Title: Pan-cancer detection and typing by mining patterns in large genome-wide cell-free DNA sequencing datasets

Running head: cfDNA data mining for cancer detection

Authors: Huiwen Che[1], Tatjana Jatsenko[1], Liesbeth Lenaerts[2], Luc Dehaspe[3], Leen Vancoillie[3], Nathalie Brison[3], Ilse Parijs[3], Kris Van Den Bogaert[3,] Daniela Fischerova[4], Ruben Heremans[5], Chiara Landolfo[6], Antonia Carla Testa[7], Adriaan Vanderstichele[6], Lore Liekens[8], Valentina Pomella[8], Agnieszka Wozniak[9], Christophe Dooms[10,11], Els Wauters[10,11], Sigrid Hatse[9,12,] Kevin Punie[12,13], Patrick Neven[6,12], Hans Wildiers[12,13], Sabine Tejpar[8], Diether Lambrechts[14], An Coosemans[15], Dirk Timmerman[5,6], Peter Vandenberghe[16,17], Frédéric Amant[2,6,18], Joris Robert Vermeesch[1,3*]

Affiliations:

[1]Department of Human Genetics, Laboratory for Cytogenetics and Genome Research, KU Leuven, Leuven, Belgium.

[2]Department of Oncology, Laboratory of Gynecological Oncology, KU Leuven, Leuven, Belgium.

[3]Centre for Human Genetics, University Hospitals Leuven, Leuven, Belgium.

[4]Department of Obstetrics and Gynaecology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic.

[5]Department of Development and Regeneration, Woman and Child, KU Leuven, Leuven, Belgium.

[6]Department of Gynecology and Obstetrics, University Hospitals Leuven, Leuven, Belgium.

[7]Department of Woman and Child Health, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Università Cattolica del Sacro Cuore Roma, Rome, Italy.

[8]Department of Oncology, Molecular Digestive Oncology, KU Leuven, Leuven, Belgium.

[9]Department of Oncology, Laboratory of Experimental Oncology, KU Leuven, Leuven, Belgium.

[10]Department of Chronic Diseases and Metabolism, Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), KU Leuven, Leuven, Belgium.

[11]Department of Pneumology, University Hospitals Leuven, Leuven, Belgium.

[12]Multidisciplinary Breast Centre, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium.

[13]Department of General Medical Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium.

[14]Department of Human Genetics, Laboratory of Translational Genetics, VIB-KU Leuven, Leuven, Belgium.

[15]Department of Oncology, Laboratory of Tumor Immunology and Immunotherapy, Leuven Cancer Institute, KU Leuven, Leuven, Belgium.

[16]Department of Human Genetics, Laboratory of Genetics of Malignant Diseases, KU Leuven, Leuven, Belgium.

[17]Department of Hematology, University Hospitals Leuven, Leuven, Belgium.

[18]Center for Gynecological Oncology Amsterdam: Academic Medical Centre Amsterdam-University of Amsterdam and The Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands.

*Corresponding author: Joris Robert Vermeesch, Katholieke Universiteit Leuven (KU Leuven) Universitaire Ziekenhuizen Leuven (University Hospitals Leuven), Human Genetics, Herestraat 49, box 602, 3000 Leuven, Belgium.

Telephone: +32 16 34 5941

51 Email: joris.vermeesch@uzleuven.be

Abstract

Background

Cell-free DNA (cfDNA) analysis holds great promise for non-invasive cancer screening, diagnosis and monitoring. We hypothesized that mining the patterns of big datasets of shallow whole genome sequencing cfDNA from cancer patients could improve cancer detection.

Methods

By applying unsupervised clustering and supervised machine learning on large shallow whole-genome sequencing cfDNA datasets from healthy individuals (n=367), patients with different hematological (n=238) and solid malignancies (n=320), we identify cfDNA signatures that enable cancer detection and typing.

Results

Unsupervised clustering revealed cancer-type-specific sub-grouping. Classification using supervised machine learning model yielded an overall accuracy of 81.62% in discriminating malignant from control samples. The accuracy of disease type prediction was 85% and 70% for the hematological and solid cancers, respectively. We demonstrate the clinical utility of our approach by classifying benign from invasive and borderline adnexal masses with an AUC of 0.8656 and 0.7388, respectively.

Conclusions

This approach provides a generic and cost-effective strategy for non-invasive pan-cancer detection.

key words: liquid biopsy, cfDNA, ctDNA, hematological malignancies, solid tumors, ovarian tumors, machine learning

Introduction

Cell-free DNA (cfDNA) is a promising non-invasive biomarker in liquid biopsy for cancer management. Shallow whole-genome sequencing (sWGS) of cfDNA can identify cancer-specific copy number aberrations (CNAs) in cancer patients (1,2). Using genome-wide cfDNA sequencing data to profile genomic imbalances, we reported that CNAs in the asymptomatic population can be indicative of incipient tumors and has potential as a cancer screening tool (3).

In addition to CNAs, sequencing of cfDNA provides a unique view on the genome-wide cfDNA fragmentation profile (4,5). CfDNA fragments carry tissue-associated nucleosome and preferred end position information (6,7), reflecting tissue-specific degradation, chromatin accessibility and nucleosome organization of its cellular origin (8,9). In healthy individuals, plasma cfDNA comprises DNA fragments that are mainly resulting from apoptotic release of DNA from the cells of hematopoietic origin (10). In plasma of cancer patients, circulating tumor DNA (ctDNA) has decreased fragment sizes and signatures of the tissue of origin (8,11). Consequently, fragmentomics is emerging as an approach to reveal cfDNA properties, broadening the potential of cfDNA as a biomarker (4,12).

Increasing availability of cfDNA sWGS data from large-scale liquid biopsy projects offer unique opportunities to explore the cfDNA profiles by machine learning. We hypothesized that mining variation between sWGS profiles may uncover distinct patterns that can be associated with different pathological or physiological states. Hence, we applied an unsupervised clustering analysis and supervised machine learning workflow, which we term GIP*Xplore*, on a large number of genome-wide sWGS cfDNA profiles from cancer patients with different hematological and solid tumors and unveiled cancer-type-specific and also shared tumor-

100 associated signatures that are absent in healthy individuals. This approach enables accurate
101 detection of different cancers and allows prediction of the cancer types.

Materials and Methods

Patients and clinical data

The study was approved by the ethical committee of the University Hospitals Leuven (S57999, S62285, S62795, S50623, S56534, S63240, S51375, S59207, S64205 and S64035). Samples and consents were obtained from healthy controls and cancer patients. Blood was collected either into Streck Cell-Free DNA BCT or Roche Cell-Free DNA Collection Tubes. Plasma was isolated via a standard, two-step centrifugation procedure and stored at -80°C. Previously published sequencing data from 260 healthy subjects (3) and 177 patients with Hodgkin's lymphoma (13) were included in the study.

sWGS Analysis

cfDNA was extracted from plasma using standard processing procedures and sWGS sequencing (14) (details described in the Supplemental Materials). Each sample ended with 57509 autosome bin features from standard processing. Principal component analysis (PCA) was used for dimension reduction to transform data from high dimension to low dimension. We performed the supervised learning on both the original data space and PCA transformed space and found marginal gains of performance in the majority of analyses with the original data space. As the computational time was much higher using the original data space, we used PCA features in the main analyses such that features being used in both unsupervised and supervised learning were consistent.

GIP*Xplore*

As illustrated in **Fig. 1**, we developed GIP*Xplore* to mine sWGS cfDNA data for identification of signatures. We utilized unsupervised clustering and supervised machine learning. For unsupervised clustering, we evaluated the variance being explained from principal components

(PCs) in the tumor data. Overall, the top 30, 50 and 100 PCs explained above 80%, 85% and 90% of the variance in the data, respectively. While there is no absolute optimal number of PCs to be used for further analysis, non-trivial components - 50 PCs (**Supplemental Materials**) were determined as a default number for downstream analyses in the results. The Euclidean metric was used to measure dissimilarity among samples for clustering analysis. Proximity matrix based on dissimilarity of samples was generated. The t-distributed stochastic neighbor embedding (tSNE) (15) was used to map high-dimensional data to two (or three) dimensions and to visualize the clusters. Due to the random process of tSNE, we applied Walktrap community (16) detection on the original proximity matrix for cluster assignments regardless of the presentation of tSNE visualization. In running tSNE, we set parameters perplexity of 15/30 and iteration of 10000 with exact tSNE for accuracy, and the process was repeated for 10 times with different seeds. For Walktrap, we used the parameters of 8 initial numbers of neighbors search and a walk step of 2. Clusters defined from the community detection were used for annotation. In supervised learning, PCA transformed genome-wide features were used in the machine learning model for training. PCA was performed on training data, and test data was projected on PCA space of training data for classification tasks. We measured performance by repeating the tenfold cross-validation 10 times and leave-one-out (LOO) procedures. For cross fold (CV) validation, the ROC curve and performance was calculated by averaging over 10 repeats. For classifiers, we used a support vector machine (SVM) and hyperparameters were chosen based on the grid search with a subset of the data. A separate model was trained to localize tissue of origin and LOO was used to evaluate performance characteristics. Weighted sample size was accounted for in the model for imbalanced classes.

Results

GIP*Xplore* detects and classifies hematological malignancies with high accuracy

To assess the potential, we applied our method on a set of cfDNA samples from healthy controls (n=260) and patients with hematological malignancies that included Hodgkin's lymphoma (HL; n=179), diffuse large B-cell lymphoma (DLBCL; n=37), multiple myeloma (MM; n=22) (**Table 1**). Walktrap community detection was performed on the dataset, and 15 clusters were defined. Visualization with the tSNE yielded separations between malignant and healthy control profiles, and the tSNE representation was largely in agreement with the clusters found by Walktrap (**Fig. 2, A**). Moreover, we observed cancer type-specific clusters. Cluster 1, 3 and 4 was exclusively composed of HL samples. Cluster 9 was enriched for DLBCL samples, and cluster 13 was specific to MM samples (**Fig. 2, B** and **Supplemental Fig. 1**).

In parallel, we benchmarked our method against the ichorCNA (17) algorithm for copy number profiling and tumor fraction (TF) estimation from sWGS data. IchorCNA utilizes the depth of coverage to evaluate the presence of large-scale copy number aberrations and the probabilistic model is used to infer copy number states and estimate fraction of tumor. Overall, only 52.95% of hematological cancer samples had detectable tumor-derived cfDNA levels (**Fig. 2, C, Supplemental Fig. 2** and **Supplemental Table 1**), using the 3% detection limit suggested in the ichorCNA for detecting the presence of tumor. The above-mentioned clusters 1, 3, and 4 consisted of profiles characterized by large chromosomal aberrations and high tumor load. Clusters 2 and 8 consisted of profiles from HL patients with both high and low tumor fractions, implying that the clustering was not completely CNA-driven. In particular, ten out of 65 (15.38%) lymphoma samples in cluster 2 with normal-like profiles (without detectable CNAs) grouped together with samples characterized by detectable CNAs. A less pronounced separation could be observed between clusters containing healthy controls and cluster 8, in

which 76.47% (26 out of 34) malignant cases had normal-like profiles with less than 3% TFs. Nine HL samples in cluster 10 showed higher bin-to-bin log2 ratio variations and were more likely to be noisy on a genome-wide scale (**Fig. 2, D** and **Supplemental Fig. 3**). The remaining malignant cases without detectable CNAs co-localized with healthy controls. To further explore whether clustering of malignant samples would be mainly CNA-driven, we performed clustering analysis using the log2 copy ratio values produced by ichorCNA. The analysis revealed that genome-wide copy number ratios alone were less informative (**Supplemental Fig. 4**). In addition, we tested whether our method could detect underlying genome-wide changes irrespective of the presence of CNAs by restricting the clustering to the cancer samples with low TF (< 3%). The separation between some malignant and healthy samples still remained (**Supplemental Fig. 5**). Collectively, the clustering analysis on genome-wide features showed separation between malignant and healthy profiles and grouping of similar cancer type-specific profiles.

The unsupervised learning delineated cancer-associated profile changes, which suggested that a more precise prediction can be made by learning representations within different tumor types using supervised classification. Therefore, we evaluated the capability to detect cancer signals and identify cancer types with supervised learning on the hematological cohort. Both leave-one-out (LOO) and repeated 10-fold cross validation (CV) was used to assess the performance of the classifier. Incorporating transformed genome-wide features, the SVM machine learning model correctly classified 220 (out of 238) malignant cases in LOO analysis, at a sensitivity of 92.44% (95% CI: 88.31% - 95.46%) and a specificity of 98.46% (95% CI: 96.11% - 99.58%), including 170 HL, 32 DLBCL and 18 MM cfDNA samples (**Supplemental Table 1**). The remaining 18 misclassified malignant samples had normal-like profiles and clustered together with healthy controls (**Supplemental Fig. 6**). The detection sensitivity was the highest for HL

(**Supplemental Table 1**). The sensitivity did not differ substantially between early (I-II) and advanced (III-IV) stages for these cancer types, though the distribution of the cases across clinical stages was unequal (**Fig. 3, A**). ROC analysis had an AUC value of 0.989 (95% CI: 0.980 – 0.998) in distinguishing malignant from healthy samples, compared to ichorCNA TF-based analysis which had an AUC of 0.929 (**Fig. 3, B)**. Repeated 10-fold CV also revealed a stable performance at an averaged AUC of 0.989 (**Supplemental Fig. 7**). As the clustering analysis demonstrated the co-localization of samples originating from the same cancer type, we then attempted to determine the accuracy of our GIP*Xplore* in cancer type classification. For this purpose, we trained the classification model using the 220 correctly predicted malignant samples. The analysis showed an overall accuracy of 85.45% (95% CI: 80.09% - 89.83%), with the highest accuracy in HL prediction (**Fig. 3, C** and **Supplemental Table 2**). Consistent with the exploratory clustering analysis, where some of the profiles from DLBCL patients colocalized together with those from HL patients, DLBCL samples were more likely to be misclassified.

GIP*Xplore* identifies and classifies different types of solid malignancies and allows disease stratification

Extending our analyses, we applied our method on a solid tumor dataset, consisting of 320 cfDNA profiles from cancer patients, and a set of 107 cfDNA profiles from healthy controls. The malignant cohort was represented by five tumor types: breast (n=46), colorectal (n=70), gastrointestinal stromal tumor (GIST; n=35), lung (n=44) and ovarian (n=125; **Table 1**). Using GIP*Xplore*, 19 clusters were identified in the solid tumor dataset (**Fig. 4, A** and **Supplemental Fig. 8**). The separations between malignant and control cfDNA profiles were less distinct compared to clustering results of the hematological cancer dataset. Clusters 4, 8, 10 and 12 were found to be cancer type-specific, in which cluster 4 was mainly enriched with ovarian

cancer samples, cluster 8 was primarily consisting of cfDNA profiles from lung cancer patients, cluster 10 was GIST-specific and cluster 13 was mainly composed of colorectal samples (**Fig. 4, B**). Cluster 2, adjacent to clusters 4 and 8, was enriched with ovarian samples, although it co-localized with other tumors. Clusters 9 (mostly ovarian cancer) and 15 (intermixed cancer types) deviated from healthy and other malignant clusters. Majority of the cfDNA profiles from breast cancer patients resembled profiles from healthy controls, while one advanced stage breast cancer sample was found in cluster 8, and 2 samples from patients with advanced stage primary metastatic disease were found in cluster 2 (**Fig. 4, A** and **B**).

Compared with the hematological cancer dataset, ctDNA levels estimated by ichorCNA were generally lower in the solid malignant cohort (**Fig. 4, C**). Tumor fraction varied among different types of cancer and increased with the stage (**Supplemental Fig. 9**). The malignant cases with detectable CNAs and therefore higher TF were more likely to separate from the healthy controls (**Supplemental Fig. 10**). Cluster 4 contained ovarian cancer samples with detectable chromosome instability. Among lung cancer profiles in cluster 8, 64.29% (9 out of 14) had detectable CNAs. Clusters 16 to 19 included four ovarian samples with high chromosomal instability that greatly deviated from other profiles. Overall, in clusters 9 and 15, profiles tended to be noisy, without clear CNAs (**Fig. 4, D**), however they deviated from healthy control and other malignant clusters (**Fig. 4, A)**. When using the log2 copy ratio profiles from the CNA analysis to investigate whether the sub-grouping of cfDNA profiles was driven by CNAs, cancer type-specific clustering patterns were diminished (**Supplemental Fig. 11**). When restricting the clustering analysis to samples with TF lower than 3%, samples from clusters 9 and 15 still showed deviations (**Fig. 4, A** clusters 8 and 9) from normal profiles (**Supplemental Fig. 12**).

We next investigated whether supervised learning using genome-wide features can enhance the detection of solid malignancy signals in sWGS cfDNA data. Classification of samples as either healthy or malignant (107 healthy controls and 320 malignancies) was performed using the SVM model, with performance estimated by LOO and repeated 10-fold CV. With an overall accuracy of 65.34%, we correctly detected 177 out of 320 cancer profiles (55.31% sensitivity, 95% CI: 49.68% - 60.84%), at a specificity of 95%. Performance in individual tumor types ranged from 15.22% (95% CI: 6.34% - 28.87%) for classifying breast cancer to 80.00% (95% CI: 63.06% - 91.56%) for GIST (**Supplemental Table 3**). Stage of the disease affected the detection, with a sensitivity of 26.17% (95% CI: 18.15% - 35.55%) in the early stage (I-II) *versus* 69.95% (95% CI: 63.31% - 76.03%) in the advanced stage (III-IV). In individual tumor types, it remained true that higher sensitivities were found for the advanced stages than for the early-stage diseases (**Fig. 5, A**). Colorectal cancer was an exception as sensitivities were almost the same for early and advanced cancer stages. Misclassified malignant samples had low tumor fraction, which potentially restricted the detection of underlying tumor-specific patterns (**Supplemental Fig. 13**). We could distinguish malignancy from healthy samples with an AUC of 0.827 (95% CI: 0.787 – 0.867), which again was superior to ichorCNA TF-based analysis (0.733 AUC, 95% CI: 0.687 – 0.780; **Fig. 5, B** and **Supplemental Fig. 14**). Subsequently, we explored the potential of our GIP*Xplore* method for tumor classification. When performing tumor type-specific prediction with the 171 correctly predicted primary tumor samples, the LOO validation resulted in a 69.01% (95% CI: 61.49% - 75.84%) overall accuracy. Highest sensitivities (>70%) were obtained for cfDNA samples from ovarian cancer and GIST patients. At the same time, ovarian and colorectal tumor cfDNA profiles were more likely to be misassigned to each other (**Fig. 5, C** and **Supplemental Table 4**).

Moreover, among this solid malignant cohort, we had nine cfDNA samples from patients with ovarian metastases, of which four patients had gastrointestinal primary site, one lymphoma, one leiomyosarcoma, one uterine origin, and the remaining two had Krukenberg tumors. Annotation of these nine cases on the tSNE plot showed that metastatic profiles could resemble profiles of either the primary tumor or the distant site (**Supplemental Fig. 15, A**). Applying the type-specific classifier to the six metastatic cases that were predicted as malignant cases by the malignancy classifier, the case with gastrointestinal origin that was co-clustered with colorectal samples was classified to be colorectal class. Two out of three metastatic cases that were identified in intermixed clusters of lung and ovarian tumors were predicted to be lung class and the other one was assigned to ovarian class. The additional two cases identified by the classifier were classified to the ovarian and colorectal classes, respectively (**Supplemental Fig. 15, B**).

Accurate classification of benign from invasive and borderline adnexal masses may improve clinical management

In addition to the invasive ovarian tumor samples, our cohort contained 160 benign and 63 borderline ovarian samples. To assess the potential utility of the method for ovarian cancer management, we analyzed the ovarian tumor cohort independently by performing clustering analysis and building the ovarian-specific classifier to differentiate benign from malignant adnexal masses. Benign and borderline samples were less likely to have detectable ctDNA levels (**Supplemental Fig. 16**). In the clustering analysis, 35 invasive samples formed a distinct group in cluster 1. In clusters 6-9, common patterns were found for invasive, benign and borderline samples, although they remained distinct from controls (**Supplemental Fig. 17**). The classification analysis exhibited an AUC of 0.8656 (95% CI: 0.7761 – 0.8689) in discriminating benign from invasive samples, and an AUC of 0.7388 (95% CI: 0.6857 –

300 0.7920) in discriminating benign from borderline and invasive samples (**Supplemental Fig. 18**

301 and **19**).

DISCUSSION

We present a generic approach for cancer identification and classification by mapping genome-wide cfDNA signatures, without prior knowledge of genetic alterations or predefined signatures in the sequencing data. The unsupervised clustering allows the discovery of hidden genome-wide patterns, and the supervised learning model can be trained to detect such underlying signatures. This method can be used to classify cfDNA samples by matching to existing datasets and has the potential to be used as a pan-cancer assay for detection and typing of multiple cancers from one blood draw.

Current sWGS cfDNA analyses mainly focus on the detection of somatic CNAs (17–19). These methods are blind to events that involve copy neutral abnormalities. Our approach also differs from the previous method that classified tumor types based on selected CNAs, and in which normal-like profiles were incapable of tumor classification (20). We demonstrate that even profiles without detectable CNAs carry informative and discriminative patterns in sWGS data. Different recent studies have utilized methylation, transcription factor binding, fragment lengths, or cfChIP-seq for cancer detection (4,12,21–25). While these studies have important implications and show cfDNA as a promising biomarker, they require more specific workup and/or deeper sequencing. In contrast, analysis of sWGS data can be easily adapted in clinical settings and complement CNA analysis. By mapping differences among the cfDNA profiles, shared abnormality patterns are captured.

To date, Liu *et al.* have reported the largest population-level cfDNA methylation study for multi-cancer detection, in which the targeted methylation analysis of cfDNA enabled detection of more than 50 cancer types at a sensitivity of 54.9% and at a specificity of 99% (21). This test was refined and validated in an independent follow-up study, with an overall sensitivity of

51.5% at 99.5% specificity was reported (26). In line with these findings, we estimate the combined sensitivity of 68.03% at above 95% specificity for the hematological and solid cancer cohorts. Performance for cancer signal detection varied among the different cancer types and stages. The prediction accuracy was highest for hematological malignancies and lowest for breast cancer. Shedding of the ctDNA from breast cancer is known to be low (27,28). Also, our cohort had an over-representation of early-stage cancers, with 50% of the samples from stage I. Apart from potential screening applications, we also demonstrated that GIP*Xplore* could be used for risk stratification and management of a specific cancer type. Discrimination between malignant, borderline and benign masses at diagnosis is of critical importance to improve patient management (29,30).

The accuracy of tumor type-specific prediction might depend on the intrinsic tumor characteristics. For example, DLBCL being more heterogeneous on molecular level (31,32), had lower classification accuracy than HL and MM. The subtype of colorectal and ovarian tumors is of similar cellular origin, and histological subtypes can be hard to distinguish (33–35), which might be a reason for misclassification amongst the two cancer types. The identification of the origin of some metastases, suggests the method may allow the identification of unknown primary cancers. The metastatic cases were classified into profiles of its primary or distant sites, possibly reflecting changes during the metastatic progression or dynamic tumor DNA shedding from tumor tissues (36–38).

Interestingly, besides tumor type- or aberration-specific subgroups, our analysis revealed the presence of additional clusters that segregated from healthy controls (**Fig. 4, B** and **Supplemental Fig. 17**). Though the origin of such segregations remains unknown, we hypothesize the method provides a system-wide insight, potentially reflecting

(patho)physiological conditions of these individuals. Dynamic cellular responses and malignant cell proliferation with active involvement of immune response during (early) carcinogenesis might lead to the observed common changes in cfDNA composition across different cancer types (39,40). Therefore, it is possible that our analysis detected tumor-driven immune or other biological responses or states.

GIP*Xplore* provides an unbiased genome-wide scan of cfDNA profiles. However, it also has some limitations. Increasing the sequencing depth might improve detection of disease-specific cfDNA patterns and improve the sensitivity of our methodology further. The data presented here has a larger proportion of HL and ovarian cancer samples and is limited in the number of different cancer types, which may affect the aggregated sensitivity and distort tumor typing accuracies. We foresee that expanding the breadth of the evaluated cancer types may improve prediction of tissue/cell origin and facilitate a deeper understanding of cfDNA in the context of tumors. Increasing the range of physiological states and diseases that are relevant for these tumor samples will be essential to fully interrogate the potential and limitations of our approach. The approach may also be further broadened to project and embed new treatment or follow-up data for cancer prognosis and monitoring.

In summary, we have extended the scope of cfDNA analysis, allowing cost-effective identification of genome-wide cancer-(type-)specific signatures from shallow sequencing data, allowing improved discrimination between profiles from cancer patients and healthy individuals. This study lays the foundation for enhanced genomic characterization of cfDNA that can be used for improved cancer management. We foresee that the method can be scaled up for detection of multiple pathological conditions.

Acknowledgements: We would like to acknowledge the patients and blood donors. We would like to thank Gitte Thirion and Annick Van den Broeck for the collection of samples and the extraction of ctDNA, Kate Stanley for helpful suggestions for the manuscript.

Funding: This study was supported by the Research Foundation-Flanders (FWO-Vlaanderen) (G080217N to FA and JRV, G0A1116N to PV), Agentschap Innoveren en Ondernemen (VLAIO; Flanders Innovation & Entrepreneurship grant HBC.2018.2108 to JRV), Kom Op Tegen Kanker (Stand Up to Cancer, the Flemish Cancer Society under grant 2016/10728/2603 to AC), Stichting tegen Kanker (FAF-C/2016/836 to PV, 2018-134 to JRV and FA) and KU Leuven funding (no C1/018 to JRV and DL).

Conflict of interest: Patent application pending on 'Method for analyzing cell-free nucleic acids' (JRV and LD).

Author contributions: HC, TJ, LL, LD, FA and JRV conceptualized and designed the study. KP, AW, EW, AC, PN, ST, DT, PV, HW and FA provided clinical samples and patient data. LL, LV, NB, IP, KVDB, CD, DF, RH, SH, CL, L.Liekens, VP, ACT, AV and AW carried out clinical sample procurement and processing. LV, NB, IP and KVDB coordinated sequencing of cell-free DNA. TJ and LL conducted project coordination and administration. HC and LD performed bioinformatics analysis of sWGS data. HC, TJ, LL, LD, KP, AW, EW, AC, DL, ST, DT, PV, FA and JRV contributed to the interpretation of results. HC, TJ and JRV wrote the manuscript; all co-authors reviewed the manuscript.

Data and materials availability: Processed alignments of sequencing data are archived to ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) with unrestricted access under accession number E-MTAB-10934. Code will be available upon request. All other materials associated with this study are present in the paper or the Supplementary Materials.

Reference

1. Vandenberghe P, Wlodarska I, Tousseyn T, Dehaspe L, Dierickx D, Verheecke M, et al. Non-invasive detection of genomic imbalances in Hodgkin/Reed-Sternberg cells in early and advanced stage Hodgkin's lymphoma by sequencing of circulating cell-free DNA: a technical proof-of-principle study. Lancet Haematol. 2015;2:e55–65.

2. Lenaerts L, Che H, Brison N, Neofytou M, Jatsenko T, Lefrère H, et al. Breast Cancer Detection and Treatment Monitoring Using a Noninvasive Prenatal Testing Platform: Utility in Pregnant and Nonpregnant Populations. Clin Chem. 2020;66:1414–23.

3. Lenaerts L, Vandenberghe P, Brison N, Che H, Neofytou M, Verheecke M, et al. Genomewide copy number alteration screening of circulating plasma DNA: potential for the detection of incipient tumors. Ann Oncol. 2019;30:85–95.

4. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. Nature Publishing Group; 2019;570:385–9.

5. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. American Association for the Advancement of Science; 2021;372. Available from: http://science.sciencemag.org/content/372/6538/eaaw3616

6. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci. National Academy of Sciences; 2018;115:E10925–33.

7. Chan KCA, Jiang P, Sun K, Cheng YKY, Tong YK, Cheng SH, et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc Natl Acad Sci U S A. 2016;113:E8159–68.

8. Mouliere F, Robert B, Peyrotte EA, Rio MD, Ychou M, Molina F, et al. High Fragmentation Characterizes Tumour-Derived Circulating DNA. PLOS ONE. Public Library of Science; 2011;6:e23418.

9. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164:57–68.

10. Jiang P, Lo YMD. The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. Trends Genet. Elsevier; 2016;32:360–71.

11. Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci. National Academy of Sciences; 2015;112:E1317–25.

12. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. American Association for the Advancement of Science; 2018;10. Available from: https://stm.sciencemag.org/content/10/466/eaat4921

13. Buedts L, Wlodarska I, Finalet-Ferreiro J, Gheysens O, Dehaspe L, Tousseyn T, et al. The landscape of copy number variations in classical Hodgkin lymphoma: a joint KU Leuven and LYSA study on cell-free DNA. Blood Adv. 2021;5:1991–2002.

14. Bayindir B, Dehaspe L, Brison N, Brady P, Ardui S, Kammoun M, et al. Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. Eur J Hum Genet. 2015;23:1286–93.

15. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

16. Pons P, Latapy M. Computing Communities in Large Networks Using Random Walks. In: Yolum pInar, Güngör T, Gürgen F, Özturan C, editors. Comput Inf Sci - ISCIS 2005. Berlin, Heidelberg: Springer; 2005. page 284–93.

17. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. Nature Publishing Group; 2017;8:1324.

18. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, et al. Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing. Sci Transl Med. American Association for the Advancement of Science; 2012;4:162ra154-162ra154.

19. Ulz P, Belic J, Graf R, Auer M, Lafer I, Fischereder K, et al. Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. Nat Commun. 2016;7:12008.

20. Molparia B, Nichani E, Torkamani A. Assessment of circulating copy number variant detection for cancer screening. PLOS ONE. Public Library of Science; 2017;12:e0180647.

21. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31:745–59.

22. Sun K, Jiang P, Chan KCA, Wong J, Cheng YKY, Liang RHS, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci. 2015;112:E5503–12.

23. Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun. Nature Publishing Group; 2019;10:4666.

24. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. Nature Publishing Group; 2018;9:5068.

25. Sadeh R, Sharkia I, Fialkoff G, Rahat A, Gutin J, Chappleboim A, et al. ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin. Nat Biotechnol. Nature Publishing Group; 2021;1–13.

26. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol. Elsevier; 2021;0. Available from: https://www.annalsofoncology.org/article/S0923-7534(21)02046-9/abstract

27. Wan JCM, Heider K, Gale D, Murphy S, Fisher E, Mouliere F, et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. Sci Transl Med. American Association for the Advancement of Science; 2020;12. Available from: http://stm.sciencemag.org/content/12/548/eaaz8084

28. Moss J, Zick A, Grinshpun A, Carmon E, Maoz M, Ochana BL, et al. Circulating breast-derived DNA allows universal detection and monitoring of localized breast cancer. Ann Oncol. Elsevier; 2020;31:395–403.

29. Vanderstichele A, Busschaert P, Smeets D, Landolfo C, Nieuwenhuysen EV, Leunen K, et al. Chromosomal Instability in Cell-Free DNA as a Highly Specific Biomarker for Detection of Ovarian Cancer in Women with Adnexal Masses. Clin Cancer Res. American Association for Cancer Research; 2017;23:2223–31.

30. Kaijser J. Towards an evidence-based approach for diagnosis and management of adnexal masses: findings of the International Ovarian Tumour Analysis (IOTA) studies. Facts Views Vis ObGyn. 2015;7:42–59.

31. Cascione L, Aresu L, Baudis M, Bertoni F. DNA Copy Number Changes in Diffuse Large B Cell Lymphomas. Front Oncol. Frontiers; 2020;10. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2020.584095/full

32. Zhang J, Grubor V, Love CL, Banerjee A, Richards KL, Mieczkowski PA, et al. Genetic heterogeneity of diffuse large B-cell lymphoma. Proc Natl Acad Sci. 2013;110:1398–403.

33. Kelemen LE, Köbel M. Mucinous carcinomas of the ovary and colorectum: different organ, same dilemma. Lancet Oncol. 2011;12:1071–80.

34. Cheasley D, Wakefield MJ, Ryland GL, Allan PE, Alsop K, Amarasinghe KC, et al. The molecular origin and taxonomy of mucinous ovarian carcinoma. Nat Commun. Nature Publishing Group; 2019;10:3935.

35. Nishizuka S, Chen S-T, Gwadry FG, Alexander J, Major SM, Scherf U, et al. Diagnostic Markers That Distinguish Colon and Ovarian Adenocarcinomas. Cancer Res. 2003;63:5243–50.

36. Wyatt AW, Annala M, Aggarwal R, Beja K, Feng F, Youngren J, et al. Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer. JNCI J Natl Cancer Inst. 2017;109. Available from: https://doi.org/10.1093/jnci/djx118

37. Wei T, Zhang J, Li J, Chen Q, Zhi X, Tao W, et al. Genome-wide profiling of circulating tumor DNA depicts landscape of copy number alterations in pancreatic cancer with liver metastasis. Mol Oncol. 2020;14:1966–77.

515 38. Cresswell GD, Nichol D, Spiteri I, Tari H, Zapata L, Heide T, et al. Mapping the breast
516 cancer metastatic cascade onto ctDNA using genetic and epigenetic clonal tracking. Nat
517 Commun. Nature Publishing Group; 2020;11:1446.

518 39. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al.
519 Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. Nature
520 Publishing Group; 2017;545:446–51.

521 40. Kustanovich A, Schwartz R, Peretz T, Grinshpun A. Life and death of circulating cell-
522 free DNA. Cancer Biol Ther. Taylor & Francis; 2019;20:1057–67.

523

524 **Table 1. Participant and characteristics**

		Stage*	Age, Mean ± SD	Female, n (%)	Total Samples
Hematological cancer dataset					
	Healthy		69 (±3)	164 (63%)	**260**
	Hodgkin's lymphoma (HL)		32 (±14)	98 (55%)	**179**
		I			10
		II			145
		III			9
		IV			15
	Diffuse large B-cell lymphoma (DLBCL)		59 (±13)	22 (60%)	**37**
		I			1
		II			5
		III			7
		IV			8
		unknown			16
	Multiple myeloma (MM)		67 (±9)	8 (36%)	**22**
		I			3
		II			7
		III			7
		unknown			5
Solid tumor dataset					
	Healthy		49 (±12)	107 (91%)	**107**
	Breast		56 (±12)	46 (100%)	**46**
		I			23
		II			12
		III			5
		IV			6
	Colorectal		66 (±12)	29 (41%)	**70**
		I			19
		II			17
		III			25
		IV			9
	Gastrointestinal stromal tumor (GIST)		64 (±11)	**N.A.**	**35**
		Advanced			35
	Lung				**44**
		Advanced	**N.A.**	**N.A.**	44
	Ovarian invasive tumors		61 (±14)	125 (100%)	**125**
		I			25
		II			11
		III			49
		IV			31
		Metastatic			9
	Ovarian Benign		49 (±16)	160 (100%)	**160**
	Ovarian Borderline		51 (±17)	63 (100%)	**63**

525 ***** Multiple myeloma stratification refers to Revised International Staging System.

Figure Captions

Fig. 1. Schematic illustration of GIP*Xplore*. Plasma cfDNA in healthy individuals (blue box) comprises short nucleosome-protected DNA fragments mainly released from the cells of hematopoietic origin. In patients with cancer (green box), cfDNA is also released from the tumor. Since cfDNA fragmentation pattern is cell- or tissue-specific, sequencing and mapping of cfDNA from a patient with cancer may have differential genome-wide distribution of DNA fragments along the genome compared to a healthy one (green and blue profiles respectively). The workflow of GIP*Xplore* combines two tasks. First, explorative analysis of the high-dimensional data is performed via unsupervised clustering. Data complexity is reduced by using the first 50 linearly transformed genome-wide coverage features (non-trivial principal components, PCs) from a large number of cfDNA profiles, which are used for dataset exploration to unveil the potential biological signals or technical confounding factors based on the sub-grouping of underlying patterns that facilitate the design of the supervised models. Concurrently, classifiers are constructed to predict disease status and identify disease type to assess the use of such transformed genome-wide features as a marker for diagnostic application.

Fig. 2. Genome-wide cfDNA profiles carry cancer type-specific patterns. A, Two-dimensional tSNE visualization of the clustering result. Sample type is annotated by point color and community detection resulted clusters are annotated by point shape. Cluster numbers are labeled in the center of the defined cluster. **B,** Sample distribution in each community detection defined cluster is shown. The upper bar plot shows the total number of samples grouped in each cluster and the lower bar plot depicts the proportion of each class of samples. **C,** Tumor fraction estimated using ichorCNA. Red horizontal line indicates a detection limit of 3% tumor fraction level. **D,** Examples of copy number profiles generated from ichorCNA for selected

clusters. In each copy number profile, color red represents copy number gains and green represents copy number losses. The color is supposed to be interpreted together with the log ratio values to pinpoint copy number gains or losses.

Fig. 3. Plasma cfDNA genome-wide signatures enable hematological malignancies detection and subtype prediction. A, Sensitivities for detection of subtypes of hematological malignancies. Performance for early and advanced stages for DLBCL and HL are shown. Three (R-ISS) stages of MM are shown. 95% confidence interval is shown as an error bar. **B,** ROC curves for performance comparison between the genome-wide feature analysis and ichorCNA tumor fraction analysis. For the genome-wide feature analysis, decision value from SVM prediction is used to build a dynamic threshold of true and false positives. Tumor fraction values were used to construct ROC for ichorCNA analysis. **C,** Confusion matrix for tissue of origin detection in hematological tumor. The color shading represents the proportion of samples being correctly localized. The labeled numbers indicate the number of samples being classified into the class.

Fig. 4. Clustering analysis elucidates profile representations in solid tumors. **A,** Two-dimensional tSNE visualization of solid tumor dataset clustering result. Sample type is annotated by point color and community detection resulted clusters are annotated by point shape. Cluster numbers are labeled in the center of the defined cluster. **B,** Sample distribution in each community detection defined cluster is shown. The upper bar plot shows the total number of samples grouped in each cluster and the lower bar plot depicts the proportion of each class of samples. **C,** Tumor fraction estimation for indicated types of solid tumors. Red horizontal line indicates a detection limit of 3% tumor fraction level. **D,** Examples of copy number profiles generated from ichorCNA for selected clusters.

576

577 **Fig. 5. Malignancy detection and typing in solid tumors. A,** Sensitivities for detection of
578 different types of solid malignancies detection. Performance for detection of early and
579 advanced stages of disease is shown. **B,** ROC curves for performance comparison between the
580 genome-wide feature analysis and ichorCNA tumor fraction analysis. **C,** Confusion matrix for
581 tissue of origin detection in solid malignancies. The color shading represents the proportion of
582 samples being correctly localized. The labeled numbers indicate the number of samples being
583 classified into the class.

Figure 1.

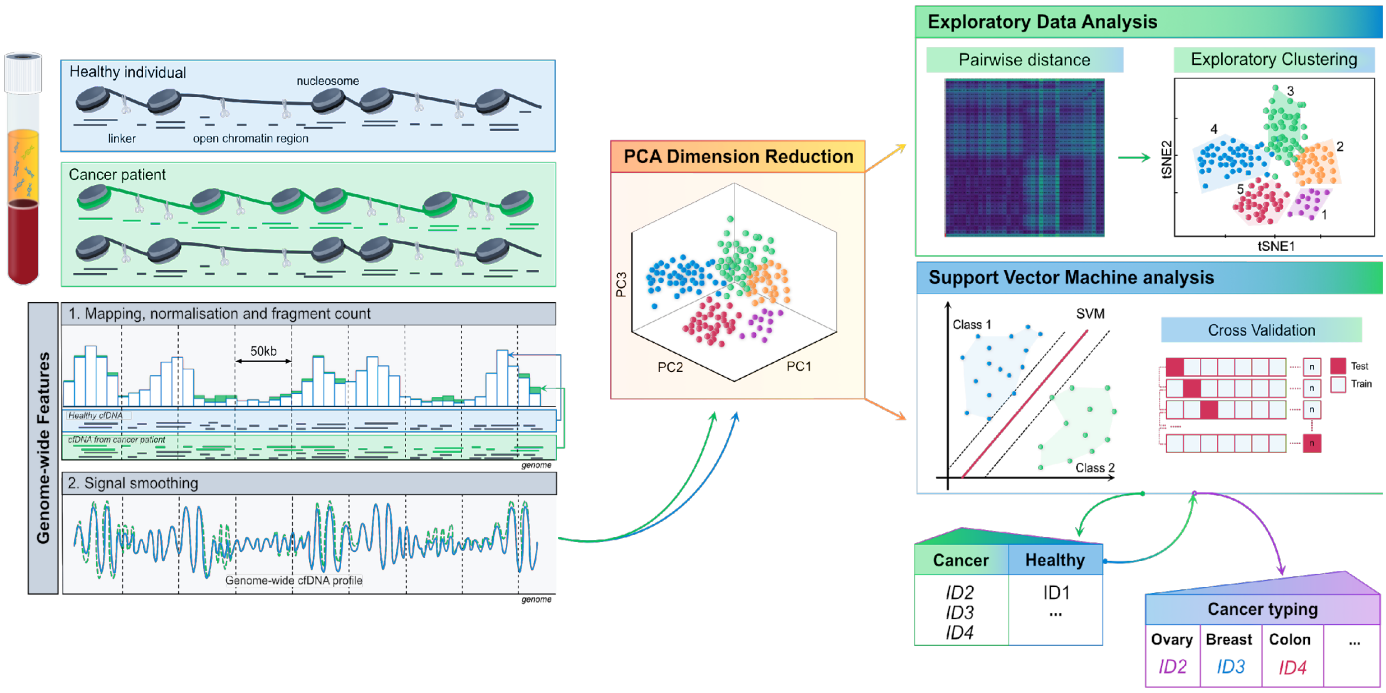

Healthy individual
nucleosome
linker
open chromatin region
Cancer patient
Genome-wide Features
1. Mapping, normalisation and fragment count
50kb
Healthy cfDNA
cfDNA from cancer patient
genome
2. Signal smoothing
Genome-wide cfDNA profile
genome
PCA Dimension Reduction
PC3
PC2
PC1
Exploratory Data Analysis
Pairwise distance
Exploratory Clustering
tSNE2
tSNE1
1
2
3
4
5
Support Vector Machine analysis
Class 1
SVM
Class 2
Cross Validation
n
Test
Train
Cancer
Healthy
ID2
ID3
ID4
ID1
...
Cancer typing
Ovary
Breast
Colon
...
ID2
ID3
ID4

Figure 2.

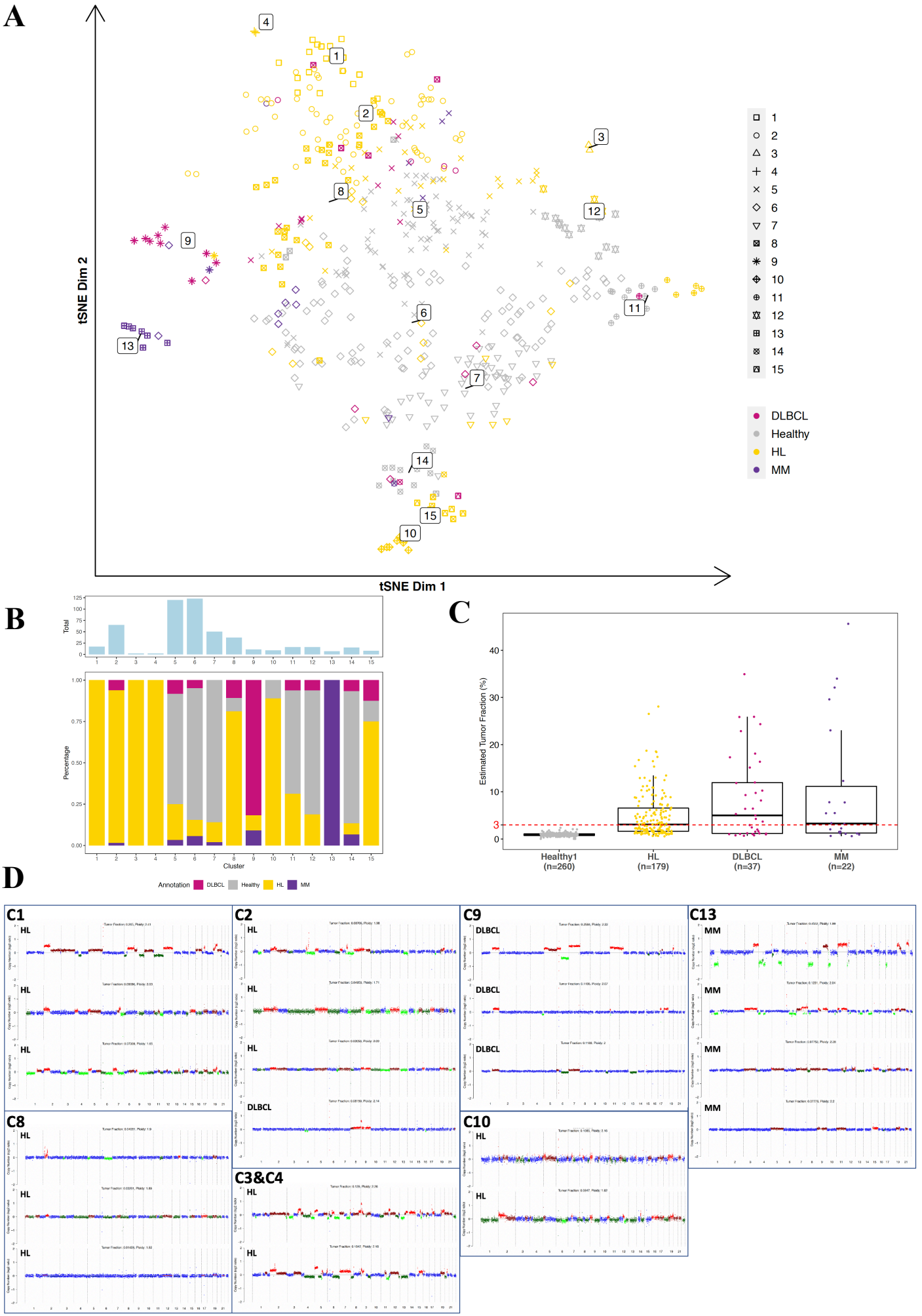
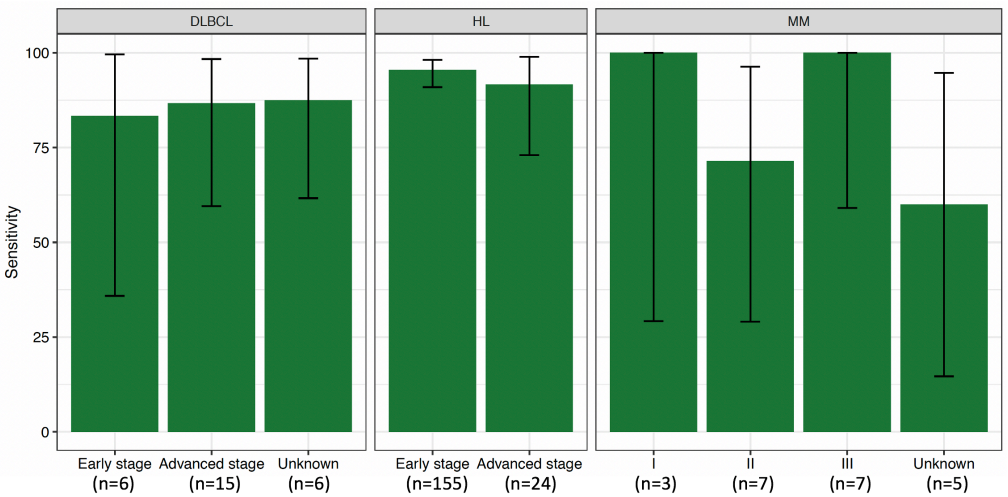
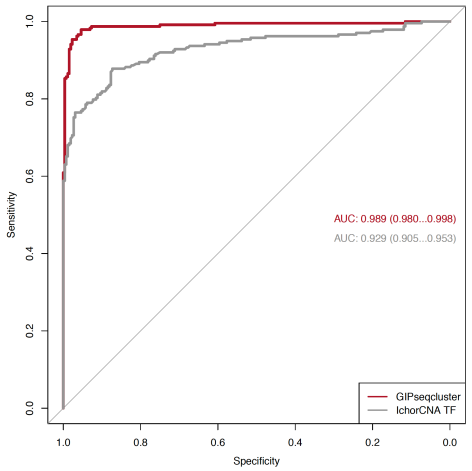

A
tSNE Dim 2
tSNE Dim 1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
DLBCL
Healthy
HL
MM
B
Total
Percentage
Cluster
Annotation
DLBCL
Healthy
HL
MM
C
Estimated Tumor Fraction (%)
Healthy1 (n=260)
HL (n=179)
DLBCL (n=37)
MM (n=22)
D
C1
C2
C9
C13
C8
C10
C3&C4

Figure 3.

A


DLBCL
HL
MM
Sensitivity
100
75
50
25
0
Early stage (n=6)
Advanced stage (n=15)
Unknown (n=6)
Early stage (n=155)
Advanced stage (n=24)
I (n=3)
II (n=7)
III (n=7)
Unknown (n=5)

B


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.989 (0.980...0.998)
AUC: 0.929 (0.905...0.953)
GIPseqcluster
IchorCNA TF

C

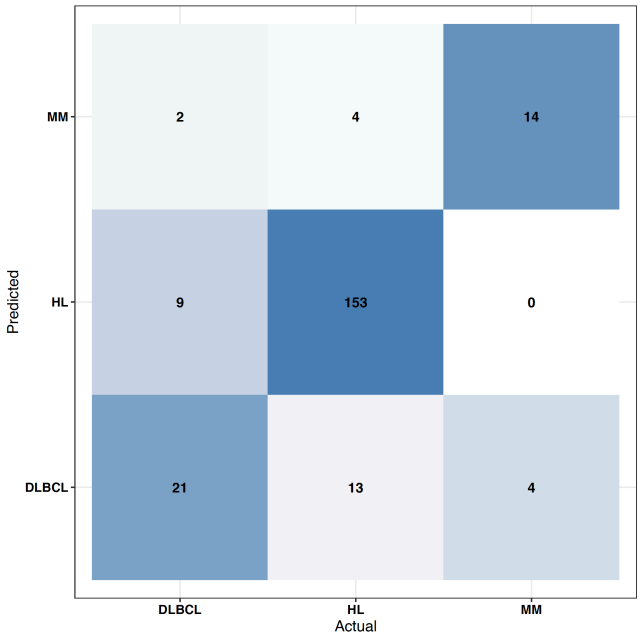

Predicted
Actual
MM
HL
DLBCL
2
4
14
9
153
0
21
13
4

Figure 4.

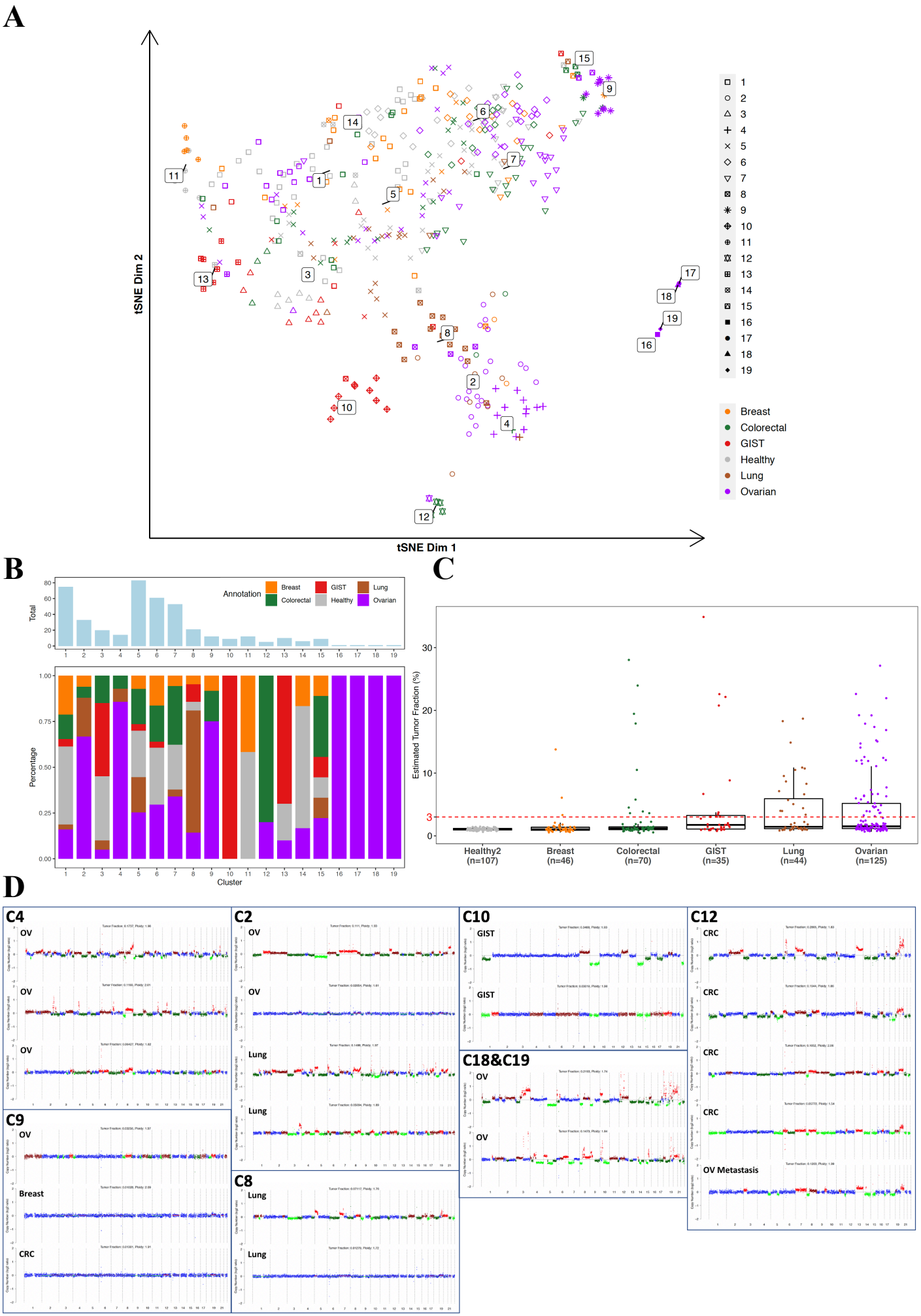
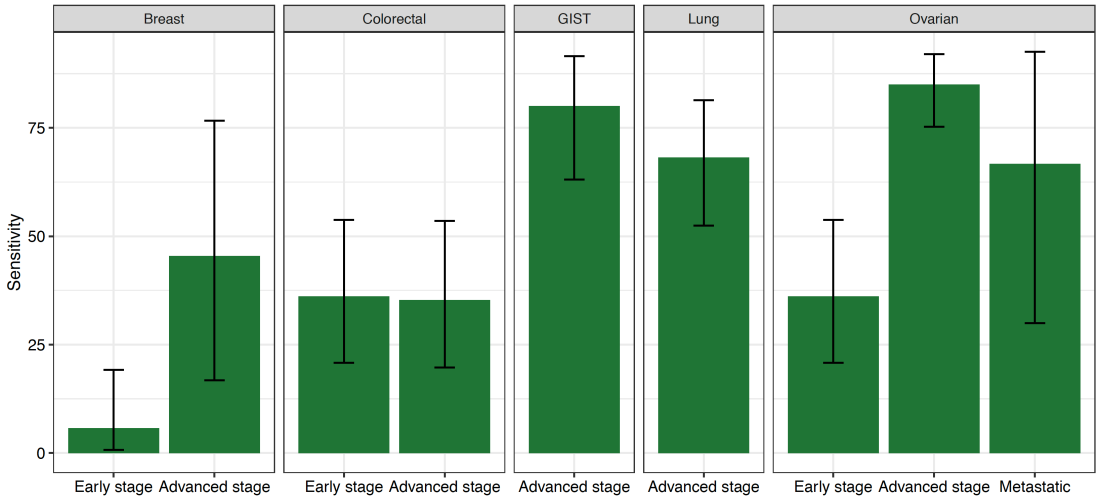
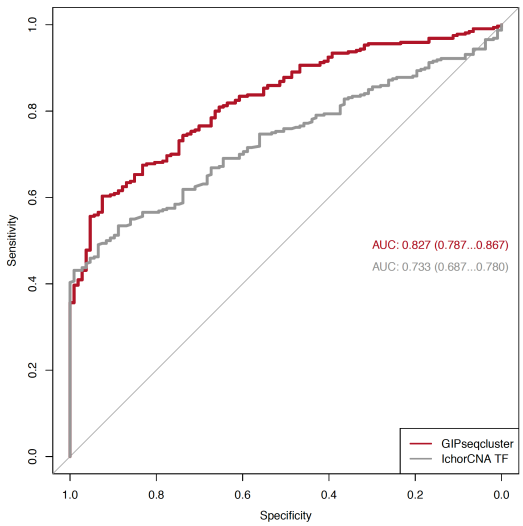

A
tSNE Dim 2
tSNE Dim 1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Breast
Colorectal
GIST
Healthy
Lung
Ovarian
B
Annotation
Total
Percentage
Cluster
C
Estimated Tumor Fraction (%)
Healthy2 (n=107)
Breast (n=46)
Colorectal (n=70)
GIST (n=35)
Lung (n=44)
Ovarian (n=125)
D
C4
C2
C10
C12
C9
C8
C18&C19
OV
Lung
GIST
CRC
Breast
OV Metastasis

Figure 5.

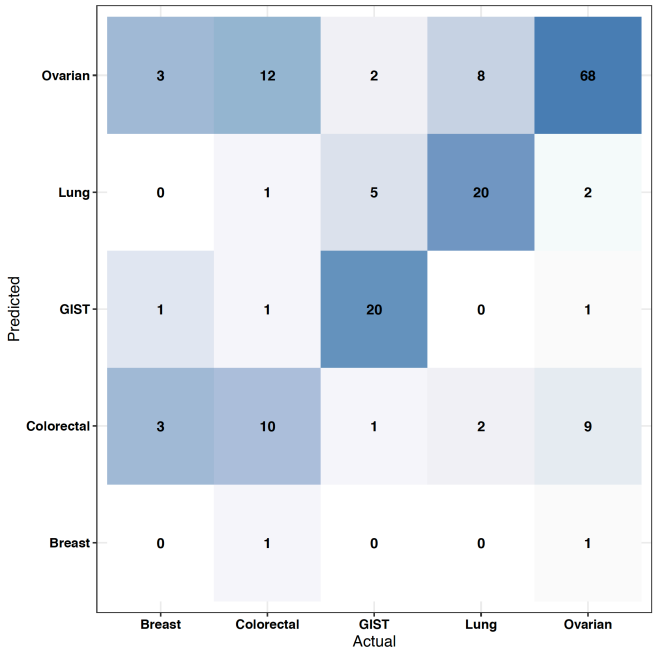
A


Breast
Colorectal
GIST
Lung
Ovarian
Sensitivity
0
25
50
75
Early stage
Advanced stage
Early stage
Advanced stage
Advanced stage
Advanced stage
Early stage
Advanced stage
Metastatic

B


Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.827 (0.787...0.867)
AUC: 0.733 (0.687...0.780)
GIPseqcluster
IchorCNA TF

C


Predicted
Ovarian
Lung
GIST
Colorectal
Breast
3 12 2 8 68
0 1 5 20 2
1 1 20 0 1
3 10 1 2 9
0 1 0 0 1
Breast
Colorectal
GIST
Lung
Ovarian
Actual