











Pan-Cancer Detection and Typing by Mining Patterns in Large Genome-Wide Cell-Free DNA Sequencing Datasets

Huiwen Che ^a Tatjana Jatsenko,^a Liesbeth Lenaerts,^b Luc Dehaspe,^c Leen Vancoillie,^c Nathalie Brison,^c Ilse Parijs,^c Kris Van Den Bogaert,^c Daniela Fischerova,^d Ruben Heremans ^e Chiara Landolfo,^f Antonia Carla Testa,^g Adriaan Vanderstichele,^f Lore Liekens,^h Valentina Pomella,^h Agnieszka Wozniak ⁱ Christophe Doms,^{j,k} Els Wauters,^{j,k} Sigrid Hatse,^{i,l} Kevin Punie ^{l,m} Patrick Neven,^{f,l} Hans Wildiers ^{l,m} Sabine Tejpar,^h Diether Lambrechts,ⁿ An Coosemans ^o Dirk Timmerman ^{e,f} Peter Vandenbergh ^{p,q} Frédéric Amant ^{b,f,r} and Joris Robert Vermeesch ^{a,c,*}

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

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

RESULTS: Unsupervised clustering revealed cancer type-specific sub-grouping. Classification using a supervised machine learning model yielded accuracies of 96% and 65% in discriminating hematological and solid malignancies from healthy controls, respectively. The accuracy of disease type prediction was 85% and 70% for the hematological and solid cancers, respectively. The

potential utility of managing a specific cancer was demonstrated by classifying benign from invasive and borderline adnexal masses with an area under the curve of 0.87 and 0.74, respectively.

CONCLUSIONS: This approach provides a generic analytical strategy for non-invasive pan-cancer detection and cancer type prediction.

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 patients with cancer (1, 2). Using genome-wide cfDNA sequencing data to profile genomic imbalances, we previously reported that CNAs in the asymptomatic population were indicative of incipient tumors and had potential as a cancer screening tool (3).

^aDepartment of Human Genetics, Laboratory for Cytogenetics and Genome Research, KU Leuven, Leuven, Belgium; ^bDepartment of Oncology, Laboratory of Gynecological Oncology, KU Leuven, Leuven, Belgium; ^cCentre for Human Genetics, University Hospitals Leuven, Leuven, Belgium; ^dDepartment of Obstetrics and Gynaecology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Prague, Czech Republic; ^eDepartment of Development and Regeneration, Woman and Child, KU Leuven, Leuven, Belgium; ^fDepartment of Gynecology and Obstetrics, University Hospitals Leuven, Leuven, Belgium; ^gDepartment of Woman and Child Health, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Università Cattolica del Sacro Cuore Roma, Rome, Italy; ^hDepartment of Oncology, Molecular Digestive Oncology, KU Leuven, Leuven, Belgium; ⁱDepartment of Oncology, Laboratory of Experimental Oncology, KU Leuven, Leuven, Belgium; ^jDepartment of Chronic Diseases and Metabolism, Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), KU Leuven, Leuven, Belgium; ^kDepartment of Pneumology, University Hospitals Leuven, Leuven, Belgium; ^lMultidisciplinary Breast Centre, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium;

^mDepartment of General Medical Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium; ⁿDepartment of Human Genetics, Laboratory of Translational Genetics, VIB-KU Leuven, Leuven, Belgium; ^oDepartment of Oncology, Laboratory of Tumor Immunology and Immunotherapy, Leuven Cancer Institute, KU Leuven, Leuven, Belgium; ^pDepartment of Human Genetics, Laboratory of Genetics of Malignant Diseases, KU Leuven, Leuven, Belgium; ^qDepartment of Hematology, University Hospitals Leuven, Leuven, Belgium; ^rDepartment of Surgery, Center for Gynecological Oncology Amsterdam, Academic Medical Centre Amsterdam-University of Amsterdam and the Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands.

*Address correspondence to this author at: Katholieke Universiteit Leuven (KU Leuven) Universitaire Ziekenhuizen Leuven (University Hospitals Leuven), Human Genetics, Herestraat 49, Box 602, 3000 Leuven, Belgium. Fax +32 16 34 60 60; E-mail: joris.vermeesch@uzleuven.be.

Received December 12, 2021; accepted April 25, 2022.
<https://doi.org/10.1093/clinchem/hvac095>

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 mainly result from apoptotic release of DNA from the cells of hematopoietic origin (10). In plasma of patients with cancer, 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 could be associated with different pathological or physiological states. To test this hypothesis, we applied an unsupervised clustering analysis and supervised machine learning workflow, which we termed *GIPXplore*, on a large number of genome-wide sWGS cfDNA profiles from patients with different hematological or solid malignancies.

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 patients with cancer. Blood was collected either into Streck Cell-Free DNA BCT or Roche Cell-Free DNA collection tubes. Plasma was isolated through a standard centrifugation procedure. Previously published sequencing data from 260 healthy subjects (3) and 177 patients with Hodgkin lymphoma (13) were included in the study.

SWGS ANALYSIS

cfDNA was extracted from plasma using standard processing procedures and sWGS sequencing (14) (see [online Supplemental Material](#)). Each sample contained 57 509 autosome bin features—normalized and smoothed bin read counts from the standard processing. Principal component analysis (PCA) was used for dimension reduction to transform these bin features 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. Since 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.

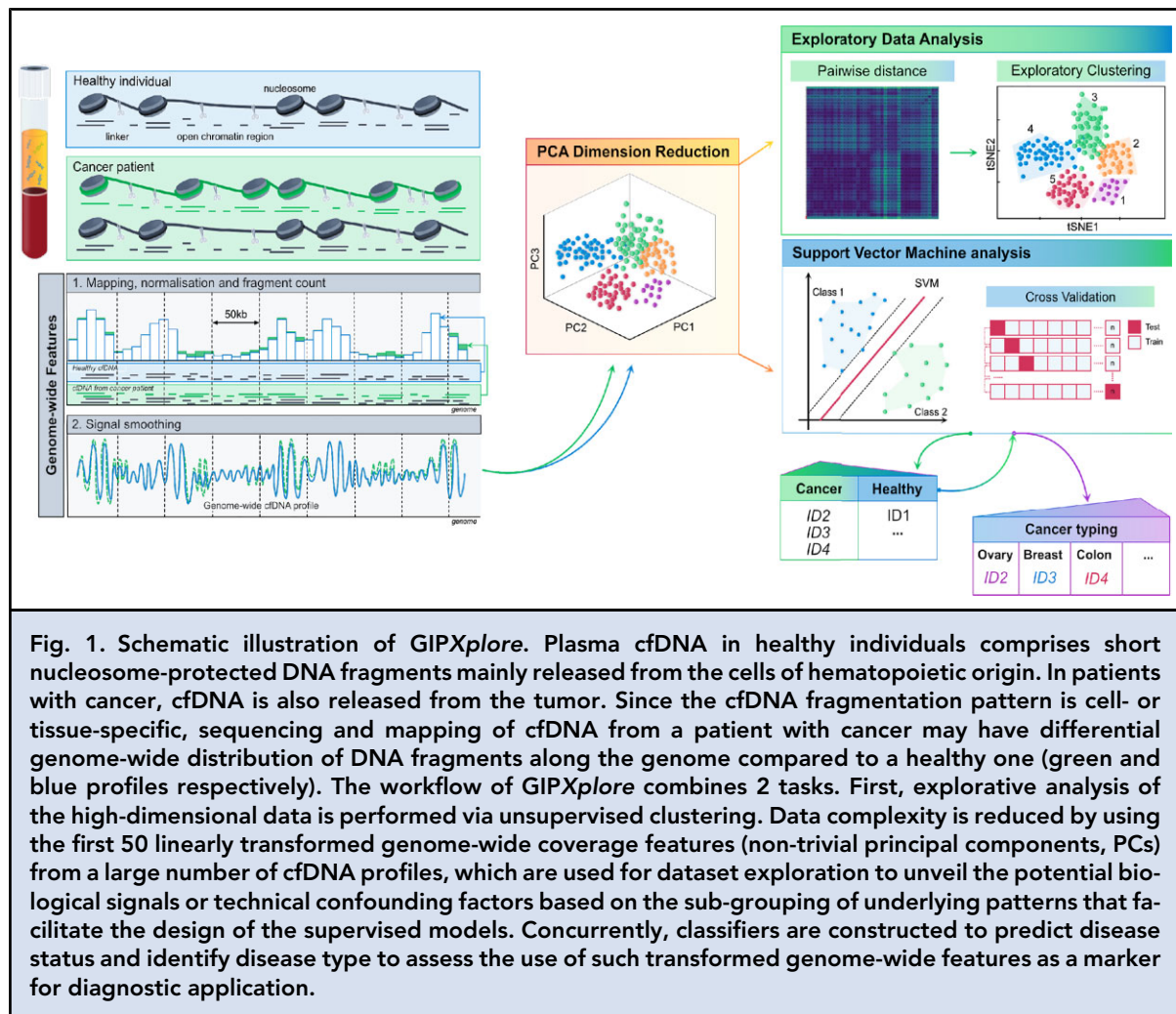
GIPXPLORE

A schematic illustration of *GIPXplore* for mining sWGS cfDNA data for identification of signatures is provided in [Fig. 1](#). We utilized unsupervised clustering and supervised machine learning. For unsupervised clustering, we evaluated the variance being explained from principal components in the tumor data. Overall, the top 30, 50, and 100 principal components explained above 80%, 85%, and 90% of the variance in the data, respectively. While there is no absolute optimal number of principal components to be used for further analysis, 50 non-trivial principal components ([online Supplemental Material](#)) 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. A 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 2 (or 3) 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 a parameter perplexity of 15/30 and the number of iterations to 10 000 with exact tSNE for accuracy, and the process was repeated 10 times with different seeds. For Walktrap, we used the parameters from 8 initial numbers of neighbors' searches 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 the PCA space of training data for classification tasks. We measured performance by repeating the 10-fold cross validation 10 times and leave-one-out (LOO) procedures. For cross validation, the receiver operating characteristic (ROC) curve and performance were calculated using the mean of 10 repeats. For classifiers, we used a support vector machine 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

GIPXPLORE DETECTS AND CLASSIFIES HEMATOLOGICAL MALIGNANCIES WITH HIGH ACCURACY

To assess the potential for identifying cancer signals in sWGS data, we applied our method on a set of



cfDNA samples from healthy controls ($n = 260$) and patients with hematological malignancies that included Hodgkin lymphoma (HL; $n = 179$), diffuse large B-cell lymphoma (DLBCL; $n = 37$), and 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. Clusters 1, 3, and 4 were 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 TF. Overall, only 52.95% of hematological cancer samples had a TF higher than 3%—the detection limit previously suggested for ichorCNA for accurate detection of the tumor presence (17) (Fig. 2, C, Supplemental Fig. 2, and Supplemental Table 1). The abovementioned clusters 1, 3, and 4 consisted of profiles characterized by large chromosomal aberrations and high tumor load. Clusters 2 and 8 consisted of profiles mainly from patients with HL with both high and low TF, implying that the clustering was not completely CNA-driven. In particular, 10 out of 65 (15%) 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% (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

Table 1 Participants and characteristics.

	Stage ^a	Age, mean (SD)	Female, n (%)	Total samples
Hematological cancer dataset				
Healthy		69 (3)	164 (63)	260
Hodgkin lymphoma		32 (14)	98 (55)	179
	I			10
	II			145
	III			9
	IV			15
Diffuse large B-cell lymphoma		59 (13)	22 (60)	37
	I			1
	II			5
	III			7
	IV			8
	unknown			16
Multiple myeloma		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
GIST tumor		64 (11)	NA	35
	Advanced			35
Lung				44
	Advanced	NA	NA	44
Ovarian invasive tumors		61 (14)	125 (100)	125
	I			25
	II			11
	III			49
	IV			31
	Metastatic			9
				Continued

Table 1 (continued)

Stage ^a	Age, mean (SD)	Female, n (%)	Total samples
Ovarian benign	49 (16)	160 (100)	160
Ovarian borderline	51 (17)	63 (100)	63

^aMultiple myeloma stratification refers to Revised International Staging System.

ratio variations and were more likely to be noise 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 (online 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 <3% TF. The separation between some malignant and healthy samples still remained (online 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 LOO and repeated 10-fold cross validation were used to assess the performance of the classifier. Incorporating transformed genome-wide features, the support vector machine learning model correctly classified 220 (out of 238) malignant cases in LOO analysis, at a clinical sensitivity of 92% (95% CI, 88%–95%) and a clinical specificity of 98% (95% CI, 96%–100%), 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 (online Supplemental Fig. 6). The detection sensitivity was the highest for HL (Supplemental Table 1). The clinical 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 yielded an AUC value of 0.99 (95% CI, 0.98–1) in distinguishing

malignant from healthy samples, compared to ichorCNA TF-based analysis, which had an AUC of 0.93 (Fig. 3, B). Repeated 10-fold cross validation also revealed a stable performance at a mean AUC of 0.99 (online Supplemental Fig. 7). Since the clustering analysis demonstrated the co-localization of samples originating from the same cancer type, we then attempted to determine the accuracy of our GIPX_{pl}ore in cancer type classification. For this purpose, we trained the classification model using the 220 correctly predicted malignant samples from the LOO analysis. The analysis showed an overall accuracy of 85% (95% CI, 80%–90%), with the highest accuracy in HL prediction (Fig. 3, C and Supplemental Table 2). Consistent with the exploratory clustering analysis, where some of the cfDNA profiles from patients with DLBCL colocalized together with those from patients with HL, DLBCL samples were more likely to be misclassified.

GIPXPLORE 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 patients with cancer, and a set of 107 cfDNA profiles from healthy controls. The malignant cohort was represented by 5 tumor types: breast (n = 46), colorectal (n = 70), gastrointestinal stromal tumor (GIST; n = 35), lung (n = 44), and ovarian (n = 125; Table 1). Using GIPX_{pl}ore, 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 enriched with a particular cancer type, 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 12 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 tumor types. Clusters 9 (mostly ovarian cancer) and 15 (intermixed cancer types) deviated from healthy and other malignant

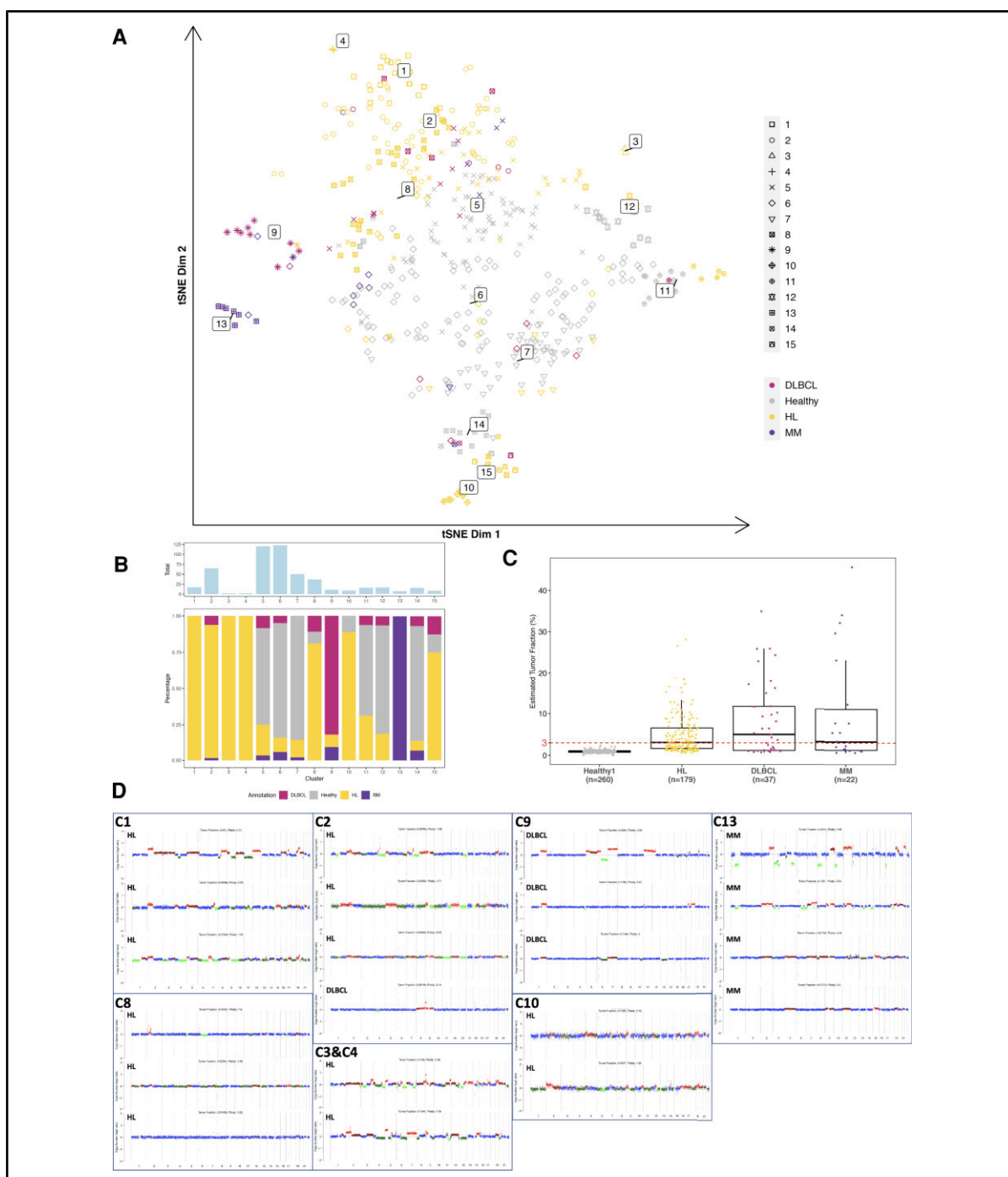
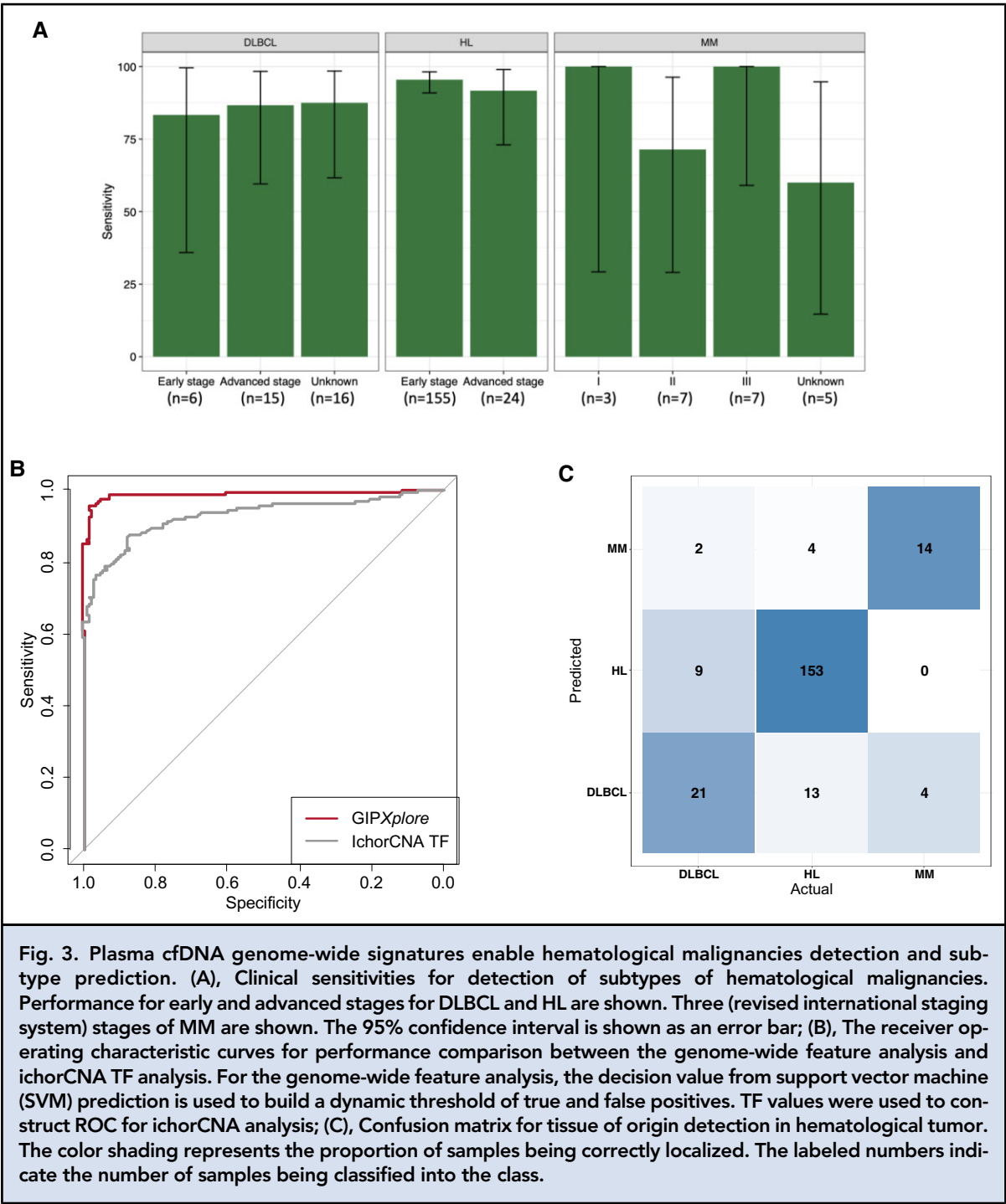
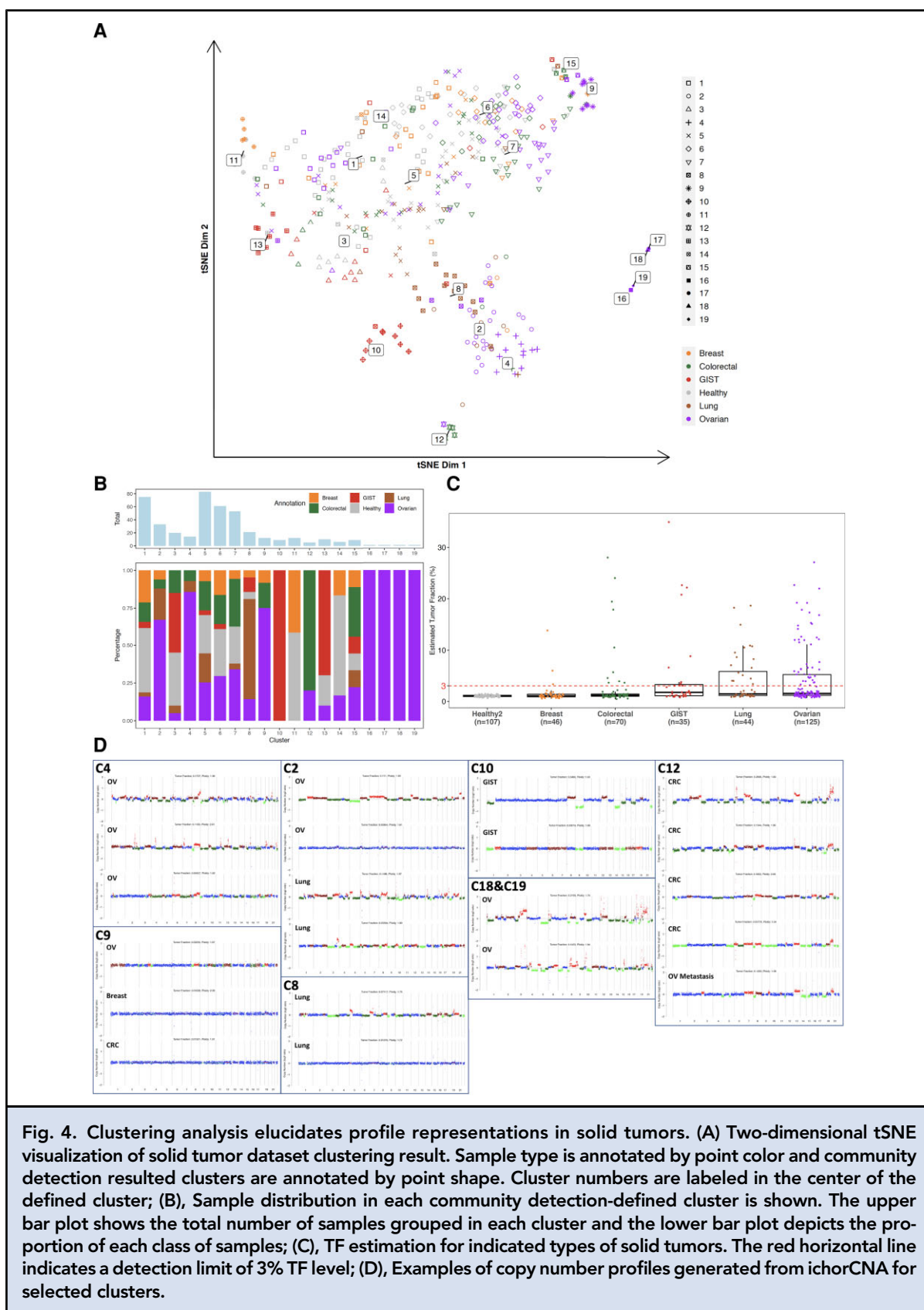


Fig. 2. Genome-wide cfDNA profiles carry cancer type-specific patterns. (A), Two-dimensional tSNE visualization of the clustering result. Sample type is annotated by point color and community detection-resulted clusters are annotated by point shape. Cluster numbers are labeled in the center of the defined 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), TF estimated using ichorCNA. The 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, 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.



clusters. The majority of the cfDNA profiles from patients with breast cancer 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 fractions estimated by ichorCNA were generally lower in the solid malignant cohort (Fig. 4, C). TF varied among different types of cancer and increased with the stage (online Supplemental Fig. 9). The malignant cases with detectable CNAs and therefore higher TF



were more likely to separate from the healthy controls (online [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 4 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 (online [Supplemental Fig. 11](#)). When restricting the clustering analysis to samples with TF <3%, samples from clusters 9 and 15 still showed deviations from normal profiles (online [Supplemental Fig. 12](#), clusters 8 and 9).

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 support vector machine model, with performance estimated by LOO and repeated 10-fold cross validation. With an overall accuracy of 65%, we correctly detected 177 out of 320 cancer profiles (55% clinical sensitivity, 95% CI, 50%–61%), at a clinical specificity of 95%. Performance in individual tumor types ranged from 15% (95% CI, 6%–29%) for classifying breast cancer to 80% (95% CI, 63%–92%) for GIST (see online [Supplemental Table 3](#)). Stage of the disease affected the detection, with a clinical sensitivity of 26% (95% CI, 18%–36%) in the early stage (I–II) vs 70% (95% CI, 63%–76%) 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 clinical sensitivities were almost the same for early and advanced cancer stages. Misclassified malignant samples had low TF, which potentially restricted the detection of underlying tumor-specific patterns (online [Supplemental Fig. 13](#)). We could distinguish malignancy from healthy samples with an AUC of 0.83 (95% CI, 0.79–0.87), which again was superior to ichorCNA TF-based analysis (0.73 AUC; 95% CI, 0.69–0.78; [Fig. 5, B](#) and [Supplemental Fig. 14](#)). Subsequently, we explored the potential of our GIPXplore 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% (95% CI, 61%–76%) overall accuracy. Highest clinical sensitivities (>70%)

were obtained for cfDNA samples from patients with ovarian cancer and GIST. 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](#)).

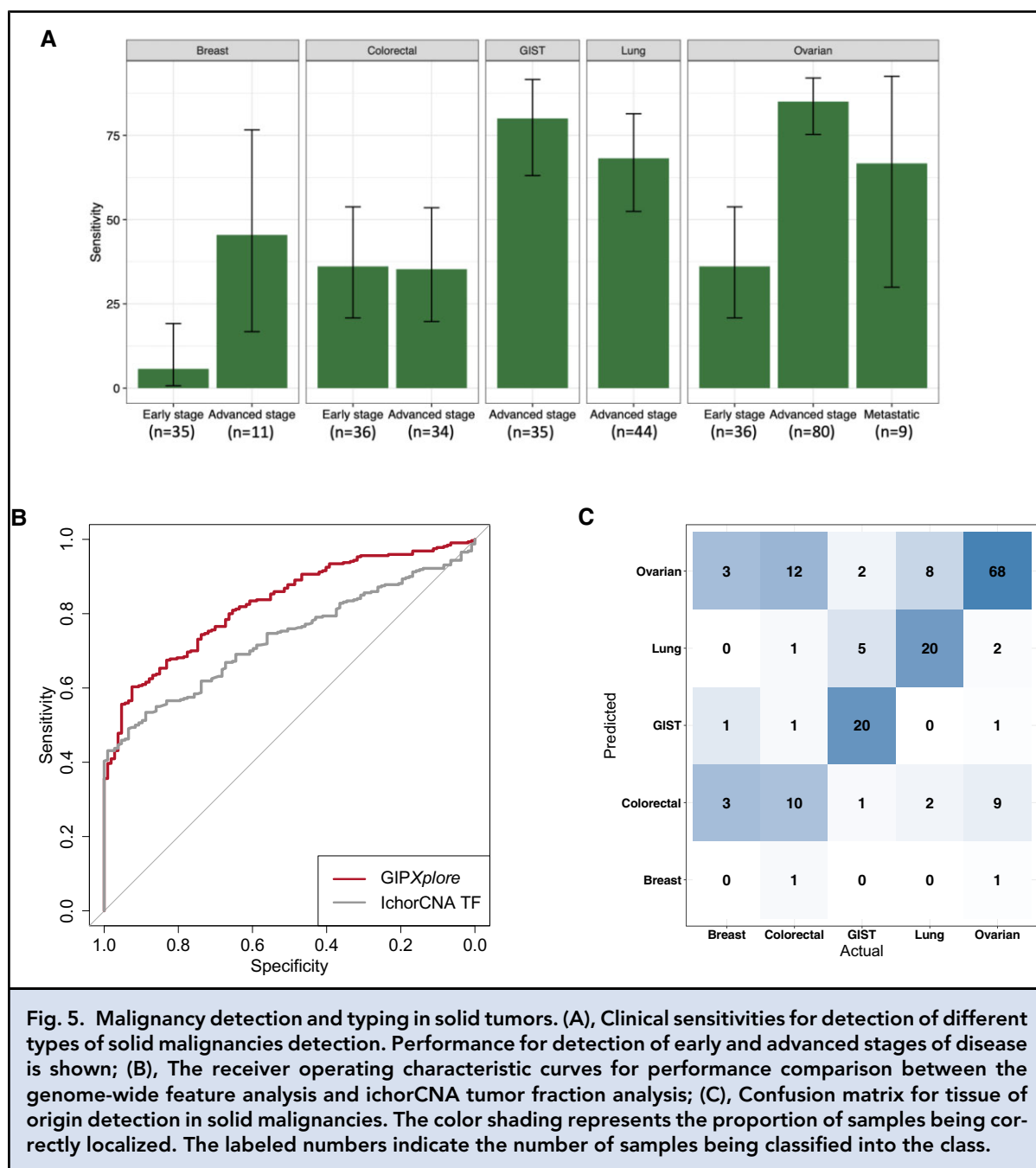
Moreover, among this solid malignant cohort, we had 9 cfDNA samples from patients with ovarian metastases, of which 4 patients had gastrointestinal primary sites, 1 lymphoma, 1 leiomyosarcoma, 1 uterine origin, and the remaining 2 had Krukenberg tumors. Annotation of these 9 cases on the tSNE plot showed that metastatic profiles could resemble profiles of either the primary tumor or the distant site (online [Supplemental Fig. 15, A](#)). Applying the type-specific classifier to the 6 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 3 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 2 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 (online [Supplemental Fig. 16](#)). In the clustering analysis, 35 invasive samples formed a distinct group in cluster 1. In clusters 6 to 9, common patterns were found for invasive, benign and borderline samples, although they remained distinct from controls (online [Supplemental Fig. 17](#)). The classification analysis exhibited an AUC of 0.85 (95% CI, 0.80–0.90) in discriminating benign from invasive samples, and an AUC of 0.74 (95% CI, 0.69–0.79) in discriminating benign from borderline and invasive samples (online [Supplemental Figs. 18 and 19](#)).

Discussion

We present here 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 chromatin immunoprecipitation of cell-free nucleosomes sequencing 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.

One prior study has 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 clinical sensitivity of 54.9% and at a clinical specificity of 99% (21). This test was refined and validated in an independent follow-up study, with an overall clinical sensitivity of 51.5% at 99.5% clinical specificity reported (26). We estimate the clinical sensitivity of 92% and 55% at above 95% clinical specificity for the hematological and solid cancer cohorts, respectively. Performance for cancer signal detection varied among the different cancer types and stages. While hematological malignancies, with tumor cells being in direct contact with blood, were more likely to be detected, the prediction accuracy of solid malignancies was lower. Though with a high proportion (72%) of early-stage diseases, the hematological malignancies showed higher overall tumor fractions and higher dispersion of cfDNA profiles. The estimated TFs of solid malignancies were lower. In addition, distributions of early and advanced stages were unequal in the solid tumor cohort: GIST and lung had only advanced cancers and showed higher clinical sensitivity. Early-stage solid malignancies had lower performance, and breast cancer, having 76% of the early-stage disease, showed the lowest clinical sensitivity among all tumor types in the study. Shedding of the ctDNA from breast cancer is known to be low (27, 28). Apart from potential screening applications, we also demonstrated that GIPXplore 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 the 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 2 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 their 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 sub-groups, 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.

GIPXplore provides an unbiased genome-wide scan of cfDNA profiles. However, it also has some limitations. Inspection of the data using clustering (online Supplemental Fig. 20, A) revealed that pre-analytical factors, i.e., usage of different library preparation kits, could result in cluster separations (Supplemental Fig. 20, B). Hence, to avoid the bias, we analyzed the data of hematological and solid tumor cohorts separately. Other potential pre-analytical factors, including sequencing batch, age, and sex, did not indicate an apparent confounding effect on the clustering of cfDNA samples (Supplemental Fig. 20, C–E). Increasing the sequencing depth might improve detection of disease-specific cfDNA patterns and improve the clinical 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. A higher proportion of advanced stage cancer samples in the solid cancer cohort may also skew the performance estimation. Further investigation on an independent cohort is required to evaluate the test performance and utility of such analysis in asymptomatic populations. 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 conclusion, we have extended the scope of cfDNA analysis, allowing affordable 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.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: cfDNA, cell-free DNA; sWGS, shallow whole-genome sequencing; CNA, copy number aberration; PCA, principal component analysis; tSNE, *t*-distributed stochastic neighbor embedding; LOO, leave-one-out; HL, Hodgkin lymphoma; DLBCL, diffuse large B-cell lymphoma; MM, multiple myeloma; TF, tumor fraction; AUC, area under the curve; GIST, gastrointestinal stromal tumor.

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 Material](#).

Author Declaration: A version of this article was previously posted as a preprint on medRxiv as <https://www.medrxiv.org/content/10.1101/2022.02.16.22268780v1>.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.

H. Che, T. Jatsenko, L. Lenaerts, L. Dehaspe, F. Amant, and J.R. Vermeesch conceptualized and designed the study. K. Punie, A. Wozniak, E. Wauters, A. Coosemans, P. Neven, S. Tejpar, D. Timmerman, P. Vandenberghe, H. Wildiers, and F. Amant provided

clinical samples and patient data. L. Lenaerts, L. Vancoillie, N. Brison, I. Parijs, K. Van Den Bogaert, C. Dooms, D. Fischerova, R. Heremans, S. Hatse, C. Landolfo, L. Liekens, V. Pomella, A.C. Testa, A. Vanderstichele, and A. Wozniak carried out clinical sample procurement and processing. L. Vancoillie, N. Brison, I. Parijs, and K. Van Den Bogaert coordinated sequencing of cfDNA. T. Jatsenko and L. Lenaerts conducted project coordination and administration. H. Che and L. Dehaspe performed bioinformatics analysis of sWGS data. H. Che, T. Jatsenko, L. Lenaerts, L. Dehaspe, K. Punie, A. Wozniak, E. Wauters, A. Coosemans, D. Lambrechts, S. Tejpar, D. Timmerman, P. Vandenberghe, F. Amant, and J.R. Vermeesch contributed to the interpretation of results. H. Che, T. Jatsenko, and J.R. Vermeesch wrote the manuscript; all co-authors reviewed the manuscript.

Authors' Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.

Consultant or Advisory Role: A. Coosemans is a contracted researcher for Oncinvent AS and Novocure and a consultant for Sotio a.s.

Stock Ownership: None declared.

Honoraria: None declared.

Research Funding: Research Foundation-Flanders (FWO-Vlaanderen; G080217N to F. Amant and J.R. Vermeesch, G0A1116N to P. Vandenberghe), Agentschap Innoveren en Ondernemen (VLAIO; Flanders Innovation & Entrepreneurship grant HBC.2018.2108 to T. Jatsenko), Kom Op Tegen Kanker (Stand Up to Cancer, the Flemish Cancer Society under grant 2016/10728/2603 to A. Coosemans), Stichting tegen Kanker (FAF-C/2016/836 to P. Vandenberghe, 2018-134 to J.R. Vermeesch and F. Amant), EASI Genomics (ZL50015800 to J.R. Vermeesch), and KU Leuven funding (no C1/018 to J.R. Vermeesch and D. Lambrechts, C3/20/100 to J.R. Vermeesch).

Expert Testimony: None declared.

Patents: Patent application pending on 'Method for analyzing cell-free nucleic acids' number 1707735.5 (J.R. Vermeesch and L. Dehaspe).

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Acknowledgments: 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, Amin Ardeshirdavani and Baran Bayindir for their involvement in mapping the clinical potential, and Kate Stanley for helpful suggestions for the manuscript.

References

- Vandenberghe P, Wlodarska I, Tousseyn T, Dehaspe L, Dierickx D, Verhecke 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–e65.
- 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.
- Lenaerts L, Vandenberghe P, Brison N, Che H, Neofytou M, Verhecke 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.
- 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* 2019;570:385–9.
- Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* 2021;372:eaaw3616.
- 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 USA* 2018;115:E10925–33.
- 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 USA* 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* 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* 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 USA* 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* 2018;10: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. van der Maaten L, 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 P, Güngör T, Gürgeç F, and Özturan C, editors. *Computer and Information Sciences – ISICS 2005*. Heidelberg (Germany): Springer Berlin; 2005. p. 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* 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* 2012;4:162ra154.
 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* 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 USA* 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* 2019;10: 4666.
 24. Moss J, Magenheimer 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* 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* 2021;39:586–98.
 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* 2021;32:1167–77.
 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* 2020;12: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* 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* 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* 2020;10: 584095.
 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 USA* 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* 2019;10:3935.
 35. Nishizuka S, Chen S-T, Gwady FG, Alexander J, Major SM, Scherf U, et al. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. *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. *J Natl Cancer Inst* 2017;109: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.
 38. Cresswell GD, Nichol D, Spiteri I, Tari H, Zapata L, Heide T, et al. Mapping the breast cancer metastatic cascade onto ctDNA using genetic and epigenetic clonal tracking. *Nat Commun* 2020;11: 1446.
 39. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* 2017;545:446–51.
 40. Kustanovich A, Schwartz R, Peretz T, Grinshpun A. Life and death of circulating cell-free DNA. *Cancer Biol Ther* 2019;20: 1057–67.