Cancer Center; Bakar Computational Health Sciences Institute; Institute for Human Genetics; University of California, San Francisco, San Francisco, CA

[7]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: Research

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

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

ABSTRACT

Background: Cancer researchers use cell lines, patient derived xenografts, engineered mice, and tumoroids 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 under 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.

Methods: We developed a machine learning based 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, 26 distinct genetically engineered mouse models, and 131 tumoroids. We validated CancerCellNet by application to independent data, and we tested several predictions with immunofluorescence.

Results: We have documented the cancer models with the greatest transcriptional fidelity to natural tumors, we have identified cancers underserved by adequate models, and we have found models with annotations that do not match their classification. By comparing models across modalities, we report that, on average, genetically engineered mice and tumoroids have higher transcriptional fidelity than patient derived xenografts and cell lines in four out of five tumor types. However, several patient derived xenografts and tumoroids have classification scores that are on par with native tumors, highlighting both their potential as faithful model classes and their heterogeneity.

Conclusions: CancerCellNet enables the rapid assessment of transcriptional fidelity of tumor models. We have made CancerCellNet available as freely downloadable software and as a web application that can be applied to new cancer models that allows for direct comparison to the cancer models evaluated here.

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], patient derived xenografts (PDXs)[3], and tumoroids[4]. 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[5], GEMMs to chart progression of molecularly defined-disease[6], and PDXs to explore patient-specific response to therapy in a physiologically relevant context[7]. More recently, tumoroids have emerged as relatively inexpensive, physiological, in vitro 3D models of tumor epithelium with applications ranging from measuring drug responsiveness to exploring tumor dependence on cancer stem cells. Models also differ in the extent to which the they represent specific aspects of a cancer type[8]. 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[9,10].

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[11]. 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[12–14]. These studies were tumor-type specific, focusing on CCLs that model, for example, hepatocellular carcinoma or breast cancer. Notably, 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[15]. Most recently, Najgebauer et al[16] and Salvadores et al[17]

112 have developed methods to assess CCLs using molecular traits such as copy number
113 alterations (CNA), somatic mutations, DNA methylation and transcriptomics. While all of these
114 studies have provided valuable information, they leave two major challenges unmet. The first
115 challenge is to determine the fidelity of GEMMs, PDXs, and tumoroids, and whether there are
116 stark differences between these classes of models and CCLs. The other major unmet challenge
117 is to enable the rapid assessment of new, emerging cancer models. This challenge is especially
118 relevant now as technical barriers to generating models have been substantially lowered[18,19],
119 and because new models such as PDXs and tumoroids can be derived on patient-specific basis
120 therefore should be considered a distinct entity requiring individual validation[4,20].
121 To address these challenges, we developed CancerCellNet (CCN), a computational tool
122 that uses transcriptomic data to quantitatively assess the similarity between cancer models and
123 22 naturally occurring tumor types and 36 subtypes in a platform- and species-agnostic manner.
124 Here, we describe CCN's performance, and the results of applying it to assess 657 CCLs, 415
125 PDXs, 26 GEMMs, and 131 tumoroids. This has allowed us to identify the most faithful models
126 currently available, to document cancers underserved by adequate models, and to find models
127 with inaccurate tumor type annotation. Moreover, because CCN is open-source and easy to
128 use, it can be readily applied to newly generated cancer models as a means to assess their
129 fidelity.
130

131 ## RESULTS

132 ### CancerCellNet classifies samples accurately across species and technologies

133 Previously, we had developed a computational tool using the Random Forest
134 classification method to measure the similarity of engineered cell populations to their *in vivo*
135 counterparts based on transcriptional profiles[21,22]. More recently, we elaborated on this
136 approach to allow for classification of single cell RNA-seq data in a manner that allows for
137 cross-platform and cross-species analysis[23]. Here, we used an analogous approach to build a

138 platform that would allow us to quantitatively compare cancer models to naturally occurring
139 patient tumors (**Fig 1A**). In brief, we used TCGA RNA-seq expression data from 22 solid tumor
140 types to train a top-pair multi-class Random forest classifier (**Fig 1B**). We combined training
141 data from Rectal Adenocarcinoma (READ) and Colon Adenocarcinoma (COAD) into one
142 COAD_READ category because READ and COAD are considered to be virtually
143 indistinguishable at a molecular level[24]. We included an 'Unknown' category trained using
144 randomly shuffled gene-pair profiles generated from the training data of 22 tumor types to
145 identify query samples that are not reflective of any of the training data. To estimate the
146 performance of CCN and how it is impacted by parameter variation, we performed a parameter
147 sweep with a 5-fold 2/3 cross-validation strategy (i.e. 2/3 of the data sampled across each
148 cancer type was used to train, 1/3 was used to validate) (**Fig 1C**). The performance of CCN, as
149 measured by the mean area under the precision recall curve (AUPRC), did not fall below 0.945
150 and remained relatively stable across parameter sets (**Supp Fig 1A**). The optimal parameters
151 resulted in 1,979 features. The mean AUPRCs exceeded 0.95 in most tumor types with this
152 optimal parameter set (**Fig 1D, Supp Fig 1B**). The AUPRCs of CCN applied to independent
153 data RNA-Seq data from 725 tumors across five tumor types from the International Cancer
154 Genome Consortium (ICGC)[25] ranged from 0.93 to 0.99, supporting the notion that the platform
155 is able to accurately classify tumor samples from diverse sources (**Fig 1E**).
156 As one of the central aims of our study is to compare distinct cancer models, including
157 GEMMs, our method needed to be able to classify samples from mouse and human samples
158 equivalently. We used the Top-Pair transform[23] to achieve this and we tested the feasibility of
159 this approach by assessing the performance of a normal (i.e. non-tumor) cell and tissue
160 classifier trained on human data as applied to mouse samples. Consistent with prior
161 applications[23], we found that the cross-species classifier performed well, achieving mean
162 AUPRC of 0.97 when applied to mouse data (**Supp Fig 1C**).

163 To evaluate cancer models at a finer resolution, we also developed an approach to
164 perform tumor subtype classifications (**Supp Fig 1D**). We constructed 11 different cancer
165 subtype classifiers based on the availability of expression or histological subtype
166 information[24,26–36]. We also included non-cancerous, normal tissues as categories for several
167 subtype classifiers when sufficient data was available: breast invasive carcinoma (BRCA),
168 COAD_READ, head and neck squamous cell carcinoma (HNSC), kidney renal clear cell
169 carcinoma (KIRC) and uterine corpus endometrial carcinoma (UCEC). The 11 subtype
170 classifiers all achieved high overall average AUPRs ranging from 0.80 to 0.99 (**Supp Fig 1E**).
171

172 **Fidelity of cancer cell lines**

173 Having validated the performance of CCN, we then used it to determine the fidelity of
174 CCLs. We mined RNA-seq expression data of 657 different cell lines across 20 cancer types
175 from the Cancer Cell Line Encyclopedia (CCLE) and applied CCN to them, finding a wide
176 classification range for cell lines of each tumor type (**Fig 2A, Supp Tab 1**). To verify the
177 classification results, we applied CCN to expression profiles from CCLE generated through
178 microarray expression profiling[37]. To ensure that CCN would function on microarray data, we
179 first tested it by applying a CCN classifier created to test microarray data to 720 expression
180 profiles of 12 tumor types. The cross-platform CCN classifier performed well, based on the
181 comparison to study-provided annotation, achieving a mean AUPRC of 0.91 (**Supp Fig 2A**).
182 Next, we applied this cross-platform classifier to microarray expression profiles from CCLE
183 (**Supp Fig 2B**). From the classification results of 571 cell lines that have both RNA-seq and
184 microarray expression profiles, we found a strong overall positive association between the
185 classification scores from RNA-seq and those from microarray (**Supp Fig 2C**). This comparison
186 supports the notion that the classification scores for each cell line are not artifacts of profiling
187 methodology. Moreover, this comparison shows that the scores are consistent between the
188 times that the cell lines were first assayed by microarray expression profiling in 2012 and by

189 RNA-Seq in 2019. We also observed high level of correlation between our analysis and the
190 analysis done by Yu et al[15](**Supp Fig 2D**), further validating the robustness of the CCN results.
191 Next, we assessed the extent to which CCN classifications agreed with their nominal
192 tumor type of origin, which entailed translating quantitative CCN scores to classification labels.
193 To achieve this, we selected a decision threshold that maximized the Macro F1 measure,
194 harmonic mean of precision and recall, across 50 cross validations. Then, we annotated cell
195 lines based their CCN score profile as follows. Cell lines with CCN scores > threshold for the
196 tumor type of origin were annotated as 'correct'. Cell lines with CCN scores > threshold in the
197 tumor type of origin and at least one other tumor type were annotated as 'mixed'. Cell lines with
198 CCN scores > threshold for tumor types other than that of the cell line's origin were annotated
199 as 'other'. Cell lines that did not receive a CCN score > threshold for any tumor type were
200 annotated as 'none' (**Fig 2B**). We found that majority of cell lines originally annotated as Breast
201 invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical
202 adenocarcinoma (CESC), Skin Cutaneous Melanoma (SKCM), Colorectal Cancer
203 (COAD_READ) and Sarcoma (SARC) fell into the 'correct' category (**Fig 2B**). On the other
204 hand, no Esophageal carcinoma (ESCA), Pancreatic adenocarcinoma (PAAD) or Brain Lower
205 Grade Glioma (LGG) were classified as 'correct', demonstrating the need for more
206 transcriptionally faithful cell lines that model those general cancer types.
207 There are several possible explanations for cell lines not receiving a 'correct'
208 classification. One possibility is that the sample was incorrectly labeled in the study from which
209 we harvested the expression data. Consistent with this explanation, we found that colorectal
210 cancer line NCI-H684[38,39], a cell line labelled as liver hepatocellular carcinoma (LIHC) by CCLE,
211 was classified strongly as COAD_READ (**Supp Tab 1**). Another possibility to explain low CCN
212 score is that cell lines were derived from subtypes of tumors that are not well-represented in
213 TCGA. To explore this hypothesis, we first performed tumor subtype classification on CCLs from
214 11 tumor types for which we had trained subtype classifiers (**Supp Tab 2**). We reasoned that if

215 a cell was a good model for a rarer subtype, then it would receive a poor general classification
216 but a high classification for the subtype that it models well. Therefore, we counted the number of
217 lines that fit this pattern. We found that of the 188 lines with no general classification, 25 (13%)
218 were classified as a specific subtype, suggesting that derivation from rare subtypes is not the
219 major contributor to the poor overall fidelity of CCLs.

220 Another potential contributor to low scoring cell lines is intra-tumor stromal and immune
221 cell impurity in the training data. If impurity were a confounder of CCN scoring, then we would
222 expect a strong positive correlation between mean purity and mean CCN classification scores of
223 CCLs per general tumor type. However, the Pearson correlation coefficient between the mean
224 purity of general tumor type and mean CCN classification scores of CCLs in the corresponding
225 general tumor type was low (0.14), suggesting that tumor purity is not a major contributor to the
226 low CCN scores across CCLs (**Supp Fig 2E**).

227

228 **Comparison of SKCM and GBM CCLs to scRNA-seq**

229 To more directly assess the impact of intra-tumor heterogeneity in the training data on
230 evaluating cell lines, we constructed a classifier using cell types found in human melanoma and
231 glioblastoma scRNA-seq data[40,41]. Previously, we have demonstrated the feasibility of using our
232 classification approach on scRNA-seq data[23]. Our scRNA-seq classifier achieved a high
233 average AUPRC (0.95) when applied to held-out data and high mean AUPRC (0.99) when
234 applied to few purified bulk testing samples (**Supp Fig 3A-B**). Comparing the CCN score from
235 bulk RNA-seq general classifier and scRNA-seq classifier, we observed a high level of
236 correlation (Pearson correlation of 0.89) between the SKCM CCN classification scores and
237 scRNA-seq SKCM malignant CCN classification scores for SKCM cell lines (**Fig 2C, Supp Fig**
238 **3C**). Of the 41 SKCM cell lines that were classified as SKCM by the bulk classifier, 37 were also
239 classified as SKCM malignant cells by the scRNA-seq classifier. Interestingly, we also observed
240 a high correlation between the SARC CCN classification score and scRNA-seq cancer

241 associated fibroblast (CAF) CCN classification scores (Pearson correlation of 0.92). Six of the
242 seven SKCM cell lines that had been classified as exclusively SARC by CCN were classified as
243 CAF by the scRNA-seq classifier (**Fig 2D, Supp Fig 3C**), which suggests the possibility that
244 these cell lines were derived from CAF or other mesenchymal populations, or that they have
245 acquired a mesenchymal character through their derivation. The high level of agreement
246 between scRNA-seq and bulk RNA-seq classification results shows that heterogeneity in the
247 training data of general CCN classifier has little impact in the classification of SKCM cell lines.
248 In contrast, we observed a weaker correlation between GBM CCN classification scores
249 and scRNA-seq GBM neoplastic CCN classification scores (Pearson correlation of 0.72) for
250 GBM cell lines (**Fig 2E, Supp Fig 3D**). Of the 31 GBM lines that were not classified as GBM
251 with CCN, 25 were classified as GBM neoplastic cells with the scRNA-seq classifier. Among the
252 22 GBM lines that were classified as SARC with CCN, 15 cell lines were classified as CAF (**Fig**
253 **2F**), 10 which were classified as both GBM neoplastic and CAF in the scRNA-seq classifier.
254 Similar to the situation with SKCM lines that classify as CAF, this result is consistent with the
255 possibility that some GBM lines classified as SARC by CCN could be derived from
256 mesenchymal subtypes exhibiting both strong mesenchymal signatures and glioblastoma
257 signatures or that they have acquired a mesenchymal character through their derivation. The
258 lower level of agreement between scRNA-seq and bulk RNA-seq classification results for GBM
259 models suggests that the heterogeneity of glioblastomas[42] can impact the classification of GBM
260 cell lines, and that the use of scRNA-seq classifier can resolve this deficiency.
261
262 **Immunofluorescence confirmation of CCN predictions**
263 To experimentally explore some of our computational analyses, we performed
264 immunofluorescence on three cell lines that were not classified as their labelled categories: the
265 ovarian cancer line SK-OV-3 had a high UCEC CCN score (0.246), the ovarian cancer line
266 A2780 had a high Testicular Germ Cell Tumors (TGCT) CCN score (0.327), and the prostate

267 cancer line PC-3 had a high bladder cancer (BLCA) score (0.307) (**Supp Tab 1**). We reasoned
268 that if SK-OV-3, A2780 and PC-3 were classified most strongly as UCEC, TGCT and BLCA,
269 respectively, then they would express proteins that are indicative of these cancer types.
270 First, we measured the expression of the uterine-associated transcription factor
271 HOXB6[43,44], and the UCEC serous ovarian tumor biomarker WT1[45] in SK-OV-3, in the OV cell
272 line Caov-4, and in the UCEC cell line HEC-59. We chose Caov-4 as our positive control for OV
273 biomarker expression because it was determined by our analysis and others[11,15] to be a good
274 model of OV. Likewise, we chose HEC-59 to be a positive control for UCEC. We found that SK-
275 OV-3 has a small percentage (5%) of cells that expressed the uterine marker HOXB6 and a
276 large proportion (73%) of cells that expressed WT1 (**Fig 3A**). In contrast, no Caov-4 cells
277 expressed HOXB6, whereas 85% of cells expressed WT1. This suggests that SK-OV-3 exhibits
278 both biomarkers of ovarian tumor and uterine tissue. From our computational analysis and
279 experimental validation, SK-OV-3 is most likely an endometrioid subtype of ovarian cancer. This
280 result is also consistent with prior classification of SK-OV-3[46], and the fact that SK-OV-3 lacks
281 p53 mutations, which is prevalent in high-grade serous ovarian cancer[47], and it harbors an
282 endometrioid-associated mutation in ARID1A[11,46,48]. Next, we measured the expression of
283 markers of OV and germ cell cancers (LIN28A[49]) in the OV-annotated cell line A2780, which
284 received a high TCGT CCN score. We found that 54% of A2780 cells expressed LIN28A
285 whereas it was not detected in Caov-4 (**Fig 3B**). The OV marker WT1 was also expressed in
286 fewer A2780 cells as compared to Caov-4 (48% vs 85%), which suggests that A2780 could be a
287 germ cell derived ovarian tumor. Taken together, our results suggest that SK-OV-3 and A2780
288 could represent OV subtypes of that are not well represented in TCGA training data, which
289 resulted in a low OV score and higher CCN score in other categories.
290 Lastly, we examined PC-3, annotated as a PRAD cell line but classified to be most
291 similar to BLCA. We found that 30% of the PC-3 cells expressed PPARG, a contributor to
292 urothelial differentiation[50] that is not detected in the PRAD Vcap cell line but is highly expressed

293 in the BLCA RT4 cell line (**Fig 3C**). PC-3 cells also expressed the PRAD biomarker FOLH1[51]
294 suggesting that PC-3 has an PRAD origin and gained urothelial or luminal characteristics
295 through the derivation process. In short, our limited experimental data support the CCN
296 classification results.
297
298 **Subtype classification of cancer cell lines**
299 Next, we explored the subtype classification of CCLs from three general tumor types in
300 more depth. We focused our subtype visualization (**Fig 4A-C**) on CCL models with general CCN
301 score above 0.1 in their nominal cancer type as this allowed us to analyze those models that fell
302 below the general threshold but were classified as a specific sub-type (**Supp Tab 1-2**).
303 Focusing first on UCEC, the histologically defined subtypes of UCEC, endometrioid and serous,
304 differ in prevalence, molecular properties, prognosis, and treatment. For instance, the
305 endometrioid subtype, which accounts for approximately 80% of uterine cancers, retains
306 estrogen receptor and progesterone receptor status and is responsive towards progestin
307 therapy[52,53]. Serous, a more aggressive subtype, is characterized by the loss of estrogen and
308 progesterone receptor and is not responsive to progestin therapy[52,53]. CCN classified the
309 majority of the UCEC cell lines as serous except for JHUEM-1 which is classified as mixed, with
310 similarities to both endometrioid and serous (**Fig 4A**). The preponderance CCLE lines of serous
311 versus endometroid character may be due to properties of serous cancer cells that promote
312 their *in vitro* propagation, such as upregulation of cell adhesion transcriptional programs[54].
313 Some of our subtype classification results are consistent with prior observations. For example,
314 HEC-1A, HEC-1B, and KLE were previously characterized as type II endometrial cancer, which
315 includes a serous histological subtype[55]. On the other hand, our subtype classification results
316 contradict prior observations in at least one case. For instance, the Ishikawa cell line was
317 derived from type I endometrial cancer (endometrioid histological subtype)[55,56], however CCN
318 classified a derivative of this line, Ishikawa 02 ER-, as serous. The high serous CCN score

319 could result from a shift in phenotype of the line concomitant with its loss of estrogen receptor
320 (ER) as this is a distinguishing feature of type II endometrial cancer (serous histological
321 subtype)[52]. Taken together, these results indicate a need for more endometroid-like CCLs.
322 Next, we examined the subtype classification of Lung Squamous Cell Carcinoma
323 (LUSC) and Lung adenocarcinoma (LUAD) cell lines (Fig 4B-C). All the LUSC lines with at least
324 one subtype classification had an underlying primitive subtype classification. This is consistent
325 either with the ease of deriving lines from tumors with a primitive character, or with a process by
326 which cell line derivation promotes similarity to more primitive subtype, which is marked by
327 increased cellular proliferation[28]. Some of our results are consistent with prior reports that have
328 investigated the resemblance of some lines to LUSC subtypes. For example, HCC-95,
329 previously been characterized as classical[28,57], had a maximum CCN score in the classical
330 subtype (0.429) . Similarly, LUDLU-1 and EPLC-272H, previously reported as classical[57] and
331 basal[57] respectively, had maximal tumor subtype CCN scores for these sub-types (0.323 and
332 0.256) (**Fig 4B, Supp Tab 2**) despite classified as Unknown**.** Lastly, the LUAD cell lines that
333 were classified as a subtype were either classified as proximal inflammation or proximal
334 proliferation (**Fig 4C**). RERF-LC-Ad1 had the highest general classification score and the
335 highest proximal inflammation subtype classification score. Taken together, these subtype
336 classification results have revealed an absence of cell lines models for basal and secretory
337 LUSC, and for the Terminal respiratory unit (TRU) LUAD subtype.
338
339 **Cancer cell lines' popularity and transcriptional fidelity**
340 Finally, we sought to measure the extent to which cell line transcriptional fidelity related
341 to model prevalence. We used the number of papers in which a model was mentioned,
342 normalized by the number of years since the cell line was documented, as a rough
343 approximation of model prevalence. To explore this relationship, we plotted the normalized
344 citation count versus general classification score, labeling the highest cited and highest

classified cell lines from each general tumor type (**Fig 4D**). For most of the general tumor types, the highest cited cell line is not the highest classified cell line except for Hep G2, AGS and ML-1, representing liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), and thyroid carcinoma (THCA), respectively. On the other hand, the general scores of the highest cited cell lines representing BLCA (T24), BRCA (MDA-MB-231), and PRAD (PC-3) fall below the classification threshold of 0.25. Notably, each of these tumor types have other lines with scores exceeding 0.5, which should be considered as more faithful transcriptional models when selecting lines for a study (**Supp Tab 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[20] to CCN. Similar to the results of CCLs, the PDXs exhibited a wide range of classification scores (**Fig 5A, 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 5B**). In contrast to CCLs, we found a higher proportion of correctly classified PDX in STAD, PAAD and KIRC (**Fig 5B**)**.** However, similar to CCLs, 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 CCLs (**Fig 5C**). Several LUSC PDXs that were classified as a subtype were also classified as Head and Neck squamous cell carcinoma (HNSC) or mix HNSC and LUSC (**Fig 5D**). This could be due to the similarity in expression profiles of basal and classical subtypes of HNSC and LUSC[28,58], which is

371 consistent with the observation that these PDXs were also subtyped as classical. No LUSC
372 PDXs were classified as the secretory subtype. In contrast to LUAD CCLs, four of the five LUAD
373 PDXs with a discernible sub-type were classified as proximal inflammatory (**Fig 5E**). On the
374 other hand, similar to the CCLs, there were no TRU subtypes in the LUAD PDX cohort. In
375 summary, we found that while individual PDXs can reach extremely high transcriptional fidelity
376 to both general tumor types and subtypes, many PDXs were not classified as the general tumor
377 type from which they originated.
378
379 **Evaluation of GEMMs**
380 Next, we used CCN to evaluate GEMMs of six general tumor types from nine studies for
381 which expression data was publicly available[59–67]. As was true for CCLs and PDXs, GEMMs
382 also had a wide range of CCN scores (**Fig 6A, Supp Tab 5**). We next categorized the CCN
383 scores based on the proportion of samples associated with each tumor type that were correctly
384 classified (**Fig 6B**)**.** In contrast to LGG CCLs, LGG GEMMs, generated by Nf1 mutations
385 expressed in different neural progenitors in combination with Pten deletion[66], consistently were
386 classified as LGG (**Fig 6A-B**)**.** The GEMM dataset included multiple replicates per model, which
387 allowed us to examine intra-GEMM variability. Both at the level of CCN score and at the level of
388 categorization, GEMMs were invariant. For example, replicates of UCEC GEMMs driven by
389 Prg(cre/+)Pten(lox/lox) received almost identical general CCN scores (**Fig 6C, Supp Tab 6**).
390 GEMMs sharing genotypes across studies, such as LUAD GEMMs driven by Kras mutation and
391 loss of p53[59,65,67], also received similar general and subtype classification scores (**Fig 6A,B,E**)**.**
392 Next, we explored the extent to which genotype impacted subtype classification in
393 UCEC, LUSC, and LUAD. Prg(cre/+)Pten(lox/lox) GEMMs had a mixed subtype classification of
394 both serous and endometrioid, consistent with the fact that Pten loss occurs in both subtypes
395 (albeit more frequently in endometrioid). We also analyzed Prg(cre/+)Pten(lox/lox)Csf3r-/-
396 GEMMs. Polymorphonuclear neutrophils (PMNs), which play anti-tumor roles in endometrioid

397 cancer progression, are depleted in these animals. Interestingly, Prg(cre/+)Pten(lox/lox)Csf3r-/-
398 GEMMs had a serous subtype classification, which could be explained by differences in PMN
399 involvement in endometrioid versus serous uterine tumor development that are reflected in the
400 respective transcriptomes of the TCGA UCEC training data. We note that the tumor cells were
401 sorted prior to RNA-seq and thus the shift in subtype classification is not due to contamination of
402 GEMMs with non-tumor components. In short, this analysis supports the argument that tumor-
403 cell extrinsic factors, in this case a reduction in anti-tumor PMNs, can shift the transcriptome of
404 a GEMM so that it more closely resembles a serous rather than endometrioid subtype.
405 The LUSC GEMMs that we analyzed were Lkb1$^{fl/fl}$ and they either overexpressed of
406 Sox2 (via two distinct mechanisms) or were also Pten$^{fl/fl}$ [65]. We note that the eight lenti-Sox2-
407 Cre-infected;Lkb1$^{fl/fl}$ and Rosa26LSL-Sox2-IRES-GFP;Lkb1$^{fl/fl}$ samples that classified as
408 'Unknown' had LUSC CCN scores only modestly lower than the decision threshold (**Fig 6D**)
409 (mean CCN score = 0.217). Thirteen out of the 17 of the Sox2 GEMMs classified as the
410 secretory subtype of LUSC. The consistency is not surprising given both models overexpress
411 Sox2 and lose Lkb1. On the other hand, the Lkb1$^{fl/fl}$;Pten$^{fl/fl}$ GEMMs had substantially lower
412 general LUSC CCN scores and our subtype classification indicated that this GEMM was mostly
413 classified as 'Unknown', in contrast to prior reports suggesting that it is most similar to a basal
414 subtype[68]. None of the three LUSC GEMMs have strong classical CCN scores. Most of the
415 LUAD GEMMs, which were generated using various combinations of activating Kras mutation,
416 loss of Trp53, and loss of Smarca4L[59,65,67], were correctly classified (**Fig 6E**). Those that were
417 not classified have modestly lower CCN score than the decision threshold (mean CCN score =
418 0.214) . There were no substantial differences in general or subtype classification across driver
419 genotypes. Although the sub-type of all LUAD GEMMs was 'Unknown', the subtypes tended to
420 have a mixture of high CCN proximal proliferation, proximal inflammation and TRU scores.
421 Taken together, this analysis suggests that there is a degree of similarity, and perhaps plasticity
422 between the primitive and secretory (but not basal or classical) subtypes of LUSC. On the other

423 hand, while the LUAD GEMMs classify strongly as LUAD, they do not have strong particular
424 subtype classification -- a result that does not vary by genotype.
425
426 **Evaluation of Tumoroids**
427 Lastly, we used CCN to assess a relatively novel cancer model: tumoroids. We
428 downloaded and assessed 131 distinct tumoroid expression profiles spanning 13 cancer
429 categories from The NCI Patient-Derived Models Repository (PDMR)[69] and from three individual
430 studies[70–72] (**Fig 7A, Supp Tab 7**). We note that several categories have three or fewer samples
431 (BRCA, CESC, KIRP, OV, LIHC, and BLCA from PDMR). Among the cancer categories
432 represented by more than three samples, only LUSC and PAAD have fewer than 50% classified
433 as their annotated label (**Fig 7B**). In contrast to GBM CCLs, all three induced pluripotent stem
434 cell-derived GBM tumoroids[72] were classified as GBM with high CCN scores (mean = 0.53). To
435 further characterize the tumoroids, we performed subtype classification on them (**Supp Tab 8**).
436 UCEC tumoroids from PDMR contains a wide range of subtypes with two endometrioid, two
437 serous and one mixed type (**Fig 7C**). On the other hand, LUSC tumoroids appear to be
438 predominantly of classical subtypes with one tumoroid classified as a mix between classical and
439 primitive (**Fig 7D**). Lastly, similar to the CCL and PDX counterparts, LUAD tumoroids are
440 classified as proximal inflammatory and proximal proliferation with no tumoroids classified as
441 TRU subtype (**Fig 7E**).
442
443 **Comparison of CCLs, PDXs, GEMMs and tumoroids**
444 Finally, we sought to estimate the comparative transcriptional fidelity of the four cancer
445 models modalities. We compared the general CCN scores of each model on a per tumor type
446 basis (**Fig 8**). In the case of GEMMs, we used the mean classification score of all samples with
447 shared genotypes. We also used mean classification of technical replicates found in LIHC
448 tumoroids[70]. We evaluated models based on both the maximum CCN score, as this represents

449 the potential for a model class, and the median CCN score, as this indicates the current overall
450 transcriptional fidelity of a model class. PDXs achieved the highest CCN scores in three (UCEC,
451 PAAD, LUAD) out of the five cancer categories in which all four modalities were available (**Fig**
452 **8**), despite having low median CCN scores. Notably, PDXs have a median CCN score above
453 the 0.25 threshold in PAAD while none of the other three modalities have any samples above
454 the threshold. In LIHC, the highest CCN score for PDX (0.9) is only slightly lower than the
455 highest CCN score for tumoroid (0.91). This suggest that certain individual PDXs most closely
456 mimic the transcriptional state of native patient tumors despite a portion of the PDXs having low
457 CCN scores. Similarly, while the majority of the CCLs have low CCN scores, several lines
458 achieve high transcriptional fidelity in LUSC, LUAD and LIHC (**Fig 8**). Collectively, GEMMs and
459 tumoroids had the highest median CCN scores in four of the five model classes (LUSC and
460 LUAD for GEMMs and UCEC and LIHC for tumoroids). Notably, both of the LIHC tumoroids
461 achieved CCN scores on par with patient tumors (**Fig 8**). In brief, this analysis indicates that
462 PDXs and CCLs are heterogenous in terms of transcriptional fidelity, with a portion of the
463 models highly mimicking native tumors and the majority of the models having low transcriptional
464 fidelity (with the exception of PAAD for PDXs). On the other hand, GEMMs and tumoroids
465 displayed a consistently high fidelity across different models.

466 Because the CCN score is based on a moderate number of gene features (i.e. 1,979
467 gene pairs consisting of 1,689 unique genes) relative to the total number of protein-coding
468 genes in the genome, it is possible that a cancer model with a high CCN score might not have a
469 high global similarity to a naturally occurring tumor. Therefore, we also calculated the GRN
470 status, a metric of the extent to which tumor-type specific gene regulatory network is
471 established[21], for all models (**Supp Fig 4**). We observed high level of correlation between the
472 two similarity metrics, which suggests that although CCN classifies on a selected set of genes,
473 its scores are highly correlated with global assessment of transcriptional similarity.

474 We also sought to compare model modalities in terms of the diversity of subtypes that
475 they represent (**Supp Fig 5**). As a reference, we also included in this analysis the overall
476 subtype incidence, as approximated by incidence in TCGA. Replicates in GEMMs and
477 tumoroids were averaged into one classification profile. In models of UCEC, there is a notable
478 difference in endometroid incidence, and the proportion of models classified as endometroid,
479 with PDX and tumoroids having any representatives (**Supp Fig 5**). All of the CCL, GEMM, and
480 tumoroid models of PAAD have an unknown subtype classification and no correct general
481 classification. However, the majority of PDXs are subtyped as either a mixture of basal and
482 classical, or classical alone. LUAD have proximal inflammation and proximal proliferation
483 subtypes modelled by CCLs and PDX (**Supp Fig 5**). Likewise, LUSC have basal, classical and
484 primitive subtypes modelled by CCLs and PDXs, and secretory subtype modelled by GEMMs
485 exclusively (**Supp Fig 5**). Taken together, these results demonstrate the need to carefully select
486 different model systems to more suitably model certain cancer subtypes.
487

488 **DISCUSSION**

489 A major goal in the field of cancer biology is to develop models that mimic naturally occurring
490 tumors with enough fidelity to enable therapeutic discoveries. However, methods to measure
491 the extent to which cancer models resemble or diverge from native tumors are lacking. This is
492 especially problematic now because there are many existing models from which to choose, and
493 it has become easier to generate new models. Here, we present CancerCellNet (CCN), a
494 computational tool that measures the similarity of cancer models to 22 naturally occurring tumor
495 types and 36 subtypes. While the similarity of CCLs to patient tumors has already been
496 explored in previous work, our tool introduces the capability to assess the transcriptional fidelity
497 of PDXs, GEMMs, and tumoroids. Because CCN is platform- and species-agnostic, it
498 represents a consistent platform to compare models across modalities including CCLs, PDXs,
499 GEMMs and tumoroids. Here, we applied CCN to 657 cancer cell lines, 415 patient derived

500 xenografts, 26 distinct genetically engineered mouse models and 131 tumoroids. Several
501 insights emerged from our computational analyses that have implications for the field of cancer
502 biology.

503 First, PDXs have the greatest potential to achieve transcriptional fidelity with three out of
504 five general tumor types for which data from all modalities was available, as indicated by the
505 high scores of individual PDXs. Notably PDXs are the only modality with samples classified as
506 PAAD. At the same time, the median CCN scores of PDXs were lower than that of GEMMs and
507 tumoroids in the other four tumor types. It is unclear what causes such a wide range of CCN
508 scores within PDXs. We suspect that some PDXs might have undergone selective pressures in
509 the host that distort the progression of genomic alterations away from what is observed in
510 natural tumor[73]. Future work to understand this heterogeneity is important so as to yield
511 consistently high fidelity PDXs, and to identify intrinsic and host-specific factors that so
512 powerfully shape the PDX transcriptome.

513 Second, in general GEMMs and tumoroids have higher median CCN scores than those
514 of PDXs and CCLs. This is also consistent with that fact that GEMMs are typically derived by
515 recapitulating well-defined driver mutations of natural tumors, and thus this observation
516 corroborates the importance of genetics in the etiology of cancer[74]. Moreover, in contrast to
517 most PDXs, GEMMs are typically generated in immune replete hosts. Therefore, the higher
518 overall fidelity of GEMMs may also be a result of the influence of a native immune system on
519 GEMM tumors[75]. The high median CCN scores of tumoroids can be attributed to several factors
520 including the increased mechanical stimuli and cell-cell interactions that come from 3D self-
521 organizing cultures[76,77].

522 Third, we have found that none of the samples that we evaluated here are
523 transcriptionally adequate models of ESCA. This may be due to an inherent lability of the ESCA
524 transcriptome that is often preceded by a metaplasia that has obscured determining its cell
525 type(s) of origin[78]. Therefore, this tumor type requires further attention to derive new models.

526 Fourth, we found that in several tumor types, GEMMs tend to reflect mixtures of
527 subtypes rather than conforming strongly to single subtypes. The reasons for this are not clear
528 but it is possible that in the cases that we examined the histologically defined subtypes have a
529 degree of plasticity that is exacerbated in the murine host environment.

530 Lastly, we recognize that many CCLs are not classified as their annotated labels. While
531 we have suggested that the lack of immune component is not a major confounder, we suspect
532 that the CCLs could undergo genetic divergence due to high number of passages,
533 chemotherapy before biopsy, culture condition and genetic instability[79–82], which could all be
534 factors that drive CCLs away from their labelled tumors.

535 Currently, there are several limitations to our CCN tool, and caveats to our analyses
536 which indicate areas for future work and improvement. First, CCN is based on transcriptomic
537 data but other molecular readouts of tumor state, such as profiles of the proteome[83],
538 epigenome[84], non-coding RNA-ome[84], and genome[74] would be equally, if not more important, to
539 mimic in a model system. Therefore, it is possible that some models reflect tumor behavior well,
540 and because this behavior is not well predicted by transcriptome alone, these models have
541 lower CCN scores. To both measure the extent that such situations exist, and to correct for
542 them, we plan in the future to incorporate other omic data into CCN so as to make more
543 accurate and integrated model evaluation possible. As a first step in this direction, we plan to
544 incorporate DNA methylation and genomic sequencing data as additional features for our
545 Random forest classifier as this data is becoming more readily available for both training and
546 cancer models. We expect that this will allow us to both refine our tumor subtype categories and
547 it will enable more accurate predictions of how models respond to perturbations such as drug
548 treatment.

549 A second limitation is that in the cross-species analysis, CCN implicitly assumes that
550 homologs are functionally equivalent. The extent to which they are not functionally equivalent
551 determines how confounded the CCN results will be. This possibility seems to be of limited

552 consequence based on the high performance of the normal tissue cross-species classifier and
553 based on the fact that GEMMs have the highest median CCN scores (in addition to tumoroids).
554 A third caveat to our analysis is that there were many fewer distinct GEMMs and
555 tumoroids than CCLs and PDXs. As more transcriptional profiles for GEMMs and tumoroids
556 emerge, this comparative analysis should be revisited to assess the generality of our results.
557 Finally, the TCGA training data is made up of RNA-Seq from bulk tumor samples, which
558 necessarily includes non-tumor cells, whereas the CCLs are by definition cell lines of tumor
559 origin. Therefore, CCLs theoretically could have artificially low CCN scores due to the presence
560 of non-tumor cells in the training data. This problem appears to be limited as we found no
561 correlation between tumor purity and CCN score in the CCLE samples. However, this problem
562 is related to the question of intra-tumor heterogeneity. We demonstrated the feasibility of using
563 CCN and single cell RNA-seq data to refine the evaluation of cancer cell lines contingent upon
564 availability of scRNA-seq training data. As more training single cell RNA-seq data accrues, CCN
565 would be able to not only evaluate models on a per cell type basis, but also based on cellular
566 composition.
567 We have made the results of our analyses available online so that researchers can
568 easily explore the performance of selected models or identify the best models for any of the 22
569 general tumor types and the 36 subtypes presented here. To ensure that CCN is widely
570 available we have developed a free web application, which performs CCN analysis on user-
571 uploaded data and allows for direct comparison of their data to the cancer models evaluated
572 here. We have also made the CCN code freely available under an Open Source license and as
573 an easily installed R package, and we are actively supporting its further development. Included
574 in the web application are instructions for training CCN and reproducing our analysis. The
575 documentation describes how to analyze models and compare the results to the panel of
576 models that we evaluated here, thereby allowing researchers to immediately compare their
577 models to the broader field in a comprehensive and standard fashion.

578

579 **Online Methods**

580 **Training General CancerCellNet Classifier**

581 To generate training data sets, we downloaded 8,991 patient tumor RNA-seq expression
582 count matrix and their corresponding sample table across 22 different tumor types from TCGA
583 using TCGAWorkflowData, TCGAbiolinks[85] and SummarizedExperiment[86] packages. We used
584 all the patient tumor samples for training the general CCN classifier. We limited training and
585 analysis of RNA-seq data to the 13,142 genes in common between the TCGA dataset and all
586 the query samples (CCLs, PDXs, GEMMs, and tumoroids). To train the top pair Random forest
587 classifier, we used a method similar to our previous method[23]. CCN first normalized the training
588 counts matrix by down-sampling the counts to 500,000 counts per sample. To significantly
589 reduce the execution time and memory of generating gene pairs for all possible genes, CCN
590 then selected *n* up-regulated genes, *n* down-regulated genes and *n* least differentially
591 expressed genes (CCN training parameter nTopGenes = *n*) for each of the 22 cancer
592 categories using template matching[87] as the genes to generate top scoring gene pairs. In short,
593 for each tumor type, CCN defined a template vector that labelled the training tumor samples in
594 cancer type of interest as 1 and all other tumor samples as 0 CCN then calculated the Pearson
595 correlation coefficient between template vector and gene expressions for all genes. The genes
596 with strong match to template as either upregulated or downregulated had large absolute
597 Pearson correlation coefficient. CCN chose the upregulated, downregulated and least
598 differentially expressed genes based on the magnitude of Pearson correlation coefficient.

599 After CCN selected the genes for each cancer type, CCN generated gene pairs among
600 those genes. Gene pair transformation was a method inspired by the top-scoring pair classifier[88]
601 to allow compatibility of classifier with query expression profiles that were collected through
602 different platforms (e.g. microarray query data applied to RNA-seq training data). In brief, the
603 gene pair transformation compares 2 genes within an expression sample and encodes the

604 “gene1_gene2” gene-pair as 1 if the first gene has higher expression than the second gene.
605 Otherwise, gene pair transformation would encode the gene-pair as 0. Using all the gene pair
606 combinations generated through the gene sets per cancer type, CCN then selected top *m*
607 discriminative gene pairs (CCN training parameter nTopGenePairs = *m*) for each category using
608 template matching (with large absolute Pearson correlation coefficient) described above. To
609 prevent any single gene from dominating the gene pair list, we allowed each gene to appear at
610 maximum of three times among the gene pairs selected as features per cancer type.
611 After the top discriminative gene pairs were selected for each cancer category, CCN
612 grouped all the gene pairs together and gene pair transformed the training samples into a binary
613 matrix with all the discriminative gene pairs as row names and all the training samples as
614 column names. Using the binary gene pair matrix, CCN randomly shuffled the binary values
615 across rows then across columns to generate random profiles that should not resemble training
616 data from any of the cancer categories. CCN then sampled 70 random profiles, annotated them
617 as “Unknown” and used them as training data for the “Unknown” category. Using gene pair
618 binary training matrix, CCN constructed a multi-class Random Forest classifier of 2000 trees
619 and used stratified sampling of 60 sample size to ensure balance of training data in constructing
620 the decision trees.
621 To identify the best set of genes and gene-pair parameters (n and m), we used a grid-
622 search cross-validation[89] strategy with 5 cross-validations at each parameter set. The specific
623 parameters for the final CCN classifier using the function “broadClass_train” in the package
624 cancerCellNet are in **Supp Tab 9**. The gene-pairs are in **Supp Tab 10.**
625
626 **Validating General CancerCellNet Classifier**
627 Two thirds of patient tumor data from each cancer type were randomly sampled as
628 training data to construct a CCN classifier. Based on the training data, CCN selected the
629 classification genes and gene-pairs and trained a classifier. After the classifier was built, 35

630 held-out samples from each cancer category were sampled and 40 “Unknown” profiles were
631 generated for validation. The process of randomly sampling training set from 2/3 of all patient
632 tumor data, selecting features based on the training set, training classifier and validating was
633 repeated 50 times to have a more comprehensive assessment of the classifier trained with the
634 optimal parameter set. To test the performance of final CCN on independent testing data, we
635 applied it to 725 profiles from ICGC spanning 6 projects that do not overlap with TCGA (BRCA-
636 KR, LIRI-JP, OV-AU, PACA-AU, PACA-CA, PRAD-FR).
637

638 **Selecting Decision Thresholds**

639 Our strategy for selecting a decision threshold was to find the value that maximizes the
640 average Macro F1 measure[90] for each of the 50 cross-validations that were performed with the
641 optimal parameter set, testing thresholds between 0 and 1 with a 0.01 increment. The F1
642 measure is defined as:

643 $$Macro\ F1 = \frac{2 \times precision \times recall}{precision + recall}$$

644 We selected the most commonly occurring threshold above 0.2 that maximized the average
645 Macro F1 measure across the 50 cross-validations as the decision threshold for the final
646 classifier (threshold = 0.25). The same approach was applied for the subtype classifiers. The
647 thresholds and the corresponding average precision, recall and F1 measures are recorded in
648 (**Supp Tab 11**).
649

650 **Classifying Query Data into General Cancer Categories**

651 We downloaded the RNA-seq cancer cell lines expression profiles and sample table
652 from (https://portals.broadinstitute.org/ccle/data), and microarray cancer cell lines expression
653 profiles and sample table from Barretina et al [37]. We extracted two WT control NCCIT RNA-seq
654 expression profiles from Grow et al[91]. We received PDX expression estimates and sample

655 annotations from the authors of Gao et al [20]. We gathered GEMM expression profiles from nine
656 different studies[59–67]. We downloaded tumoroid expression profiles from The NCI Patient-
657 Derived Models Repository (PDMR)[69] and from three individual studies[70–72]. To use CCN
658 classifier on GEMM data, the mouse genes from GEMM expression profiles were converted into
659 their human homologs. The query samples were classified using the final CCN classifier. Each
660 query classification profile was labelled as one of the four classification categories: "correct",
661 "mixed", "none" and "other" based on classification profiles. If a sample has a CCN score higher
662 than the decision threshold in the labelled cancer category, we assigned that as "correct". If a
663 sample has CCN score higher than the decision threshold in labelled cancer category and in
664 other cancer categories, we assigned that as "mixed". If a sample has no CCN score higher
665 than the decision threshold in any cancer category or has the highest CCN score in 'Unknown'
666 category, then we assigned it as "none". If a sample has CCN score higher than the decision
667 threshold in a cancer category or categories not including the labelled cancer category, we
668 assigned it as "other". We analyzed and visualized the results using R and R packages
669 pheatmap[92] and ggplot2[93].
670

671 **Cross-Species Assessment**

672 To assess the performance of cross-species classification, we downloaded 1003
673 labelled human tissue/cell type and 1993 labelled mouse tissue/cell type RNA-seq expression
674 profiles from Github (https://github.com/pcahan1/CellNet). We first converted the mouse genes
675 into human homologous genes. Then we found the intersecting genes between mouse
676 tissue/cell expression profiles and human tissue/cell expression profiles. Limiting the input of
677 human tissue RNA-seq profiles to the intersecting genes, we trained a CCN classifier with all
678 the human tissue/cell expression profiles. The parameters used for the function
679 "broadClass_train" in the package cancerCellNet are in **Supp Tab 9.** 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 12**). We found the intersecting genes between the microarray profiles and TCGA patient RNA-seq profiles. Limiting the input of RNA-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 9**. 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 2A**. The same CCN classifier was used to assess microarray CCL samples **Supp Fig 2B**.

Training and validating scRNA-seq Classifier

We extracted labelled human melanoma and glioblastoma scRNA-seq expression profiles[40,41], and compiled the two datasets excluding 3 cell types T.CD4, T.CD8 and Myeloid 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 9.** 25 cells from each of the 11 cell types from the held-out data were selected to assess the single cell classifier**.** Using maximization of average Macro F1 measure, we selected the decision threshold of 0.255. The gene-pairs that were selected to construct the classifier are in **Supp Tab 10**. To assess the cross-technology capability of applying scRNA-seq classifier to bulk RNA-seq, we downloaded 305 expression

705 profiles spanning 4 purified cell types (B cells, endothelial cells, monocyte/macrophage,
706 fibroblast) from https://github.com/pcahan1/CellNet.
707
708 **Training Subtype CancerCellNet**
709 We found 11 cancer types (BRCA, COAD, ESCA, HNSC, KIRC, LGG, PAAD, UCEC,
710 STAD, LUAD, LUSC) which have meaningful subtypes based on either histology or molecular
711 profile and have sufficient samples to train a subtype classifier with high AUPR. We also
712 included normal tissues samples from BRCA, COAD, HNSC, KIRC, UCEC to create a normal
713 tissue category in the construction of their subtype classifiers. Training samples were either
714 labelled as a cancer subtype for the cancer of interest or as "Unknown" if they belong to other
715 cancer types. Similar to general classifier training, CCN performed gene pair transformation and
716 selected the most discriminate gene pairs for each cancer subtype. In addition to the gene pairs
717 selected to discriminate cancer subtypes, CCN also performed general classification of all
718 training data and appended the classification profiles of training data with gene pair binary
719 matrix as additional features. The reason behind using general classification profile as additional
720 features is that many general cancer types may share similar subtypes, and general
721 classification profile could be important features to discriminate the general cancer type of
722 interest from other cancer types before performing finer subtype classification. The specific
723 parameters used to train individual subtype classifiers using "subClass_train" function of
724 CancerCellNet package can be found in **Supp Tab 9** and the gene pairs are in **Supp Tab 10**.
725
726 **Validating Subtype CancerCellNet**
727 Similar to validating general class classifier, we randomly sampled 2/3 of all samples in
728 each cancer subtype as training data and sampled an equal amount across subtypes in the 1/3
729 held-out data for assessing subtype classifiers. We repeated the process 20 times for more
730 comprehensive assessment of subtype classifiers.

731 **Classifying Query Data into Subtypes**

732 We assigned subtype to query sample if the query sample has CCN score higher than
733 the decision threshold. The table of decision threshold for subtype classifiers are in **Supp Tab**
734 **11**. If no CCN scores exceed the decision threshold in any subtype or if the highest CCN score
735 is in 'Unknown' category, then we assigned that sample as 'Unknown'. Analysis was performed
736 in R and visualizations were generated with the ComplexHeatmap package[94].

737

738 **Cells culture, Immunohistochemistry and histomorphometry**

739 Caov-4 (ATCC® HTB-76™), SK-OV-3(ATCC® HTB-77™), RT4 (ATCC® HTB-2™), and
740 NCCIT(ATCC® CRL-2073™) cell lines were purchased from ATCC. HEC-59 (C0026001) and
741 A2780 (93112519-1VL) were obtained from Addexbio Technologies and Sigma-Aldrich. Vcap
742 and PC-3. SK-OV-3, Vcap, and RT4 were cultured in Dulbecco's Modified Eagle Medium
743 (DMEM, high glucose, 11960069, Gibco) with 1% Penicillin-Streptomycin-Glutamine (
744 10378016, Life Technologies); Caov-4, PC-3, NCCIT, and A2780 were cultured using RPMI-
745 1640 medium (11875093, Gibco) while HEC-59 was in Iscove's Modified Dulbecco's Medium
746 (IMDM, 12440053, Gibco). Both media were supplemented with 1% Penicillin-Streptomycin
747 (15140122, Gibco). All medium included 10% Fetal Bovine Serum (FBS).

748 Cells cultured in 48-well plate were washed twice with PBS and fixed in 10% buffered
749 formalin for 24 hrs at 4 °C. Immunostaining was performed using a standard protocol. Cells
750 were incubated with primary antibodies to goat HOXB6 (10 µg/mL, PA5-37867, Invitrogen),
751 mouse WT1(10 µg/mL, MA1-46028, Invitrogen), rabbit PPARG (1:50, ABN1445, Millipore),
752 mouse FOLH1(10 µg/mL, UM570025, Origene), and rabbit LIN28A (1:50, #3978, Cell Signaling)
753 in Antibody Diluent (S080981-2, DAKO), at 4 °C overnight followed with three 5 min washes in
754 TBST. The slides were then incubated with secondary antibodies conjugated with fluorescence
755 at room temperature for 1 h while avoiding light followed with three 5 min washes in TBST and

nuclear stained with mounting medium containing DAPI. Images were captured by Nikon EcLipse Ti-S, DS-U3 and DS-Qi2.

Histomorphometry was performed using ImageJ (Version 2.0.0-rc-69/1.52i). % N.positive cells was calculated by the percentage of the number of positive stained cells divided by the number of DAPI-positive nucleus within three of randomly chosen areas. The data were expressed as means ± SD.

Tumor Purity Analysis

We used the R package ESTIMATE[95] 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[95]:

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

Extracting Citation Counts

We used the R package RISmed[96] 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.

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

GRN construction and GRN Status

GRN construction was extended from our previous method[21]. 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

780 differentially expressed genes per each cancer type and extracting the subnetwork using those
781 genes.

782 To extend the original GRN status algorithm[21] across different platforms and species, we
783 devised a rank-based GRN status algorithm. Like the original GRN status, rank based GRN
784 status is a metric of assessing the similarity of cancer type specific GRN between training data
785 in the cancer type of interest and query samples. Hence, high GRN status represents high level
786 of establishment or similarity of the cancer specific GRN in the query sample compared to those
787 of the training data. The expression profiles of training data and query data were transformed
788 into rank expression profiles by replacing the expression values with the rank of the expression
789 values within a sample (highest expressed gene would have the highest rank and lowest
790 expressed genes would have a rank of 1). Cancer type specific mean and standard deviation of
791 every gene's rank expression were learned from training data. The modified Z-score values for
792 genes within cancer type specific GRN were calculated for query sample's rank expression
793 profiles to quantify how dissimilar the expression values of genes in query sample's cancer type
794 specific GRN compared to those of the reference training data:

795 $$Zscore(gene\ i)_{mod} = \begin{cases} 0,\ if\ Zscore\ is\ positive\ and\ the\ gene\ is\ found\ to\ be\ upregulated \\ 0,\ if\ Zscore\ is\ negative\ and\ the\ gene\ is\ found\ to\ be\ downregulated \\ abs(Zscore),\ otherwise \end{cases}$$

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

801 GRN status for query sample is calculated as the weighted mean of the
802 $(1000 - Zscore(gene\ i)_{mod})$ 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. Finally, the GRN status gets normalized with 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 lowest average GRN status in the cancer type of interest. $avg(GRN\ status_{cancer\ type\ interest})$ 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

Acknowledgements

This work was supported by the National Institutes of Health NCI Ovarian Cancer SPORE P50CA228991 via a Development Research Program award to PC. FWH was supported by a Prostate Cancer Foundation Young Investigator Award, Department of Defense W81XWH-17-PCRP-HD (F.W.H.), the National Institutes of Health/National Cancer Institute P20 CA233255-

826 01 (F.W.H.) U19 CA214253 (F.W.H.). We would like to thank John Powers, Hao Zhu, Tian-Li
827 Wang, Charles Eberhart, and Kaloyan Tsanov for comments on the manuscript and helpful
828 discussions. Some figures were created in part with Biorender.com.
829

830 **FIGURE LEGENDS**

831 **Fig. 1 CancerCellNet (CCN) workflow, training, and performance. (A)** Schematic of CCN
832 usage. CCN was designed to assess and compare the expression profiles of cancer models
833 such as CCLs, PDXs, GEMMs, and tumoroids with native patient tumors. To use trained
834 classifier, CCN inputs the query samples (e.g. expression profiles from CCLs, PDXs, GEMMs,
835 tumoroids) and generates a classification profile for the query samples. The column names of
836 the classification heatmap represent sample annotation and the row names of the classification
837 heatmap represent different cancer types. Each grid is colored from black to yellow representing
838 the lowest classification score (e.g. 0) to highest classification score (e.g. 1). **(B)** Schematic of
839 CCN training process. CCN uses patient tumor expression profiles of 22 different cancer types
840 from TCGA as training data. First, CCN identifies *n* genes that are upregulated, *n* that are
841 downregulated, and *n* that are relatively invariant in each tumor type versus all of the others.
842 Then, CCN performs a pair transform on these genes and subsequently selects the most
843 discriminative set of *m* gene pairs for each cancer type as features (or predictors) for the
844 Random forest classifier. Lastly, CCN trains a multi-class Random Forest classifier using gene-
845 pair transformed training data. **(C)** Parameter optimization strategy. 5 cross-validations of each
846 parameter set in which 2/3 of TCGA data was used to train and 1/3 to validate was used search
847 for the values of *n* and *m* that maximized performance of the classifier as measured by area
848 under the precision recall curve (AUPRC). **(D)** Mean and standard deviation of classifiers based
849 on 50 cross-validations with the optimal parameter set. **(E)** AUPRC of the final CCN classifier
850 when applied to independent patient tumor data from ICGC.
851

852 **Fig. 2 Evaluation of cancer cell lines. (A)** General classification heatmap of CCLs extracted
853 from CCLE. Column annotations of the heatmap represent the labelled cancer category of the
854 CCLs given by CCLE and the row names of the heatmap represent different cancer categories.
855 CCLs' general classification profiles are categorized into 4 categories: correct (red), correct
856 mixed (pink), no classification (light green) and other classification (dark green) based on the
857 decision threshold of 0.25. **(B)** Bar plot represents the proportion of each classification category
858 in CCLs across cancer types ordered from the cancer types with the highest proportion of
859 correct and correct mixed CCLs to lowest proportion. **(C)** Comparison between SKCM general
860 CCN scores from bulk RNA-seq classifier and SKCM malignant CCN scores from scRNA-seq
861 classifier for SKCM CCLs. **(D)** Comparison between SARC general CCN scores from bulk RNA-
862 seq classifier and CAF CCN scores from scRNA-seq classifier for SKCM CCLs. **(E)** Comparison
863 between GBM general CCN scores from bulk RNA-seq classifier and GBM neoplastic CCN
864 scores from scRNA-seq classifier for GBM CCLs. **(F)** Comparison between SARC general CCN
865 scores and CAF CCN scores from scRNA-seq classifier for GBM CCLs. The green lines
866 indicate the decision threshold for scRNA-seq classifier and general classifier.
867
868 **Fig. 3 Immunofluorescence of selected cell lines. (A)** Classification profiles (left) and IF
869 expression (middle) of Caov-4 (OV positive control), HEC-59 (UCEC positive control) and SK-
870 OV-3 for WT1 (OV biomarker) and HOXB6 (uterine biomarker). The bar plots quantify the
871 average percentage of positive cells for WT1 (top-right) and HOXB6 (bottom-right). **(B)**
872 Classification profiles (left) and IF expression (middle) of Caov-4, NCCIT (germ cell tumor
873 positive control) and A2780 for WT1 and LIN28A (germ cell tumor biomarker). Classification of
874 NCCIT were performed using RNA-seq profiles of WT control NCCIT duplicate from Grow et
875 al[91]. The bar plots quantify the average percentage of positive cells for WT1 (top-right) and
876 LIN28A (bottom-right). **(C)** Classification profiles (left) and IF expression (middle) of Vcap
877 (PRAD positive control), RT4 (BLCA positive control) and PC-3 for FOLH1 (prostate biomarker)

878 and PPARG (urothelial biomarker). The bar plots quantify the average percentage of positive
879 cells for FOLH1 (top-right) and PPARG (bottom-right).
880
881 **Fig. 4 Subtype classification of CCLs and CCL prevalence.** The heatmap visualizations
882 represent subtype classification of **(A)** UCEC CCLs**, (B)** LUSC CCLs and **(C)** LUAD CCLs. Only
883 samples with CCN scores > 0.1 in their nominal tumor type are displayed. **(D)** Comparison of
884 normalized citation counts and general CCN classification scores of CCLs. Labelled cell lines
885 either have the highest CCN classification score in their labelled cancer category or highest
886 normalized citation count. Each citation count was normalized by number of years since first
887 documented on PubMed.
888
889 **Fig. 5 Evaluation of patient derived xenografts. (A)** General classification heatmap of PDXs.
890 Column annotations represent annotated cancer type of the PDXs, and row names represent
891 cancer categories. **(B)** Proportion of classification categories in PDXs across cancer types is
892 visualized in the bar plot and ordered from the cancer type with highest proportion of correct and
893 mixed correct classified PDXs to the lowest. Subtype classification heatmaps of **(C)** UCEC
894 PDXs, **(D)** LUSC PDXs and **(E)** LUAD PDXs. Only samples with CCN scores > 0.1 in their
895 nominal tumor type are displayed.
896
897 **Fig. 6 Evaluation of genetically engineered mouse models. (A)** General classification
898 heatmap of GEMMs. Column annotations represent annotated cancer type of the GEMMs, and
899 row names represent cancer categories. **(B)** Proportion of classification categories in GEMMs
900 across cancer types is visualized in the bar plot and ordered from the cancer type with highest
901 proportion of correct and mixed correct classified GEMMs to the lowest. Subtype classification
902 heatmap of **(C)** UCEC GEMMs, **(D)** LUSC GEMMs and **(E)** LUAD GEMMs. Only samples with
903 CCN scores > 0.1 in their nominal tumor type are displayed.

904

905 **Fig. 7 Evaluation of tumoroid models. (A)** General classification heatmap of tumoroids.
906 Column annotations represent annotated cancer type of the tumoroids, and row names
907 represent cancer categories. **(B)** Proportion of classification categories in tumoroids across
908 cancer types is visualized in the bar plot and ordered from the cancer type with highest
909 proportion of correct and mixed correct classified tumoroids to the lowest. Subtype classification
910 heatmap of **(C)** UCEC tumoroids, **(D)** LUSC tumoroids and **(E)** LUAD tumoroids. Only samples
911 with CCN scores > 0.1 in their nominal tumor type are displayed.

912

913 **Fig. 8 Comparison of CCLs, PDXs, and GEMMs.** Box-and-whiskers plot comparing general
914 CCN scores across CCLs, GEMMs, PDXs of five general tumor types (UCEC, PAAD, LUSC,
915 LUAD, LIHC).

916

917 **Supplementary Information**

918 **Supplementary Figure 1** Assessment of CCN general classifier and subtype classifier. **(A)**
919 Mean AUPRC of repeated grid-search cross-validation for each parameter grid. **(B)** Mean and
920 range of CCN classifier's PR curves from 50 cross validations based on the optimal feature
921 selection parameters *n* and *m*. **(C)** AUPRC of CCN human tissue classifier when applied to
922 mouse tissue data. **(D)** The schematic of training a subtype classifier in CCN. CCN uses patient
923 tumor expression profiles from cancer of interest as training data. CCN performs gene-pair
924 transformation and selects the most discriminative gene pairs among the cancer subtypes from
925 training data as features. CCN then applies the general classification on training data and uses
926 the general classification profile as features in addition to gene pairs for training a Random
927 Forest classifier. The weight of the general classification profiles as features can be tuned to
928 improve AUPRC. **(E)** The mean and standard deviation of AUPRC for 11 subtype classifiers
929 based on 20 iterations of random sampling of training and held-out data, training subtype

930 classifier using training data, classification of held-out data, and calculation of recall and
931 precision.
932
933 **Supplementary Figure 2** Further validation of CCN and classification results. To validate the
934 cross-platform classification performance of CCN, a new classifier specifically trained to classify
935 microarray data was trained using RNA-seq data from TCGA as training data and intersecting
936 genes between RNA-seq data and microarray data. **(A)** AUPRC of CCN classifier when applied
937 to tumor profiles assayed on microarrays. **(B)** Classification heatmap of CCLs using microarray
938 expression data. **(C)** Pearson correlation between CCN scores of CCLE lines generated from
939 RNA-seq data and microarray data. **(D)** Comparison between CCLs' CCN scores and the
940 similarity metric from Yu et al[15], median correlations of transcriptional profiles between CCLs
941 and TCGA tumors from CCLs' labelled cancer category. **(E)** Comparison of mean tumor purity
942 of training data and mean CCN scores of CCLs for each cancer category.
943
944 **Supplementary Figure 3** Single-cell classification of SKCM and GBM cell lines. **(A)** AUPRC of
945 the single-cell classifier when applied to scRNA-seq held-out data. **(B)** AUPRC of the scRNA-
946 seq classifier when applied to purified bulk RNA samples. **(C)** Single-cell classification of SKCM
947 CCLs. Red bar-plot (top) represents general CCN scores in SARC and blue bar-plot (bottom)
948 represents general CCN scores in SKCM. **(D)** Single-cell classification of GBM CCLs. Red bar-
949 plot (top) represents general CCN scores in SARC and yellow bar-plot (bottom) represents
950 general CCN scores in GBM.
951
952 **Supplementary Figure 4** Correlation between cancer type specific network GRN status and
953 general CCN scores.
954
955
956 **Supplementary Figure 5** Proportion of cancer subtypes in different cancer models and TCGA
957 tumor data across 11 general cancer types.
958

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 General classification profiles of tumoroids.

Supplementary Table 8 Subtype classification profiles of tumoroids.

Supplementary Table 9 Specific parameters used for training of all classifiers.

Supplementary Table 10 Gene-pairs selected for final training of CCN general, subtype classifiers and single-cell classifier.

Supplementary Table 11 Decision thresholds and the corresponding precision and recall for the general classifier and subtype classifier.

Supplementary Table 12 Accessions of tumor microarray data used in validation.

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. Drost, J. & Clevers, H. Organoids in cancer research. *Nat. Rev. Cancer* **18,** 407–418 (2018).
5. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33,** 306–312 (2015).
6. Koren, S. *et al.* PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. *Nature* **525,** 114–118 (2015).
7. 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).

8. 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).
9. 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).
10. Stuckelberger, S. & Drapkin, R. Precious GEMMs: emergence of faithful models for ovarian cancer research. *J. Pathol.* **245,** 129–131 (2018).
11. 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).
12. Jiang, G. *et al.* Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. *BMC Genomics* **17 Suppl 7,** 525 (2016).
13. 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).
14. 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).
15. Yu, K. *et al.* Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10,** 3574 (2019).
16. Najgebauer, H. *et al.* CELLector: Genomics-Guided Selection of Cancer In Vitro Models. *Cell Syst.* **10,** 424–432.e6 (2020).
17. Salvadores, M., Fuster-Tormo, F. & Supek, F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Sci. Adv.* **6,** (2020).
18. Guernet, A. & Grumolato, L. CRISPR/Cas9 editing of the genome for cancer modeling. *Methods* **121-122,** 130–137 (2017).
19. Gargiulo, G. Next-Generation in vivo Modeling of Human Cancers. *Front. Oncol.* **8,** 429 (2018).
20. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21,** 1318–1325 (2015).
21. Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158,** 903–915 (2014).
22. Radley, A. H. *et al.* Assessment of engineered cells using CellNet and RNA-seq. *Nat. Protoc.* **12,** 1089–1102 (2017).
23. 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).
24. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
25. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford)* **2011,** bar026 (2011).
26. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
27. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

1051 28. Wilkerson, M. D. *et al.* Lung squamous cell carcinoma mRNA expression subtypes
1052 are reproducible, clinically important, and correspond to normal cell types. *Clin.*
1053 *Cancer Res.* **16,** 4864–4875 (2010).
1054 29. Cancer Genome Atlas Research Network. Electronic address:
1055 andrew_aguirre@dfci.harvard.edu & Cancer Genome Atlas Research Network.
1056 Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer*
1057 *Cell* **32,** 185–203.e13 (2017).
1058 30. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization
1059 of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
1060 31. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization
1061 of oesophageal carcinoma. *Nature* **541,** 169–175 (2017).
1062 32. Cancer Genome Atlas Network. Comprehensive genomic characterization of head
1063 and neck squamous cell carcinomas. *Nature* **517,** 576–582 (2015).
1064 33. Cancer Genome Atlas Research Network. Comprehensive molecular
1065 characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
1066 34. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant
1067 subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR,
1068 and NF1. *Cancer Cell* **17,** 98–110 (2010).
1069 35. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of
1070 lung adenocarcinoma. *Nature* **511,** 543–550 (2014).
1071 36. Hu, B. *et al.* Gastric cancer: Classification, histology and application of molecular
1072 pathology. *J. Gastrointest. Oncol.* **3,** 251–261 (2012).
1073 37. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling
1074 of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
1075 38. Medico, E. *et al.* The molecular landscape of colorectal cancer cell lines unveils
1076 clinically actionable kinase targets. *Nat. Commun.* **6,** 7002 (2015).
1077 39. Park, J.-G. *et al.* Characteristics of Cell Lines Established from Human Colorectal
1078 Carcinoma. *Cancer Res.* (1987).
1079 40. Jerby-Arnon, L. *et al.* A cancer cell program promotes T cell exclusion and
1080 resistance to checkpoint blockade. *Cell* **175,** 984–997.e24 (2018).
1081 41. Darmanis, S. *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at
1082 the Migrating Front of Human Glioblastoma. *Cell Rep.* **21,** 1399–1410 (2017).
1083 42. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in
1084 primary glioblastoma. *Science* **344,** 1396–1401 (2014).
1085 43. Xu, B. *et al.* Regulation of endometrial receptivity by the highly expressed HOXA9,
1086 HOXA11 and HOXD10 HOX-class homeobox genes. *Hum. Reprod.* **29,** 781–790
1087 (2014).
1088 44. Raines, A. M. *et al.* Recombineering-based dissection of flanking and paralogous
1089 Hox gene functions in mouse reproductive tracts. *Development* **140,** 2942–2952
1090 (2013).
1091 45. Netinatsunthorn, W., Hanprasertpong, J., Dechsukhum, C., Leetanaporn, R. &
1092 Geater, A. WT1 gene expression as a prognostic marker in advanced serous
1093 epithelial ovarian carcinoma: an immunohistochemical study. *BMC Cancer* **6,** 90
1094 (2006).
1095 46. Kelly, Z. *et al.* The prognostic significance of specific HOX gene expression patterns
1096 in ovarian cancer. *Int. J. Cancer* **139,** 1608–1617 (2016).

1097 47. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian
1098 carcinoma. *Nature* **474,** 609–615 (2011).
1099 48. Wiegand, K. C. *et al.* ARID1A mutations in endometriosis-associated ovarian
1100 carcinomas. *N. Engl. J. Med.* **363,** 1532–1543 (2010).
1101 49. Murray, M. J. *et al.* LIN28 Expression in malignant germ cell tumors downregulates
1102 let-7 and increases oncogene levels. *Cancer Res.* **73,** 4872–4884 (2013).
1103 50. Biton, A. *et al.* Independent component analysis uncovers the landscape of the
1104 bladder tumor transcriptome and reveals insights into luminal and basal subtypes.
1105 *Cell Rep.* **9,** 1235–1245 (2014).
1106 51. Fair, W. R., Israeli, R. S. & Heston, W. D. Prostate-specific membrane antigen.
1107 *Prostate* **32,** 140–148 (1997).
1108 52. Black, J. D., English, D. P., Roque, D. M. & Santin, A. D. Targeted therapy in
1109 uterine serous carcinoma: an aggressive variant of endometrial cancer. *Womens*
1110 *Health (Lond. Engl.)* **10,** 45–57 (2014).
1111 53. Yang, S., Thiel, K. W. & Leslie, K. K. Progesterone: the ultimate endometrial tumor
1112 suppressor. *Trends Endocrinol. Metab.* **22,** 145–152 (2011).
1113 54. Huszar, M. *et al.* Up-regulation of L1CAM is linked to loss of hormone receptors and
1114 E-cadherin in aggressive subtypes of endometrial carcinomas. *J. Pathol.* **220,** 551–
1115 561 (2010).
1116 55. Kozak, J., Wdowiak, P., Maciejewski, R. & Torres, A. A guide for endometrial
1117 cancer cell lines functional assays using the measurements of electronic
1118 impedance. *Cytotechnology* **70,** 339–350 (2018).
1119 56. Korch, C. *et al.* DNA profiling analysis of endometrial and ovarian cell lines reveals
1120 misidentification, redundancy and contamination. *Gynecol. Oncol.* **127,** 241–248
1121 (2012).
1122 57. Wu, D. *et al.* Gene-expression data integration to squamous cell lung cancer
1123 subtypes reveals drug sensitivity. *Br. J. Cancer* **109,** 1599–1608 (2013).
1124 58. Walter, V. *et al.* Molecular subtypes in head and neck cancer exhibit distinct
1125 patterns of chromosomal gain and loss of canonical cancer genes. *PLoS One* **8,**
1126 e56823 (2013).
1127 59. Adeegbe, D. O. *et al.* BET Bromodomain Inhibition Cooperates with PD-1 Blockade
1128 to Facilitate Antitumor Response in Kras-Mutant Non-Small Cell Lung Cancer.
1129 *Cancer Immunol Res* **6,** 1234–1245 (2018).
1130 60. Blaisdell, A. *et al.* Neutrophils oppose uterine epithelial carcinogenesis via
1131 debridement of hypoxic tumor cells. *Cancer Cell* **28,** 785–799 (2015).
1132 61. Fitamant, J. *et al.* YAP inhibition restores hepatocyte differentiation in advanced
1133 HCC, leading to tumor regression. *Cell Rep.* **10,** 1692–1707 (2015).
1134 62. Jia, D. *et al.* Crebbp loss drives small cell lung cancer and increases sensitivity to
1135 HDAC inhibition. *Cancer Discov.* **8,** 1422–1437 (2018).
1136 63. Kress, T. R. *et al.* Identification of MYC-Dependent Transcriptional Programs in
1137 Oncogene-Addicted Liver Tumors. *Cancer Res.* **76,** 3463–3472 (2016).
1138 64. Li, L. *et al.* GKAP acts as a genetic modulator of NMDAR signaling to govern
1139 invasive tumor growth. *Cancer Cell* **33,** 736–751.e5 (2018).
1140 65. Mollaoglu, G. *et al.* The Lineage-Defining Transcription Factors SOX2 and NKX2-1
1141 Determine Lung Cancer Cell Fate and Shape the Tumor Immune
1142 Microenvironment. *Immunity* **49,** 764–779.e9 (2018).

1143 66. Pan, Y. *et al.* Whole tumor RNA-sequencing and deconvolution reveal a clinically-
1144 prognostic PTEN/PI3K-regulated glioma transcriptional signature. *Oncotarget* **8,**
1145 52474–52487 (2017).
1146 67. Lissanu Deribe, Y. *et al.* Mutations in the SWI/SNF complex induce a targetable
1147 dependence on oxidative phosphorylation in lung cancer. *Nat. Med.* **24,** 1047–1057
1148 (2018).
1149 68. Xu, C. *et al.* Loss of Lkb1 and Pten leads to lung squamous cell carcinoma with
1150 elevated PD-L1 expression. *Cancer Cell* **25,** 590–604 (2014).
1151 69. NCI-Frederick, Frederick, MD. National Laboratory for Cancer Research. The NCI
1152 Patient-Derived Models Repository (PDMR). (2019). at <https://pdmr.cancer.gov/>
1153 70. Broutier, L. *et al.* Human primary liver cancer-derived organoid cultures for disease
1154 modeling and drug screening. *Nat. Med.* **23,** 1424–1435 (2017).
1155 71. Lee, S. H. *et al.* Tumor Evolution and Drug Response in Patient-Derived Organoid
1156 Models of Bladder Cancer. *Cell* **173,** 515–528.e17 (2018).
1157 72. Ogawa, J., Pao, G. M., Shokhirev, M. N. & Verma, I. M. Glioblastoma model using
1158 human cerebral organoids. *Cell Rep.* **23,** 1220–1229 (2018).
1159 73. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor
1160 evolution. *Nat. Genet.* **49,** 1567–1575 (2017).
1161 74. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,**
1162 719–724 (2009).
1163 75. Balkwill, F. R., Capasso, M. & Hagemann, T. The tumor microenvironment at a
1164 glance. *J. Cell Sci.* **125,** 5591–5596 (2012).
1165 76. Lancaster, M. A. & Knoblich, J. A. Organogenesis in a dish: modeling development
1166 and disease using organoid technologies. *Science* **345,** 1247125 (2014).
1167 77. Bregenzer, M. E. *et al.* Integrated cancer tissue engineering models for precision
1168 medicine. *PLoS One* **14,** e0216564 (2019).
1169 78. Wang, D. H. & Souza, R. F. Biology of Barrett's esophagus and esophageal
1170 adenocarcinoma. *Gastrointest Endosc Clin N Am* **21,** 25–38 (2011).
1171 79. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and
1172 EGF more closely mirror the phenotype and genotype of primary tumors than do
1173 serum-cultured cell lines. *Cancer Cell* **9,** 391–403 (2006).
1174 80. Wenger, S. L. *et al.* Comparison of established cell lines at different passages by
1175 karyotype and comparative genomic hybridization. *Biosci. Rep.* **24,** 631–639 (2004).
1176 81. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug
1177 response. *Nature* **560,** 325–330 (2018).
1178 82. Cooke, S. L. *et al.* Genomic analysis of genetic heterogeneity and evolution in high-
1179 grade serous ovarian carcinoma. *Oncogene* **29,** 4905–4913 (2010).
1180 83. Hristova, V. A. & Chan, D. W. Cancer biomarker discovery and translation:
1181 proteomics and beyond. *Expert Rev Proteomics* **16,** 93–103 (2019).
1182 84. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy.
1183 *Cell* **150,** 12–27 (2012).
1184 85. Silva, T. C. *et al.* TCGA Workflow: Analyze cancer genomics and epigenomics data
1185 using Bioconductor packages. [version 2; peer review: 1 approved, 2 approved with
1186 reservations]. *F1000Res.* **5,** 1542 (2016).
1187 86. Morgan, M., Obenchain, V., Hester, J. & Pag`es, H. *SummarizedExperiment:*
1188 *SummarizedExperiment container.* (2018).

87. Pavlidis, P. & Noble, W. S. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol.* **2,** RESEARCH0042 (2001).
88. 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).
89. Krstajic, D., Buturovic, L. J., Leahy, D. E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **6,** 10 (2014).
90. Lipton, Z. C., Elkan, C. & Naryanaswamy, B. Optimal Thresholding of Classifiers to Maximize F1 Measure. *Mach. Learn. Knowl. Discov. Databases* **8725,** 225–239 (2014).
91. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522,** 221–225 (2015).
92. Kolde, R. *pheatmap: Pretty Heatmaps*. (CRAN, 2019).
93. Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis* . (Springer-Verlag New York, 2016). doi:10.1007/978-0-387-98141-3
94. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).
95. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
96. Kovalchik, S. *RISmed: Download Content from NCBI Databases*. (CRAN.R-project, 2017).

Figure 1

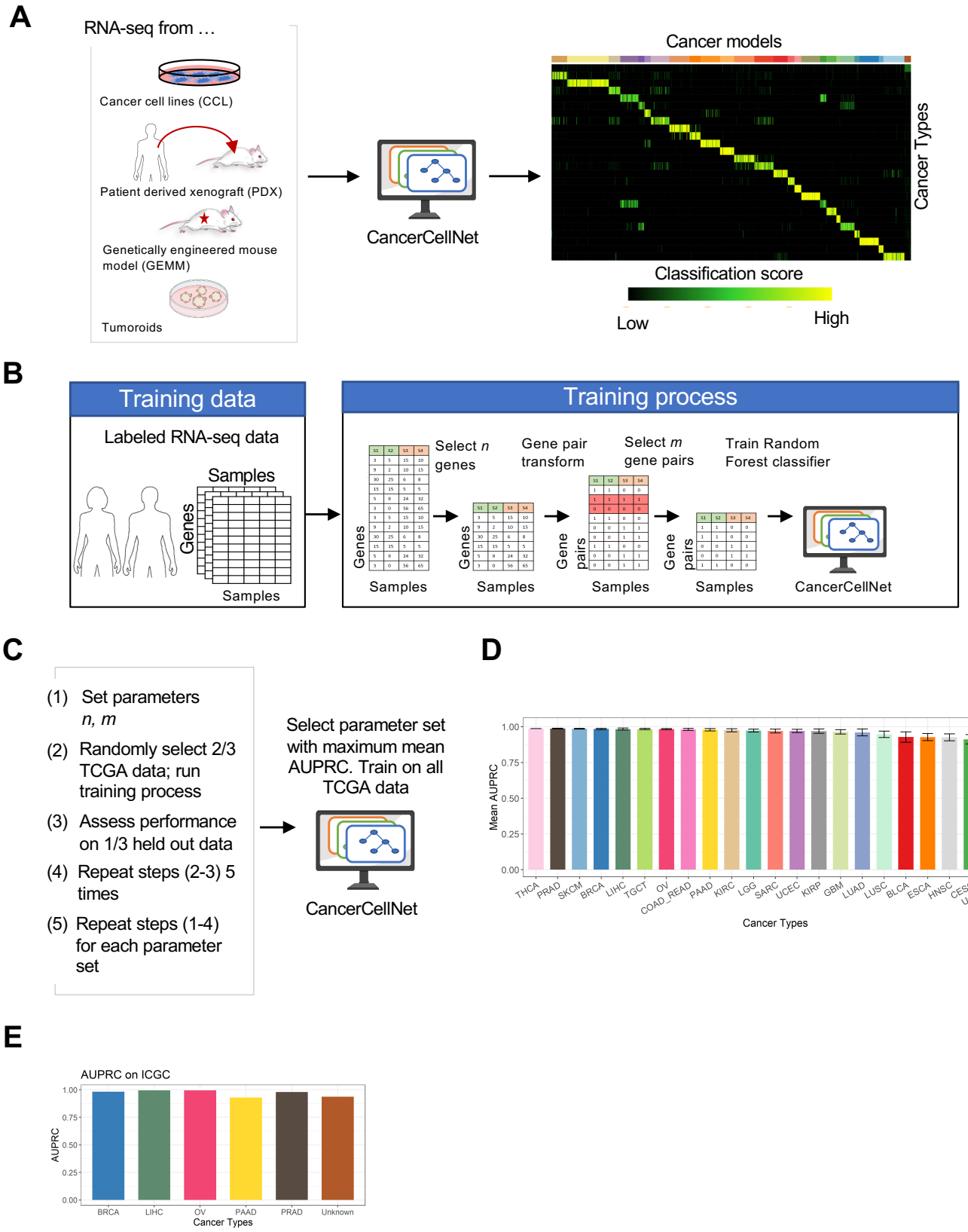

A
RNA-seq from …
Cancer cell lines (CCL)
Patient derived xenograft (PDX)
Genetically engineered mouse model (GEMM)
Tumoroids
CancerCellNet
Cancer models
Cancer Types
Classification score
Low
High
B
Training data
Labeled RNA-seq data
Samples
Genes
Samples
Training process
Select n genes
Gene pair transform
Select m gene pairs
Train Random Forest classifier
Genes
Samples
Genes
Samples
Gene pairs
Samples
Gene pairs
Samples
CancerCellNet
C
(1) Set parameters n, m
(2) Randomly select 2/3 TCGA data; run training process
(3) Assess performance on 1/3 held out data
(4) Repeat steps (2-3) 5 times
(5) Repeat steps (1-4) for each parameter set
Select parameter set with maximum mean AUPRC. Train on all TCGA data
CancerCellNet
D
Mean AUPRC
1.00
0.75
0.50
0.25
0.00
THCA
PRAD
SKCM
BRCA
LIHC
TGCT
OV
COAD_READ
PAAD
KIRC
LGG
SARC
UCEC
KIRP
GBM
LUAD
LUSC
BLCA
ESCA
HNSC
CESC
Unknown
STAD
Cancer Types
E
AUPRC on ICGC
AUPRC
1.00
0.75
0.50
0.25
0.00
BRCA
LIHC
OV
PAAD
PRAD
Unknown
Cancer Types

Figure 2

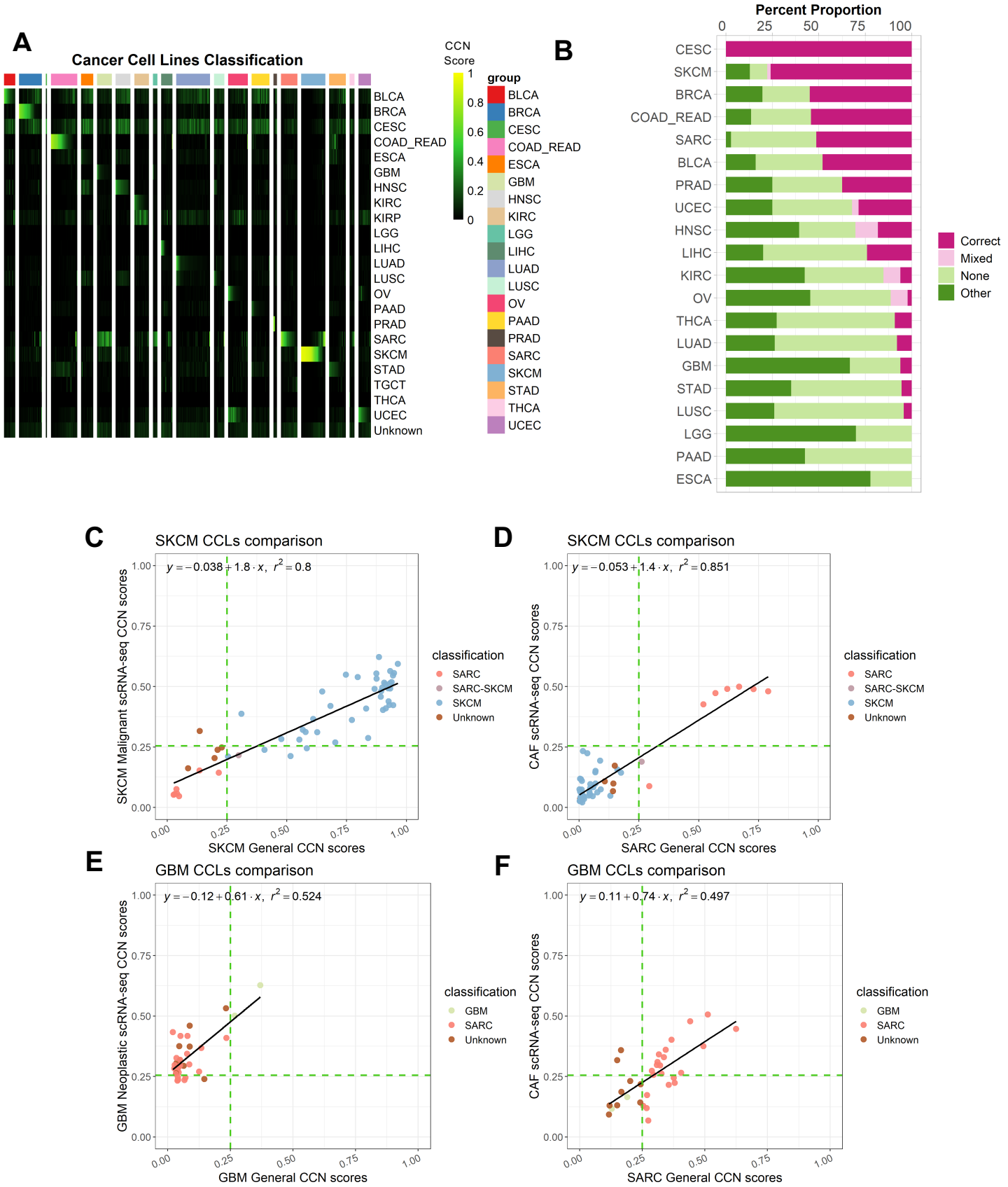

A
Cancer Cell Lines Classification
CCN Score
1
0.8
0.6
0.4
0.2
0
group
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
B
Percent Proportion
0
25
50
75
100
CESC
SKCM
BRCA
COAD_READ
SARC
BLCA
PRAD
UCEC
HNSC
LIHC
KIRC
OV
THCA
LUAD
GBM
STAD
LUSC
LGG
PAAD
ESCA
Correct
Mixed
None
Other
C
SKCM CCLs comparison
$y = -0.038 + 1.8 \cdot x,\ r^2 = 0.8$
SKCM Malignant scRNA-seq CCN scores
SKCM General CCN scores
classification
SARC
SARC-SKCM
SKCM
Unknown
D
SKCM CCLs comparison
$y = -0.053 + 1.4 \cdot x,\ r^2 = 0.851$
CAF scRNA-seq CCN scores
SARC General CCN scores
classification
SARC
SARC-SKCM
SKCM
Unknown
E
GBM CCLs comparison
$y = -0.12 + 0.61 \cdot x,\ r^2 = 0.524$
GBM Neoplastic scRNA-seq CCN scores
GBM General CCN scores
classification
GBM
SARC
Unknown
F
GBM CCLs comparison
$y = 0.11 + 0.74 \cdot x,\ r^2 = 0.497$
CAF scRNA-seq CCN scores
SARC General CCN scores
classification
GBM
SARC
Unknown

Figure 3

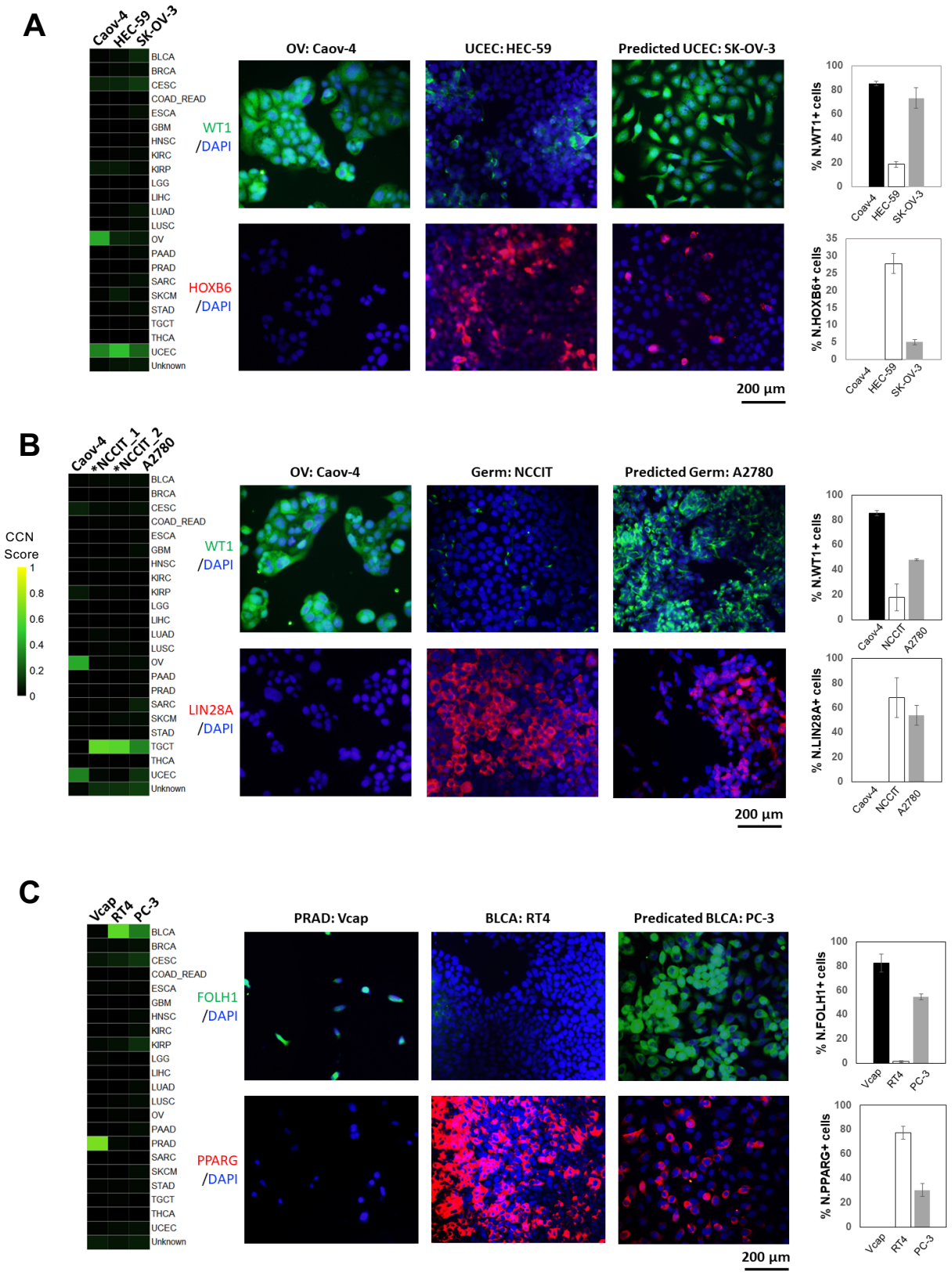

A
Caov-4
HEC-59
SK-OV-3
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
WT1
/DAPI
HOXB6
/DAPI
OV: Caov-4
UCEC: HEC-59
Predicted UCEC: SK-OV-3
200 µm
% N.WT1+ cells
% N.HOXB6+ cells
B
Caov-4
*NCCIT_1
*NCCIT_2
A2780
CCN
Score
WT1
/DAPI
LIN28A
/DAPI
OV: Caov-4
Germ: NCCIT
Predicted Germ: A2780
200 µm
% N.WT1+ cells
% N.LIN28A+ cells
C
Vcap
RT4
PC-3
FOLH1
/DAPI
PPARG
/DAPI
PRAD: Vcap
BLCA: RT4
Predicated BLCA: PC-3
200 µm
% N.FOLH1+ cells
% N.PPARG+ cells

Figure 4

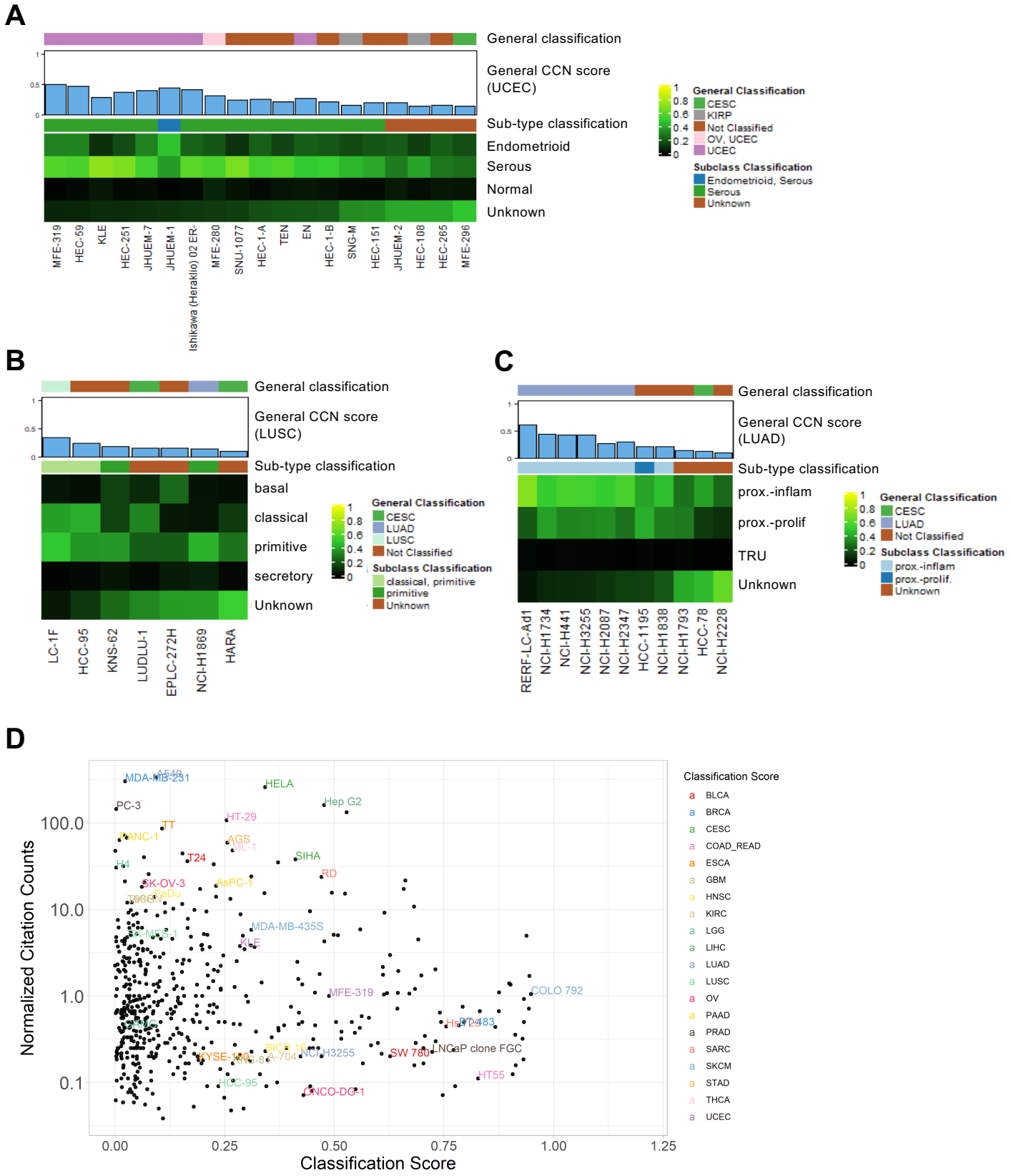

A
General classification
General CCN score (UCEC)
Sub-type classification
Endometrioid
Serous
Normal
Unknown
MFE-319
HEC-59
KLE
HEC-251
JHUEM-7
JHUEM-1
Ishikawa (Heraklio) 02 ER-
MFE-280
SNU-1077
HEC-1-A
TEN
EN
HEC-1-B
SNG-M
HEC-151
JHUEM-2
HEC-108
HEC-265
MFE-296
General Classification
CESC
KIRP
Not Classified
OV, UCEC
UCEC
Subclass Classification
Endometrioid, Serous
Serous
Unknown
B
General classification
General CCN score (LUSC)
Sub-type classification
basal
classical
primitive
secretory
Unknown
LC-1F
HCC-95
KNS-62
LUDLU-1
EPLC-272H
NCI-H1869
HARA
General Classification
CESC
LUAD
LUSC
Not Classified
Subclass Classification
classical, primitive
primitive
Unknown
C
General classification
General CCN score (LUAD)
Sub-type classification
prox.-inflam
prox.-prolif
TRU
Unknown
RERF-LC-Ad1
NCI-H1734
NCI-H441
NCI-H3255
NCI-H2087
NCI-H2347
HCC-1195
NCI-H1838
NCI-H1793
HCC-78
NCI-H2228
General Classification
CESC
LUAD
Not Classified
Subclass Classification
prox.-inflam
prox.-prolif.
Unknown
D
Normalized Citation Counts
Classification Score
MDA-MB-231
A549
HELA
Hep G2
PC-3
HT-29
TT
PANC-1
AGS
T24
SIHA
H4
RD
SK-OV-3
AsPC-1
FaDu
MDA-MB-435S
KLE
MFE-319
COLO 792
RT-4
LNCaP clone FGC
NCI-H3255
SW 780
HT55
HCC-95
ONCO-DG-1
Classification Score
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
THCA
UCEC

Figure 5

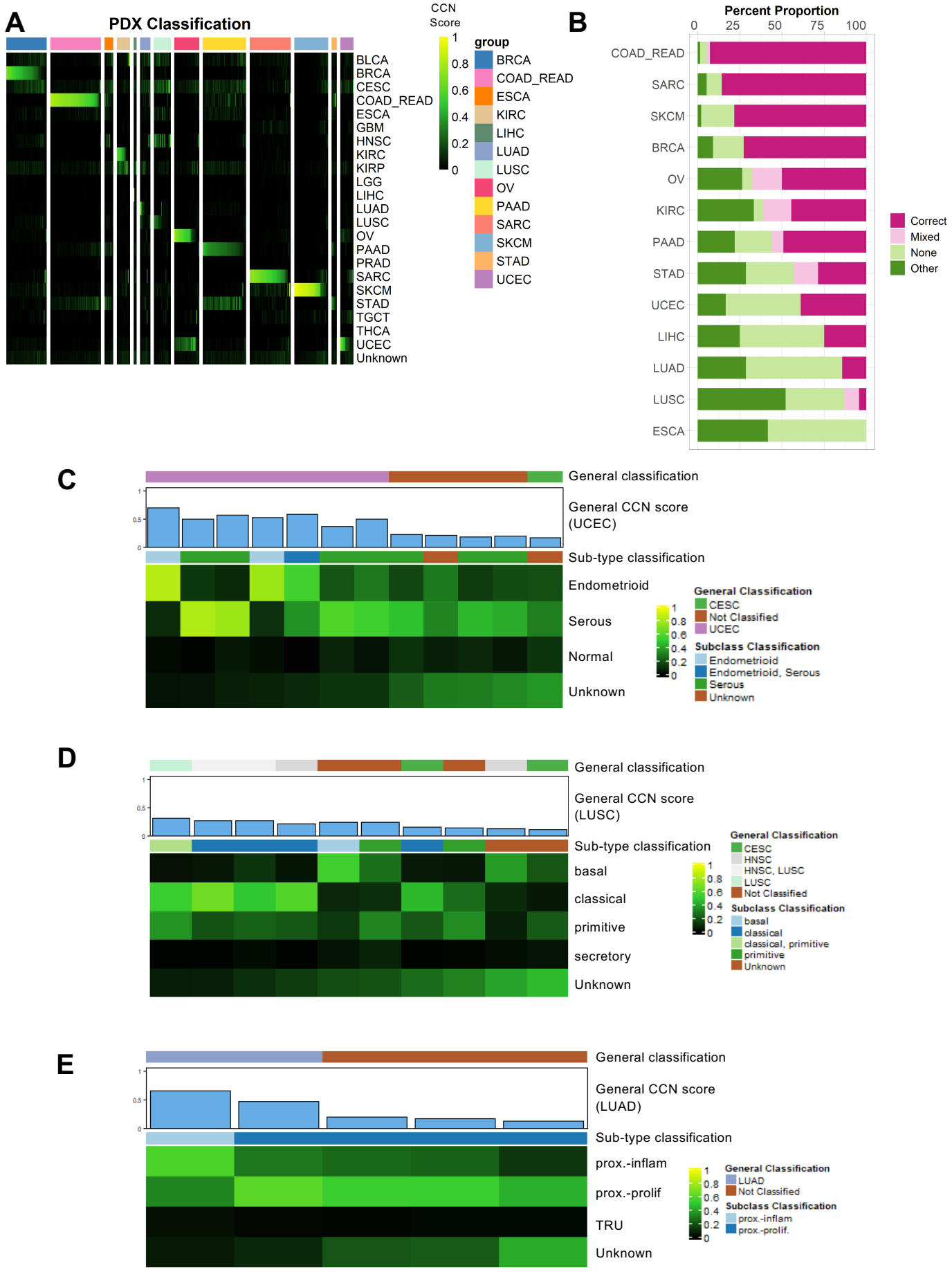

A
PDX Classification
CCN Score
1
0.8
0.6
0.4
0.2
0
group
BRCA
COAD_READ
ESCA
KIRC
LIHC
LUAD
LUSC
OV
PAAD
SARC
SKCM
STAD
UCEC
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
B
Percent Proportion
0
25
50
75
100
COAD_READ
SARC
SKCM
BRCA
OV
KIRC
PAAD
STAD
UCEC
LIHC
LUAD
LUSC
ESCA
Correct
Mixed
None
Other
C
General classification
General CCN score (UCEC)
Sub-type classification
Endometrioid
Serous
Normal
Unknown
General Classification
CESC
Not Classified
UCEC
Subclass Classification
Endometrioid
Endometrioid, Serous
Serous
Unknown
D
General classification
General CCN score (LUSC)
Sub-type classification
basal
classical
primitive
secretory
Unknown
General Classification
CESC
HNSC
HNSC, LUSC
LUSC
Not Classified
Subclass Classification
basal
classical
classical, primitive
primitive
Unknown
E
General classification
General CCN score (LUAD)
Sub-type classification
prox.-inflam
prox.-prolif
TRU
Unknown
General Classification
LUAD
Not Classified
Subclass Classification
prox.-inflam
prox.-prolif.

Figure 6

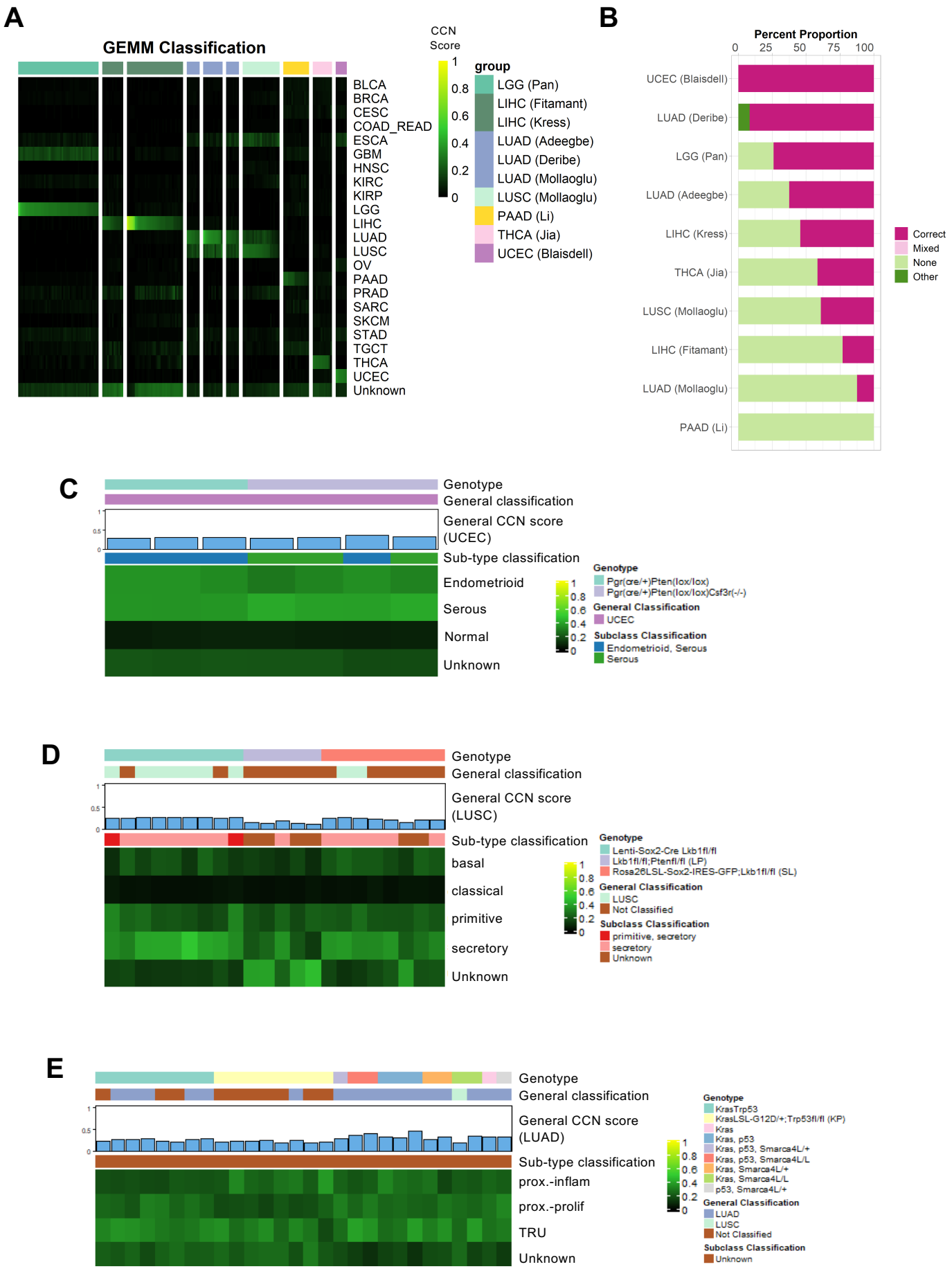

A
GEMM Classification
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
CCN Score
1
0.8
0.6
0.4
0.2
0
group
LGG (Pan)
LIHC (Fitamant)
LIHC (Kress)
LUAD (Adeegbe)
LUAD (Deribe)
LUAD (Mollaoglu)
LUSC (Mollaoglu)
PAAD (Li)
THCA (Jia)
UCEC (Blaisdell)
B
Percent Proportion
0
25
50
75
100
UCEC (Blaisdell)
LUAD (Deribe)
LGG (Pan)
LUAD (Adeegbe)
LIHC (Kress)
THCA (Jia)
LUSC (Mollaoglu)
LIHC (Fitamant)
LUAD (Mollaoglu)
PAAD (Li)
Correct
Mixed
None
Other
C
Genotype
General classification
General CCN score (UCEC)
Sub-type classification
Endometrioid
Serous
Normal
Unknown
Genotype
Pgr(cre/+)Pten(lox/lox)
Pgr(cre/+)Pten(lox/lox)Csf3r(-/-)
General Classification
UCEC
Subclass Classification
Endometrioid, Serous
Serous
D
Genotype
General classification
General CCN score (LUSC)
Sub-type 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
primitive, secretory
secretory
Unknown
E
Genotype
General classification
General CCN score (LUAD)
Sub-type classification
prox.-inflam
prox.-prolif
TRU
Unknown
Genotype
KrasTrp53
KrasLSL-G12D/+;Trp53fl/fl (KP)
Kras
Kras, p53
Kras, p53, Smarca4L/+
Kras, p53, Smarca4L/L
Kras, Smarca4L/+
Kras, Smarca4L/L
p53, Smarca4L/+
General Classification
LUAD
LUSC
Not Classified
Subclass Classification
Unknown

Figure 7

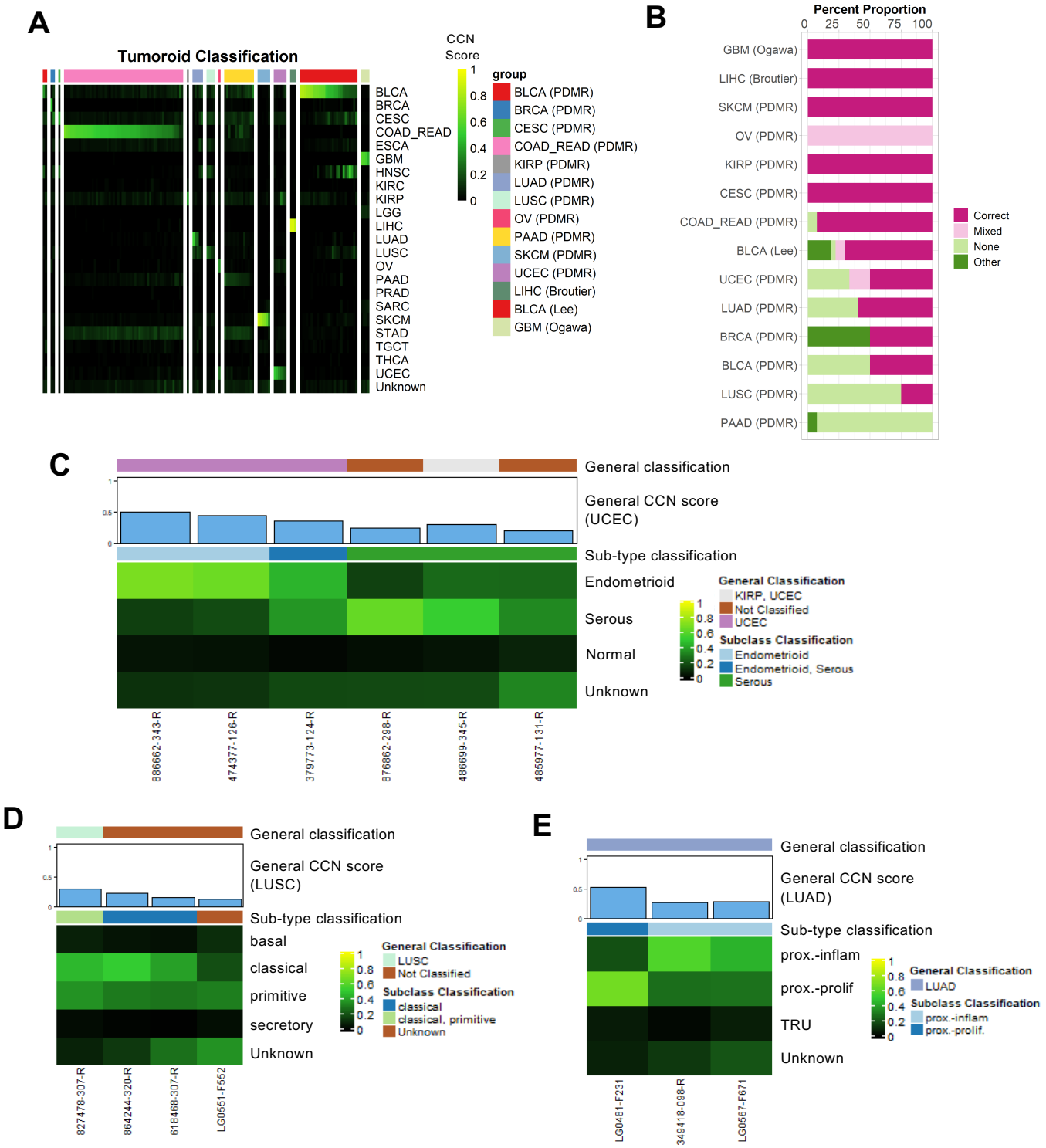
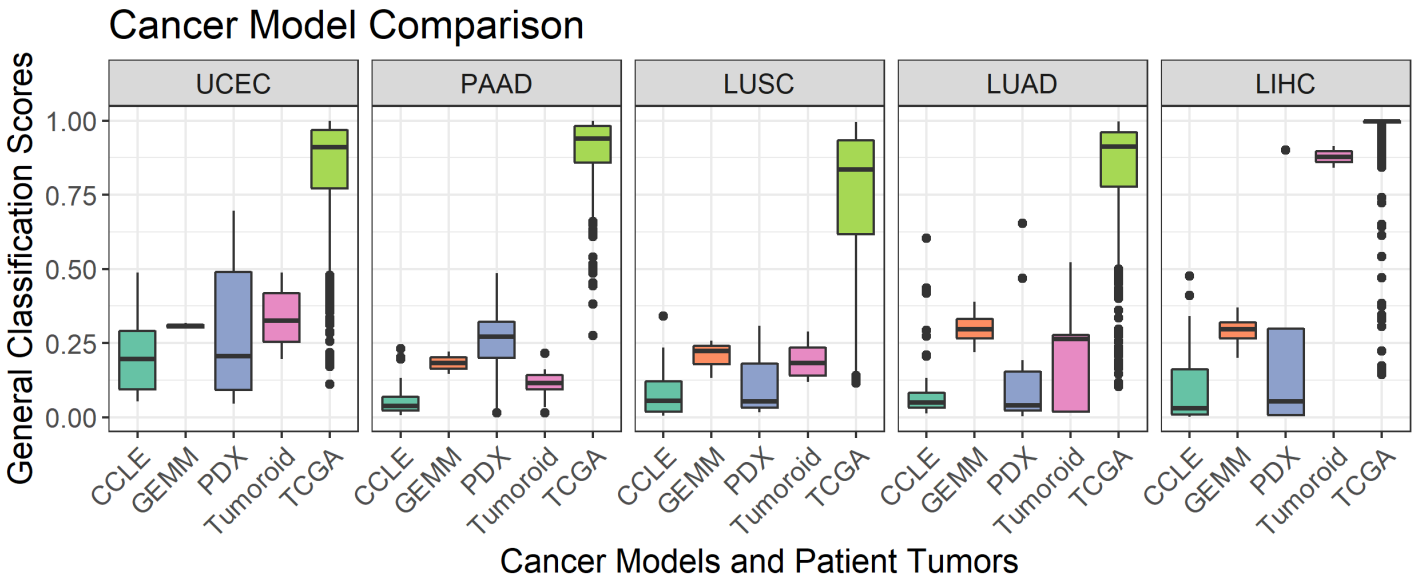
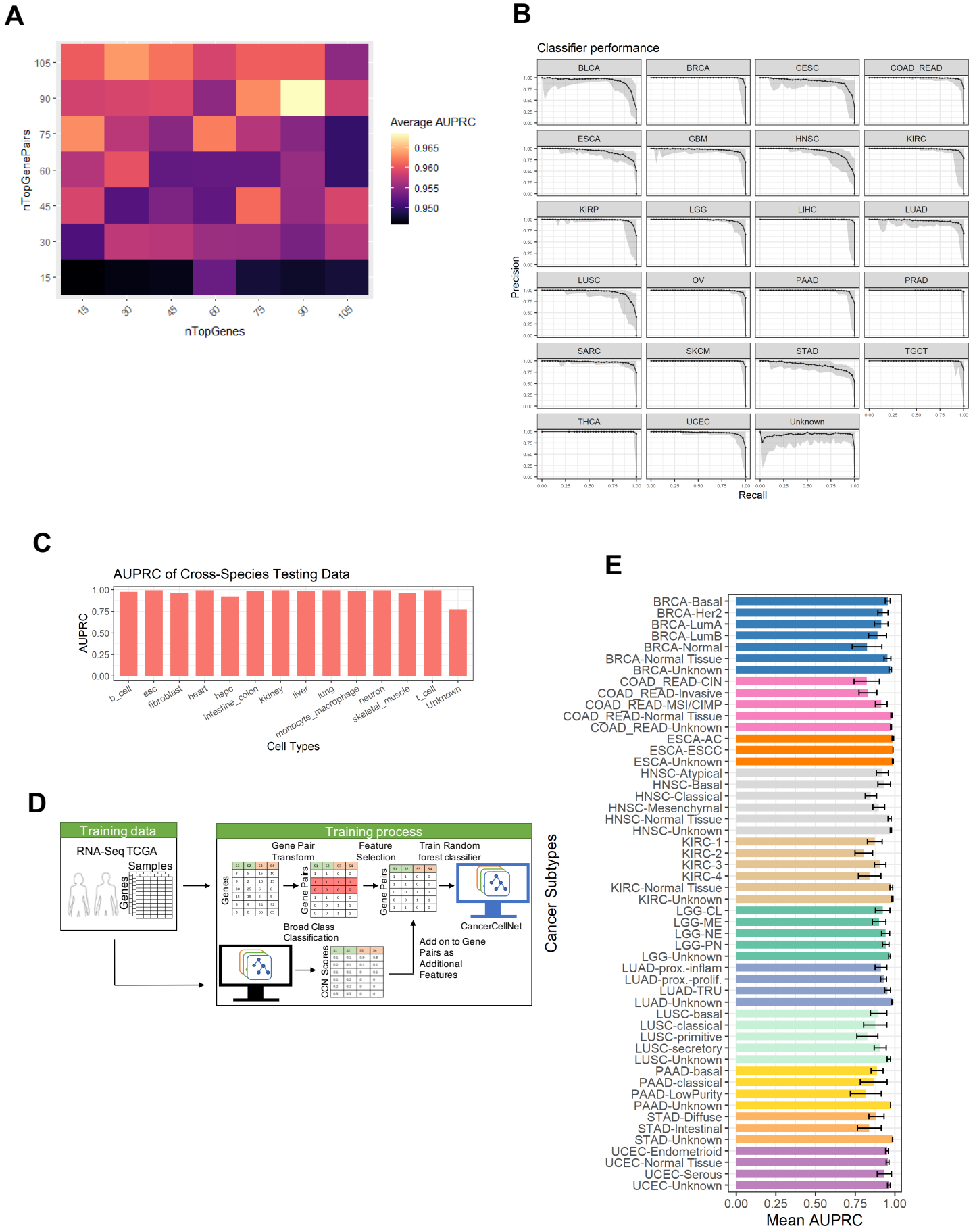

A
Tumoroid Classification
CCN Score
group
BLCA (PDMR)
BRCA (PDMR)
CESC (PDMR)
COAD_READ (PDMR)
KIRP (PDMR)
LUAD (PDMR)
LUSC (PDMR)
OV (PDMR)
PAAD (PDMR)
SKCM (PDMR)
UCEC (PDMR)
LIHC (Broutier)
BLCA (Lee)
GBM (Ogawa)
BLCA
BRCA
CESC
COAD_READ
ESCA
GBM
HNSC
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
TGCT
THCA
UCEC
Unknown
B
Percent Proportion
0 25 50 75 100
Correct
Mixed
None
Other
C
General classification
General CCN score (UCEC)
Sub-type classification
Endometrioid
Serous
Normal
Unknown
General Classification
KIRP, UCEC
Not Classified
UCEC
Subclass Classification
Endometrioid
Endometrioid, Serous
Serous
886662-343-R
474377-126-R
379773-124-R
876862-298-R
486699-345-R
485977-131-R
D
General classification
General CCN score (LUSC)
Sub-type classification
basal
classical
primitive
secretory
Unknown
General Classification
LUSC
Not Classified
Subclass Classification
classical
classical, primitive
Unknown
827478-307-R
864244-320-R
618468-307-R
LG0551-F552
E
General classification
General CCN score (LUAD)
Sub-type classification
prox.-inflam
prox.-prolif
TRU
Unknown
General Classification
LUAD
Subclass Classification
prox.-inflam
prox.-prolif.
LG0481-F231
349418-098-R
LG0567-F671

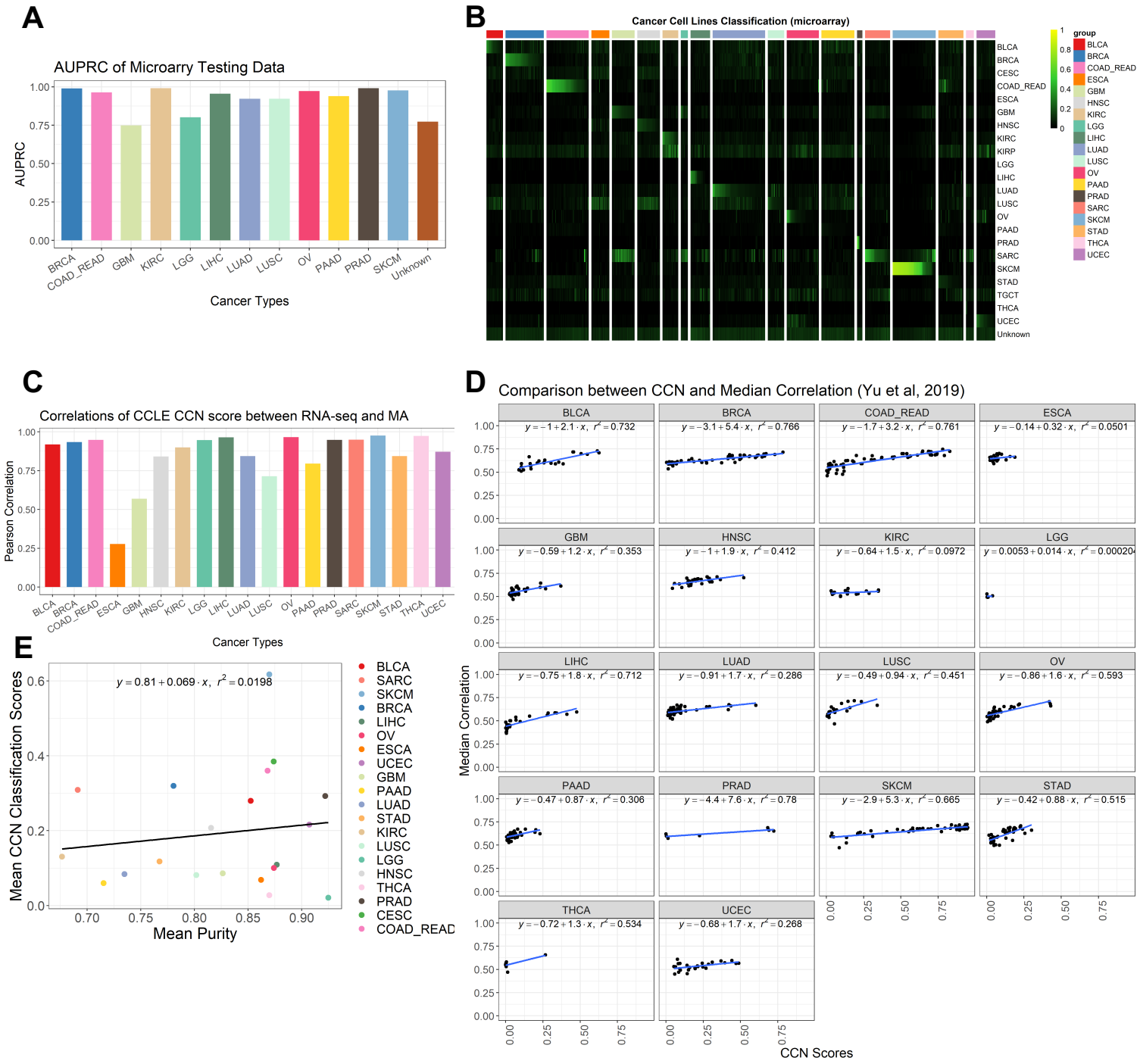
Figure 8


Cancer Model Comparison
UCEC
PAAD
LUSC
LUAD
LIHC
General Classification Scores
1.00
0.75
0.50
0.25
0.00
CCLE
GEMM
PDX
Tumoroid
TCGA
Cancer Models and Patient Tumors

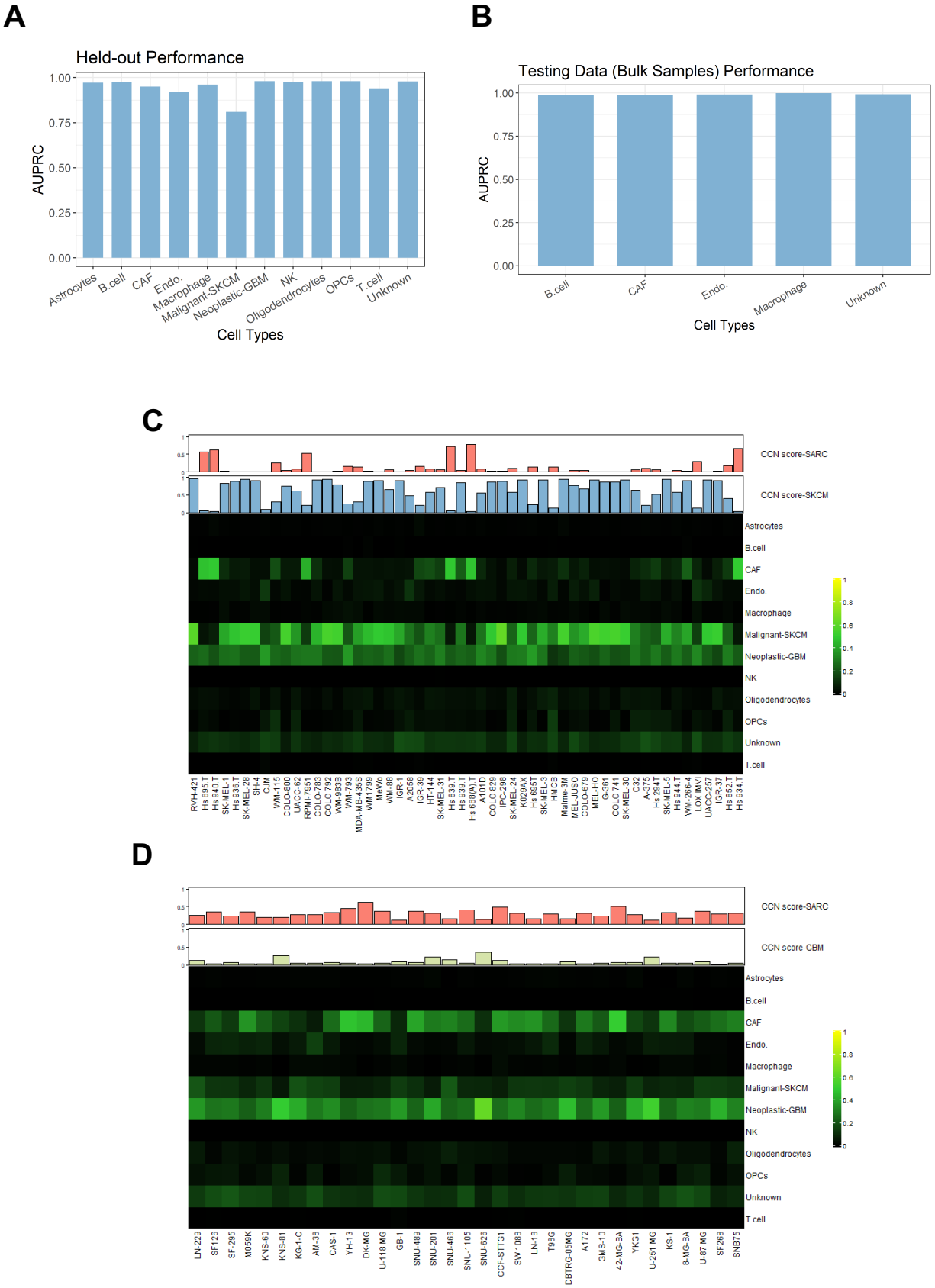
Supplemental Figure 1



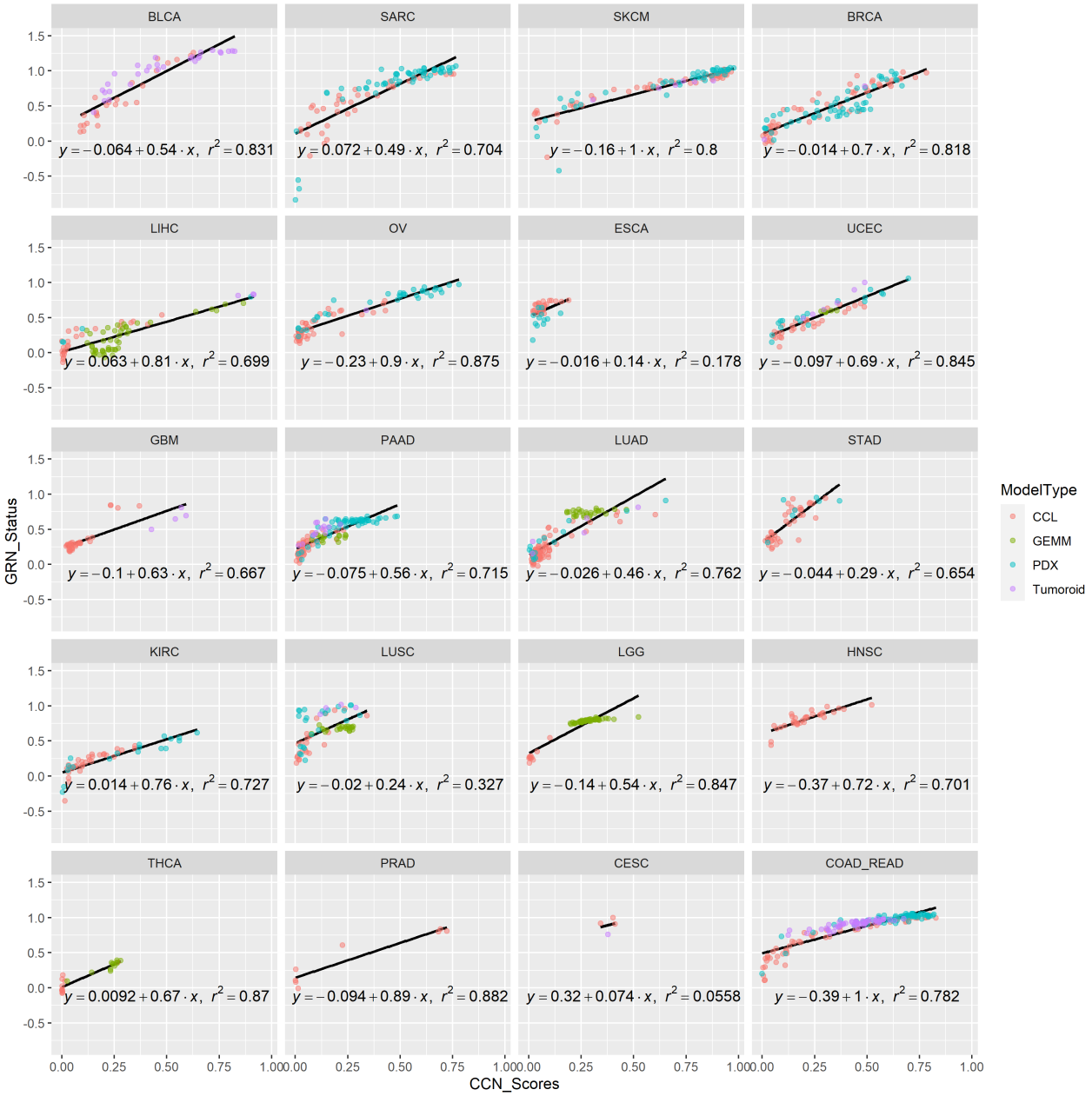
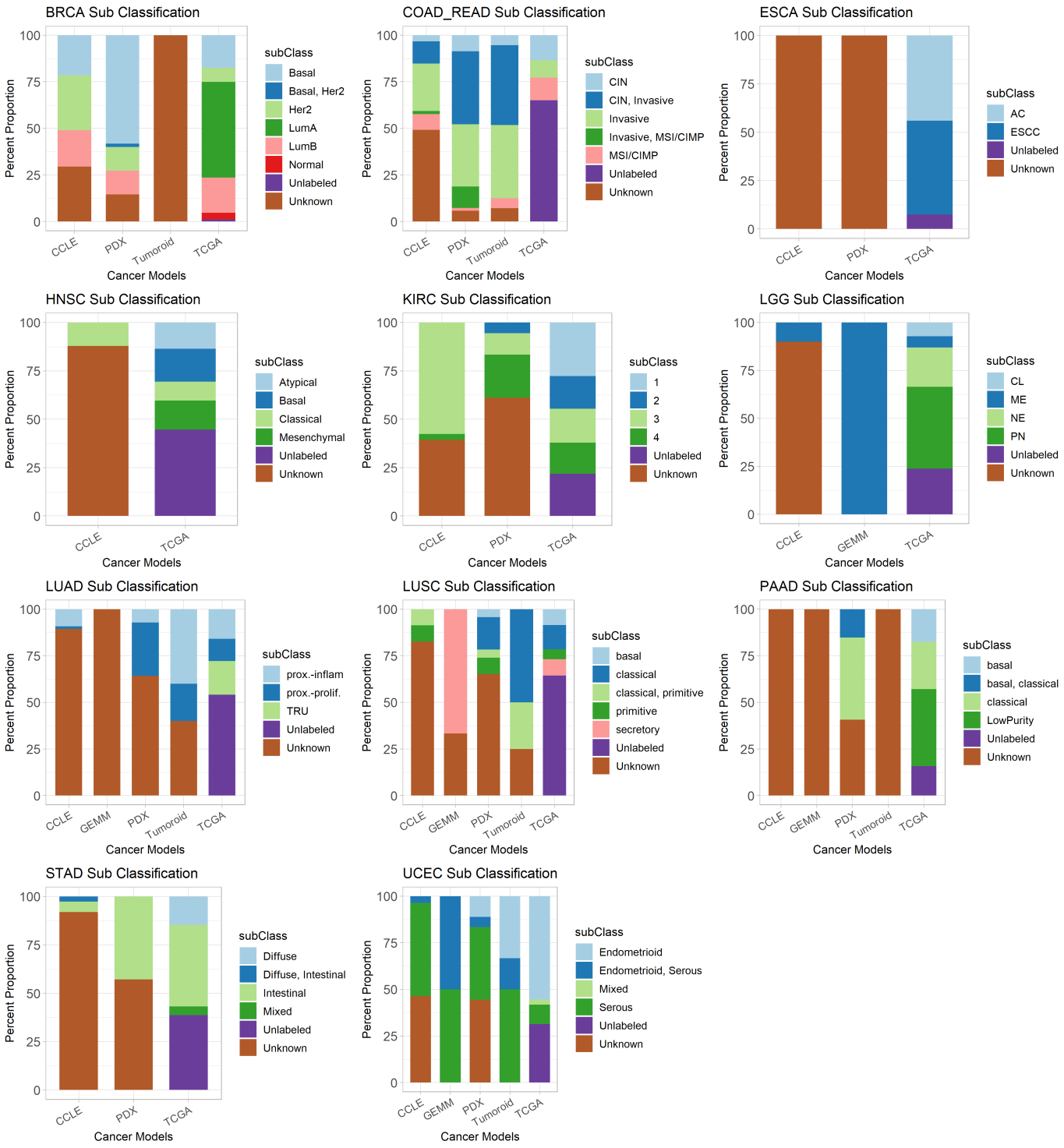
Supplemental Figure 2



Supplemental Figure 3



Supplemental Figure 4



Supplemental Figure 5