A novel machine learning algorithm picks proteome signature to specifically identify cancer exosomes

Bingrui Li[1], Fernanda G. Kugeratski[1], Raghu Kalluri[1, 2, 3]

[1]*Department of Cancer Biology, University of Texas MD Anderson Cancer Center, Houston, TX*

[2]*Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX*

[3]*Department of Bioengineering, Rice University, Houston, TX*

Corresponding author:

Raghu Kalluri, MD, PhD

Email: rkalluri@mdanderson.org

Abstract

Non-invasive early cancer diagnosis remains challenging due to the low sensitivity and specificity of current diagnostic approaches. Exosomes are membrane-bound nanovesicles secreted by all cells that contain DNA, RNA, and proteins that are representative of the parent cells. This property, along with the abundance of exosomes in biological fluids makes them compelling candidates as biomarkers. However, a rapid and flexible exosome-based diagnostic method to distinguish human cancers across cancer types in diverse biological fluids is yet to be defined. Here, we describe a novel machine learning-based computational method to distinguish cancers using a panel of proteins associated with exosomes. Employing datasets of exosome proteins from human cell lines, tissue, plasma, serum and urine samples from a variety of cancers, we identify Clathrin Heavy Chain (CLTC), Ezrin, (EZR), Talin-1 (TLN1), Adenylyl cyclase-associated protein 1 (CAP1) and Moesin (MSN) as highly abundant universal biomarkers for exosomes and define three panels of pan-cancer exosome proteins that distinguish cancer exosomes from other exosomes and aid in classifying cancer subtypes employing random forest models. All the models using proteins from plasma, serum, or urine-derived exosomes yield AUROC scores higher than 0.91 and demonstrate superior performance compared to Support Vector Machine, K Nearest Neighbor Classifier and Gaussian Naive Bayes. This study provides a reliable protein biomarker signature associated with cancer exosomes with scalable machine learning capability for a sensitive and specific non-invasive method of cancer diagnosis.

36 **Introduction**

37

38 Tissue biopsies have traditionally been a definitive way to diagnose and stage cancer,
39 however, a biopsy may not be easily accessible for many tumors such as those in the pancreas,
40 lung and brain[1-5]. In addition, the small amount of biopsied tissue does not represent the entire
41 heterogeneous pathological profile of the tumor[6]. In recent years, liquid biopsy has emerged as a
42 plausible diagnostic and monitoring approach with the capability to detect tumor biomarkers in
43 more accessible biological fluids such as plasma, serum and urine[7]. Detectable tumor biomarkers
44 can include circulating tumor DNA (ctDNA), circulating tumor cells (CTCs) and exosomes[8].

45 Exosomes are extracellular vesicles of endosomal origin that are between ~40-180 nm in
46 diameter and have been shown to mediate intercellular communication in health and disease[8-10].
47 They can contain a variety of biomolecules including DNA, RNA, proteins, lipids, metabolites
48 and other materials representative of the parent cell[8]. Exosomes are present at high
49 concentrations in biological fluids, which is a potential advantage as a biomarker[11, 12]. Exosomal
50 mRNAs and miRNAs have been investigated intensively as diagnostic biomarkers, and mounting
51 evidence suggests that exosomal proteins circulating in biological fluids could be used for cancer
52 diagnosis and monitoring cancer progression[7, 13]. Challenges that remain include standardization
53 of methods for consistent exosome isolation from various tissues, identification of biomarkers
54 that distinguish cancer and normal exosomes across different cancer types, and the identification
55 of biomarkers that are unique to specific biological fluids (e.g., plasma, serum, urine, etc.).

56 In addition to accurately identifying exosomal protein biomarkers, it is also challenging
57 to successfully utilize them to diagnose cancers due to intratumor and interpatient
58 heterogeneity[14]. Conventional diagnostic approaches predominantly rely on a single biomarker,
59 which is often not specific or sensitive[15, 16]. However, recent advances in machine learning

60 algorithms connected to artificial intelligence provide an opportunity to construct a classifier that
61 identifies a panel of exosomal protein biomarkers that would possess a more comprehensive
62 ability to reflect the complex disease status of different patients and distinguish cancer samples
63 from normal samples with significantly improved sensitivity and specificity.
64

65 **Results**
66
67 **Unbiased proteomics analysis of exosomes identifies 18 abundant plasma membrane**
68 **protein markers for various human cell lines**

69 To identify universal exosomal protein biomarkers for differentiating cancer from non-cancer
70 exosomes, we analyzed protein abundance data from 228 cancer and 57 control cell line-derived
71 exosomes, representing various cancer types (**Figure 1A; Supplementary Table S1**). Because
72 studies employ distinct isolation and mass spectrometry quantification techniques, the number of
73 identified proteins is different among the studies. To overcome the bias caused by such technical
74 factors, we examined the proteins common to all studies and identified 1124 overlapping
75 proteins (**Figure 2A**). To determine the heterogeneity among cancer and control cell line-derived
76 exosomes, we performed principal component analysis (PCA) using these 1124 proteins in 285
77 different studies with cancer and control cell line-derived exosomes (**Figure 2B**). The PCA
78 indicated that the exosomes derived from cancer and control cell lines are heterogeneous and
79 show significant variation in protein expression across cell lines.

80 We next investigated the frequency of the proteins detected in the exosomes from all cell
81 lines, only the cancer cell lines and just the control cell lines. Commonly used exosome
82 biomarkers (e.g., CD9, tetraspanins) were examined first, and from the 12 traditional exosome
83 markers[17-19], only eight proteins were detected among the 1124 overlapping proteins from all cell

84 lines. Six of eight proteins were detected in at least 90% of all samples, with CD9 and HSPA4
85 detected with the least frequency (**Figure 2C**). In addition, the frequency of FLOT1, FLOT2 and
86 TSG101 proteins was higher in cancer cell line-derived exosomes when compared to control cell
87 line-derived exosomes (**Figure 2C**).

88 To identify biomarkers detectable at a high frequency in all cancer and control cell line-
89 derived exosomes, we searched for the proteins that were detected in ≥ 90% of all samples
90 (**Supplementary Table S2**). We annotated these proteins using IPA and found that 78.0% of
91 these proteins localize to the cytoplasm and 13.5% of proteins are associated with the plasma
92 membrane (**Figure 2D**). Gene Ontology (GO) analysis revealed the enrichment of proteins from
93 pathways related to vesicle-mediated transport, secretory vesicles, exocytosis, endocytosis, and
94 other exosome-related pathways (**Figure 2E**). To further explore the utility of the proposed
95 biomarkers, we examined the proteins located on the plasma membrane that met the threshold of
96 detection in ≥ 90% of all samples (**Figure 2F**). Clathrin Heavy Chain (CLTC) was ranked as the
97 top plasma membrane protein detected in 99.6% of all samples and 100% of control samples
98 (**Figure 2F**). In addition, the scaffolding protein, Syntenin-1(SDCBP) was detected at a high
99 frequency of 97.9% of all samples, corroborating previous findings[20]. Next, we sought to identify
100 unique markers that can identify cancer cell-derived exosomes (cancer exosomes) by filtering out
101 the proteins present in ≤ 10% of 57 control cell line-derived exosomes. Interestingly, Ataxin 2
102 Like (ATXN2L), which has been reported to promote cancer cell invasiveness and resistance to
103 chemotherapy, was uniquely detected in the cancer cell-derived exosomes (**Figure 3A**)[21]. In
104 total, we identified a set of 18 exosome protein markers that are present at a high abundance in
105 all the cancer exosomes examined (**Figure 2F**).
106

107 **Comparison of exosomal proteins derived from cell lines and tissues identified five**
108 **universal plasma membrane protein markers**

109 We next sought to investigate common biomarkers for the tissue-derived exosomes across cancer
110 types. We calculated the detection frequency of commonly used exosome markers in the 157
111 samples (101 cancers; 56 controls). Two established exosome markers, CD63 and TSG101, were
112 only detected in 33.1% and 45.9% of all samples, respectively (**Figure 3B**). To identify high-
113 frequency biomarkers for exosomes derived from both cell lines and tissues, we examined the
114 overlapping proteins that met a threshold of ≥ 90% in all samples for exosomes derived from cell
115 lines and tissues and found 31 common proteins (**Figure 3C**). Among the 31 proteins, there were
116 five proteins that were detected in over 90% of all cell line and tissue-derived exosomes (**Figure**
117 **3D**). These include Clathrin Heavy Chain (CLTC), Ezrin, (EZR), Talin-1 (TLN1), Adenylyl
118 cyclase-associated protein 1 (CAP1) and Moesin (MSN).

119

120 **An exosome proteome signature of 18 proteins can differentiate cancer exosomes from non-**
121 **cancer exosomes across multiple cancer types**

122 Plasma or serum is the most readily accessible source for non-invasive biopsies. We next sought
123 to identify if exosome proteins in plasma and serum could differentiate cancer exosomes from
124 non-cancer exosomes across multiple cancer types. We pooled exosome proteomics data derived
125 from plasma or serum for 205 cancer and 51 control samples from five different studies[22-26],
126 which included breast cancer, colorectal cancer, glioblastoma, lung carcinoma, liver cancer,
127 neuroblastoma, and pancreatic cancer (**Supplementary Table S1)**. To account for differences in
128 the methodologies of these studies, we first reviewed the proteins and identified 46 proteins that
129 were detected in all the studies (**Figure 4A**). We then examined their abundances in the 205
130 cancer and 51 control samples (**Figure 4B**). Although we could detect differences between the

131 exosomes derived from cancer cell and non-cancer cells, it was difficult to classify them robustly
132 based on PCA (**Supplementary Figure S1A**). Therefore, we sought to employ the advanced
133 machine learning algorithm to differentiate cancer exosomes from non-cancer exosomes. We
134 first calculated the mutual information score (MI) for each protein and trained the random forest
135 classifier using the different numbers of top proteins according to their MI score to determine the
136 best set of proteins we should include in the classifier. We found that the model performed best
137 using 18 proteins, where the performance was evaluated by the area under the curve of the
138 receiver operating characteristic curve (AUROC) (**Figure 4C**). Several key cancer-associated
139 proteins were included among these 18 proteins. For example, apolipoprotein C1 (APOC1),
140 which was ranked as the top protein, is significantly decreased in samples of cancer patients
141 (**Supplementary Figure S1B**) and has been previously reported to be down-regulated in the
142 non-small lung cancer, colorectal cancer, papillary thyroid carcinoma and pediatric
143 nephroblastoma[27-30].
144 Employing all 18 proteins with 5-fold cross-validation, we constructed a random forest
145 classifier[31] to distinguish cancer and control samples and compared it with multiple popular
146 machine learning models, including Support Vector Machine (SVM)[32], K Nearest Neighbor
147 Classifier (K-NN)[33] and Gaussian Naive Bayes[34]. Our random forest classifier demonstrated the
148 highest AUROC (**Figure 4D**). More specifically, it yielded an AUROC of 0.96, with an accuracy
149 of 0.92, a precision of 0.94, and a recall of 0.96 (**Figure 4E& F**). When applied to the
150 independent test set, the model yielded an AUROC of 0.99, an accuracy of 0.95, a precision of
151 0.96 and a recall of 0.98 (**Figure 4G& H**). Importantly, only one sample was misclassified in the
152 independent test set, and 51 cancer samples were correctly classified. Taken together, our results

153 showed the advantage and clinical potential of applying the random forest classifier model to
154 plasma or serum exosome protein based liquid biopsy for cancer diagnosis.
155
156 **Five plasma/serum exosomal proteins can reliably differentiate five common cancer types**
157 We next sought to further enhance the clinical utility of the exosomes for differentiating between
158 cancer types. We analyzed proteomics data from plasma or serum derived exosomes from
159 patients with five common cancer types, including breast cancer, colorectal cancer, glioma, lung
160 cancer and pancreatic cancer. Initial PCA revealed differences in exosome levels among cancer
161 patients but failed to distinguish the five cancer types (**Figure 5A**). We determined the crucial
162 features for cancer-type classification by computing mutual information scores for 46 common
163 proteins and built a random forest classifier to determine the optimal number of features to
164 include in the final classifier. We ultimately selected a set of five proteins based on AUROC
165 scores (**Figure 5B**). We then increased the independent testing data size by utilizing 40% of the
166 total samples and used the remaining 60% as the training set to minimize overfitting issues.
167 Employing the five proteins with 5-fold cross-validation to train the random forest classifier, the
168 model achieved a very high accuracy of 0.99 (**Figure 5C**), and when applied to the independent
169 test set, the model consistently yielded a high accuracy of 0.94 (**Figure 5D**). The abundance of
170 the five protein features varied across the five cancer types, reflecting the potential roles of these
171 proteins in specific cancers (**Figure 5E**). For example, Histidine-rich glycoprotein (HRG),
172 which was abundant in the colorectal cancer plasma-derived exosomes, has been reported to
173 promote the tumor migration of colorectal cancer patients[35]. Overall, our results demonstrated
174 that this exosome protein-based classification model can reliably differentiate between cancer
175 types and further enhance the diagnostic value of our approach.

176 **A urinary exosome proteome signature consisting of 17 proteins detects cancer exosomes**
177 **across multiple cancer types**
178 Urine is emerging as a superior non-invasive marker for urologic cancers, as its composition
179 directly reflects the physiological changes in the urogenital system[36]. To test the use of urinary
180 exosome proteins in cancer diagnosis, we collected data from 261 cancer patient samples and
181 124 control samples from four studies, including bladder cancer, prostate cancer, and renal
182 cancer, lung cancer, cervical cancer, colorectal cancer, esophageal and gastric cancer
183 (**Supplementary Table S1**)[36-39]. Upon examination of proteins common across all four studies,
184 229 proteins were identified (**Figure 6A**). PCA revealed the variance between samples but failed
185 to differentiate between cancer and control samples (**Figure 6B**). As described earlier, we next
186 employed the random forest classifier method to differentiate cancer and control samples based
187 on their exosomal proteomic profiles. To reduce the feature space and select the most relevant
188 features, we utilized the mutual information score to rank the 229 protein features and then
189 trained the random forest classifier using varying numbers of the top-ranking proteins (**Figure**
190 **6C**). Based on the AUROC scores of including the different number of features, we selected 17
191 features that resulted in the highest AUROC score (**Figure 6C**). A majority of these 17 proteins
192 displayed significant variations in abundance between cancer and control samples (**Figure 6D**).
193 By training a random forest classifier with the 17 protein features and 5-fold cross-validation, the
194 model achieved an AUROC of 0.96, an accuracy of 0.90, a precision of 0.92, and a recall of 0.93
195 (**Figure 6E& F**). When tested on an independent set, the model produced an AUROC of 0.91, an
196 accuracy of 0.82, a precision of 0.83, and a recall of 0.92 (**Figure 6G& H**). To summarize, our
197 findings indicated the promising clinical potential of using urinary exosome proteins for the
198 diagnosis of urologic cancers as well as other non-urologic cancer types.

199
200 **Discussion**
201
202 Liquid biopsy has numerous benefits in the early detection of cancer, categorizing cancer types,
203 tracking cancer progression, and monitoring response to treatment[40]. Exosomes found in
204 biological fluids can provide a forensic view of their cells of origin. Despite previous studies
205 proposing common protein biomarkers to identify exosomes, a comprehensive set of exosome
206 biomarkers derived from different biological materials has not been established, owing to
207 limitations in isolation and quantification methods[41-44]. Additionally, a reliable diagnostic tool
208 based on proteins associated with exosomes that can be applied across all cancers is yet to be
209 identified.
210 Here, we generate a comprehensive proteomics profile of exosomes derived from cell
211 lines, tissues, plasma, serum, and urine from 1083 cancer and control samples. An extensive
212 analysis of these samples showed that several widely used exosome markers, such as CD63,
213 CD81, HSP70, and HSPA8, are absent in exosomes derived from a subset of cell lines. Further,
214 FLOT1, FLOT2, TSG101, and CD63 are present at low levels (< 60%) in all exosomes derived
215 from tissues, indicating a need for the identification of additional universal markers for both cell
216 line- and tissue-derived exosomes. Our study identifies five highly abundant universal exosome
217 biomarkers-CLTC, EZR, TLN1, CAP1, and MSN- that were present in over 90% of all cell line-
218 and tissue-derived samples. Additionally, we found that ATXN2L was only present in cancer
219 exosomes and absent in non-cancer exosomes. In this regard, GPC1 which was identified in
220 many studies identified as cancer exosomes specific marker, was not identified here as many of
221 databases did not pick up this protein in analysis.
222 Here we describe a novel computational approach using the random forest classifier
223 method to define exosome protein panels that serve as effective biomarkers specifically for

224 plasma, serum or urine across cancer types. By training the random forest model and testing with
225 independent datasets, our model yields excellent scores in AUROC, sensitivity, and specificity
226 and differentiates cancer exosomes from non-cancer exosomes. We show that this approach can
227 also be used to classify, with high accuracy, five common cancer types based on their exosome
228 protein signatures.

229 A majority of the protein makers identified in this study have demonstrable biological
230 relevance in cancer. As an example, ITIH3, which was identified among the protein features for
231 the plasma or serum-based classifier and highly abundant in cancer samples, was reported to be
232 more highly expressed in the plasma of gastric cancer samples compared to the control[45] and
233 increased with tumor staging of clear cell renal cell carcinoma patients [46]. Importantly, the
234 biomarker panels identified for cell lines, tissues, plasma/serum, and urine overcome the bias
235 associated with the isolation and quantification method as well as the inter-patient variability
236 inherent to the complex process of cancer. These panels are designed to be applied to a variety of
237 cancers.

238 Collectively, our results demonstrate that exosome protein features can be utilized as
239 reliable biomarkers for the early detection of cancer, classification of cancer types, and
240 potentially for diagnosing tumors of undetermined origin. These results have the potential to
241 advance the development and standardization of innovative and optimized methods for the
242 isolation of exosomes and the implementation of routine plasma, serum- and urine-based
243 exosome screening in clinical settings.

244

Methods

Public exosome proteomics data

We collected publicly available exosome protein data from cell lines, plasma, serum, and urine from previous studies as summarized in **Supplementary Table S1**[20, 22-26, 36-39, 43, 47]. We obtained the raw spectral count and intensity data from the original paper or directly from the authors through communication. The obtained data was log normalized accordingly.

Feature selection and machine learning algorithm

We employed mutual information scores to evaluate the importance of protein features for each prediction, quantifying the amount of uncertainty reduction for one variable given the knowledge of another variable. We calculated the mutual information score for each protein with a target label (cancer/control) using SelectKBest from the scikit-learn 1.1.1 Python library. According to the sample size and characteristics of the data collected for plasma and urine-derived exosomes, we selected a customized number of best features for each prediction. To select the optimum number of proteins to include in the prediction, we ranked the protein features based on the mutual information score and built a random forest classifier to evaluate performance on including a range of number of features. The random forest model can decrease the probability of over-fitting and enhances the resilience towards outliers and input data noise. The area under the curve of the receiver operating characteristic curve (AUROC) was employed to evaluate the performance of the classifier. We also calculated accuracy, precision and recall for comprehensive evaluation. All models were evaluated using 5-fold cross-validation with stratified train-test splits that preserved the percentage of samples for the prediction target. We also tested the performance of alternate machine learning algorithms including support vector classifier, K nearest neighbor classifier and gaussian naive bayes. Overall, the random forest

classifier achieved the best performance in our analysis (**Supplementary Figure S1B**). To visualize high-dimensional datasets, kernel principal component analysis (PCA) from scikit-learn 1.1.1 Python library was employed to perform PCA and plots were generated using the scatterplot function from seaborn 0.11.2 Python library.

Gene ontology and pathway enrichment analysis

The WebGestalt 2019[48] online tool was used to perform the gene ontology and pathway enrichment analysis of the selected proteins. Pathways with FDR < 0.05 were considered significant.

Statistical analysis

All statistical analyses were conducted using R 4.2.1 software. Significance was determined by the Wilcoxon rank-sum test unless specified otherwise. Significance was concluded if the p-value was <0.05, while in pathway analysis, significance was concluded if the FDR was <0.05 after correction for multiple comparisons.

Disclosure of conflict of interest

UT MD Anderson Cancer Center and R.K. hold patents in the area of exosome biology licensed to Codiak Biosciences, Inc. UT MD Anderson Cancer Center and R.K. are stock equity holders in Codiak Biosciences, Inc. R.K. is a consultant and scientific adviser for Codiak Biosciences, Inc. The remaining authors declare no competing interests.

Acknowledgments

The EV work in the Kalluri lab is supported by MD Anderson Cancer Center, NIH R35CA263815, and NIH P40OD024628 and gifts from Fifth Generation (Love, Tito's), Lyda Hill Philanthropies, and Bosarge Family Trust.

Data availability

The public datasets used in this study were described in the Methods sections.

Contributions

B.L. and R.K. conceived and designed the study. B.L. developed the computational algorithm and performed the analysis with supervision from R.K. F.G.K. generated the proteomics data of the cell lines. B.L. and R.K. wrote the manuscript. All the authors read the manuscript and discussed the results.

References

1. Distler, M., Aust, D., Weitz, J., Pilarsky, C. & Grutzmann, R. Precursor lesions for sporadic pancreatic cancer: PanIN, IPMN, and MCN. *Biomed Res Int* **2014**, 474905 (2014).
2. Morris, J.P.t., Cano, D.A., Sekine, S., Wang, S.C. & Hebrok, M. Beta-catenin blocks Kras-dependent reprogramming of acini into pancreatic cancer precursor lesions in mice. *J Clin Invest* **120**, 508-520 (2010).
3. Suram, A. et al. Oncogene-induced telomere dysfunction enforces cellular senescence in human cancer precursor lesions. *EMBO J* **31**, 2839-2851 (2012).
4. Mullerat, J., Deroide, F., Winslet, M.C. & Perrett, C.W. Proliferation and p53 expression in anal cancer precursor lesions. *Anticancer Res* **23**, 2995-2999 (2003).
5. Hammes, L.S. et al. Up-regulation of VEGF, c-fms and COX-2 expression correlates with severity of cervical cancer precursor (CIN) lesions and invasive disease. *Gynecol Oncol* **110**, 445-451 (2008).
6. Oshi, M. et al. Urine as a Source of Liquid Biopsy for Cancer. *Cancers (Basel)* **13** (2021).
7. Hu, C. et al. Potentiality of Exosomal Proteins as Novel Cancer Biomarkers for Liquid Biopsy. *Front Immunol* **13**, 792046 (2022).
8. Kalluri, R. & LeBleu, V.S. The biology, function, and biomedical applications of exosomes. *Science* **367** (2020).
9. Johnstone, R.M., Adam, M., Hammond, J.R., Orr, L. & Turbide, C. Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes). *J Biol Chem* **262**, 9412-9420 (1987).
10. Maas, S.L.N., Breakefield, X.O. & Weaver, A.M. Extracellular Vesicles: Unique Intercellular Delivery Vehicles. *Trends Cell Biol* **27**, 172-188 (2017).
11. Colombo, M., Raposo, G. & Thery, C. Biogenesis, secretion, and intercellular interactions of exosomes and other extracellular vesicles. *Annu Rev Cell Dev Biol* **30**, 255-289 (2014).
12. Valencia, K. & Montuenga, L.M. Exosomes in Liquid Biopsy: The Nanometric World in the Pursuit of Precision Oncology. *Cancers (Basel)* **13** (2021).
13. Chen, I.H. et al. Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. *Proc Natl Acad Sci U S A* **114**, 3175-3180 (2017).
14. Muluneh, M., Shang, W. & Issadore, D. Track-etched magnetic micropores for immunomagnetic isolation of pathogens. *Adv Healthc Mater* **3**, 1078-1085 (2014).
15. Ballehaninna, U.K. & Chamberlain, R.S. Serum CA 19-9 as a Biomarker for Pancreatic Cancer-A Comprehensive Review. *Indian J Surg Oncol* **2**, 88-100 (2011).
16. Singh, S., Tang, S.J., Sreenarasimhaiah, J., Lara, L.F. & Siddiqui, A. The clinical utility and limitations of serum carbohydrate antigen (CA19-9) as a diagnostic tool for pancreatic cancer and cholangiocarcinoma. *Dig Dis Sci* **56**, 2491-2496 (2011).
17. Thery, C., Amigorena, S., Raposo, G. & Clayton, A. Isolation and characterization of exosomes from cell culture supernatants and biological fluids. *Curr Protoc Cell Biol* **Chapter 3**, Unit 3 22 (2006).
18. Kowal, J. et al. Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes. *Proc Natl Acad Sci U S A* **113**, E968-977 (2016).

19. Jeppesen, D.K. et al. Reassessment of Exosome Composition. *Cell* **177**, 428-445 e418 (2019).
20. Kugeratski, F.G. et al. Quantitative proteomics identifies the core proteome of exosomes with syntenin-1 as the highest abundant protein and a putative universal biomarker. *Nat Cell Biol* **23**, 631-641 (2021).
21. Lin, L. et al. ATXN2L upregulated by epidermal growth factor promotes gastric cancer cell invasiveness and oxaliplatin resistance. *Cell Death Dis* **10**, 173 (2019).
22. Hoshino, A. et al. Extracellular Vesicle and Particle Biomarkers Define Multiple Human Cancers. *Cell* **182**, 1044-1061 e1018 (2020).
23. Vykoukal, J. et al. Plasma-derived extracellular vesicle proteins as a source of biomarkers for lung adenocarcinoma. *Oncotarget* **8**, 95466-95480 (2017).
24. Li, S. et al. Proteomic Landscape of Exosomes Reveals the Functional Contributions of CD151 in Triple-Negative Breast Cancer. *Mol Cell Proteomics* **20**, 100121 (2021).
25. Lin, K. et al. Comprehensive proteomic profiling of serum extracellular vesicles in patients with colorectal liver metastases identifies a signature for non-invasive risk stratification and early-response evaluation. *Mol Cancer* **21**, 91 (2022).
26. Hallal, S. et al. A Comprehensive Proteomic SWATH-MS Workflow for Profiling Blood Extracellular Vesicles: A New Avenue for Glioma Tumour Surveillance. *Int J Mol Sci* **21** (2020).
27. Zhang, Q., Wang, J., Dong, R., Yang, S. & Zheng, S. Identification of novel serum biomarkers in child nephroblastoma using proteomics technology. *Mol Biol Rep* **38**, 631-638 (2011).
28. Jin, Y. et al. Identification a novel clinical biomarker in early diagnosis of human non-small cell lung cancer. *Glycoconj J* **36**, 57-68 (2019).
29. Engwegen, J.Y. et al. Detection of Colorectal Cancer by Serum and Tissue Protein Profiling: A Prospective Study in a Population at Risk. *Biomark Insights* **3**, 375-385 (2008).
30. Fan, Y. et al. Discovery and identification of potential biomarkers of papillary thyroid carcinoma. *Mol Cancer* **8**, 79 (2009).
31. Ho, T.K. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition* **1**, 278-282 vol.271 (1995).
32. Cortes, C.a.V., Vladimir Support-vector networks. *Machine learning* **20**, 273-297 (1995).
33. Fix, E.a.H., Joseph l. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique* **57**, 238-247 (1989).
34. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).
35. Bogoevska, V. et al. HRG/HER2/HER3 signaling promotes AhR-mediated Memo-1 expression and migration in colorectal cancer. *Oncogene* **36**, 2394-2404 (2017).
36. Dhondt, B. et al. Unravelling the proteomic landscape of extracellular vesicles in prostate cancer by density-based fractionation of urine. *J Extracell Vesicles* **9**, 1736935 (2020).
37. Zhang, C. et al. Urine Proteome Profiling Predicts Lung Cancer from Control Cases and Other Tumors. *EBioMedicine* **30**, 120-128 (2018).
38. Overbye, A. et al. Identification of prostate cancer biomarkers in urinary exosomes. *Oncotarget* **6**, 30357-30376 (2015).

39. Suh, J. et al. Next-generation Proteomics-Based Discovery, Verification, and Validation of Urine Biomarkers for Bladder Cancer Diagnosis. *Cancer Res Treat* **54**, 882-893 (2022).
40. Nikanjam, M., Kato, S. & Kurzrock, R. Liquid biopsy: current technology and clinical applications. *J Hematol Oncol* **15**, 131 (2022).
41. Castillo, J. et al. Surfaceome profiling enables isolation of cancer-specific exosomal cargo in liquid biopsies from pancreatic cancer patients. *Ann Oncol* **29**, 223-229 (2018).
42. Gangoda, L. et al. Proteomic Profiling of Exosomes Secreted by Breast Cancer Cells with Varying Metastatic Potential. *Proteomics* **17** (2017).
43. Hurwitz, S.N. et al. Proteomic profiling of NCI-60 extracellular vesicles uncovers common protein cargo and cancer type-specific biomarkers. *Oncotarget* **7**, 86999-87015 (2016).
44. Ji, H. et al. Proteome profiling of exosomes derived from human primary and metastatic colorectal cancer cells reveal differential expression of key metastatic factors and signal transduction components. *Proteomics* **13**, 1672-1686 (2013).
45. Chong, P.K. et al. ITIH3 is a potential biomarker for early detection of gastric cancer. *J Proteome Res* **9**, 3671-3679 (2010).
46. Chang, Q.H. et al. Pan-cancer analysis identifies ITIH1 as a novel prognostic indicator for hepatocellular carcinoma. *Aging (Albany NY)* **13**, 11096-11119 (2021).
47. Rontogianni, S. et al. Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping. *Commun Biol* **2**, 325 (2019).
48. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* **47**, W199-W205 (2019).

Figure legends

Figure 1. Overview of the study.

Figure 2. Proteomic characterization of exosomes derived from 285 cell lines from four studies. (A) Overlapped proteins from four different studies of cell line-derived exosomes. **(B)** PCA plot of cancer and control cell line-derived exosomes. **(C)** Positivity for 8 commonly used exosomal protein biomarkers in various cell lines. The percentage of samples expressing each protein is shown in the boxes. Darker red indicates a higher percentage. **(D)** Annotation of the proteins detected in more than 90% of all samples. **(E)** GO and KEGG pathway enrichment analysis of the proteins detected in more than 90% of all samples. **(F)** Plasma membrane proteins detected in more than 90 % of all samples.

430 **Figure 3. Proteomic characterization of exosomes derived from cell lines and tissues. (A)**
431 Proteins detected at higher frequency in cancer cell line-derived exosomes. **(B)** Positivity for 11
432 commonly used exosomal protein biomarkers in various tissues. **(C)** Overlapping proteins (>90%
433 frequency) between cell line- and tissue-derived exosomes. **(D)** Positivity of five plasma
434 membrane proteins detected in more than 90% of both cell line- and tissue-derived exosomes.
435
436 **Figure 4. Identification of the signature proteins of plasma or serum-derived exosomes and**
437 **the evaluation of random forest classifier. (A)** Overlapping exosome proteins detected in the
438 plasma and serum of 205 cancer and 51 control samples from five studies. **(B)** Heat map of 46
439 overlapping exosome proteins in cancer and control plasma or serum samples. **(C)** AUROC
440 score of the random forest classifier on including various numbers of protein features. **(D)**
441 AUROC of different models in comparison. **(E)** Classification error matrix of the 75% training
442 set using a random forest classifier for the 18 selected proteins. The number of samples is
443 indicated in each box. **(F)** AUROC score of the random forest classifier trained using 75% of the
444 dataset. Other metrics are indicated on right. **(G)** Classification error matrix of 25% testing set
445 using a random forest classifier for the 18 selected proteins. The number of samples is indicated
446 in each box. **(H)** AUROC score of the random forest classifier tested using 25% of the dataset.
447 Other metrics are indicated on right.
448
449 **Figure 5. Identification of signature proteins expressed by plasma or serum-derived**
450 **exosomes for classifying five common cancer types and evaluation of random forest**
451 **classifier. (A)** PCA plot of plasma or serum-derived exosomes from five cancer types. **(B)**
452 AUROC score of the random forest classifier by including various number of protein features.

453 **(C-D)** Classification error matrix of a 60% training set and 40% testing set to classify the five
454 cancer types using a random forest classifier for the 5 selected proteins. The number of samples
455 is indicated in each box. **(E)** Protein abundance of five selected protein features in five cancer
456 types.
457
458 **Figure 6. Identification of signature proteins expressed by urine-derived exosomes and**
459 **evaluation of random forest classifier. (A)** Overlapping exosome proteins detected in the urine
460 from 261 cancer and 124 control samples from four studies. **(B)** PCA plot of cancer and control
461 urine-derived exosomes. **(C)** AUROC score of the random forest classifier by including a various
462 number of protein features. **(D)** Protein abundance of 17 selected protein features in cancer- and
463 control urine-derived exosomes. **(E)** Classification error matrix of 75% training set using a
464 random forest classifier for the 17 selected proteins. The number of samples is indicated in each
465 box. **(F)** AUROC score of the random forest classifier trained using 75% of the dataset. Other
466 metrics are indicated on right. **(G)** Classification error matrix of 25% testing set using a random
467 forest classifier for the 17 selected proteins. The number of samples is indicated in each box. **(H)**
468 AUROC score of the random forest classifier tested using 25% of the dataset. Other metrics are
469 indicated on right.
470
471 **Supplementary Figure S1. Machine learning models for plasma-derived exosomes. (A)** PCA
472 plot of cancer and control plasma or serum-derived exosomes. **(B)** Protein abundance of 18
473 selected protein features in cancer- and control plasma or serum-derived exosomes.
474
475

Figure 1

A

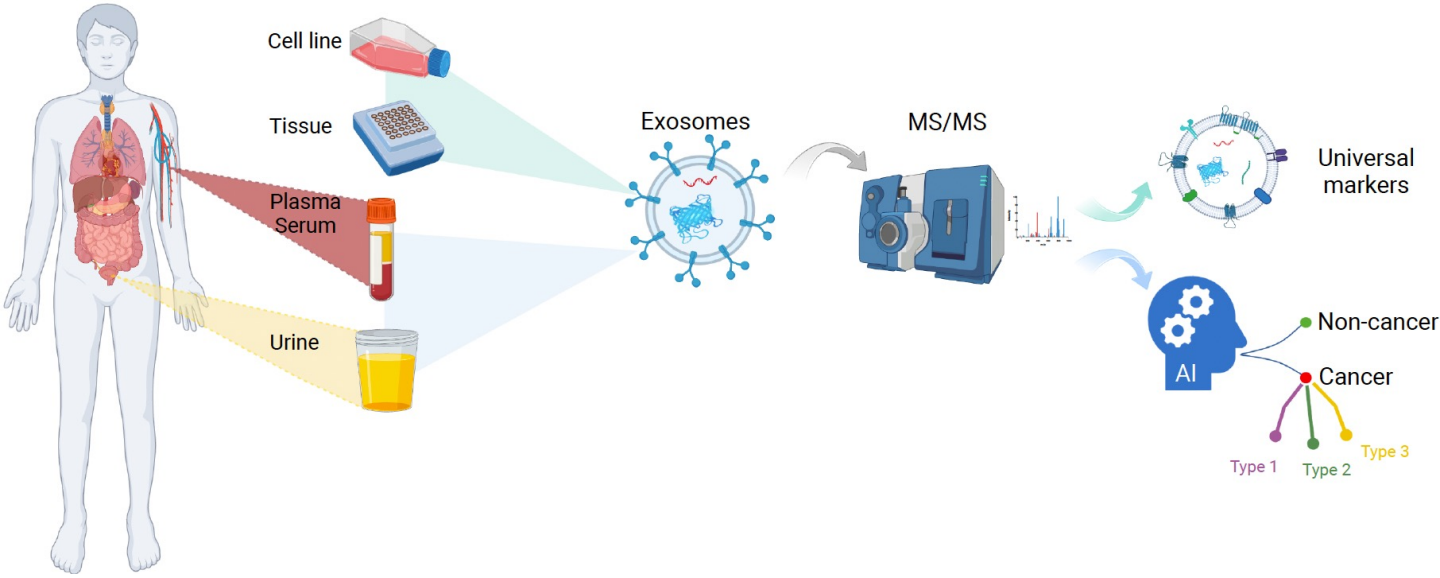

Cell line
Tissue
Plasma
Serum
Urine
Exosomes
MS/MS
Universal
markers
AI
Non-cancer
Cancer
Type 1
Type 2
Type 3

Figure 2

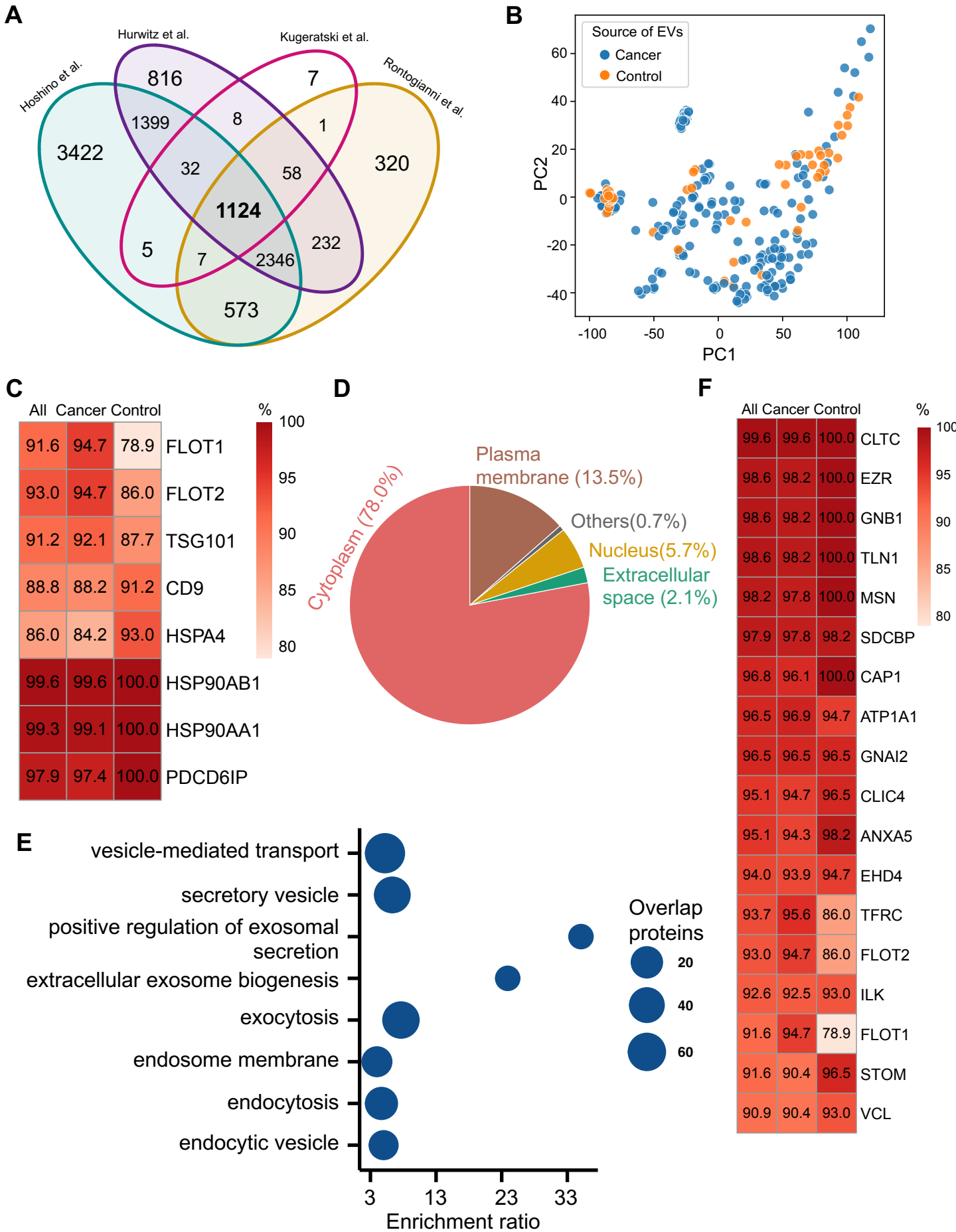

A
Hurwitz et al.
Kugeratski et al.
Hoshino et al.
Rontogianni et al.
816
7
1399
8
1
3422
32
58
320
1124
5
232
7
2346
573
B
Source of EVs
Cancer
Control
PC2
PC1
60
40
20
0
-20
-40
-100
-50
0
50
100
C
All Cancer Control
%
100
95
90
85
80
91.6 94.7 78.9 FLOT1
93.0 94.7 86.0 FLOT2
91.2 92.1 87.7 TSG101
88.8 88.2 91.2 CD9
86.0 84.2 93.0 HSPA4
99.6 99.6 100.0 HSP90AB1
99.3 99.1 100.0 HSP90AA1
97.9 97.4 100.0 PDCD6IP
D
Plasma membrane (13.5%)
Others(0.7%)
Nucleus(5.7%)
Extracellular space (2.1%)
Cytoplasm (78.0%)
E
vesicle-mediated transport
secretory vesicle
positive regulation of exosomal secretion
extracellular exosome biogenesis
exocytosis
endosome membrane
endocytosis
endocytic vesicle
3
13
23
33
Enrichment ratio
Overlap proteins
20
40
60
F
All Cancer Control
%
100
95
90
85
80
99.6 99.6 100.0 CLTC
98.6 98.2 100.0 EZR
98.6 98.2 100.0 GNB1
98.6 98.2 100.0 TLN1
98.2 97.8 100.0 MSN
97.9 97.8 98.2 SDCBP
96.8 96.1 100.0 CAP1
96.5 96.9 94.7 ATP1A1
96.5 96.5 96.5 GNAI2
95.1 94.7 96.5 CLIC4
95.1 94.3 98.2 ANXA5
94.0 93.9 94.7 EHD4
93.7 95.6 86.0 TFRC
93.0 94.7 86.0 FLOT2
92.6 92.5 93.0 ILK
91.6 94.7 78.9 FLOT1
91.6 90.4 96.5 STOM
90.9 90.4 93.0 VCL

Figure 3

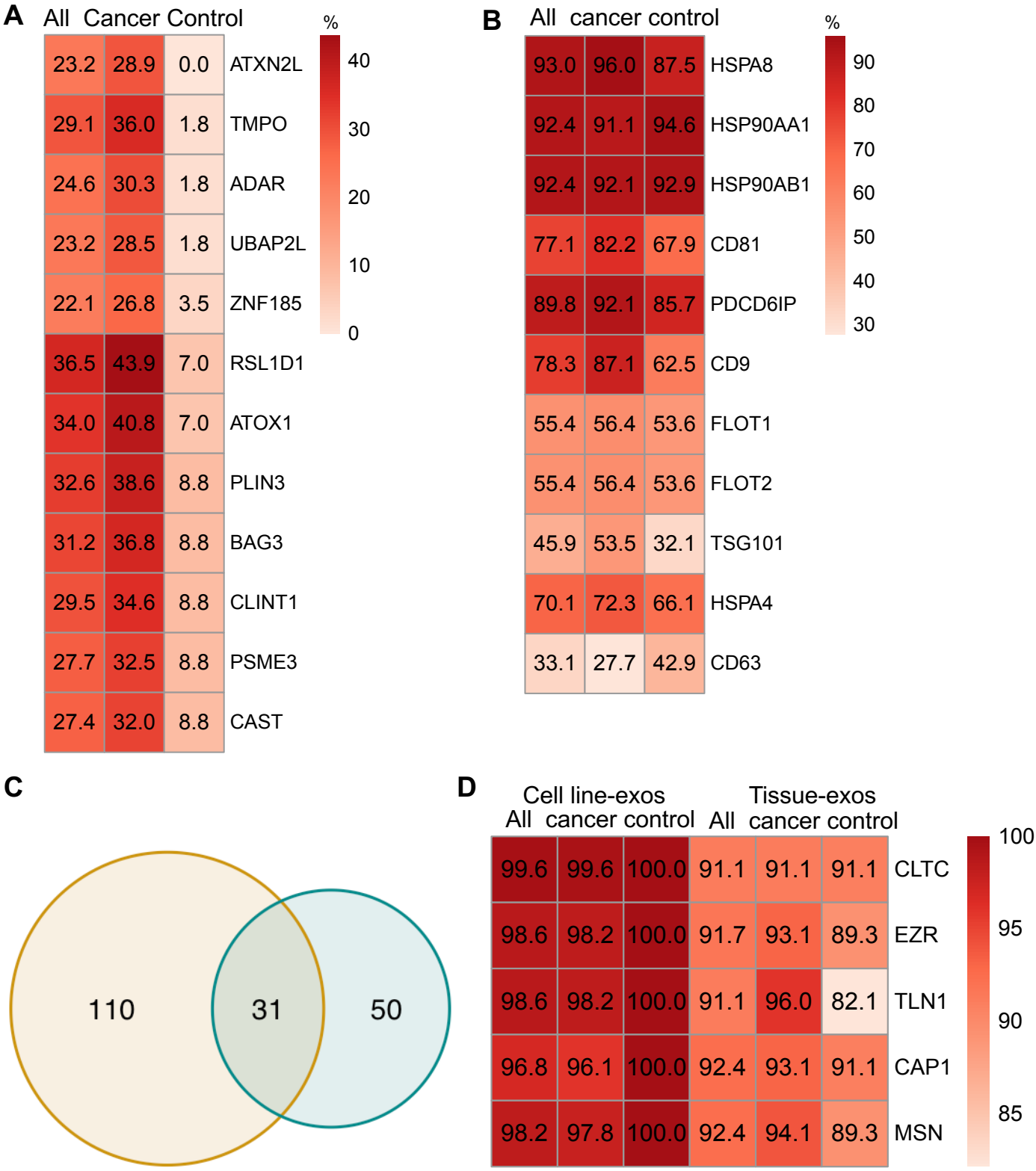

A
All Cancer Control
%
23.2 28.9 0.0 ATXN2L
29.1 36.0 1.8 TMPO
24.6 30.3 1.8 ADAR
23.2 28.5 1.8 UBAP2L
22.1 26.8 3.5 ZNF185
36.5 43.9 7.0 RSL1D1
34.0 40.8 7.0 ATOX1
32.6 38.6 8.8 PLIN3
31.2 36.8 8.8 BAG3
29.5 34.6 8.8 CLINT1
27.7 32.5 8.8 PSME3
27.4 32.0 8.8 CAST
40
30
20
10
0
B
All cancer control
%
93.0 96.0 87.5 HSPA8
92.4 91.1 94.6 HSP90AA1
92.4 92.1 92.9 HSP90AB1
77.1 82.2 67.9 CD81
89.8 92.1 85.7 PDCD6IP
78.3 87.1 62.5 CD9
55.4 56.4 53.6 FLOT1
55.4 56.4 53.6 FLOT2
45.9 53.5 32.1 TSG101
70.1 72.3 66.1 HSPA4
33.1 27.7 42.9 CD63
90
80
70
60
50
40
30
C
110
31
50
D
Cell line-exos
All cancer control
Tissue-exos
All cancer control
99.6 99.6 100.0 91.1 91.1 91.1 CLTC
98.6 98.2 100.0 91.7 93.1 89.3 EZR
98.6 98.2 100.0 91.1 96.0 82.1 TLN1
96.8 96.1 100.0 92.4 93.1 91.1 CAP1
98.2 97.8 100.0 92.4 94.1 89.3 MSN
100
95
90
85

Figure 4

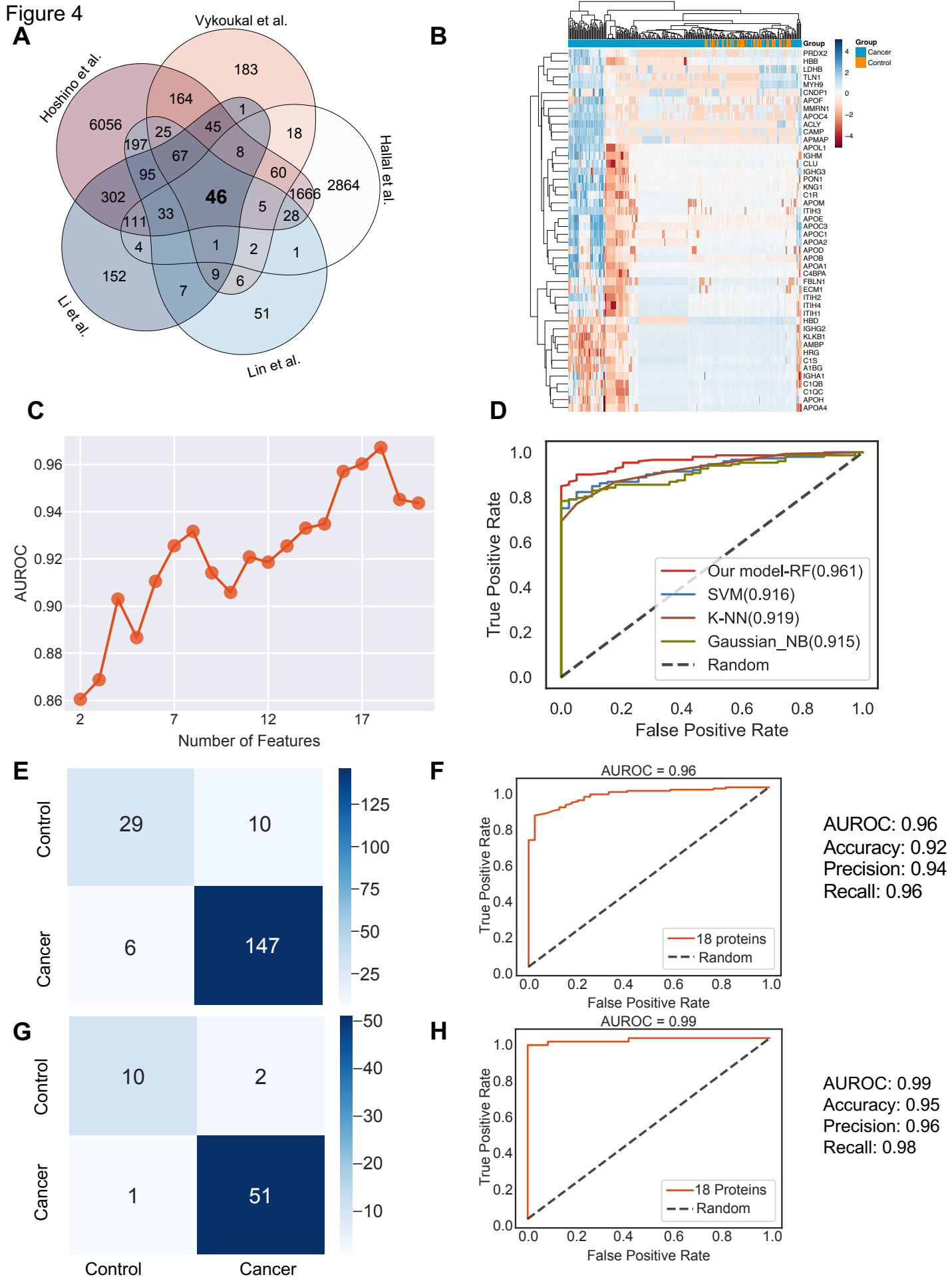

A
Vykoukal et al.
Hoshino et al.
Hallal et al.
Li et al.
Lin et al.
183
164
1
6056
25
45
18
197
67
8
95
60
2864
302
46
1666
5
33
28
111
4
1
2
1
152
9
6
7
51
B
Group
Cancer
Control
PRDX2
HBB
LDHB
TLN1
MYH9
CNDP1
APOF
MMRN1
APOC4
ACLY
CAMP
APMAP
APOL1
IGHM
CLU
IGHG3
PON1
KNG1
C1R
APOM
ITIH3
APOE
APOC3
APOC1
APOA2
APOD
APOB
APOA1
C4BPA
FBLN1
ECM1
ITIH2
ITIH4
ITIH1
HBD
IGHG2
KLKB1
AMBP
HRG
C1S
A1BG
IGHA1
C1QB
C1QC
APOH
APOA4
C
AUROC
Number of Features
D
True Positive Rate
False Positive Rate
Our model-RF(0.961)
SVM(0.916)
K-NN(0.919)
Gaussian_NB(0.915)
Random
E
Control
Cancer
29
10
6
147
F
AUROC = 0.96
True Positive Rate
False Positive Rate
18 proteins
Random
AUROC: 0.96
Accuracy: 0.92
Precision: 0.94
Recall: 0.96
G
Control
Cancer
10
2
1
51
H
AUROC = 0.99
True Positive Rate
False Positive Rate
18 Proteins
Random
AUROC: 0.99
Accuracy: 0.95
Precision: 0.96
Recall: 0.98

Figure 5

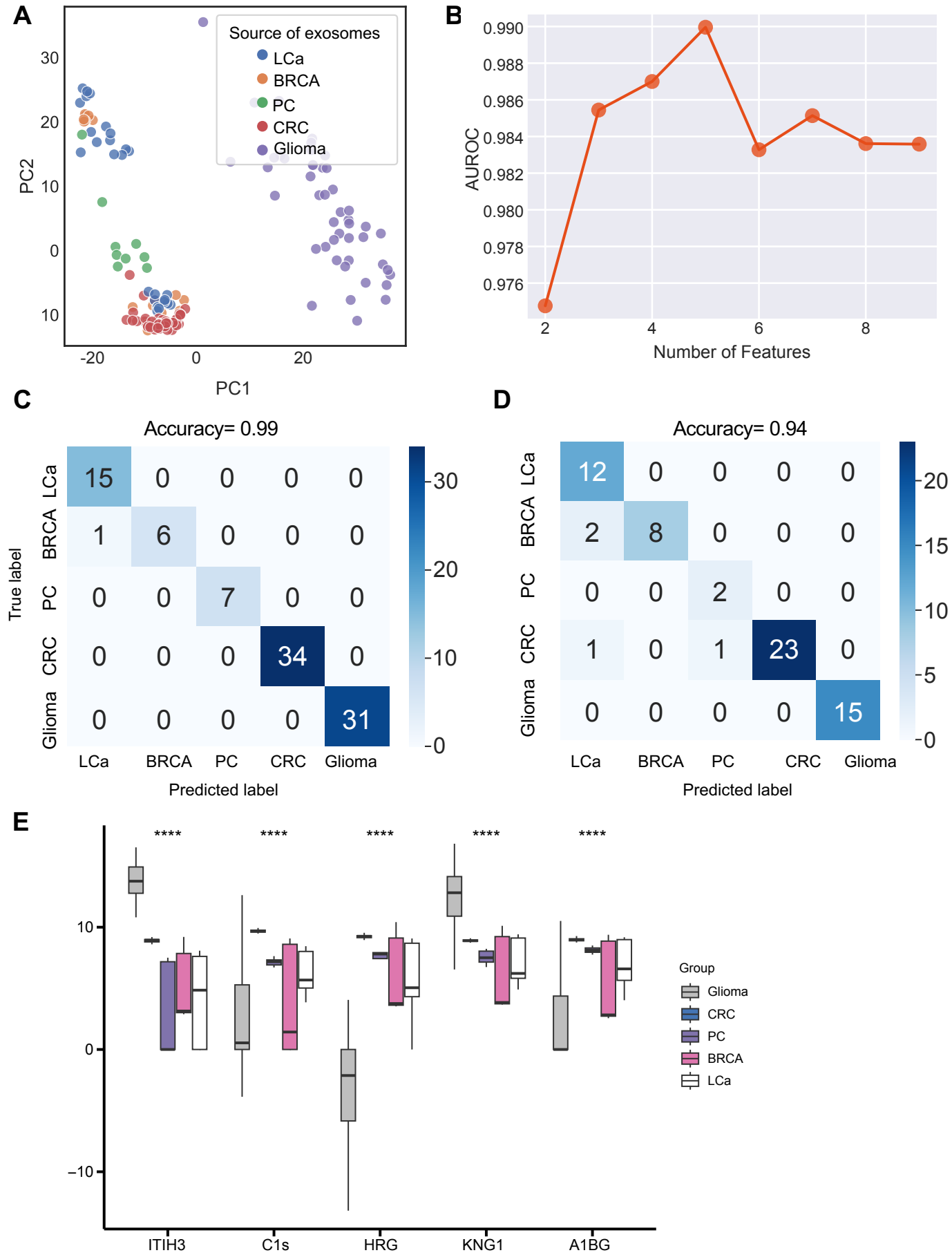
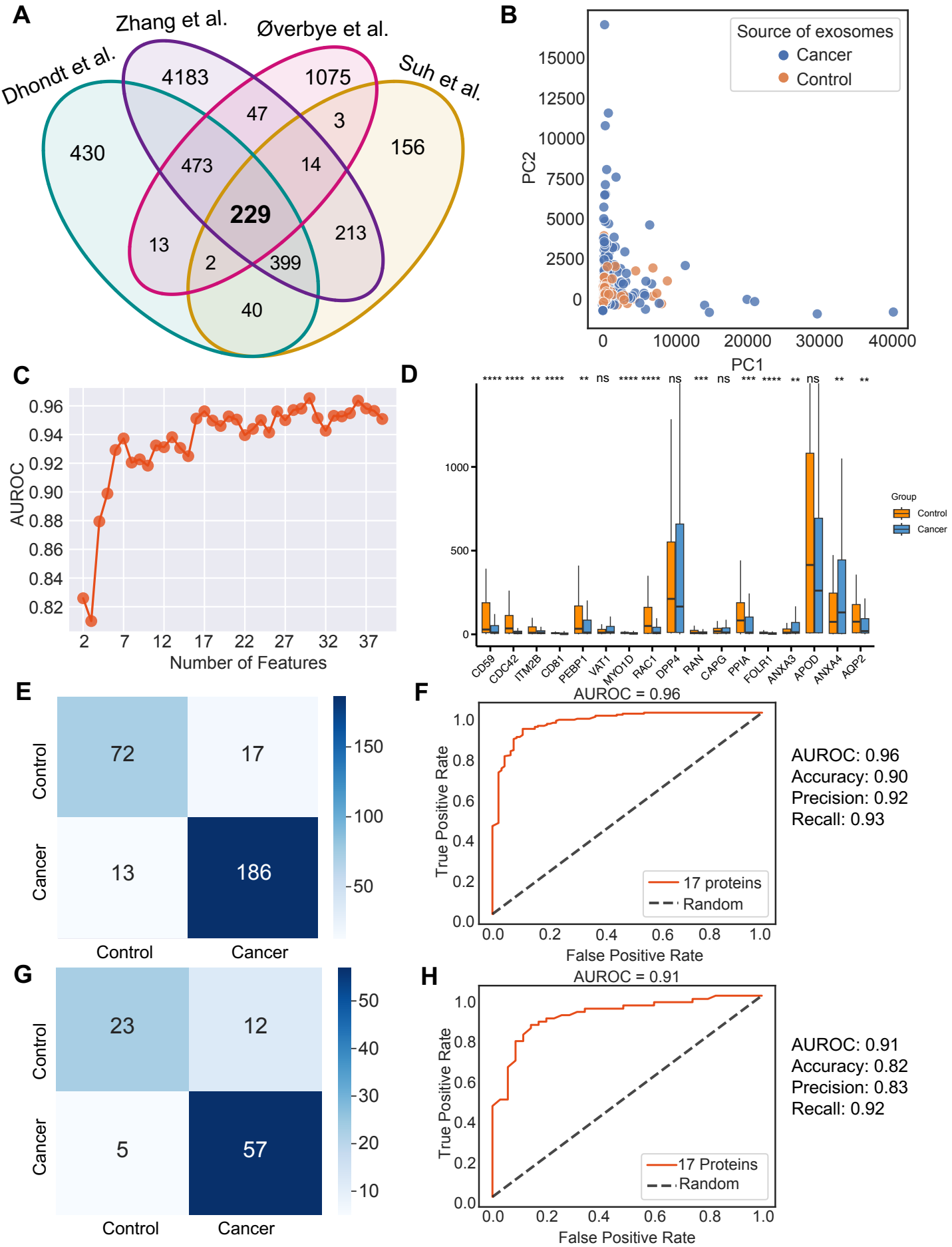

A
Source of exosomes
LCa
BRCA
PC
CRC
Glioma
PC2
PC1
B
AUROC
Number of Features
C
Accuracy= 0.99
True label
Predicted label
LCa
BRCA
PC
CRC
Glioma
D
Accuracy= 0.94
E
Group
Glioma
CRC
PC
BRCA
LCa
ITIH3
C1s
HRG
KNG1
A1BG

Figure 6


A
Zhang et al.
Øverbye et al.
Dhondt et al.
Suh et al.
4183
1075
47
3
430
473
14
156
229
13
213
2
399
40
B
Source of exosomes
Cancer
Control
PC2
PC1
C
AUROC
Number of Features
D
Group
Control
Cancer
CD59
CDC42
ITM2B
CD81
PEBP1
VAT1
MYO1D
RAC1
DPP4
RAN
CAPG
PPIA
FOLR1
ANXA3
APOD
ANXA4
AQP2
E
Control
Cancer
72
17
13
186
F
AUROC = 0.96
True Positive Rate
False Positive Rate
17 proteins
Random
AUROC: 0.96
Accuracy: 0.90
Precision: 0.92
Recall: 0.93
G
23
12
5
57
H
AUROC = 0.91
17 Proteins
Random
AUROC: 0.91
Accuracy: 0.82
Precision: 0.83
Recall: 0.92