

# SCIENTIFIC REPORTS



OPEN

## High-density single cell mRNA sequencing to characterize circulating tumor cells in hepatocellular carcinoma

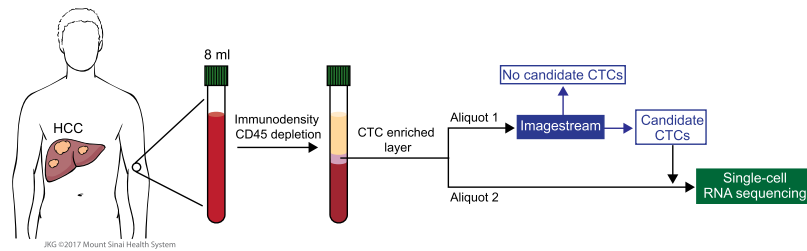
Delia D'Avola<sup>1,2</sup>, Carlos Villacorta-Martin<sup>1</sup>, Sebastiao N. Martins-Filho<sup>1,3</sup>, Amanda Craig<sup>1</sup>, Ismail Labgaa<sup>1,4</sup>, Johann von Felden<sup>1,5</sup>, Allette Kimaada<sup>6</sup>, Antoinette Bonaccorso<sup>7</sup>, Parissa Tabrizian<sup>7</sup>, Boris M. Hartmann<sup>1,8</sup>, Robert Sebra<sup>5,9</sup>, Myron Schwartz<sup>7</sup> & Augusto Villanueva<sup>1,10</sup>

Patients with hepatocellular carcinoma (HCC) release tumor cells to the bloodstream, which can be detected using cell surface markers. Despite numerous reports suggest a direct correlation between the number of circulating tumor cells (CTCs) and poor clinical outcomes, few studies have provided a thorough molecular characterization of CTCs. Due to the limited access to tissue samples in patients at advanced stages of HCC, it is crucial to develop new technologies to identify HCC cancer drivers in routine clinical conditions. Here, we describe a method that sequentially combines image flow cytometry and high density single-cell mRNA sequencing to identify CTCs in HCC patients. Genome wide expression profiling of CTCs using this approach demonstrates CTC heterogeneity and helps detect known oncogenic drivers in HCC such as IGF2. This integrated approach provides a novel tool for biomarker development in HCC using liquid biopsy.

The concept of liquid biopsy, which entails the analysis of tumor components released to the bloodstream (i.e., DNA and tumor cells), has revolutionized oncology<sup>1</sup>. The main advantages of liquid biopsy are: (1) the technique is minimally invasive, which decreases the financial costs and potential complications of tissue biopsies, and (2) it is easy to iterate during patients' follow-up. Both features are crucial considering how cancer cells adapt to pharmacological pressures, and acquire new molecular alterations not detected at baseline<sup>2</sup>. To account for this, we would be required to perform repeated invasive tissue biopsies upon every treatment, which is difficult to implement in routine clinical conditions. Thus, to develop new biomarkers using liquid biopsy strategies is imperative. This is even more critical in hepatocellular carcinoma (HCC) because these patients are generally diagnosed using imaging techniques<sup>3</sup>, so unlike most solid tumors, access to tissue to conduct biomarker studies in HCC is significantly more difficult. In the United States, liver cancer (including both HCC and intrahepatic cholangiocarcinoma) is currently the fastest growing malignancy both in terms of incidence and mortality<sup>4</sup>.

Most of the studies exploring circulating tumor cells (CTCs) in HCC have shown a direct correlation between higher CTC number and poor clinical outcomes<sup>5</sup>. In most of them, CTC detection relied on the analysis of a small number of surface markers. Indeed, the only FDA-approved device to detect CTCs is based on the expression of the surface marker EPCAM. These methods do not include isolation of CTCs to conduct a higher resolution

<sup>1</sup>Division of Liver Diseases, Department of Medicine, Liver Cancer Program, Tisch Cancer Institute, Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, USA. <sup>2</sup>LiverUnit and Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Clínica Universidad de Navarra, Pamplona, Spain. <sup>3</sup>Departamento de Patología, Faculdade de Medicina FMUSP, Universidade de Sao Paulo, Sao Paulo, Brazil. <sup>4</sup>Department of Visceral Surgery, Lausanne University Hospital CHUV, Lausanne, Switzerland. <sup>5</sup>Department of Internal Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. <sup>6</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA. <sup>7</sup>Department of Surgery, Icahn School of Medicine at Mount Sinai, New York, USA. <sup>8</sup>Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, USA. <sup>9</sup>Sema4, a Mount Sinai Venture, Stamford, CT, USA. <sup>10</sup>Division of Hematology and Medical Oncology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, USA. Correspondence and requests for materials should be addressed to A.V. (email: [augusto.villanueva@mssm.edu](mailto:augusto.villanueva@mssm.edu))



**Figure 1.** Summary of study design and sample workflow. Illustration by Jill Gregory. Printed with permission from Mount Sinai Health System, licensed under CC BY-ND (<https://creativecommons.org/licenses/by-nd/4.0/>).

molecular characterization, which is critical to use CTCs analysis as a proxy to interrogate dominant tumor molecular alterations. This concept was recently shown in a patient with metastatic breast cancer using whole genome sequencing (WGS) in CTCs<sup>6</sup>, which allowed detecting known driver mutations in the CTC compartment. Similarly, WGS of CTCs isolated from a patient with metastatic prostate cancer revealed driver mutations which could be tracked to corresponding tumor tissue<sup>7</sup>.

There are few data on whole-genome transcriptome profiling of CTCs using single-cell technologies<sup>8</sup>. This was due to the relative low throughput of single-cell isolation techniques, which allowed to sequence in parallel relatively low numbers of cells. Considering the rarity of CTCs in the blood (estimated 0–10 CTCs out of 7,000,000 nucleated cells)<sup>9</sup>, extensive enrichment for CTCs was required before single-cell sequencing. Studies using scRNAseq to analyze CTCs leveraged different approaches to isolate CTCs using robotic or microfluidic devices<sup>10,11</sup>. CTCs isolation heavily relies on the identification of specific markers such as EpCAM in prostate cancer or NG2 in melanoma<sup>10,11</sup>. These studies confirm the ability of genome-wide expression analyses of isolated CTCs to detect de-regulation of genes found in tissue counterparts<sup>10</sup>. Other CTC isolation methods such as those based on cell-surface vimentin expression enabled the identification of PD-L1 as a prognostic marker in patients with colorectal or prostate cancer<sup>12</sup>. Alternative approaches favored the use of PCR-based technologies to derive CTC expression signatures analyzing a limited number of genes, as recently shown in HCC<sup>13</sup>. However, there is limited data on whole-genome transcriptomic profiling of CTCs from HCC patients. Recent technological developments including novel single-cell technologies prompted us to conduct a dense transcriptomic profiling of CTCs using a sequential protocol that maximizes CTC detection rate. We provide the proof-of-principle to further develop single-cell sequencing for biomarker development in HCC.

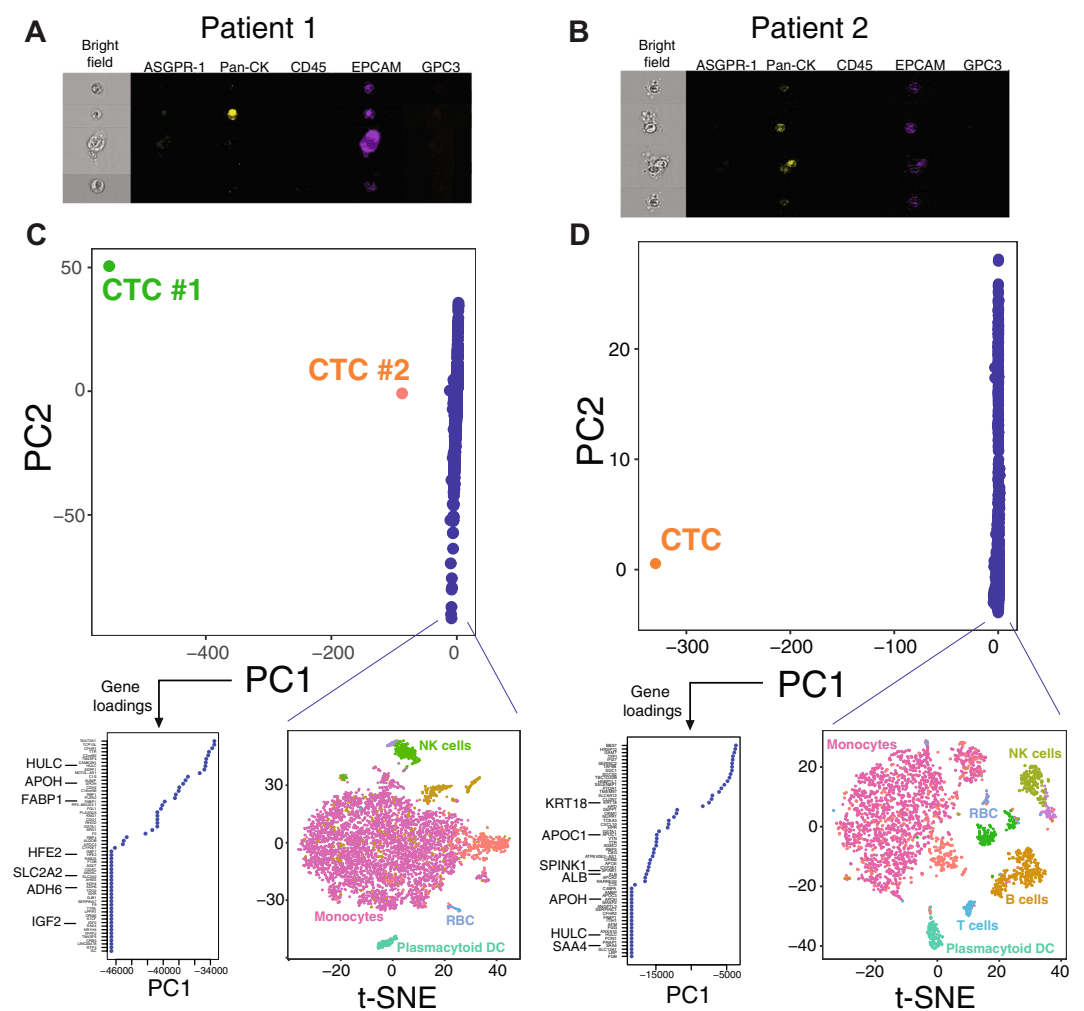
## Results

**A novel pipeline to detect CTCs using whole transcriptome data.** To enable the comprehensive characterization of CTC transcriptome, we developed an analytical pipeline that combines Imaging Flow Cytometry (IFC) and Single-cell RNA sequencing (scRNA-seq). Our protocol uses 8 mL of whole blood and consists of 3 sequential steps: (1) CD45 negative enrichment, (2) selection of patients using IFC, and (3) scRNAseq using 10X Genomics Chromium<sup>®</sup> (Fig. 1). This enables mRNA sequencing in up to 10,000 cells with a capture efficiency up to 65% for high-throughput expression profiling using unique molecular identifiers. Considering the variable range of CTCs in HCC<sup>5</sup>, our approach maximizes the odds of CTC detection. In this proof-of-principle study, we included 6 HCC patients and 1 control subject (Table 1). Four out of the six HCC patients had potential CTCs on IFC, as defined by a combination of factors: CD45 negative, at least 1 positive marker among pan-CK, GPC3, ASGPR1, or EPCAM, and compatible shape by bright field morphologic analysis. In accordance with previous reports<sup>14</sup>, the control subject had no candidate CTCs. As a pilot study, once IFC selection was optimized, we selected two out of the four patients with candidate CTCs, those with higher number of CD45 negatively enriched cells, for further analysis with scRNAseq (Fig. 2A,B).

Due to the rarity of CTCs in the blood, we adjusted the analytical scRNAseq pipeline to be specifically sensitive to detect outliers. We sequenced a total of 7,104 and 3,130 cells in patients 1 and 2, respectively. Median number of RNA molecules and genes detected per cell in patient 1 were 3,046 (range 1,004–21,169) and 1,098 (range 202–4,109). In patient 2, we detected a median of 2,552 RNA molecules (range 1,147–27,950) and 933 genes (range 202–4,484) per cell. We used the most variable genes to compute the main sources of variability in the dataset, as depicted in a principal component (PC) analysis (Fig. 2C,D). This approach identified two CTCs from patient 1 and one CTC from patient 2 with gene expression profiles dramatically different from the other clonal populations in the blood. These candidate CTCs aligned along the first PC, emphasizing the strong source of variation captured by this PC. In both cases, the genes that defined PC1, which captures the maximum variance, were highly weighted on liver specific genes (STable 1). As predicted, when comparing the scaled expression values of the 3 candidate CTCs from the 2 patients with the rest of the cells, we also found numerous liver-related genes among the 100 top differentially expressed genes (Fig. 3A and STable 2). These genes are involved in liver physiology including apolipoproteins, coagulation factors, alcohol dehydrogenases, cytochromes, etc. The 3 candidate CTCs isolated from both patients shared 12% of their top 100 expressed genes (*TTR*, *FABP1*, *GSTA1*, *APOH*, *FGB*, *HULC*, *ALB*, *APOA2*, *SEPP1*, *APOC1*, *HPD* and *ORM1*), which increased our confidence of efficient CTC detection. To further support a hepatic lineage of these outlier cells, we performed gene set enrichment analysis (GSEA) on their scaled expression as compared to all the other cells. Bile acid and xenobiotic metabolism were among the top positively enriched gene sets in these cells (Fig. 3B). We next sought to characterize the rest

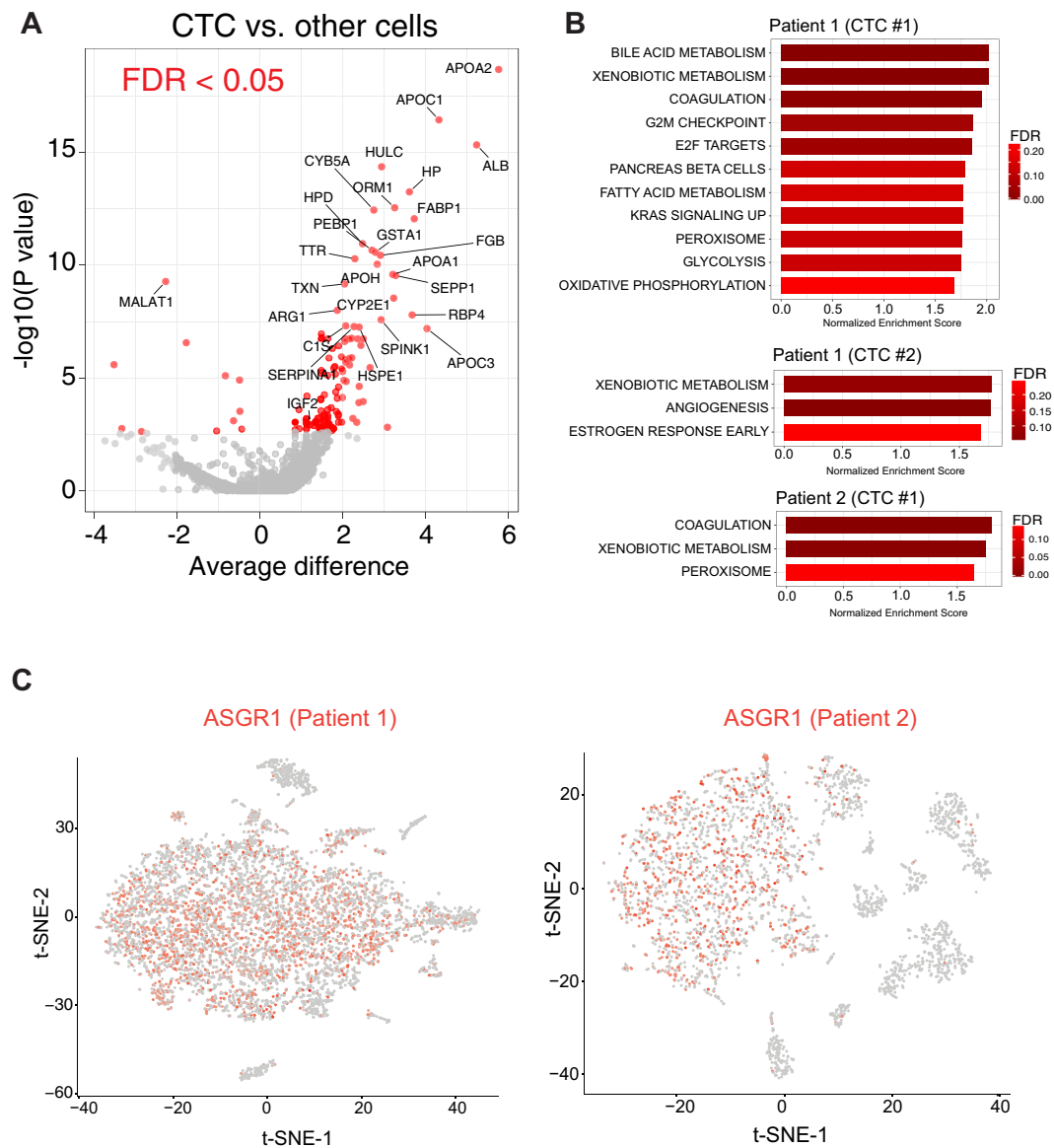
Patient	Gender	Age	Etiology	Cirrhosis	BCLC stage	AFP	Albumin (g/dl)	Bilirubin (mg/dl)	INR	Platelet (10E9/L)	Imagestream data					scRNAseq
											CTC count	EPCAM+	CK+	ASGPR1+	GPC3+	
Patient 1	M	68	HCV	Yes	C	448	2.7	1.4	1	209	5	5	3	2	0	Yes
Patient 2	M	77	HCV/Alcohol	Yes	C	460	1.9	1.8	1.2	116	5	5	5	0	0	Yes
Patient 3	M	64	Cryptogenic	No	C	140	3.2	1.1	1	200	11	11	11	4	0	No
Patient 4	M	59	HCV/Alcohol	Yes	B	26	3.4	0.8	1	116	0	0	0	0	0	No
Patient 5	M	63	HCV	No	A	2	3.5	0.2	1.1	342	4	4	0	0	0	No
Patient 6	M	44	HBV	Yes	A	1	3.6	0.5	1.2	114	0	0	0	0	0	No
Patient 7*	F	56	—	No	NA	NA	4.2	0.6	1	222	0	0	0	0	0	No

**Table 1.** Clinical characteristics and Imagestream analysis of the patients analyzed. \*Subject without cancer; NA: Not applicable; HCV: Hepatitis C virus; HBV: Hepatitis B virus; M: Male; F: Female; AFP: Alpha-fetoprotein; BCLC: Barcelona Clinic Liver Cancer.



**Figure 2.** Identification of CTCs using single-cell RNAseq. Top panels show candidate CTCs identified by Imagestream in patients 1 (A) and 2 (B). (C,D) PCA plots after outlier-detection normalization for each patient, including the loadings for the top marker genes of PC1, and a t-SNE plot of non-CTCs cells with predicted clusters based on the top-ranked marker genes (RBC: red blood cells, NK: Natural Killer, DC: dendritic cells).

of the blood cells based on their gene expression profiles. Using the marker genes that define each of these clusters (Fig. 2C,D and STable 3) we identified cells compatible with monocytes, NK cells, B and T cells, plasmacytoid dendritic cells and red blood cells.



**Figure 3.** Characterization of CTCs and identification of potential HCC driver genes. (A) Volcano plot of differential gene expression between the 3 CTCs found in the 2 patients and the rest of blood cells. Red dots denote comparisons with an FDR < 0.05. (B) Bar size represent normalized enrichment score values and FDR (i.e., red gradient) for gene sets significantly enriched in each CTC resulting from GSEA. (C) Expression of ASGR1 on scRNAseq in non-CTC blood cells. Scaled expression depicted as a red gradient (grey denotes no expression detected for ASGR1).

As predicted, our CD45 negative selection increased the ratio of CTCs to non-CTCs up to approximately 1 CTC per ~3,000 nucleated cells. In this context, high-density scRNA-seq is essential to accurately discriminate CTCs using whole transcriptome data. Unexpectedly, when we checked the expression of ASGPR1, a marker used to detect CTC in HCC<sup>15</sup>, we found that it was expressed in a significant proportion of non-CTCs, mainly monocytes (Fig. 3C). Non-hepatic expression of ASGPR1 has been recently described<sup>16</sup>, which further underscores the added value of genome wide scRNAseq for accurate CTC detection. Altogether, these observations provide further evidence of the limitations of single (or targeted) marker approaches for CTC detection. In summary, scRNAseq detected cells in the blood of HCC patients with a genomic profile highly suggestive of an hepatic lineage.

**scRNAseq identifies HCC driver genes and suggest molecular heterogeneity in CTCs.** Single-cell RNAseq was very suggestive of a hepatic origin of the outlier cells, but the low-coverage and sparse nature of these data precluded an evaluation of unequivocally malignant traits such as somatic mutations or chromosomal aberrations. Sequencing depth was suboptimal to confidently call mutations in well-known HCC hotspots. However, in both patients, candidate CTCs expressed Hepatocellular Carcinoma Up-Regulated Long Non-Coding RNA (HULC), while absent in the other blood cells. HULC is a lncRNA with known oncogenic properties in HCC<sup>17</sup>. We

also detected other genes frequently up-regulated in HCC such as SPINK1, IGF2, SPP1 or BRIC5 (STable 2). This, added to the fact that circulating cells of hepatic lineage are neither detected in the bloodstream of healthy donors nor in patients with non-malignant chronic liver diseases<sup>13</sup>, supports the malignant nature of our candidate CTCs.

Despite the mutational landscape of HCC is fairly known<sup>18</sup>, the most prevalent HCC mutations are undruggable. Indeed, HCC ranks second to last in terms of number of druggable mutations<sup>19</sup>, which emphasizes the need to develop new approaches to detect druggable events in HCC patients. Among the differentially expressed genes in our CTCs, we detected overexpression of IGF2 (Fig. 3A and STable 2), a gene overexpressed in 20% of early HCCs and recently characterized as an HCC epidriver<sup>20</sup>. Indeed, selective blockade of IGF2 using monoclonal antibodies has shown anti-tumoral effects in experimental models of HCC. Notably, IGF2 de-regulation cannot be assessed by merely probing the mutation landscape of the tumor, which renders scRNAseq of CTCs an outstanding tool to detect non-mutated druggable genomic aberrations such as IGF2 overexpression.

In addition to IGF2, many genes were differentially expressed between the two candidate CTCs detected in patient 1, as captured by their distribution along PC1. Also, GSEA revealed differences in the top ranked gene sets enriched in the 2 candidate CTCs found in this patient, including features suggestive of a more aggressive biological behavior in CTC #1 (e.g., E2F targets, G2M checkpoint, KRAS signaling, Fig. 3B). These data are indicative of transcriptome heterogeneity in the CTC compartment, and adds credence to our hypothesis that high-density expression profiling using scRNAseq could help trace cancer heterogeneity in CTCs. Unfortunately, these patients had advanced HCC and were diagnosed using imaging techniques as per clinical practice guidelines, which prevented us to access tumor tissue to recapitulate our findings on CTCs in corresponding tissue. In summary, we have applied scRNAseq to identify and characterize CTCs in patients with HCC. Our proof-of-principle study demonstrates the advantages of genome-wide transcriptome profiling to confidently detect CTCs, its potential role to monitor HCC heterogeneity and its ability to detect HCC driver genes, which will ultimately help customize therapeutic interventions in these patients.

## Experimental Procedures

**Human samples, cell sorting and imaging flow cytometry.** All patients were enrolled at a single institution (Icahn School of Medicine at Mount Sinai (ISMMS)). The study was approved by the ISMMS Program for Protection of Human Subjects/Institutional Review Board with protocol number HS-15-00540. All patients provided informed consent to participate in this study. All methods were performed in accordance with the relevant guidelines and regulations for conducting research in human subjects. Patients were diagnosed with hepatocellular carcinoma (HCC) based on the EASL guidelines (3), except one who was enrolled as a control subject (i.e., no tumor disease). Blood samples were collected in 8 ml heparin tubes (Vacutainer<sup>®</sup>, BD Bioscience) and processed within 6 hours after collection. Depletion of haematopoietic cells (red cells and CD45-positive cells) was performed by an immunodensity assay (Rosette Depletion Kit<sup>®</sup>, Stem Cells and Ficoll-Paque Plus, GE Healthcare). The enriched cell layer at the interphase between plasma and the density medium was collected and washed with 2%FBS-PBS for platelet depletion by centrifugation according to the manufacturer. Vital cells were counted with a Neubauer haemocytometer and trypan blue staining and resuspended in 2% FBS-PBS. To account for red cell contamination, a red cell lysis step was performed (Red cell lysis buffer<sup>®</sup>, eBioscience). One aliquot of 10,000 cells was collected and stored on ice. The remaining cells were stained with the following conjugated antibodies ASGPR-1-FITC (sc-52623 FITC; Santacruz Biotech), pan-CK-PE (Clone-C11, 10478; Cayman Chemicals), Glypican-3-APC (FAB2119A; R&D systems), EPCAM-VioBlue (CD326(EpCAM)VioBlue; Mylteni), CD45-Cy5-PE (PE-Cy<sup>™</sup>7 Mouse Anti-Human CD45; BD Pharmingen) and analyzed by Imaging Flow Cytometry (IFC, Imagestream X Amnis, Millipore) following a similar approach as previously described<sup>14</sup>. Of the twelve channels available, channel 1 was reserved for bright field images, whereas channels 2, 3, 6, 7 and 11 were used for the detection of the fluorochromes. Compensation beads were used to provide single color reference samples for each fluorochrome. A compensation matrix was built with these reference samples to remove the spectral overlap to adjacent channels from each detection channel. CD45-negative cells showing positivity for at least 1 marker of our antigen panel (i.e., ASGPR1, pan-CK, GPC3, EPCAM) and with a bright field image allowing the identification of their morphology were defined as potential CTC.

**Single cell RNA sequencing.** For those samples with potential CTCs based on Imagestream analysis, the second aliquot was immediately used for single-cell RNA sequencing (scRNAseq) using Chromium (10X Genomics) and Illumina NGS sequencing technologies. The cell density was determined using a C-Chip DHC-N01 disposable hemocytometer. A mixture of 5  $\mu$ l of 0.04% trypan blue (BioRad) and 5  $\mu$ l of the single cell suspension was loaded onto the hemocytometer and examined at 10X magnification, using the Invitrogen Evos FL Cell Imaging System-digital inverted microscope. The single-cell chip loading, Gel bead in Emulsion (GEM) generation & barcoding, post GEM-RT & cDNA amplification, and library construction were performed according to the Chromium<sup>™</sup> Single Cell 3' Protocol - Chemistry v2. For GEM generation an input of 10,000 cells total, at 1,000 cells/ $\mu$ L density, was targeted for each sample, with a target cell recovery of 6,000 cells. Library construction, enzymatic fragmentation, End-repair and A-tailing were performed as follows: pre-cool block at 4 °C hold, fragmentation at 32 °C for 5 minutes; End repair and A-tailing 65 °C for 30 minutes and held at 4 °C. Post reaction cleanup was performed, followed by adaptor ligation. Adaptor ligation incubation was done at 20 °C for 15 minutes. Post adaptor ligation cleanup was then performed, followed by sample index PCR with the following parameters: 98 °C for 45 seconds; followed by 14 cycles: 98 °C for 20 seconds; 54 °C for 30 seconds; and 72 °C for 20 seconds; followed by 72 °C for 1 minute and held at 4 °C. Quantification of the constructed libraries was evaluated using QubitdsDNA HS Assay Kit (Thermo Fisher), Agilent cDNA High Sensitivity Kit, and Kapa DNA Quantification Kit for Illumina platforms, following the manufacturer's instructions. Generated libraries were sequenced on the Illumina HiSeq 2500, using the paired-end 2  $\times$  125 bp sequencing protocol. Sequencing run parameters were set up according to version 2 chemistry, the number of cycles for each read as follows: Read 1: 26 cycles, i7 index: 8 cycles, i5 index: 0 cycles and Read 2: 98 cycles.

**Data analyses.** We used the package Seurat<sup>21</sup> for data analysis using several adjustments to specifically increase the sensitivity to outliers. These tuned parameters and filtering involve the consideration of all genes as input regardless of the percentage of cells expressing them. No upper-bound limits were enforced when scaling the genes. The dataset was analyzed adjusting for the effect of the total amount of expression, or unique molecular identifiers (nUMIs), and the percentage of mitochondrial reads by means of linear regression. Subsequently, a Principal Component Analysis (PCA) was performed using only the genes identified as highly variable (1,223 and 1,054 for patient 1 and 2 respectively) for computing the principal components. These highly variable genes are selected among those with the highest variation among each of the thirty bins in which the genes are categorized according to their normalized strength of expression. No lower limits regarding strength of expression were imposed in order to capture the variation derived from the subtle expression signal of CTCs. Subsequently, we visualized this data in a two-dimensional PCA space to identify the outliers. The marker genes for these candidate CTCs were manually curated, and GSEA was run on scaled fold-changes of each CTC vs. all others cells. The loadings of the principal component (PC) 1 rank the genes that more accurately discriminate all the elements along this PC. As a proxy for expression differences in individual CTCs versus non tumoral cells, and given the limitations for differential expression testing in individual samples (dropout and lack of replicates to estimate variance), we used the residuals of expression after normalization (i.e., scaling, centering) and regression.

For the analysis of the rest of the blood cells we used a robust method with upper-limit scaling followed by the linear regression of the undesired effect of nUMIs and percentage of mitochondrial reads. This analysis was less contingent on outlier detection, so we followed a conventional approach using non-linear clustering, filtering out those genes which are expressed in less than 5 cells, and capping scaled maximum expression. Next, we performed dimensionality reduction with PCA on all genes, and later we clustered the cells with a graph-based algorithm for modularity optimization on principal component space. We added to this robust clustering method the supervised labeling of the outliers identified by the outlier-focused analysis (i.e., the CTCs. See above). After this, we computed marker genes<sup>22</sup> for the blood cell clusters modules (not CTCs). Subsequently we applied manifold learning techniques (non-linear dimensionality reduction) to gain insight on the structure of the data and clusters. In particular, we used tSNE<sup>23</sup> on PC space and used it to visualize the clusters in the blood.

In order to gain power to calculate marker genes for each cluster, we performed a combined analysis of the two single cell datasets of both patients. For the combined analysis, we selected the union of the top 2,000 genes with the highest dispersion (variance to mean ratio) from both datasets. We then ran a canonical correlation analysis (CCA) to identify common sources of variation between the two datasets. The result is, effectively, a form of dimensionality reduction, where 20 canonical vectors were calculated. The distributions of the cells' embeddings for these canonical vectors were aligned in order to remove batch effects (i.e. aligning the CCA subspaces). Next, on the combined dataset, we ran maker selection algorithms; first using Area Under the ROC Curve to rank the genes with most predictive power for each cluster. We used as well a bimodal (discrete and continuous) test<sup>22</sup> for single cell differential expression. Analyses were performed using R and raw sequencing data are deposited in NCBI's Gene Expression Omnibus under accession number GSE107747.

## References

1. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.* **10**, 472–484 (2013).
2. Amirouchene-Angelozzi, N., Swanton, C. & Bardelli, A. Tumor evolution as a therapeutic target. *Cancer Discovery* **7**, 805–817 (2017).
3. Liver, E. A. F. T. S. O. T. & Cancer, E. O. F. R. A. T. O. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. *J. Hepatol.* **56**, 908–943 (2012).
4. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975–2014, Featuring Survival. *Journal of the National Cancer Institute* **109** (2017).
5. Labgaa, I. & Villanueva, A. Liquid biopsy in liver cancer. *Discov. Med.* **19**, 263–273 (2015).
6. Gulbahce, N. *et al.* Quantitative whole genome sequencing of circulating tumor cells enables personalized combination therapy of metastatic cancer. *Cancer Res.* **77**, 4530–4541 (2017).
7. Jiang, R. *et al.* A comparison of isolated circulating tumor cells and tissue biopsies using whole-genome sequencing in prostate cancer. *Oncotarget* **6**, 44781–44793 (2015).
8. Miyamoto, D. T. *et al.* RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science (80-)* **349**, 1351–1356 (2015).
9. Kelley, R. K. *et al.* Circulating tumor cells in hepatocellular carcinoma: a pilot study of detection, enumeration, and next-generation sequencing in cases and controls. *BMC Cancer* **15**, 206 (2015).
10. Cann, G. M. *et al.* mRNA-Seq of Single Prostate Cancer Circulating Tumor Cells Reveals Recapitulation of Gene Expression and Pathways Found in Prostate Cancer. *Plos One* **7** (2012).
11. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
12. Satelli, A. *et al.* Potential role of nuclear PD-L1 expression in cell-surface vimentin positive circulating tumor cells as a prognostic marker in cancer patients. *Sci. Rep.* **6** (2016).
13. Kalinich, M. *et al.* An {RNA-based} signature enables high specificity detection of circulating tumor cells in hepatocellular carcinoma. *Proc. Natl. Acad. Sci.* **114**, 1123–1128 (2017).
14. Ogle, L. F. *et al.* Imagestream detection and characterisation of circulating tumour cells - A liquid biopsy for hepatocellular carcinoma? *J. Hepatol.* **65**, 305–313 (2016).
15. Li, J. *et al.* Detection of circulating tumor cells in hepatocellular carcinoma using antibodies against asialoglycoprotein receptor, carbamoyl phosphate synthetase 1 and pan-cytokeratin. *Plos One* **9**, e96185 (2014).
16. Harris, R. L., van den Berg, C. W. & Bowen, D. J. {ASGR1} and {ASGR2}, the Genes that Encode the Asialoglycoprotein Receptor (Ashwell Receptor), Are Expressed in Peripheral Blood Monocytes and Show Interindividual Differences in Transcript Profile. *Mol. Biol. Int.* **2012**, 283974 (2012).
17. Yu, X., Zheng, H., Chan, M. T. V. & Wu, W. K. K. {HULC}: an oncogenic long non-coding {RNA} in human cancer. *J. Cell. Mol. Med.* **21**, 410–417 (2017).
18. Schulze, K., Nault, J.-C. & Villanueva, A. Genetic profiling of hepatocellular carcinoma using next-generation sequencing. *J. Hepatol.* **65**, 1031–1042 (2016).

19. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
20. Martinez-Quetglas, I. *et al.* {IGF2} Is Up-regulated by Epigenetic Mechanisms in Hepatocellular Carcinomas and Is an Actionable Oncogene Product in Experimental Models. *Gastroenterology* **151**, 1192–1205 (2016).
21. Butler, A. & Satija, R. *Integrated analysis of single cell transcriptomic data across conditions, technologies, and species.* *bioRxiv* (2017).
22. McDavid, A. *et al.* Data exploration, quality control and testing in single-cell {qPCR-based} gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).
23. Pezzotti, N. *et al.* Approximated and User Steerable {tSNE} for Progressive Visual Analytics. *IEEE Trans. Vis. Comput. Graph.* **23**, 1739–1752 (2017).

## Acknowledgements

A.V. is the recipient of the American Association for the Study of Liver Diseases Foundation (AASLDF) Alan Hofmann Clinical and Translational Research Award, and he is supported by the U.S. Department of Defense (CA150272P3) and the Tisch Cancer Institute (Cancer Center Grant P30 CA196521). A.J.C. is supported by the National Cancer Institute Ruth L. Kirschstein NRSA Institutional Research Training Grant (CA078207). I.L. is supported by a grant from the Swiss National Science Foundation, from Foundation Roberto & Gianna Gonella and Foundation SICPA. D.D. is supported by the Grant for Studies Broadening from the Spanish Association for the Study of the Liver (Asociación Española para el Estudio del Hígado, AEEH) and the Cancer Research Grant from NuovoSoldati Foundation. J.v.F. is supported by Deutsche Forschungsgemeinschaft (German Research Foundation). B.M.H. is supported by PRIME (Program for Research on Immune Modeling and Experimentation), an NIAID-funded Modeling Immunity for Biodefense Center (Grant U19 AI117873). The authors thank the office of Scientific Computing and the Genomics Core Facility at the Icahn School of Medicine at Mount Sinai (ISMMS) for providing computational resources and staff expertise, the Flow Cytometry Core at the ISMMS and Yolanda Garcia (Department of Immunology, ISMMS) for technical support.

## Author Contributions

Study concept and design: D.D., A.V. Acquisition of data: D.D., S.M., I.L., M.S., P.T., A.C., A.B., A.K., R.S., S.M. Analysis and interpretation of data: D.D., C.V.M., A.V., B.H., J.V.F. Drafting of the manuscript: D.D., C.V.M., A.V. Critical revision of the manuscript for important intellectual content: D.D., A.V., C.V.M., J.V.F.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-30047-y>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018