Transcriptomic Profiling of Plasma Extracellular Vesicles Enables Reliable Annotation of the Cancer-specific Transcriptome and Molecular Subtype

Vahid Bahrambeigi[1,2]*, Jaewon J. Lee[1,2,3,8]*#, Vittorio Branchi[1,2], Jason T. Henry[5], Kimal I. Rajapakshe[1,2], Bret M. Stephens[1,2], Sarah Dhebat[1,2], Mark W. Hurd[1], Ryan Sun[4], Peng Yang [5,9], Wenyi Wang[5], Scott E. Kopetz[6], Anirban Maitra[1,2,7], Paola A. Guerrero[1,2]#

*equal contribution

#corresponding author

[1]Sheikh Ahmed Center for Pancreatic Cancer Research

[2]Department of Translational Molecular Pathology

[3]Department of Surgical Oncology

[4]Department of Biostatistics

[5]Department Bioinformatics and Computational Biology

[6]Department of GI Medical Oncology

[7]Department of Pathology

The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[8]Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[9]Department of Statistics Rice University, Houston, TX, USA

Correspondence

Paola A. Guerrero, PhD

Department of Translational Molecular Pathology

Sheikh Ahmed Center for Pancreatic Cancer Research

The University of Texas MD Anderson Cancer Center

6565 MD Anderson Blvd

Houston, TX 77030

Email: paguerrero@mdanderson.org

Jaewon J. Lee, MD

Department of Surgery

Cedars-Sinai Medical Center

8700 Beverly Blvd.

Los Angeles, CA 90048

Email: Jaewon.Lee@cshs.org

Conflict of interest

A.M. receives royalties for a pancreatic cancer biomarker test from Cosmos Wisdom Biotechnology, and this financial relationship is managed and monitored by the UTMDACC Conflict of Interest Committee. A.M. is also listed as an inventor on a patent that has been licensed by Johns Hopkins University to ThriveEarlier Detection. A.M. serves as a consultant for Freenome and Tezcat Biotechnology. A.M. is supported by the MD Anderson Pancreatic Cancer

Moon Shot Program, the Sheikh Khalifa Bin Zayed Al-Nahyan Foundation and NIH (U01CA196403 and U01CA200468). S.E.K. and A.M. are supported by NIH (R01CA218230, P50CA221707 and P30CA016672), P.A.G. is supported by Break Through Cancer, J. J. Lee by NIH (T32CA009599), R. S. is supported by NIH (5R01CA218230 and 5P50CA221707) and has received consulting fees from Boehringer Ingelheim and W. W. is supported by NIH (R01CA239342 and R01CA268380).

Acknowledgments

We thank our patients and their families.

Abstract

Longitudinal monitoring of patients with advanced cancers is crucial to evaluate both disease burden and treatment response. Current liquid biopsy approaches mostly rely on the detection of DNA-based biomarkers. However, plasma RNA analysis can unleash tremendous opportunity for tumor state interrogation and molecular subtyping. Through the application of deep learning algorithms to the deconvolved transcriptomes of plasma extracellular vesicles (EVs) RNA, we successfully predict consensus molecular subtypes in metastatic colorectal cancer (mCRC) patients and analogous RNA-based subtypes in patients with pancreatic ductal adenocarcinoma (PDAC). We further demonstrate the ability monitor changes in transcriptomic subtype under treatment selection pressure. Our approach also identified gene fusions and neoepitopes from plasma EVs, which were validated in matched tissue samples. These results demonstrate the feasibility of transcriptomic-based liquid biopsy platforms for precision oncology approaches, spanning from the longitudinal monitoring of tumor subtype changes to identification of expressed fusions and neoantigens as potential therapeutic targets, *sans* the need for tissue-based sampling.

Statement of significance

Direct analysis of tumor-encoded transcriptomes in liquid biopsies can unveil new opportunities for cancer detection, classification, and identification of therapeutic targets. We developed an approach to interrogate tumor molecular subtypes and their changes, as well as gene fusions and neoepitopes through the analysis of bulk RNA transcriptomic data from plasma EVs.

Introduction

Liquid biopsies can detect cancer-derived materials without the need to perform an invasive tissue biopsy. Most liquid biopsy assays designed for therapeutic assignment and response monitoring in cancer are currently based on circulating cell-free DNA (cfDNA), such as the detection of mutations, fragmentation patterns or methylation signatures [1-4]. Recent studies have made initial forays into inferring tumor transcriptomes indirectly via the use of fragmented cfDNA profiles in plasma samples, and identifying pertinent biological pathways based on the anatomical location of the primary cancer [5, 6]. Nonetheless, the direct interrogation of tumor transcriptomes in plasma could enable significant insights into the cancer molecular landscape beyond what genomic alterations alone would provide. The comprehensive assessment of cancer-derived RNA in plasma has been challenging due to the presence of circulating RNAases, which cleave coding RNAs, confounding their assessment with next generation sequencing (NGS) platforms, and restricting most studies of this nature to the non-coding transcriptome [7, 8]. A second major challenge of a circulating RNA-based approach is the reliable separation of tumor-derived transcriptomic signatures from the "background" RNA released by non-neoplastic cell types. Among liquid biopsy substrates, extracellular vesicles (EVs) represent a facile compartment for querying tumor transcriptomes as they are actively released from cancer cells and contain intact RNA cargoes (evRNA) [9]. While we and others have previously demonstrated the ability to map the DNA landscapes of advanced cancers reliably with EVs [10, 11], the feasible extrapolation of this to cancer-derived evRNA has not been documented.

In this paper, we describe our pipeline for isolating, sequencing, and interrogating the whole transcriptomes of advanced colorectal cancer (CRC) and pancreatic ductal adenocarcinoma

(PDAC) using plasma EVs. Notably, the use of evRNA cargo overcomes the current limitation of sequencing extensively fragmented RNAs in circulation, while our machine learning algorithm facilitates deconvolution of cancer-specific transcriptome data from the cumulative "pool" of plasma EVs. We demonstrate our ability to reliably assign current widely accepted RNA-based tumor subtype signatures typically obtained via tissue samples – namely, Consensus Molecular Subtypes (CMS) for CRC and Moffitt classification for PDAC, respectively in baseline plasma samples [12, 13]. Notably, in the CRC cohort, we further demonstrate the ability of our pipeline to longitudinally monitor changes in evRNA-based CMS subtype under treatment selection, which was clinically associated with disease progression. While such "subtype switching" as a basis for treatment resistance is well established [14], it is based entirely on serial sampling of tumor tissue, a requirement that can now be precluded with longitudinal monitoring of evRNAs. Finally, our evRNA-based profiling identifies expressed gene fusions and neoantigens in the circulation of CRC patients, which has significant implications for monitoring those on targeted and immunotherapies. The pipeline described herein should be applicable across cancer types and provide an avenue for leveraging the full potential of mapping tumor transcriptomes sans the requirement for repeated tissue biopsies.

Results

EVs transcriptomic profiles recapitulate the transcriptomic profiles of the cells of origin in vitro

Previous studies have shown that plasma EVs contain mRNA [15, 16]. However, it is still unclear if the transcriptomic profiles of cancer-derived EVs correspond to those of the cancer cells *in vivo*. To address this question, we performed RNA sequencing (RNA-seq) on total cellular RNA (cellRNA) and RNA from EVs (evRNA) isolated from different CRC cell lines, which were previously classified according to the four-tier Consensus Molecular Subtype (CMS) classification system (CMS1-4, respectively) [17, 18]. The expression profiles of customized CMS template genes [19, 20] were similar between matched cellRNA and evRNA (**Figure S1a**). Furthermore, based on nearest template prediction utilizing the CMS template genes, we observed a 100% CMS concordance between EVs and corresponding cells of origin across all cell lines (**Figures S1a-b**).

A deconvolution pipeline identifies tumor-specific evRNA in plasma even at low circulating proportions.

Plasma samples contain both tumor and non-tumor derived EVs mixed in unknown proportions. To overcome this intrinsic limitation of the evRNA analyte, we designed a series of experiments to bioinformatically deconvolve evRNA into tumor and non-tumor-derived transcripts and then classify the latter through a trainable multilayer artificial neural network (ANN) [21] (**Figure 1a**).

As a proof of concept, we applied a computational deconvolution pipeline on *in vitro* cellRNA mixtures containing known proportions of cancer and normal cells to infer cancer-derived RNA proportions and test whether the cancer-specific transcriptome can be used for downstream

analyses. Mixture experiments have been successfully used to standardize RNA-seq and scRNA analysis [22]. To cross-validate our approach, two different deconvolution methods were tested. The first method (CIBERSORTx) applies a signature matrix consisting of genes that can discriminate each compartment of interest (cancer and non-cancer) to infer compartment proportion and expression signatures. The second method (DeMixT) only requires the expression profiles of the non-cancer compartment as signature matrix [23, 24]. After data deconvolution, we applied the ANN-based CMS classifier on deconvolved transcriptomes, and we then compared the results from the same cell mixtures that were not subject to deconvolution. Both methods were able to accurately predict the proportion of cancer cell RNA in the cellRNA mixtures (**Figure S1c**). Mixtures containing as low as 1% cancer cell RNA could be stratified into the correct CMS after deconvolution (**Figure S1d**).

An *in silico* mixing experiment was then performed to assess the accuracy of the ANN-based CMS classifier on a larger deconvolved transcriptomic dataset. First, CRC tumor bulk RNA data from The Cancer Genome Atlas (TCGA) data were mixed *in silico* with bulk RNA-seq data derived from plasma EVs of healthy individuals. Deconvolution was then separately performed with CIBERSORTx and DeMixT. The performance of both deconvolution methods was high in predicting the proportion of cancer-derived RNA. Next, the ANN-based CMS classifier was applied to the deconvolved data, which showed that the assigned CMS subtype of the original TCGA sample could be successfully predicted after deconvolution (**Figures S2a** and **S2b**). Mixture samples with lower proportions of cancer RNA were more likely to have discordant subtypes between the original sample and the deconvolved mixture (**Figure S2c**).

We then applied the EVs RNA deconvolution pipeline on 46 baseline liquid biopsies samples from mCRC patients. In this dataset, the interquartile range of imputed proportions of cancer-derived RNA was 2.31 (median 3.48%) after deconvolution with DeMixT and 0.48 (median 1.52%), after deconvolution with CIBERSORTx. Receiver operating characteristic (ROC) curve for prediction of proportion of cancer-derived RNA showed an area under the curve (AUC) of 0.958 with specificity of 0.9 and sensitivity of 0.93 at imputed cancer proportion threshold of 2.07% for the DeMixT method. For the CIBERSORTx method AUC was 0.961 with specificity of 0.96 and sensitivity of 0.9 at imputed cancer proportion threshold of 0.97 (**Figure S3**). These results suggest that our EVs RNA deconvolution pipeline can accurately detect the presence of cancer even if the proportion of cancer-derived transcripts in the circulating evRNA transcriptome is as low as 1%.

Plasma EVs transcriptome can be used to predict tumor subtype and track longitudinal subtype changes

RNA based subtypes have been well characterized in PDAC, where the classical subtype is associated with better response to therapy compared to the basal-like subtype [13, 25-27]. In CRC, the tumor expression based CMS classification system has become widely accepted and has shown prognostic as well as predictive potential [28]. Both these classification systems are based on tissue-derived transcriptome data. To test if the deconvolution pipeline can be used to predict tumor subtype through liquid biopsy, we collected 7 plasma samples from PDAC patients in addition to the aforementioned 46 plasma samples from mCRC patients with matching tissue samples. First, PDAC molecular subtypes signatures were selected according to Raphael et al to create a signature matrix [29]. The signature matrix was then applied to TCGA data to create a

training model for DeepCC. Finally, we evaluated paired plasma samples from a cohort of four fine needle aspirates (FNA) and three core biopsies that were previously analyzed by single cell RNA sequencing by our group (**Figure 2a**) [30]. Our deconvolution analysis confirms that the molecular subtype based on Moffitt et. al. [13] as predicted in plasma evRNA correlates with the major subtype identified in tissue with a 100% concordance (**Figure 2b**). We then applied our evRNA deconvolution pipeline on the CRC liquid biopsies samples, and we were able to correctly predict CMS of the tumor of origin in 65% (30/46) of cases. This concordance was significantly higher than the random concordance rate of 22% (**Figure 2c**). Interestingly, the concordance between tumor tissue and liquid biopsy was as high as 75% (24/32) in patients with tumor purity higher than 30% (**Figure 2c**). Of note, most tissue samples with tumor purity less than 30% were classified as CMS4. Overall, in our evRNA cohort, 4.3% (2/46) of the patients' tumors were classified as CMS1, 56.5% (26/46) as CMS2, 10.8% (5/46) as CMS3, and 28.3% (13/46) as CMS4 (**Figure 2c**).

We then applied our evRNA deconvolution pipeline to 11 longitudinal samples collected from three mCRC patients (**Figure 3**). In patient ID 1, an evRNA liquid biopsy performed after neoadjuvant treatment for a single metachronous metastasis assigned the tumor to subtype CMS2. After liver resection, both CT scans and cfDNA liquid biopsies were negative. However, evRNA liquid biopsy was positive 3 months before radiological manifestation of disease recurrence and our algorithm initially assigned the recurrence to CMS2. At the time of radiological progression, however, a switch to CMS4 was observed, while cfDNA liquid biopsies continued to be negative. (**Figure 3a**). A similar switch from CMS2 to CMS4 was detected at the time of radiological progression in patient ID 2. Interestingly, the CMS switch in

this patient was associated with the identification of mutant *KRAS* in cfDNA. (**Figure 3b**). Patient ID 3 presented with multiple synchronous liver metastases. After neoadjuvant therapy and partial metastasectomy combined with radiofrequency ablation, evRNA liquid biopsy initially assigned a CMS2. The subsequent clinical course continued to demonstrate a relatively stable CMS2 subtype on repeated sampling. Nonetheless, a switch to CMS 1 was observed only few weeks preceding a CT scan that demonstrated a substantial progression in the liver. (**Figure 3c**). Thus, in all three cases, we observe a "subtype switch" in evRNA CMS assignment which predates the appearance of therapeutic resistance and radiological progression.

Deconvolution of EV transcriptome can unmask gene fusions and neoepitopes

Three bioinformatic pipelines (Arriba, Pizzly pair and Fusioncatcher, see Methods) were utilized to detect gene fusions on evRNA and matched tissue samples. Three gene fusions, DOCK4::IMMP2L, SLC6A6::XPC and TFG::ADGRG7, were detected independently in three samples, both in evRNA and paired tissues (**Figure 4a-c**). The gene dfusion DOCK4::IMMP2L is of particular interest, since *IMMP2L* somatic breakpoints have been reported in inflammatory bowel disease-associated CRC, and somatic rearrangements have been also described for this gene [31, 32]. The gene fusion GATC::COX6A1 could be identified in two independent evRNA samples but not in the matched tissues. This could be ascribable to the limited cancer region sampled for sequencing. (**Figure 4d**). Furthermore, we investigated the presence of neoepitopes from gene fusions, insertions/deletions (INDELs) and exitrons. Through the pipeline NeoFuse we were able to predicted the peptide HQDQAVSL from the gene fusion GATC::COX6A1 (**Table S1**). ScanNeo was utilized to predict INDELs-derived neoepitopes. With this pipeline we predicted matched plasma and tissue neoepitopes in 15/46 patients, and a total of 466 paired

neoepitopes with a <500mM mutant binding affinity were found. Among these, 181 neoepitopes from 9 patients were predicted to have <100nM binding affinity and 158 from 10 patients showed a <50nM binding (**Figure S3c, Table S1**). We then applied ScanExitron, a computational pipeline to identify exitron spicing events. We were able to identify seven exitrons in both plasma and matched tumor sample. Among these, two exitrons contained one and two predicted neoantigens, respectively (**Table S1**).

Discussion

In this study, we coupled low-input RNA-seq data obtained from circulating EVs with a bioinformatic deconvolution pipeline and deep learning algorithms in order to interrogate multiple tumor-specific transcriptomic features *sans* the requirement for tissue biopsies. Shi and colleagues are credited with the first demonstration that transcriptomic analyses of circulating EVs can be used to predict the response of advanced melanomas to immune checkpoint inhibitor (ICI) immunotherapy. Notably, response to ICIs was associated with an enrichment for transcripts related to immune pathways within post-treatment EV samples, and the latter were principally contributed by non-tumor sources (i.e., immune cells). Although this prior study principally focused on the non-tumor EV transcriptome, it laid the foundational basis for exploring the possibility that evRNAs might represent a rich resource for mining cancer-specific information, so long as the cancer-specific compartment could be reliably extracted from the composite pool of circulating EVs.

Our study demonstrates that bioinformatic deconvolution of cancer-specific RNAs in plasma EVs is indeed feasible, to as low a fraction as only 1% of cancer-derived transcripts in the circulating evRNA transcriptome. While the current limit of detection in our algorithm is at 1% of cancer-derived transcripts within the evRNA pool, ongoing improvements in the pipeline is likely to further reduce this threshold, enabling us to track patients in the minimal residual disease (MRD) setting. Using our deconvolution pipeline, we were able to successfully predict tumor molecular subtype in the majority of CRC and PDAC cases analyzed, using tissue-based RNA profiling as the “ground truth”. The discrepancies identified between blood and tissue in the case of CRC are likely the combined result of underlying heterogeneity, where region-

specific sampling of the tissue impacts the predicted subtype, and the impact of non-tumor components (stroma and immune cells) to the subtype annotation. Importantly, our pipeline used evRNA signatures not only for classifying tumors into established molecular subtypes at baseline, but also for longitudinal monitoring of changes in aforementioned molecular subtypes over time. For example, longitudinal monitoring of evRNAs in a cohort of CRC patients was able to identify changes in molecular subtype prior to onset of imaging-based progression, underscoring the importance of this approach to act as a sentinel for emergence of treatment resistance. While the association between tumor subtype switching and treatment resistance is well recognized [14], such monitoring has mostly been restricted to serial tissue samples, which is neither universally feasible nor cost effective. Liquid biopsies may facilitate the ability to detect these CMS changes without the need of invasive biopsies. Interestingly, a switch to CMS3 after NED and before radiological progression in one patient was associated with a newly detected mutant *KRAS* in cfDNA. In addition, a switch from CMS2 to CMS1 was detected before radiological progression in case 3, thus opening new possibilities for potential immune therapy intervention for this patient.

Finally, the identification of gene fusions and neoepitopes in blood brings a unique opportunity for precision oncology for advanced cancer patients. In both CRC and PDAC, expressed fusions are often actionable [33-35], but their identification can be limited in tissue-based assays. A blood-based assay provides redundant opportunities for assessing fusions, including over the course of time. Moreover, the identification of potential neoepitopes without the prerequisite for tissue biopsies is of considerable translational relevance in patients receiving neoantigen targeted therapies (such as vaccines or engineered T cell receptors), where both *de novo* epitopes and

newly emerged candidates can be tracked longitudinally over time and the therapeutic modulated accordingly.

Our data underpins the utility of plasma evRNA as a manifold useful liquid biopsy analyte in cancer patients, complementing the well-established of DNA-based profiling. Transcriptomic subtyping of tumors through liquid biopsy can be of particular interest when routine tissue biopsy is not feasible. This approach can be used to predict molecular subtypes of solid tumors and identify potential targets for immunotherapy. Additional studies are needed to further dissect the contribution of tumor microenvironment and tumor heterogeneity to the transcriptomic profiles of liquid biopsies.

Methods

Cell lines and culturing

Ten established CRC cell lines, LOVO (CMS1), COLO201 (CMS1), LS1034 (CMS2), SNUC1 (CMS2), HT29 (CMS3), LS513 (CMS3), WiDr (CMS3), HCT116 (CMS4), LS123 (CMS4), SW480 (CMS4) from primary or metastatic CRC tissue were used in this study. A CRC neoplastic cell line, HCC2998, with mixed molecular subtypes and 481CoN, a normal colon cell line, were used as controls. All cell lines were maintained either in RPMI-1640 or DMEM medium with 10% FBS, 100 units/mL penicillin and 100 µg/mL streptomycin (Gibco, Life Technologies, Carlsbad, CA, USA) at 37°C and 5% CO2 in a humidified incubator. Before EV isolation, cell lines were cultured in HYPERflasks (Corning Inc, Corning, NY, USA) in medium with 10% exosome-free FBS for at least 48 hours and conditioned media was then harvested. Next, 500 mL of media per cell line was centrifuged serially at 1000 rpm for 10 minutes and 5000 rpm for 5 minutes at 4°C to remove all cellular debris. The media was then passed through a 400-nm filter to remove any remaining debris and transferred to 50-mL tubes. The media was then ultra-centrifuged at 36,000 rpm at 4°C overnight. The resulting pellet was resuspended in PBS with a subsequent ultra-centrifugation for 2-4 hours at 36,000 rpm. The EV pellet was resuspended in 600 uL PBS.

Patients samples

Forty-six patients with histologically confirmed CRC who were diagnosed and treated at the University of Texas MD Anderson Cancer Center (UTMDACC)b etween 8/8/2014 and 8/29/2019 were enrolled for this study Written informed consent was obtained following protocol LAB10-0982 in accordance with standard ethical guidelines approved by the

institutional review board (IRB) and the Declaration of Helsinki. Blood samples were collected preoperatively on the day of surgery, tissue sample where collected at the time of surgical resection. Blood samples from fifty-nine healthy individuals were collected (matched based on age/gender).

Next-generation sequencing library construction for tumor tissues

Following RNA extraction, all tissue samples were processed using the KAPA Stranded RNA-Seq kit with RiboErase (HMR) (KAPA, Roche, cat. # 08098140702). Libraries were constructed according to the manufacturer protocol and were then quantified using an Agilent 2100 Tapestation. Libraries were then sequenced using 150-bp paired-end runs on a NextSeq 500 sequencer (Illumina, San Diego, CA, USA).

EV Isolation from human plasma samples

Three 8ml EDTA Vacutainer tubes (BD) of blood were collected from each case. Plasma was separated by centrifuging blood samples at 2500 rpm for 10 minutes at room temperature (RT), followed by an additional 5-minute 5000 rpm centrifugation at 4°C to remove cellular debris. The plasma was ultracentrifugated overnight at 36,000 rpm. The EV pellet was then washed with PBS and centrifuged at 36,000 rpm for 2-4 hours. The supernatant was then discarded, and the EV pellets were resuspended in 600ul PBS for long term storage at -80°C.

RNA extraction and next generation sesquencing

200 μL of resuspended EV were used for RNA extraction using the total exosome RNA & Protein Isolation kit (Invitrogen, cat. # 4478545) following the manufacturer's protocol. Small

RNA species such as miRNAs were removed during the washing steps. cDNA amplification was performed using the SMART-Seq® v4 Ultra® low input RNA kit (Takara, cat. # 634891) with sixteen cycles of PCR. cDNA samples were then diluted to 0.2-0.3 ng per uL and 1 ng of cDNA was carried over into library preparation using the Nextera XT DNA Library Preparation Kit (Illumina, cat. # FC-131-1096) following the manufacturer's protocol. Libraries were quantified with an Agilent 2100 Tapestation and sequenced in a NextSeq 500 sequencer. Samples with fewer than 1 million coding reads and high percentage of globin RNA (gRNA) and ribosomal RNA (rRNA), library prep was repeated with an additional globin RNA and ribosomal RNA depletion step using riboPOOLs gRNA/rRNA depletion kit (siTOOLsBiotech, cat. # dp-K096-001002).

d

Creation of custom gene set for CMS classification of CRC cell line EV

The use of the CMS gene set provided in the 'CMScaller' package [18] led to inaccurate classification in two of eleven cell line EV (data not shown). In order to improve the performance of nearest template prediction for EV, the preexisting gene list was subset by taking the genes that are highly expressed in the cells and EV of each CMS group. For cell lines and EV of each CMS group, differentially expressed genes compared to cell lines and EV of other CMS groups were obtained using the 'limma' package. Only those genes in the 'CMScaller' gene set that were upregulated in cell lines and EV of each CMS group with fold change of 2 or greater and adjusted P-value < 0.05 were included in the new gene set.

Artificial mixing experiments

Synthetic mixtures from CRC TCGA RNA sequencing data were generated for performance assessment of deconvolution methods. We first evaluated the correlation between the observed

and expected cancer proportions. We then assessed the correlation between observed cancer proportions and molecular subtypes concordance between liquid biopsy and tissue samples. To generate synthetic mixture, the average read counts from normal plasma evRNA samples were computationally spiked in with RNA reads from CRC TCGA data using predefined vectors to obtain mixtures with different cancer portion ratios ranging from 0.01% to 10%. Multiple linear regression was used to assess the linear relationship between the observed and expected cancer levels. We then assigned each deconvolved expression profile to a molecular subtype by using pure TCGA expression data as the reference for ANN-based classification.

RNA-Seq data analysis, deconvolution and molecular classification

QC was performed on FASTQ files using MultiQC. Raw reads were aligned to the human genome reference (hg19) using STAR. Deep cancer subtype classification (DeepCC) [21] was used to classify the transcriptomic profiles of tumor tissues. In summary, we trained a DeepCC classifier using a TCGA CRC data set ($n = 456$) and a multilayer artificial neural network (ANN) as previously described by Gao et al [21]. We then obtained the functional spectra from tumor expression profiles and used them as input to predict molecular subtypes for tumor samples.

To predict the molecular subtypes from liquid biopsy samples, we first bioinformatically deconvolved the bulk gene expression profiles in order to exclude the transcriptomic profile of normal EV. We applied two deconvolution methods CIBERSORTx [23] and DeMixT [24] to the normalized reads to impute proportions of cancer in the plasma of CRC patients and to assign molecular subtypes to the gene expression profile of the cancer-specific compartment. Both deconvolution methods were tested, by first *in silico* experiments and then deconvoluting cellular

RNA mixtures with known proportions (**Figure S1 and Figure S2**). CMS subtyping was performed by two different classifiers, one CMScaller (R package CMScaller) which uses CMS genes predicted by the nearest template prediction method and another by ANN-based classification [18] [21]. For the latter, gene expression profiles of CRC-stratified cell lines were used to train a DeepCC classifier (**Figure S1**).

Next, we performed deconvolution of bulk RNA isolated from the plasma EV of CRC patients (**Figure 1**). Deconvolution was first used for the assessment of cancer evRNA abundance (**Figure S3**). To evaluate and compare cancer evRNA levels between CRC and normal plasma samples, samples from healthy individuals were randomly divided into two groups. A group of 30 normal samples was used as the predefined normal group for both deconvolution methods. For the accurate assessment of cancer evRNA levels, 29 CRC plasma samples were selected (based on sample quality) to compare with the remaining normal samples (29 samples, control group) (**Figures S3**).

We then used the cancer-specific compartment of the deconvolved transcriptomic profiles to predict the CMS class by ANN-based classification. CRC TCGA RNA sequencing data was used to train a classifier as described for the tumor tissue CMS classification (**Figures 1b**). For the deconvolution by CIBERSORTx, a signature matrix, which is comprised of genes that are enriched in each group of interest, expression profiles of TCGA RNA sequencing data and 59 normal plasma EV were used (**Figure 1a**). High-resolution expression deconvolution was then applied to 46 EV samples from the same patients with tumor samples (**Figure 1a**). By comparison, DeMixT, which is based on differential gene expression and gene-enrichment analyses, does not require a signature matrix to impute the gene expression profile of the cancer-specific compartment. Therefore, we only used CRC plasma samples and 59 normal plasma EV

to extract the expression profiles of cancer EV in bulk plasma EV for the molecular classification (**Figure 1b**). Paired Student's *t* test was used to calculate *P*-values when comparing imputed cancer portions.

Gene fusions and neoantigens prediction

Gene fusions were analyzed from 48 tissue samples, matched liquid biopsies and 59 healthy controls. Three different methods were used to detect gene fusions and settings criteria was kept as: a) Arriba [36], confidence of medium or high, discordant_mates ≥1, split reads_reads ≥2, b) Pizzly [37], paircount ≥1 and split count ≥2 and c) FusionCatcher [38], spanning_pairs ≥1 and spanning_unique_reads ≥2. Gene fusions present in healthy controls were removed from the analysis. Neoantigens were detected by NeoFuse [39] and only cases with binding affinity < 100nM were reported (**Figure 2 and Table S1**).

Neoepitopes derived from INDELs and Exitrons

INDELs and exitrons were analyzed from 46 tissue samples, matched liquid biopsies and 59 healthy controls. INDELs were called by TransIndel [40] followed by neoepitopes analysis by ScanNeo [41] and calls present in healthy controls were removed. Exitrons were called by ScanExitron [42] and ScanNeo further predicted neoantigens (**Table S1**).

Figure Legends

Figure 1. Performance of CMS subtyping of liquid biopsies from patients with mCRC. A visual overview of workflow. RNA-seq was performed on tumor samples and plasma EV of cancer patients as well as plasma EV of healthy controls. For deconvolution with CIBERSORTx, a signature matrix was created using genes that are enriched in the tumor samples and in the plasma EV of healthy controls. For deconvolution with DeMixT, only the expression profile of healthy controls was used. Deconvolution could impute the proportion of cancer present in bulk plasma evRNA, which also allowed for generation of ROC curve based on whether cancer RNA was present in the sample. Deconvolved plasma EV profiles were also used in artificial neural network to assign the CMS groups, which were then compared to the subtypes of the tumor samples.

Figure 2. Application of RNA deconvolution in plasma from PDAC and CRC patients

a, Molecular classifications of PDAC epithelial single cells published to evaluate tumor heterogeneity using CMScaller and Moffitt template. **b,** Comparison of classical and basal-like subtype classification between pseudo-bulk transcriptome of all single cells and liquid biopsies in seven patients. **c,** Summary of transcriptomic subtyping of liquid biopsy samples and their corresponding tumors. Each column represents a patient and columns are ordered by tumor purity. Top row represents the CMS group of each solid tumor, the middle row represents the prediction for deconvolved plasma EV, and the bottom row represents the prediction for bulk plasma EV. If the predicted subtype of the liquid biopsy is concordant with the tumor sample, the column is marked by black box.

Figure 3. Longitudinal monitoring of CMS changes in mCRC patients

Timeline showing disease history of three mCRC patients. (**a** and **b**) In patients one and two, EVs-based CMS was determined at baseline, before and at the time of radiological progression. **c,** patient three, EVs-based CMS was determined at baseline and at five additional time points. Disease responses according to RECIST 1.1 [43] criteria are shown in the y axis. An upward segment between two time points represents recurrence/progressive disease (Rec/PD), a flat segment represents stable disease (SD) or non-evidence of disease (NED) and a downward segment represents complete response/partial response (CR/PR).

Figure 4. Prediction of gene fusions in plasma derived EV. Plasma, paired tumor and healthy controls were analyzed for gene fusions using Arriba, Pizzly and FusionCatcher.

Structural representation of gene fusion and circle plot depicting main chromosomal aberrations. **a,** DOCK4:IMMP2L fusion was detected in patient ID 21 and supported by arriba: pair=4, reads=4, pizzly:pair=15, reads=11, and fusioncatcher: pair=7, reads=6. **b,** SLC6A6:XPC and antisense SLC6A6:XPC-AS1 fusions were detected in patient ID 24 and supported by arriba: pair=13,reads=24 and fusioncatcher: pair=16,reads=7 respectively. **c,** TFG:ADGRG7 fusion was detected in patient ID 27 and was supported by fusioncatcher: pair=2, reads=4 and pizzly: pair=1, reads=3. **d,** GATC:COX6A1 was detected in patients IDs 11 and 40 and was supported by arriba: pair =3, reads =2 and arriba: pair=7, reads =26, respectively.

Figure S1. CRC cell lines EV transcriptome from Consensus molecular subtypes of lines and their EV are concordant. a, Heatmap representing the expression of CMS genes in cRNA and evRNA from 10 different colorectal cancer cell lines. Rows are ordered by genes in each CMS class and columns are ordered by CMS subtype and the P-value of nearest template prediction (white bar). **b,** Heatmap representation of CMS group of each cell line and its EV, with top row representing previously published subtypes. **c,** Scatter plot of actual proportion of cancer cellRNA against imputed proportion after deconvolution using DeMixT and CIBERSORTx. **d,** Heatmap representation of CMS group of each cellRNA mixture for cell lines of known molecular subtype before and after deconvolution and different subtyping methods.

Figure S2. Performance of deconvolution methods on artificial mixing experiments using TCGA CRC gene expression data. a, Correlation plot of actual proportion of cancer RNA and imputed cancer RNA using CIBERSORTx (top) and DeMixT (bottom). **b,** Sankey diagram demonstrating the CMS group of original cancer samples on the left and the CMS group of deconvolved expression data from artificial mixtures (right) using CIBERSORTx (top) and DeMixT (bottom). **c,** Violin plots of imputed cancer proportion in deconvolved expression data comparing discordant and concordant samples using CIBERSORTx (top) and DeMixT (bottom).

Figure S3. Analysis of cancer proportions. Receiver Operating Characteristic (ROC) curve analysis for the imputed cancer portions was performed in order to estimate the area under the curve (AUC) for **a,** CIBERSORTx and **b,** DeMIxT. **c,** predicted neoepitopes identified in paired plasma and tissue samples and grouped by binding affinity.

Table S1. Predicted matched plasma and tissue neoantigens identified in gene fusions, INDELs and exitrons with a mutant binding affinity <100nM.

References

1. Cohen, J.D., et al., *Detection and localization of surgically resectable cancers with a multi-analyte blood test.* Science, 2018. **359**(6378): p. 926-930.
2. Shen, S.Y., et al., *Sensitive tumour detection and classification using plasma cell-free DNA methylomes.* Nature, 2018. **563**(7732): p. 579-583.
3. Cristiano, S., et al., *Genome-wide cell-free DNA fragmentation in patients with cancer.* Nature, 2019. **570**(7761): p. 385-389.
4. Liu, M.C., et al., *Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA.* Ann Oncol, 2020. **31**(6): p. 745-759.
5. Larson, M.H., et al., *A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection.* Nat Commun, 2021. **12**(1): p. 2357.
6. Esfahani, M.S., et al., *Inferring gene expression from cell-free DNA fragmentation profiles.* Nat Biotechnol, 2022. **40**(4): p. 585-597.
7. Roskams-Hieter, B., et al., *Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies.* NPJ Precis Oncol, 2022. **6**(1): p. 28.
8. Lin, Y., et al., *A Plasma Long Noncoding RNA Signature for Early Detection of Lung Cancer.* Transl Oncol, 2018. **11**(5): p. 1225-1231.
9. O'Brien, K., et al., *RNA delivery by extracellular vesicles in mammalian cells and its applications.* Nat Rev Mol Cell Biol, 2020. **21**(10): p. 585-606.
10. Bernard, V., et al., *Circulating Nucleic Acids Are Associated With Outcomes of Patients With Pancreatic Cancer.* Gastroenterology, 2019. **156**(1): p. 108-118 e4.
11. Moding, E.J., et al., *Detecting Liquid Remnants of Solid Tumors: Circulating Tumor DNA Minimal Residual Disease.* Cancer Discov, 2021. **11**(12): p. 2968-2986.
12. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
13. Moffitt, R.A., et al., *Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma.* Nat Genet, 2015. **47**(10): p. 1168-78.
14. Woolston, A., et al., *Genomic and Transcriptomic Determinants of Therapy Resistance and Immune Landscape Evolution during Anti-EGFR Treatment in Colorectal Cancer.* Cancer Cell, 2019. **36**(1): p. 35-50 e9.
15. Skog, J., et al., *Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers.* Nat Cell Biol, 2008. **10**(12): p. 1470-6.
16. Valadi, H., et al., *Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells.* Nat Cell Biol, 2007. **9**(6): p. 654-9.
17. Sveen, A., et al., *Colorectal Cancer Consensus Molecular Subtypes Translated to Preclinical Models Uncover Potentially Targetable Cancer Cell Dependencies.* Clin Cancer Res, 2018. **24**(4): p. 794-806.
18. Eide, P.W., et al., *CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models.* Sci Rep, 2017. **7**(1): p. 16618.
19. Hoshida, Y., *Nearest template prediction: a single-sample-based flexible class prediction*

with confidence assessment. PLoS One, 2010. **5**(11): p. e15543.
20. Hoshida, Y., et al., *Gene expression in fixed tissues and outcome in hepatocellular carcinoma.* N Engl J Med, 2008. **359**(19): p. 1995-2004.
21. Gao, F., et al., *DeepCC: a novel deep learning-based framework for cancer molecular subtype classification.* Oncogenesis, 2019. **8**(9): p. 44.
22. Tian, L., et al., *Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments.* Nat Methods, 2019. **16**(6): p. 479-487.
23. Newman, A.M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry.* Nat Biotechnol, 2019. **37**(7): p. 773-782.
24. Wang, Z., et al., *Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration.* iScience, 2018. **9**: p. 451-460.
25. Collisson, E.A., et al., *Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy.* Nat Med, 2011. **17**(4): p. 500-3.
26. Bailey, P., et al., *Genomic analyses identify molecular subtypes of pancreatic cancer.* Nature, 2016. **531**(7592): p. 47-52.
27. Aung, K.L., et al., *Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial.* Clin Cancer Res, 2018. **24**(6): p. 1344-1354.
28. Dienstmann, R., et al., *Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer.* Nat Rev Cancer, 2017. **17**(4): p. 268.
29. Cancer Genome Atlas Research Network. Electronic address, a.a.d.h.e. and N. Cancer Genome Atlas Research, *Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma.* Cancer Cell, 2017. **32**(2): p. 185-203 e13.
30. Lee, J., et al., *Elucidation of Tumor-Stromal Heterogeneity and the Ligand-Receptor Interactome by Single-CellTranscriptomics in Real-world Pancreatic Cancer Biopsies.* Clin Cancer Research, 2021.
31. Rajamaki, K., et al., *Genetic and Epigenetic Characteristics of Inflammatory Bowel Disease-Associated Colorectal Cancer.* Gastroenterology, 2021. **161**(2): p. 592-607.
32. Bass, A.J., et al., *Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion.* Nat Genet, 2011. **43**(10): p. 964-968.
33. Fusco, M.J., et al., *Identification of Targetable Gene Fusions and Structural Rearrangements to Foster Precision Medicine in KRAS Wild-Type Pancreatic Cancer.* JCO Precis Oncol, 2021. **5**.
34. Jones, M.R., et al., *NRG1 Gene Fusions Are Recurrent, Clinically Actionable Gene Rearrangements in KRAS Wild-Type Pancreatic Ductal Adenocarcinoma.* Clin Cancer Res, 2019. **25**(15): p. 4674-4681.
35. Heining, C., et al., *NRG1 Fusions in KRAS Wild-Type Pancreatic Cancer.* Cancer Discov, 2018. **8**(9): p. 1087-1095.
36. Uhrig, S., et al., *Accurate and efficient detection of gene fusions from RNA sequencing data.* Genome Res, 2021. **31**(3): p. 448-460.
37. Melsted, P., et al., *Modular and efficient pre-processing of single-cell RNA-seq.* BioRxiv, 2019.
38. Nicorici, D., et al., *FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data.* BioRxiv, 2014.
39. Fotakis, G., et al., *NeoFuse: predicting fusion neoantigens from RNA sequencing data.* Bioinformatics, 2020. **36**(7): p. 2260-2261.

40. Yang, R., J.L. Van Etten, and S.M. Dehm, *Indel detection from DNA and RNA sequencing data with transIndel.* BMC Genomics, 2018. **19**(1): p. 270.
41. Wang, T.Y., et al., *ScanNeo: identifying indel-derived neoantigens using RNA-Seq data.* Bioinformatics, 2019. **35**(20): p. 4159-4161.
42. Wang, T.Y. and R. Yang, *Integrated protocol for exitron and exitron-derived neoantigen identification using human RNA-seq data with ScanExitron and ScanNeo.* STAR Protoc, 2021. **2**(3): p. 100788.
43. Eisenhauer, E.A., et al., *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).* Eur J Cancer, 2009. **45**(2): p. 228-47.

Figure 1

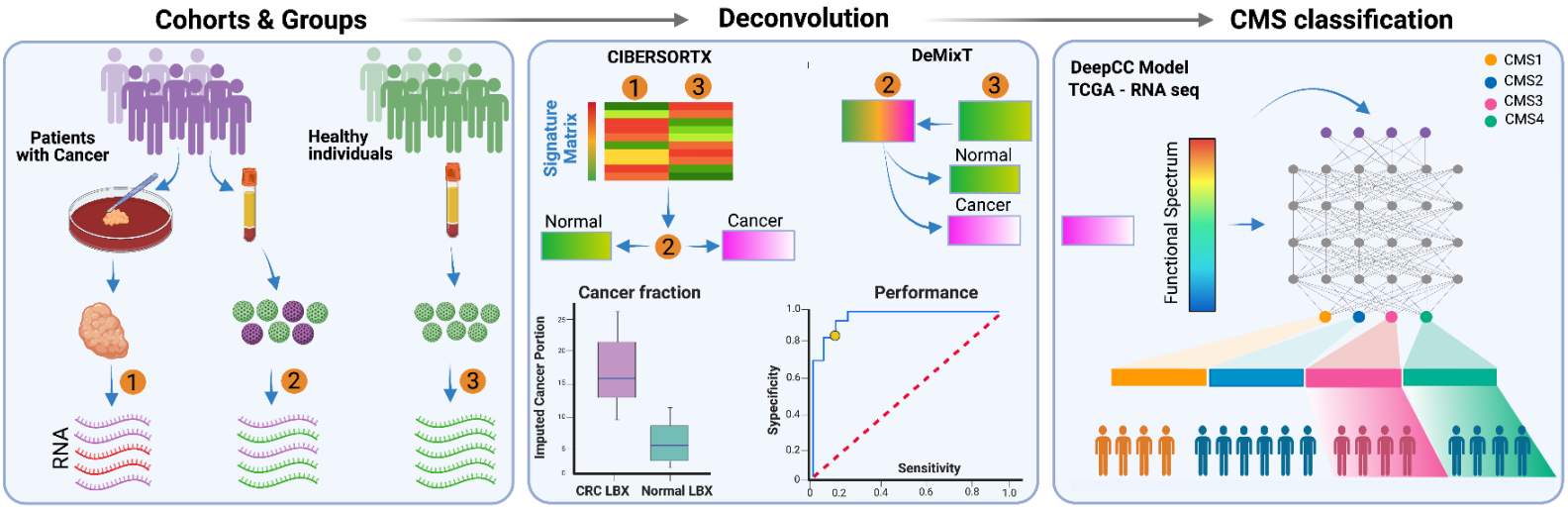

Cohorts & Groups
Deconvolution
CMS classification
Patients with Cancer
Healthy individuals
RNA
CIBERSORTX
Signature Matrix
Normal
Cancer
DeMixT
Normal
Cancer
Cancer fraction
Imputed Cancer Portion
CRC LBX
Normal LBX
Performance
Sypecificity
Sensitivity
DeepCC Model
TCGA - RNA seq
Functional Spectrum
CMS1
CMS2
CMS3
CMS4

Figure 2

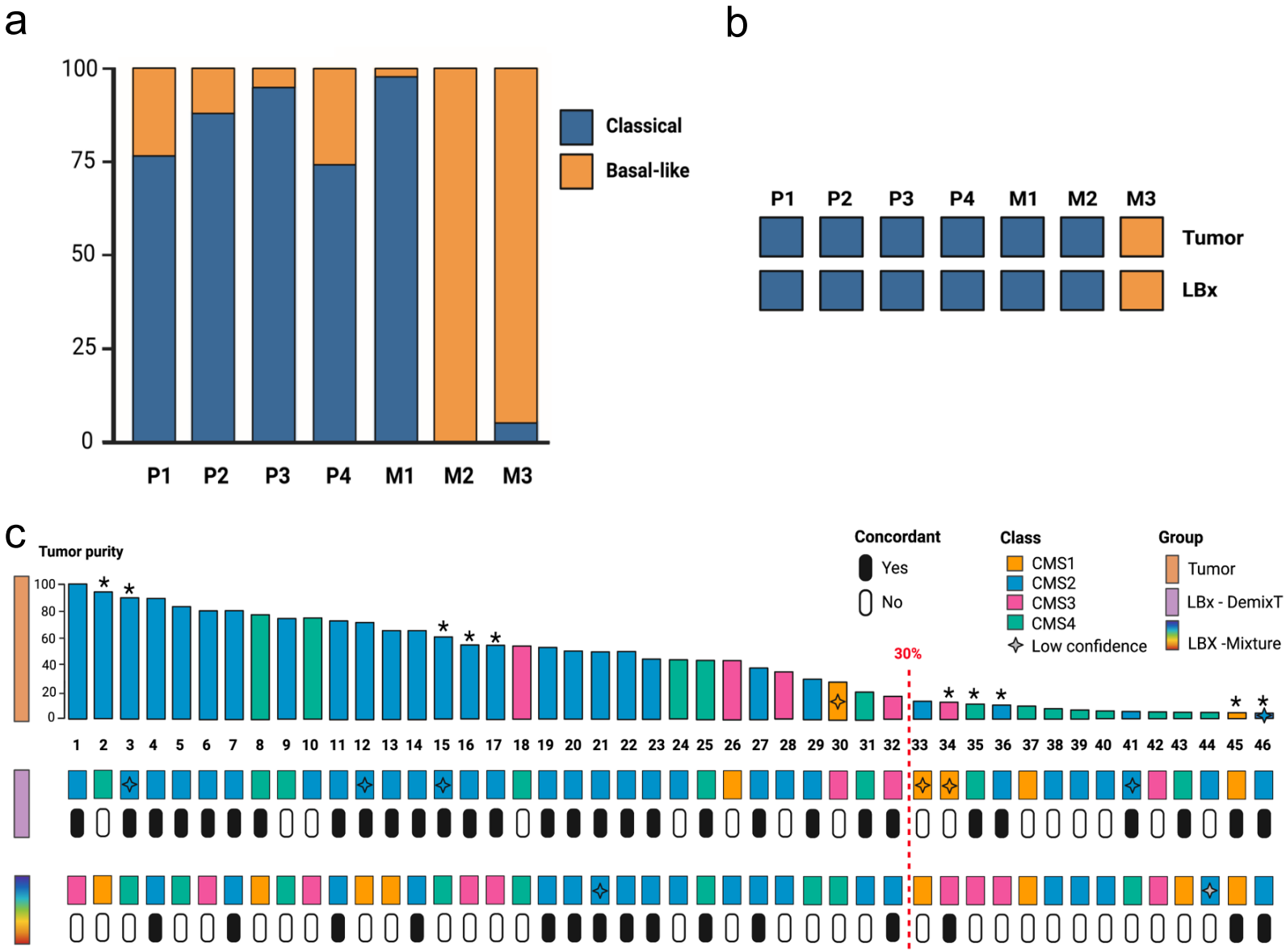

a
100
75
50
25
0
P1
P2
P3
P4
M1
M2
M3
Classical
Basal-like
b
P1
P2
P3
P4
M1
M2
M3
Tumor
LBx
c
Tumor purity
100
80
60
40
20
0
Concordant
Yes
No
Class
CMS1
CMS2
CMS3
CMS4
Low confidence
Group
Tumor
LBx - DemixT
LBX -Mixture
30%

Figure 3

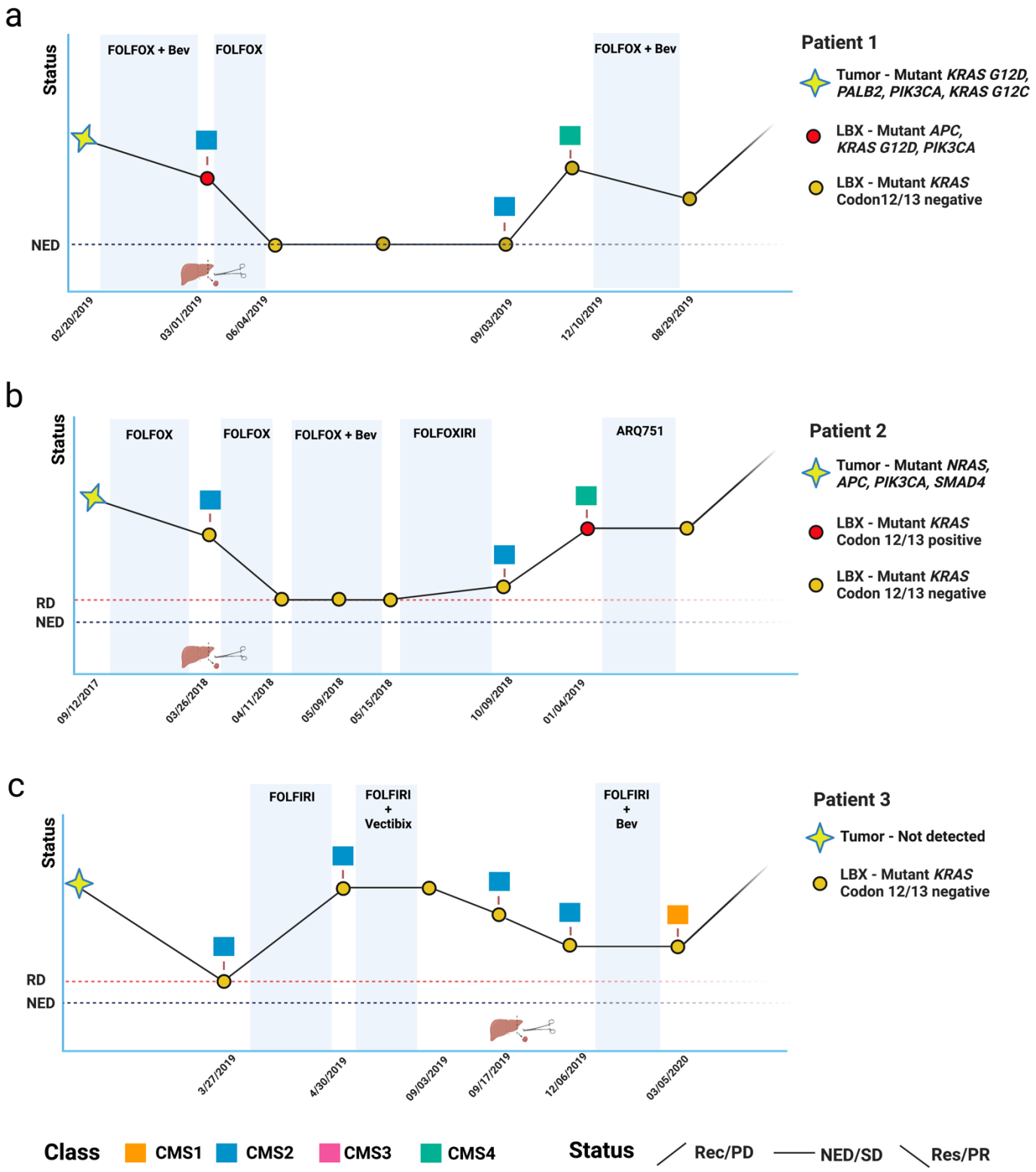
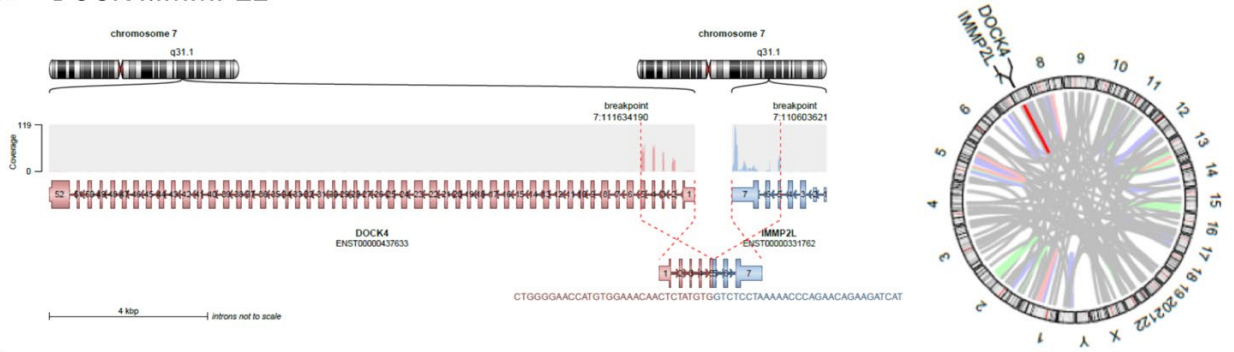
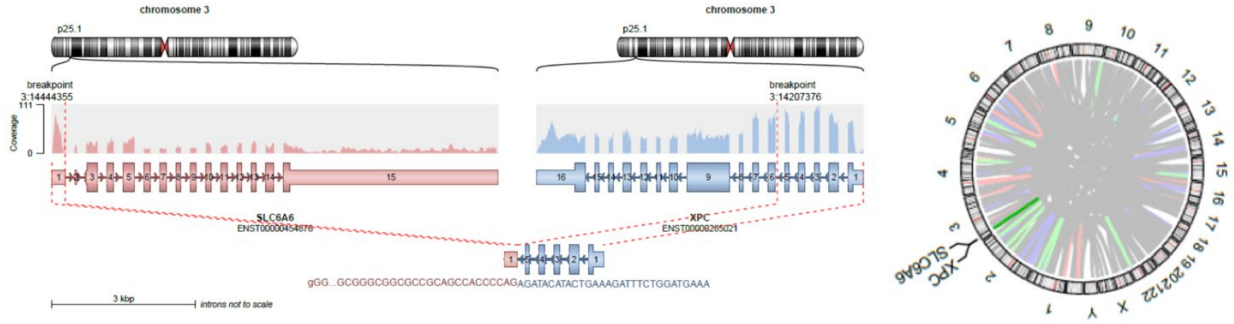

a
Status
FOLFOX + Bev
FOLFOX
FOLFOX + Bev
NED
02/20/2019
03/01/2019
06/04/2019
09/03/2019
12/10/2019
08/29/2019
Patient 1
Tumor - Mutant KRAS G12D, PALB2, PIK3CA, KRAS G12C
LBX - Mutant APC, KRAS G12D, PIK3CA
LBX - Mutant KRAS Codon12/13 negative
b
Status
FOLFOX
FOLFOX
FOLFOX + Bev
FOLFOXIRI
ARQ751
RD
NED
09/12/2017
03/26/2018
04/11/2018
05/09/2018
05/15/2018
10/09/2018
01/04/2019
Patient 2
Tumor - Mutant NRAS, APC, PIK3CA, SMAD4
LBX - Mutant KRAS Codon 12/13 positive
LBX - Mutant KRAS Codon 12/13 negative
c
Status
FOLFIRI
FOLFIRI + Vectibix
FOLFIRI + Bev
RD
NED
3/27/2019
4/30/2019
09/03/2019
09/17/2019
12/06/2019
03/05/2020
Patient 3
Tumor - Not detected
LBX - Mutant KRAS Codon 12/13 negative
Class
CMS1
CMS2
CMS3
CMS4
Status
Rec/PD
NED/SD
Res/PR

Figure 4

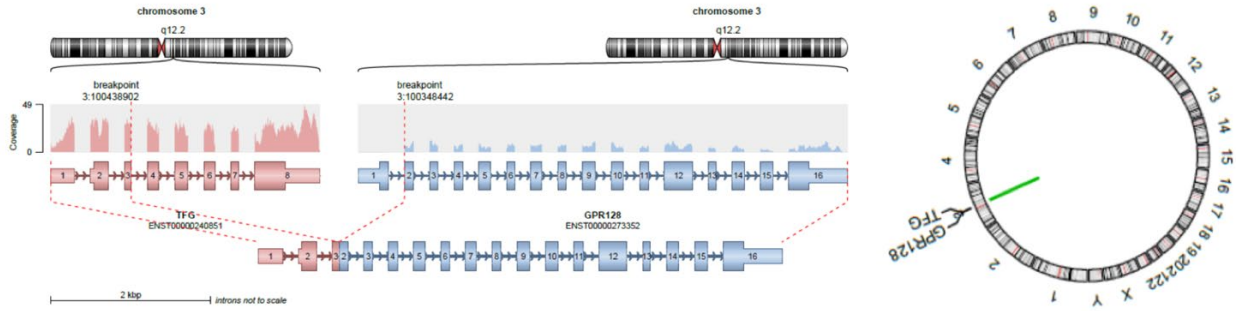
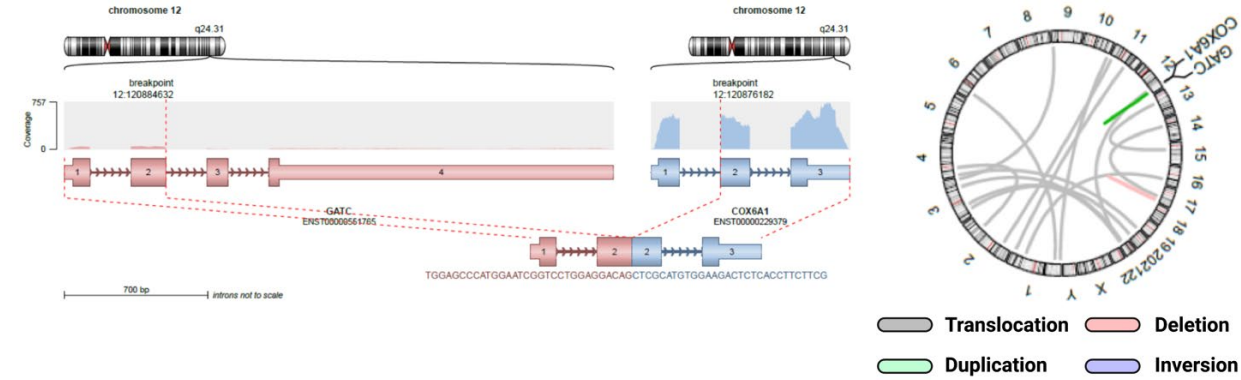
a DOCK4:IMMP2L



b SLC6A6:XPC



c TFG:ADGRG7



d GATC:COX6A1