Supplementary Materials for

Tumor-derived circulating endothelial cell clusters in colorectal cancer

Igor Cima, Say Li Kong, Debarka Sengupta, Iain B. Tan, Wai Min Phyo, Daniel Lee, Min Hu, Ciprian Iliescu, Irina Alexander, Wei Lin Goh, Mehran Rahman, Nur-Afidah Mohamed Suhaimi, Jess H. Vo, Joyce A. Tai, Joanna H. Tan, Clarinda Chua, Rachel Ten, Wan Jun Lim, Min Hoe Chew, Charlotte A.E. Hauser, Rob M. van Dam, Wei-Yen Lim, Shyam Prabhakar, Bing Lim, Poh Koon Koh, Paul Robson, Jackie Y. Ying, Axel M. Hillmer, Min-Han Tan*

*Corresponding author. Email: mhtan@ibn.a-star.edu.sg

DOI: 10.1126/scitranslmed.aad7369

The PDF file includes:

- Materials and Methods
- Fig. S1. Device setup, retrieval efficiency, and sample output purity.
- Fig. S2. scrmPCR methodology and proof of principle in cell lines.
- Fig. S3. CD45^- clusters express EMT markers, do not share common mutations with the primary tumor, and have normal chromosomal structures.
- Fig. S4. Cytomorphology of CD45^- clusters used for RNA-seq.
- Fig. S5. Cell-type inference workflow from RNA-seq data.
- Fig. S6. Validation of the cell-type inference algorithm.
- Fig. S7. Comparison between endothelial clusters (this study) and CTC clusters in the study of Aceto et al. (9).
- Fig. S8. Circulating endothelial clusters are tumor-derived.
- Fig. S9. tEC clusters in a mouse model of CRC.
- Fig. S10. tEC clusters do not correlate with selected CRC tumor and patient characteristics or comorbidities.

Legends for tables S1 to S14

References (38–42)

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencetranslationalmedicine.org/cgi/content/full/8/345/345ra89/DC1)

- Table S1 (Microsoft Excel format). scrmPCR primers used in this study.
Table S2 (Microsoft Excel format). WGA false-positives are detected only at VAF < 10%.
Table S3 (Microsoft Excel format). Circulating CD45− clusters do not mirror matching primary tumor mutations.
Table S4 (Microsoft Excel format). Sporadic mutations in circulating CD45− clusters are not detected in matching primary tumor tissues.
Table S5 (Microsoft Excel format). Germline variants in tumor tissues and circulating CD45− cell clusters (coverage and zygosity).
Table S6 (Microsoft Excel format). RNA-seq data: number of uniquely mapped reads to hg19 exons.
Table S7 (Microsoft Excel format). RNA-seq data: processed data (FPKM) of all samples described in the study.
Table S8 (Microsoft Excel format). Definition of the cell types used for the inference algorithm.
Table S9 (Microsoft Excel format). Top 80 genes ordered from highest to lowest specificity index (S) for each cell type selected for the inference algorithm.
Table S10 (Microsoft Excel format). Positive published RNA-seq control samples used to validate the inference algorithm.
Table S11 (Microsoft Excel format). Characteristics of tEC clusters isolated in this study.
Table S12 (Microsoft Excel format). Circulating endothelial cell cluster count before and after surgery.
Table S13 (Microsoft Excel format). Baseline patients’ and healthy donors’ characteristics.
Table S14 (Microsoft Excel format). CD45− cluster count for each baseline sample type and number of single clusters analyzed for each technique in the corresponding samples.
Materials and Methods

Clinical samples processing and data collection
All samples were collected in EDTA Vacutainer tubes (Becton-Dickinson), and processed within 6 hr at IBN. Two CRC cases were excluded from the analysis because of a technical issue with the microfiltration. Wherever available, matched tumor and metastatic samples, along with normal mucosa sampled at least 5 cm away from the tumor, were immediately frozen after resection, and stored at −80°C until use. Clinicopathologic data for participating subjects are described in table S11, and were collected retrospectively after completion of circulating clusters counts. Furthermore, table S12 includes circulating clusters counts for each baseline sample and includes number of single clusters analyzed for each technique and patient throughout the study. Clinical data was collected without prior knowledge of clusters counts. Similarly, clinical data for CRC patients were not known when determining clusters counts, except for the diagnosis and preoperative status of all FSH samples.

Device assembly
The retrieval device was assembled as follows. Firstly, the rubber plug of a 3-ml syringe plunger was removed, and a 5.5 mm-diameter hole was created using a punch cutter. The perforated rubber plug was placed in the 3-ml syringe. Next, an O-ring was placed in the slot of the sleeve insert, followed by the microsieve and another O-ring. Finally, the sleeve insert, together with the microsieve and O-rings were placed in the 3-ml syringe above the perforated rubber plug. This arrangement enabled the microfiltration of cells by size from the whole blood, and the subsequent retrieval of captured cells from the upper surface of the microsieve.

Microfiltration optimization
To optimize blood microfiltration with microsieve, cells labeled with 5 µM CellTracker (Life Technologies) were added to donor blood at 10–50 cells per ml of whole blood. Blood was microfiltered at various flow rates by means of a peristaltic pump (Ismatec). After 6 washes with PBS, 0.5% BSA and 2 mM EDTA, cells were resuspended in culture medium. Cell nuclei were then stained using Hoechst 33342 (Life Technologies), and cells were retrieved to determine retrieval efficiency and fold-depletion of contaminating WBCs. In some experiments, we also counted CellTracker-positive cells remaining on the microsieve. Percent retrieval efficiency was calculated as follows:

\[
\text{% Retrieval Efficiency} = \frac{(\text{Retrieved cells}) \times 100}{(\text{Spiked Cells})}
\]

Fold depletion from 1 ml whole blood was calculated as follows:

\[
\text{Fold Depletion} = \frac{(\text{WBCs in Whole Blood})}{(\text{WBCs in Microfiltrate})}
\]

WBC count in the microfiltrate was defined as the number of any Hoechst 33342 positive, CellTracker-negative event in the case of experimental enrichment, or as any CD45-positive event in the case of clinical samples. All clinical samples were immediately processed in downstream applications using the optimized parameters described in Figure 1C and D. To estimate the ideal target WBC depletion, we performed single-cell micromanipulation on serial dilution of PBMCs containing 50 CellTracker-positive HCT 116 cells. Micromanipulation of pure HCT 116 cells without contaminating white blood cells was easily performed after a 5,000-fold dilution of PBMCs. The ideal target retrieval efficiency was based on a literature search of label-free CTC isolation devices (17). Microfiltration of clinical samples was performed using 2 ml of whole blood for each microsieve device and the optimized microfiltration conditions.

**Single cell and single cluster micromanipulation**
Cells and clusters were identified using bright field or phase contrast imaging, nuclear staining and specific fluorescence signals. Target single cells or clusters were then manually micropipetted in a 10-µl droplet of wash buffer using a mouth pipette operated through a 25-ml syringe. Subsequently, single cells or clusters were deposited in 0.2-ml PCR tubes containing an appropriate buffer: 5 µl of 2× Reaction buffer (CellsDirect One-Step qRT-PCR Kit, Life Technologies) for scrmPCR; 2 µl of PBS for whole genome amplification; or 2 µl of SuperBlock buffer (Thermo Scientific) for low-input RNA-Seq. Cells were immediately stored at −80 °C until use. In some cases, the complete microfiltrate was spun down without micromanipulation and stored at −80 °C until further use.

**Cell lines and culture**

HCT 116, COLO 201, SW480, SW620, DLD-1, RKO, and CT26 colorectal cancer cell lines and BJ-5ta immortalized human foreskin fibroblasts were purchased from ATCC and cultured in DMEM (Life Technologies) supplemented with 10% FBS. HUVECs were purchased from ATCC, used at passage 1 to 3, and cultured in EGM-2 medium (Lonza). Human dermal and prostate microvascular cells (HDMECs and HPrECs, respectively) were purchased from ScienCell, cultured in Endothelial Cell Medium (ScienCell) on fibronectin-coated dishes, and used at passage 1 for RNA-Seq analysis and at passage 5 for the FOLH1 analysis. iPSCs were donated by Y. Chouduri (Institute of Bioengineering and Nanotechnology). All cells were maintained in a humidified incubator at 37 °C in the presence of 5% CO₂.

**Reagents for on-sieve immunofluorescence**

Following fluorescence-labeled antibodies were used: anti-CD45 1:200 (clone 2D1; eBioscience), anti-EpCAM 1:20 (9C4, BioLegend), anti-CD31 1:20 (WM59, BioLegend), anti-
CD144 1:10 (55-7H1, BD Biosciences), anti-CD41 1:20 (HIP8, BioLegend), anti-CD42B 1:20 (HIP1, BioLegend), anti-CD133 1:20 (AC133, BD Biosciences). Live iPSCs in suspension were used as positive controls for the CD133 staining. Fibrin was stained using anti-Fibrin 1:200 (Clone UC45, Pierce antibodies) and secondary goat anti-mouse IgG/IgM conjugated to Alexa 488 1:1000 (Pierce antibodies). Fibrin-positive controls included fibrin gels obtained by mixing 20 µl of fibrinogen droplets (20 mg/ml in PBS, pH = 7.2, Sigma-F3879) deposited on a well of a 96-well plate, mixed with 2 µl of thrombin (50 U/ml, Sigma-T4648) and incubated for 10 min at room temperature. For intracellular antigens, the Inside Stain Kit (Miltenyi Biotec) comprising a fixative and a permeabilization solution was used together with human FcR Blocking Reagent with the following antibodies: anti-VWF 1:200 (rabbit polyclonal A 0082, DAKO, conjugated in-house to Alexa 488 or Alexa 555 using Life Technologies APEX Antibody Labelling Kit), anti-Vimentin (V9, Santa Cruz Biotechnology), and anti-pan Cytokeratin (C11, Cell Signaling Technology). Nuclei were stained using Hoechst 33342 (Life Technologies). In some experiments, Calcein AM (Life Technologies) was used to identify live cells.

**Nucleic acid extraction**

Total RNA from tissues was isolated using the RNeasy mini kit (Qiagen). DNA from tissues was isolated using the DNeasy mini kit (Qiagen). Complete microfiltrates or laser-capture microdissected (LCM) samples were subjected to RNA extraction using the RNAqueous-Micro Total RNA Isolation Kit (Ambion) following the manufacturer’s instructions. Single cells and single clusters did not undergo nucleic acid extraction, but were lysed and used directly in downstream assays as described for the scrmPCR method, DNA sequencing and aCGH and RNA-Seq.
This method can simultaneously detect and quantify RNA transcripts and sequence DNA hotspots in the same cell. Briefly, single-cell RNA transcripts were reverse transcribed at 50 °C for 30 min using SuperScript III Reverse Transcriptase (Invitrogen) and a mix of 500 nM target reverse primers. A preamplification round was then performed using Platinum Taq DNA polymerase (Invitrogen) by adding a matching mix of forward primers to the transcript-specific reverse primers and primer pairs for targeted genomic regions. Preamplification cycling was conducted by alternating annealing and denaturation steps without extension as follows: 6 cycles at 60 °C, 4 min, 95 °C, 1 min; 6 cycles at 55 °C, 4 min, 95 °C, 1 min; 6 cycles at 50 °C, 4 min, 95 °C, 1 min. In some cases, we used 18 cycles at 60 °C, 4 min, 95 °C, 1 min. Primer clean-up was performed using the Axyprep PCR Clean-up Kit (Axygen) or using Exonuclease I enzymatic digestion (Thermo Fisher). Samples were diluted 1/20 and stored at −20 °C until further use.

To quantify RNA transcripts, quantitative PCR was performed on a ViiA7 Instrument (Applied Biosystems) using 2 µl of the preamplification mixture, semi-nested primer pairs appropriate for the target transcript (table S1) and the SensiFAST SYBR Lo-ROX Kit (Bioline) according to the manufacturer’s protocol. Relative gene expression was normalized using ACTB as the reference gene. Single cells or single clusters that did not show ACTB expression quantifiable in < 25 cycle threshold (Ct) values were discarded from the analysis. To analyze selected DNA mutational hotspots, PCR was performed using 2 µl of the preamplification mixture, nested PCR primer pairs (table S1) and a master mix containing a proof-reading polymerase (KOD Hot Start Master Mix, EMD Millipore) in line with the manufacturer’s instructions. For KRAS exon 2 sequencing of tumor and normal tissues shown in Fig. 2C, PCR amplification was performed using a primer pair different from the ones indicated in table S1):
TTTGTATTTAAAGGTACTGGTGAG as forward primer, and CCTTTATCTGTATAAGGAATGGTC as reverse primer. PCR products were separated on agarose gel; specific bands were excised and sequenced using the Sanger method.

**Digital droplet PCR (ddPCR)**

ddPCR was performed using the BioRad QX200 system and manufacturer instructions. Briefly, primers and probes were designed to target variants detected in three clusters from three different patients (fig. S3, B and C). For each sample, the PCR reaction was performed in duplicate using 450 nM forward and reverse primers, 250 nM probes targeting the mutant and the wild type alleles, 1x of ddPCR Supermix and 5ng of HindIII-digested DNA template using the following thermal cycling protocol: 95°C for 10 min; 40 cycles at 94°C, 30 s and 63°C, 1 min; 98°C for 10 min. Results were analyzed on the BioRad QX200 Droplet Reader using the QuantaSoft software. Following primers and probes were used: *AKT* c.G27A: Forward primer TACTACCCCCATCTCTCCCT, reverse primer CTGAGAGGAGCGTGAG, probe 1 (mutant) FAM-ACCTCGTTTGTGCAGCCAACCTTCCT-BLACK HOLE, probe 2 (wildtype) HEX-ACCTCGTTTGTGCAGCCAACCTTCCT-BLACK HOLE; *KRAS* c.G196A: Forward primer ATACACAAAGAAAGCCCTCC, reverse primer CAGACTGTGTTTCTCCCTTC, probe 1 (mutant) FAM-CCCTCATTGTACTGTACTCCTCTTTGACCTG-BLACK HOLE, probe 2 (wild type) HEX-CCCTCATTGTACTGTACTCCTCTTTGACCTG-BLACK HOLE; *NRAS* c.G97A: Forward primer TGGGTAAAGATGATCGACA, reverse primer GACTGAGTACA AACTGGTGG, probe 1 (mutant) FAM-CTGGGCCTCACCTCTATGGGATGATATATAT- BLACK HOLE, probe 2 (wild type) HEX-CTGGGCCTCACCTCTATGGGATGATATAT- BLACK HOLE.
**EPC assay**

Colony-forming EPC assays were performed as previously described (22). Live endothelial clusters were counted in 2-ml microfiltrates after CD144 and Calcein AM fluorescence staining. Unstained microfiltrates from 2 ml of blood from a second device were then placed in culture on a 96-well plate coated with fibronectin (1 µg/cm²) (Sigma-Aldrich) in the presence of EGM-2 cell culture medium (Lonza). The presence of clusters was confirmed by bright field microscopy before incubation. HUVECs were used as positive controls after spiking 10,000 HUVECs into 2 ml of donor blood and isolating them by microfiltration using two microsieve devices. In one device, retrieved HUVECs were quantified by CD144 and Calcein AM staining. HUVECs retrieved from the other device were seeded at 5, 10, 20, 40, 80, or 160 cells in octuplicate wells. After 2 days, the medium was changed and cell cultures were maintained for a total of 30 days, with half of the medium changed every other day. The presence and viability of colonies were monitored every week under bright field microscopy. After 30 days, cells were stained using CD144 antibodies, Calcein AM and Hoechst 33342, and quantified under an IX81 (Olympus) inverted fluorescence microscope.

**D-dimer ELISA**

D-dimer ELISA was performed using a RayBio human D-dimer ELISA kit (Raybiotech) with manufacturer instructions. Plasma samples were diluted 1:2,000,000 in the provided buffer and incubated overnight at 4°C.

**Isolation of endothelial cells from fresh tissues**

Tissues were minced before digestion with collagenase, dispase and DNaseI for 60 min at 37°C. After Ficoll-Paque density centrifugation, a two-step magnetic selection was performed using
MACS reagents and materials (Miltenyi Biotec) in line with the manufacturer’s instructions. First, CD45-expressing cells were depleted by retention on LD columns after labeling the cells with anti-CD45 magnetic beads and human FcR Blocking Reagent. Next, the CD45-depleted fraction was collected and re-labeled by adding anti-CD31 magnetic beads and human FcR Blocking Reagent. A fraction enriched for CD31+CD45− cells was harvested on MS columns before elution and storage at −80 °C until further use.

Laser capture microdissection of endothelial cells from archival frozen tissues
Eight-µm tissue sections were cut from fresh frozen samples of tumor and matched normal tissues, and placed on PEN-LCM membrane slides (Thermo Fisher). Slides were washed twice in buffer (PBS/EDTA/1%BSA/RNase inhibitor). After a 10-min peroxidase blocking step using Peroxidase Blocking Reagent (DAKO) and three washes in buffer, slides were stained for 20 min with anti-CD31 antibodies at 1 µg/ml in buffer (WM59, BioLegend). After three washing steps, bound antibodies were visualized using the REAL EnVision Detection System (Dako Cat. 5007) in line with the manufacturer’s instructions. Laser capture microdissection was immediately performed on the dry slides using the Arcturus LCM system (Thermo Fisher) following manufacturer’s instructions. Endothelia were captured on adhesive CapSure caps (Thermo Fisher) and stored at −80°C until further use. RNA was isolated using RNAqueous-Micro Total RNA Isolation Kit (Ambion) and stored at −80°C until further use. Low-input RNA-Seq libraries and sequencing were performed as described using 10 pg of RNA from each sample.

Cell type inference method
First, we generated a map of specific genes for each cell type of interest. The primary cell atlas data set (GSE49910) (38) was used for this purpose. From this data set we extracted 298
experiments of interest, corresponding to \( n = 36 \) different cell types or ‘lineages’ (table S8) without reprocessing of the data. Technical replicates were averaged. For each gene, \( g \), in each lineage, \( l \), a ‘specificity index’, \( S \), was calculated based on Shannon entropy and the \( Q \) statistics introduced by (39),

\[
S_{(l|g)} = - \sum_{i=1}^{N} p_{(l|g)} \cdot \log_2(p_{(l|g)}) - \log_2(p_{(l|g)})
\]

where \( p_{(l|g)} \) is the relative expression of gene \( g \) in lineage \( l \). Gene specificity was confirmed by visualizing gene expression data with high specificity indices using the primary cell line database from BioGPS (40) (this data base is also derived from GSE49910) (fig. S6A). For each cell type, the top 80 genes with the highest specificity index (‘specific genes’) were selected and were reported in table S9, which represents the map of specific genes for each lineage of interest. Next, for each RNA-Seq query sample, we set a predefined threshold to define expressed genes. In our dataset, we used FPKM > 0 as expressed genes. For the positive controls in figs. S6 and S7, we used the cut-offs reported in the corresponding studies. Next, in the query RNA-Seq list of expressed genes, we counted the occurrences of the top 80 specific genes for each cell type present in our map of specific genes (table S9). To determine if the number of enriched genes was different from enrichment by chance, we generated 1,000 lists of 80 randomly selected genes from the Affymetrix HG-U133_Plus_2 gene symbol list (‘random genes’) and counted the average number of genes present by chance in each experimental RNA-Seq profile for each lineage. Lastly, for each cell type in each experimental sample, a Fisher’s exact test was applied to determine whether the number of enriched specific genes was equal to the number of randomly enriched genes. The odds ratios for each test were mean-centered, scaled and visualized in a heat map comprising all tested cell types. The final results were used to generate hypotheses of cell types based on the distribution of the normalized odds ratios. The algorithm
was validated using lists of expressed genes from published RNA-Seq data from various cell types and tissues (fig. S6C).

**Principal component analysis on RNA-Seq data**

Principal component analysis (PCA) was performed in the R environment (version 3.1.0) (32) on RNA-Seq data set of single clusters and tissue samples (table S7). The *prcomp* function was used with Log$_2$(FPKM+1) data and with default parameters after excluding variables (genes) with 0 standard deviation and quantile normalization of the samples.

**Differential gene expression (DGE) and linear discriminant analysis (LDA) of RNA-Seq data**

DGE analysis and LDA of RNA-Seq data from laser-dissected nECs and tECs was performed by ‘NOISeq’ (41) and the *lda* function of the MASS package respectively, in the R environment (version 3.1.0). For DGE, input data were log$_2$(FPKM+1) normalized before analysis. To obtain differentially expressed genes, we applied the *noiseqbio* function with default parameters, except for using $r = 100$ permutations and setting the smoothing parameter $adj$ to 0.5. The list of genes presented in fig. S8D was selected using NOISeq probability >0.8 (corresponding to a q-value <0.2) and no filter on the log$_2$ fold change (log$_2$FC) data. For LDA, the data corresponding to the list of differentially expressed genes in the LCM samples were extracted from the log$_2$ normalized RNA-Seq dataset. Variables with >80% missing values [log$_2$(FPKM+1) = 0] in the clusters data were excluded from the analysis (*CDH2* and *ACAT*). A linear discriminant model using the remaining 6 variables, *ST6GALNAC2*, *WDR36*, *ARMC10*, *PSMB2*, *STOM*, and *NUCKS1* was computed on the nECs and tECs data using the *lda* function of the ‘MASS’ package, after verifying the absence of multicollinearity (Pearson’s r < 0.8 for each variable pair).
and the assumptions of normality and homogeneity of the covariance matrices by the Box’s M test \((P = 0.55)\). The model accuracy was tested using leave-one-out cross-validation (LOOCV) on the same data set and was reported in Fig. 4A. Finally, the model was applied to classify each single cluster. To calculate the expected probability of cluster classification by chance, we generated a series of 1000 random 6-gene signatures with Box’s M test \(P\) values > 0.1 and identical missingness probabilities, and tested each signature using the workflow described for the 6-gene signature. The exact binomial test was used to compare the observed probability \((P_o)\) of classified clusters as tECs with the mean expected probability \((P_e)\) of the randomly generated signatures classified as tECs.

**Microvessel analysis and lumen count**

Briefly, after scanning of the tissue sections using an Ariol SL-50 scanning platform Leica Biosystems, the tumor regions for all sections were marked and quantified. Within these regions, all areas with the greatest number of distinctly highlighted micro-vessels (hotspots) were marked and quantified. Next, nine images covering all hotspot areas were analyzed at 20× magnification for microvessel density (MVD), lumen counts and lumen size. MVD was determined using an automated ImageJ macro as follows. Ruifrok’s color deconvolution was applied to extract the CD31-positive signal. After image binarization, a common threshold level was chosen to represent correct vascular morphology and minimize background levels across all slides. Morphological post-processing steps were applied to join nearby objects, split objects and close gaps within objects similarly as described (35). Objects larger than a minimal size (default: 105 pixels, corresponding to 50 µm²) were counted as microvessels. Smaller size objects were considered as noise and excluded from the analysis. The ImageJ macro was validated by comparing it with manually counted images (Pearson’s \(r = 0.951, n = 30\) images). Lumen count
and perimeters were determined manually for each image using ImageJ measurement and count features. Mosaic and damaged vessels were counted in the tumor area and in available normal tissues present on the same slide by screening each section using a 20× magnification. False-positive stains (e.g. hematopoietic cells positive for CD31) were excluded from the analysis. These cells were identified in each tissue by morphology and lower staining intensity compared to microvessels. Mean values of 9 images per section were used for statistical assessments. Patient’s IDs and clusters counts were blinded during the analysis to avoid subjective bias during acquisition of the data.

Tail vein injection of tumor-derived endothelial cells in experimental mice. CT26 cells (500,000 in 100 µl of sterile PBS) were subcutaneously injected into the left flank of 7-week-old Balb/c mice. At 14 days after injection, mice were euthanized and the tumors (about 2 cm-diameter) were isolated for downstream processing. Tumor samples were embedded in optimal cutting temperature compound (OCT, Sakura Finetek). The remaining tumor tissues were minced using scalpel blades and digested with a solution of collagenase and dispase (Liberase DH, Roche) for 2 hr at 37 °C in cell culture medium with occasional stirring. After two washing steps, the tissues were incubated with DNase I (Thermo Fisher) for 15 min, and filtered through a 40-µm cell strainer to obtain single-cell suspensions. Cells were stained for 45 min at 4 °C using 1 µg/ml of allophycocyanin (APC)-labeled anti-CD309 antibodies (Avas12a1, eBioscience), diluted in PBS/EDTA/1%BSA or matching Rat IgG2a K Isotype control (eBioscience), in the presence of anti-mouse FC Receptor Block (eBioscience). After two washing steps, cells were stained with Propidium Iodide (PI) Readyprobes (Thermo Fisher). Cell suspensions were then sorted using a MoFlo XDP Cell Sorter to obtain a total of 500,000 CD309⁺PI⁻ single cells. Cells were subsequently divided into two equal fractions. One fraction was labeled using Celltracker
Orange (Invitrogen), while the other fraction was labeled using Celltracker Green (Invitrogen). Cell suspensions were mixed and 50,000 cells were injected into the tail vein of nine 7-week-old Balb/c mice. Mice were euthanized at the indicated time points, and the presence of single fluorescence events (single cells) or dual fluorescence events (clusters) were analyzed in blood samples by a FACS Calibur instrument (BD Biosciences), after lysis of the erythrocytes fraction using BD Pharm Lyse (BD Biosciences) and following manufacturer’s instructions. The acquired data were analyzed using FlowJo version 10.0.7 (Tree Star).

**Tumor-derived endothelial cell clusters in experimental mice.** CT26 cells (500,000 in 100 µl of sterile PBS) were subcutaneously injected into the left flank of two 6-week-old L2G85.BALB/c mice (The Jackson Laboratory). This mouse strain carries the GFP transgene under the ubiquitous CAG promoter. After 14 days, mice were euthanized and tumors were subcutaneously transplanted into five wild-type 6-week-old Balb/c mice. Fourteen days later, the mice were euthanized, and the blood samples were immediately collected in EDTA Vacutainer tubes (Becton-Dickinson) by cardiac puncture. Tumor tissues were collected and dissected in three segments. Each segment was tested for the presence of the GFP transgene by genotyping its genomic DNA with the recommended protocols and the following primers: oIMR0872 AAGTTTCATCTGCACCACCG as GFP forward primer, oIMR1416 TCCTTGAAAGATGGTGCG as GFP reverse primer, oIMR7338 CTAGGCCACAGAATTGAAAGATCT as internal positive control forward primer, and oIMR7339 GTAGGTGGAAATTCTAGCATCATCC as internal positive control reverse primer. One mouse was excluded from subsequent analysis because of the absence of GFP signal in any one of the three segments. Blood samples were processed within 2 h of collection. Erythrocytes were lysed using BD Pharm Lyse (BD Biosciences) and following manufacturer’s instructions.
Cells were washed in buffer (PBS/EDTA/1%BSA), and stained using 1 µg/ml of anti-CD309-APC (Avas12a1, eBioscience) or matching rat IgG2a K Isotype control (eBioscience) and Hoechst 33342, in the presence of anti-mouse FC Receptor Block (eBioscience). Stained cells were analyzed using a LSR II instrument (BD Biosciences) with independent laser sources for each fluorescent signal. Acquired data were analyzed using FlowJo version 10.0.7 (Tree Star). Blood from three control Balb/c mice was used as negative control.
**Supplementary Figures:**

**Fig. S1.** Device setup, retrieval efficiency, and sample output purity. (A) Device setup. Microfiltration devices enclosing a silicon microsieve (inset scale bar, 10 µm), connected to a peristaltic pump. (B) Capture efficiency of SW620 cells from whole blood, indicating percentage of captured cells on the microsieve that could be retrieved for downstream assays, or that were lost because the cells adhered to the microsieve. Four independent experiments are shown. (C) Size distribution of SW620, HCT 116, SW480, and COLO201 cells ($n = 50$ cells each). Arrows indicate median size of WBCs and CTCs isolated from colorectal, prostate and breast cancer patients respectively, reported in (14). (D) Contaminating nucleated cells after microfiltration of clinical samples using the optimized protocol shown in Fig. 1, B and C ($n = 13$ blood samples). (E) Retrieval efficiency from additional cell lines using optimized protocol shown in Fig. 1, B and C. Data are means ± SEM ($n = 3$ experiments for each cell line).
Fig. S2. scrmPCR methodology and proof of principle in cell lines. (A) scrmPCR workflow for single cells or single clusters. (B) Concurrent DNA amplification (gel electrophoresis) and RNA quantitation (heat map) for the indicated genes and representative single or bulk cells. (C) Sequence chromatograms from DNA hotspots of known mutant alleles in RKO and DLD-1 single cells. (D) Melting curve plots for the indicated genes from qPCR reactions. sc, single cell; NTC, no-template control; Rn, normalized relative to reporter signal.
**Fig. S3.** CD45− clusters express EMT markers, do not share common mutations with the primary tumor, and have normal chromosomal structures. (A) Representative four-color immunofluorescence staining for CD45, vimentin (VIM), pan-cytokeratin (CK), and DAPI, indicating heterogeneous mesenchymal and epithelial marker expression by 2 clusters from two
different patients. (B) Representative 2-D plot of droplet digital PCR confirming the absence of matching mutations between clusters and tumor tissues. Green dots indicate droplets containing PCR products from wild-type alleles; red and blue dots indicate droplets containing PCR products with mutant allele amplification; gray dots are no amplification. (C) Fractional abundance of each replicate and tissue sample or clusters assayed by ddPCR on 5 ng DNA template. Red values indicate positive signals; values <0.2% were below the assay maximum sensitivity. (D) Control experiment to assess the impact of whole genome amplification (WGA) for aCGH experiments. (E) Array comparative genomic hybridization (aCGH) of single clusters and matched tissues for three patients, in addition to patient 10 in Fig. 2E.
**Fig. S4. Cytomorphology of CD45− clusters used for RNA-seq.** Phase contrast and Hoechst 33342 images of clusters used for RNA-Seq for each patient. No images are available for patient P1 because the cells were directly deposited in RNA-Seq buffer without imaging. Scale bars, 10 µm.
Fig. S5. Cell-type inference workflow from RNA-seq data.
Fig. S6. Validation of the cell-type inference algorithm. (A) Selected genes with the highest specificity index for representative cell types were verified for specificity using BioGPS. (B) Gene expression levels for markers commonly used in CTC research to define epithelial cells. Note KRT18 expression in the endothelial lineages, and EPCAM expression in embryonic stem.
cells and hematopoietic cells. Data are represented as horizontal barplots derived from the output of the primary cell atlas dataset of BioGPS (C) Heat map comparing the number of genes enriched for each published positive control sample (rows) and cell type (columns) over random enrichment. Each colored box represents a normalized odds ratio of the respective Fisher’s exact test ranging from 0 (black) to 1 (red). (D) IDs for each cell types reported in table S8. Order of cell types corresponds to the heatmap columns shown in (C).
Fig. S7. Comparison between endothelial clusters (this study) and CTC clusters in the study of Aceto et al. (9). Selected breast cancer cell lines with epithelial and mesenchymal lineage profiles and primary endothelial cells were used as positive controls for epithelial, mesenchymal stem cells and endothelial lineages. Lineage inference of CTC clusters reported in (9) shows the presence of epithelial-derived cell clusters. Platelet and red blood cell signals were removed from this analysis to highlight cell types belonging to the above-mentioned lineages of interest.
Fig. S8. Circulating endothelial clusters are tumor-derived. (A) Principal component analysis reveals an association of endothelial clusters with tumor tissues. Left panel, principal component 1 (PC1) vs PC2. PC1 correlates with the standard deviation of inputed variables for single clusters ($r = -0.455$) and not does distinguish tumor and normal tissues (batch effect). Right
panel, PC2 vs PC3 indicating association of endothelial clusters with the tumor tissues but not with the normal tissues. Colored areas represent k-means clustering. (B) Cell type inference of whole normal and tumor tissues and matched laser-captured endothelial cells, showing enrichment of the target cell population. (C) Unsupervised hierarchical clustering of endothelial cell populations show significant association of circulating endothelial clusters with intestinal endothelial cells. nECs, colon-derived normal endothelial cells; tECs, CRC-derived tumor endothelial cells. (D) Genes differentially expressed between nECs and tECs. $P_{NOI}$ probability of differential expression as computed by NOISeq. Log$_2$FC, log$_2$(fold change). (E) Heatmap showing the expression of published tumor endothelial markers (23, 25, 42) in single circulating endothelial clusters, medians of normal tissue, tumor tissue and matched normal and tumor laser-capture dissected endothelial cells. Rows of log$_2$(FPKM+1) values were normalized by scaling between 0 and 1. Values are shown along with color-coding of the scale.
**Fig. S9. tEC clusters in a mouse model of CRC.** (A) Experimental approach. (B) GFP genotyping in CT26 xenografts at passage 0 and passage 1. Mouse 2 was excluded from subsequent analysis. (C) Gating strategy for the analysis of tumor-derived (GFP+CD309+) endothelial cell clusters. Biex, Biexponential scale. (D) Immunofluorescence staining of tumor vasculature using CD309+ antibodies at passage 0 and passage 1. Scale bar, 50 μm. (E) Boxplot represent quantification of CD309+GFP+ single and clustered cells in the blood of P1 xenografts and controls. Data are individual animals.
Fig. S10. tEC clusters do not correlate with selected CRC tumor and patient characteristics or comorbidities. Association of circulating endothelial cell cluster counts with selected tumor characteristics (tumor location or size, immune cells, blood markers) (A), general population characteristics (age, BMI, gender) (B), and comorbidities (hypertension, diabetes, heart disease) (C). Data refer to patients, for which all available clinical data were used in the analysis. Correlations were tested using the Kendall’s Tau method. Unpaired samples were tested using a two-tailed Wilcoxon-Mann-Whitney U test.
SUPPLEMENTARY TABLES

Tables S1-S14 are provided in one Supplementary Excel spreadsheet, with captions below.

Table S1. scrmPCR primers used in this study.

Table S2: WGA false-positives are detected only at VAF < 10%. WGA and unamplified DNA derived from the same sample (Patient 15) were sequenced to detect false positive mutations arising from the amplification procedure.

Table S3. Circulating CD45− clusters do not mirror matching primary tumor mutations. Targeted high-throughput sequencing of single CD45− clusters. Numbers in colored boxes indicate variant allele frequency (VAF) (coverage). Cl, cluster; ND, not detected; NA, not available.

Table S4. Sporadic mutations in circulating CD45− clusters are not detected in matching primary tumor tissues. Blue rows indicate variants tested by digital PCR and reported in fig S3. Numbers in colored boxes indicate variant allele frequency (VAF) and coverage. ND, not detected.

Table S5. Germline variants in tumor tissues and circulating CD45− cell clusters (coverage and zygosity). NA, not available; ND, not detected.

Table S6. RNA-seq data: number of uniquely mapped reads to hg19 exons.

Table S7. RNA-seq data: processed data (FPKM) of all samples described in the study.

Table S8. Definition of the cell types used for the inference algorithm.

Table S9. Top 80 genes ordered from highest to lowest specificity index (S) for each cell type selected for the inference algorithm.

Table S10. Positive published RNA-seq control samples used to validate the inference algorithm.

Table S11. Characteristics of tEC clusters isolated in this study. Staining intensity: −, negative; +, low; ++, average; ++++, strong.

Table S12. Circulating endothelial cell cluster count before and after surgery. Data are from Fig. 4B. Pre: Clusters count in blood 0-24 hrs before surgery. Post: Clusters count in blood 24-72 hrs after surgery.

Table S13. Baseline patients’ and healthy donors’ characteristics.
Table S14. CD45<sup>-</sup> cluster count for each baseline sample type and number of single clusters analyzed for each technique in the corresponding samples. Sources: FORTIS, Fortis Surgical Hospital, Singapore; NCC, National Cancer Center, Singapore; IBN, Institute of Bioengineering and Nanotechnology, Singapore; SPHS, Singapore Population Health Studies. CRC, colorectal cancer.