Received: 28 February 2024 | Revised: 29 May 2024 | Accepted: 10 June 2024

DOI: 10.1002/ijc.35079

RESEARCH ARTICLE

Innovative Tools and Methods

IJC INTERNATIONAL JOURNAL of CANCER | UICC

High-confidence calling of normal epithelial cells allows identification of a novel stem-like cell state in the colorectal cancer microenvironment

Tzu-Ting Wei[1] | Eric Blanc[1] | Stefan Peidli[2,3] | Philip Bischoff[2,4,5] | Alexandra Trinks[6] | David Horst[2,5] | Christine Sers[2,5] | Nils Blüthgen[2,3,5] | Dieter Beule[1] | Markus Morkel[2,3,6] | Benedikt Obermayer[1]

[1]Core Unit Bioinformatics, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

[2]Institute of Pathology, Charité – Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[3]Institute of Biology, Humboldt University of Berlin, Berlin, Germany

[4]BIH Biomedical Innovation Academy, BIH Charité Clinician Scientist Program, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

[5]German Cancer Consortium Partner Site Berlin, German Cancer Research Center, Heidelberg, Germany

[6]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Core Unit Bioportal Single Cells, Berlin, Germany

Correspondence
Markus Morkel, Institute of Pathology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany.
Email: markus.morkel@charite.de

Benedikt Obermayer, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Core Unit Bioinformatics, Charitéplatz 1, 10117 Berlin, Germany.
Email: benedikt.obermayer@bih-charite.de

Present address
Stefan Peidli, European Molecular Biology Laboratory, Heidelberg, Germany.

Funding information
BIH-funded PeDiOn and Clinical Scientist programs; Deutsche Forschungsgemeinschaft, Grant/Award Number: RTG CompCancer GRK2424/1

Abstract

Single-cell analyses can be confounded by assigning unrelated groups of cells to common developmental trajectories. For instance, cancer cells and admixed normal epithelial cells could adopt similar cell states thus complicating analyses of their developmental potential. Here, we develop and benchmark CCISM (for Cancer Cell Identification using Somatic Mutations) to exploit genomic single nucleotide variants for the disambiguation of cancer cells from genomically normal non-cancer cells in single-cell data. We find that our method and others based on gene expression or allelic imbalances identify overlapping sets of colorectal cancer versus normal colon epithelial cells, depending on molecular characteristics of individual cancers. Further, we define consensus cell identities of normal and cancer epithelial cells with higher transcriptome cluster homogeneity than those derived using existing tools. Using the consensus identities, we identify significant shifts of cell state distributions in genomically normal epithelial cells developing in the cancer microenvironment, with immature states increased at the expense of terminal differentiation throughout the colon, and a novel stem-like cell state arising in the left colon. Trajectory analyses show that the new cell state extends the pseudo-time range of normal colon stem-like cells in a cancer context. We identify cancer-associated fibroblasts as sources of WNT and BMP ligands potentially contributing to increased plasticity of stem cells in the cancer microenvironment. Our analyses advocate careful interpretation of cell

heterogeneity and plasticity in the cancer context and the consideration of genomic information in addition to gene expression data when possible.

KEYWORDS
cellular heterogeneity, single-cell genomics, somatic variants

What's New?

Single-cell transcriptomics is a standard means of assessing cell heterogeneity and cell hierarchies in cancer tissues. However, single-cell datasets are complex and contain cancer and non-cancer lineage cells. Here, the authors compared different strategies to analyze gene expression and genomic information and retrace the origins of epithelial cell transcriptomes in colorectal cancer (CRC) cells. Haplotype-aware copy number inference combined with a novel method to assess somatic single nucleotide variants exhibited high accuracy in differentiating between cancerous and genetically normal cells found within cancer tissue. The findings offer a novel approach to account for biological and genetic features of CRC.

1 | INTRODUCTION

Cancer cells mix and interact with their microenvironment.[1,2] In colorectal carcinoma (CRC) and in other epithelial cancers, transformed cells intermingle with non-cancer epithelial cells in areas known as the invasive front (IF).[3] Furthermore, normal tissues adjacent to tumors are re-shaped beyond the cancer's boundary, influenced by local immune responses and inflammation,[4] paracrine signals,[5] and genetic aberrations preceding malignant transformation,[6] as has been shown by multiplexed tissue imaging,[7] single-cell[8] and bulk transcriptomics.[9] This gradual change in cell composition from normal to cancer poses challenges for single-cell transcriptomics, as it is not immediately apparent from the transcriptome whether certain cells arise from malignant or normal lineages.

In CRC, single-cell transcriptome analyses revealed two overarching intrinsic consensus molecular subtypes (iCMS), termed iCMS2 and iCMS3.[10] These transcriptome subtypes are linked to patient characteristics such as localization of cancer, and to molecular features such as microsatellite stability, mutational burden, the extent of copy number aberrations, and patterns of driver mutations.[11–14] That means, left-sided tumors frequently arise due to the loss of the tumor suppressor gene APC and additionally harbor mutations in KRAS, SMAD4, and TP53; these mutational patterns lead to WNT and MYC signaling pathway activation. Furthermore, CRCs in this context are most frequently microsatellite-stable (MSS), display extensive copy number aberrations and gene expression patterns characteristic of intrinsic molecular subtype iCMS2. In contrast, CRCs progressing via serrated precursors are found mainly in the right colon, carry mutations in KRAS or BRAF, display activation of the TGF-beta signaling pathway, can be microsatellite-instable (MSI) or MSS, have a higher mutational burden but fewer copy-number changes, and show gene expression patterns of metaplasia and intrinsic molecular subtype iCMS3. We expect that the different cancer cell characteristics could also lead to a variable accuracy of cell type calling in single-cell analysis.

Numerous studies have conducted single-cell level analyses of CRC.[15–17] These investigations were either performed under the assumption that all epithelial cells derived from the cancer tissue samples are bona fide cancer cells, or they have relied solely on transcriptome-derived characteristics to differentiate between cancer and normal epithelial cells. Broadly applicable and robust methods to confidently distinguish genomically normal epithelial cells from genomically aberrant cancer cells remain elusive, especially for datasets derived from regions where both types of cells coexist, such as at the IF.

Here, we use different computational tools to disambiguate cancer and non-cancer epithelial cells in single-cell transcriptome data of 10 CRC patients across a range of clinical and molecular characteristics, using additional information derived from associated whole-genome sequencing data. Analysis of consensus sets of cancer and normal cells shows that genomically normal epithelial cells adjacent to the cancer can adopt cell states that are unlike those of epithelial cell populations in normal tissue. Developmental trajectories of non-cancer epithelium were altered in the cancer neighborhood, as stem-like and immature differentiation states were overrepresented among genomically normal cells in cancer tissue samples. We identify multiple new paracrine interactions potentially modulating normal cell development in the tumor microenvironment, including cancer-specific fibroblasts as a source of the key stemness factor WNT.

2 | MATERIALS AND METHODS

2.1 | Sample collection and data preprocessing

The sample collection and experimental processing of the clinical specimen for single-cell RNA sequencing data has been described before[16] and the new data for patient P35 was collected and processed using the same protocols. In short, tissues were processed using the Miltenyi Human Tumor Dissociation Kit

(Miltenyi, no. 130-095-929) and a Miltenyi gentleMACS Tissue Dissociator (Miltenyi, no. 130-096-427), using program 37C_h_TDK_1 for 30–45 min. Single-cell libraries were generated using the Chromium Single-Cell 3′Reagent Kits v3 and the Chromium Controller (10× Genomics). Libraries were sequenced on a HiSeq 4000 Sequencer (Illumina) at 200–400 Mio reads per library.

Whole genome sequencing (WGS) data was performed using genomic DNA isolated from microdissected material of snap-frozen (−80°C) CRC tissue, adjacent to material used for single-cell sequencing. DNA was isolated using Qiagen Allprep Kits and sequenced on the Illumina NovaSeq 6000 platform using 2 × 150 bp reads. Between 230 and 360 m reads were generated per sample. Reads were mapped using bwa-mem[18] version 0.7.17 against release GRCh38 of the human genome with decoys and virus sequences. For single-cell RNA sequencing data, UMIs were quantified using CellRanger 3.0.2[19] with reference transcriptome GRCh38. See Table S1 for sequencing statistics.

2.2 | Single-cell data quality control

All analyses on single-cell data were conducted with Python 3.9.10, Scanpy 1.8.0,[20] Numpy seed set at 123, R 4.1.2, and Seurat 4.1.1,[21] if not specifically mentioned. CellBender v0.2.2[22] was used to remove ambient RNA with default parameters, 5000 expected cells, and FDR rate at 0.01. We used Scrublet[23] for doublet removal and chose the score threshold at 0.3 after inspecting the observed and simulated doublet scores distributions of all the samples. The detected doublet rates ranged from 0.7% to 2.9%. For quality control, cells with min_counts <1000, min genes <500, or mitochondrial percentage >80% were removed, resulting in a total number of 73,294 cells. The count matrix was then normalized and log1p transformed. The top 2000 highly variable genes (HVGs) were identified with "patient" as the batch key. Principal component analysis (PCA) was conducted, and we calculated a UMAP using 50 neighbors and 20 principal components.

2.3 | Somatic variant calling in WGS and genotyping of single-cell RNA-seq

Somatic variants in whole genome sequencing data were called by Mutect2 from GATK version 4.2.0.0[24] using default parameters. The GATK public resources were used for germline variant loci, common biallelic loci were used to estimate possible contamination, and for the panel of normals. CellSNP-lite[25] 1.2.2 was used to count somatic variants in single-cell RNA sequencing data against WGS filtered.vcf files with parameters --genotype -p 22 --minMAF 0.001 --minCOUNT 1.

2.4 | CCISM model and data simulation

Cancer Cell Identification using Somatic Mutations (CCISM) is a tool for the classification of single-cell expression data based on the expectation–maximization method in Cardelino.[26] Given the total number d_{ij} of (UMI-collapsed) reads covering variant i in cell j (reference and variant allele), and the number a_{ij} of UMIs supporting the alternative allele, we evaluate the likelihood $p_{T,j}$ that cell j is a tumor cell using a binomial model:

$$p_{T,j} \propto \prod_i \binom{d_{ij}}{a_{ij}} \theta_T^{a_{ij}} (1-\theta_T)^{d_{ij}-a_{ij}}.$$

Here, θ_T is the "success probability" for the somatic variants, measuring how likely it is to observe UMIs supporting the variant allele. Similarly, we compute $p_{N,j}$ as the likelihood that cell j is normal, with a fixed nonzero parameter $\theta_N = 0.01$ allowing for sequencing errors and uncertainties in the variant calls. We calculate $p_{T,j}$ and $p_{N,j}$ in the E-step and estimate the parameter θ_T in the M-step as weighted sum over the counts d_{ij} and a_{ij}:

$$\theta_T = \frac{\sum_j (1 + p_{N,j}/p_{T,j})^{-1} \sum_i a_{ij}}{\sum_j (1 + p_{N,j}/p_{T,j})^{-1} \sum_i d_{ij}}.$$

E- and M-steps are iterated until convergence of the likelihood

$$\ln \mathscr{L} = \sum_j \ln(p_{T,j} + p_{N,j}).$$

Finally, the likelihoods are normalized to give the posterior cancer cell assignment of a particular cell $p_j = p_{T,j}/(p_{T,j} + p_{N,j})$ and a cutoff $p_j > .5$ is used to define likely cancer cells.

For the benchmark simulations (see also McCarthy et al.[26]), we take the matrix d_{ij} from a given dataset and simulate values a_{ij} using a binomial distribution with parameters $\theta_T = 0.4$ and $\theta_N = 0.0001$ for randomly assigned tumor and normal cell identity, respectively. We used the R package cardelino (v0.6.5) and the BinomMixtureVB function from the vireoSNP package (v0.5.6) for comparison.

2.5 | Methodology for consensus cancer calls and trajectory assignments

Epithelial, immune, and stromal cell identity was scored and assigned using previously published cell type markers.[27] We ran a separate PCA for the epithelial cell compartment and chose 20 neighbors and 15 PCs for the UMAP visualization.

Copy number inference from gene expression profile was performed using inferCNV v1.3.3[28] with default parameters on all the epithelial cells with CellBender-processed[22] counts (filtered_h5). The input gene expression profiles were smoothed with a window of 101 genes. The generated dendrograms were cut at $k = 2$ for each patient, and clones were assigned as copy number-aberrant if their averaged smoothed gene expression profile deviated by more than 3 SD from that of clones containing cells of normal samples. Numbat[29] v1.0.3 was run with the epithelial cells from the matched

normal samples and using default parameters, which included cellSNP-lite v1.2.2 for pile up and Eagle v2.4.1 for phasing the reads. The four samples from P09 (n1, n2, t1, t2) were piled up and phased together, and P26t and P35t were piled-up and phased separately as there were no matched normal samples. The rest of the samples were processed as paired normal and tumor samples.

For iCMS label transfer, we downloaded the CellRanger-processed count matrix ("Epithelial_Count_matrix.h5"), and the cell-level metadata ("Epithelial_metadata.csv") from the source data[10] (Synapse accession code: syn26844071, https://www.synapse.org/#!Synapse:syn26844071/), filtered by min_genes = 500 and min_counts = 1000, and concatenated this count matrix with ours. The resulting matrix was integrated by scVI with data source as covariate and passed to scANVI to learn the iCMS labels. We found that learning with only the Joanito et al.[10] gene list (1318 genes including a signature for normal cells obtained by personal request from the authors) was suboptimal since it only captured a small proportion of gene expression variance. Therefore, we used the union of all highly variable genes in either dataset and the iCMS signature genes. The resulting matrix was integrated by scVI with data source as covariate and passed to scANVI to learn the iCMS labels.

For the consensus cell identity assignment, we extracted the assignment probability from the outputs of Numbat (p_cnv) and CCISM (CCISM_p), and assigned the cell identity by the following rules: A cell is annotated as genomically cancer cell if (1) p_cnv and CCISM_p are both >0.5; or (2) CCISM_p = 0.5 and p_cnv >0.5; or (3) p_cnv >0.5 in MSS samples; or (4) CCISM_p >0.5 in MSI samples. A cell is annotated as genomically normal cell if p_cnv and CCISM_p are both <0.5. A cell that does not fit into any of the categories above is annotated as "unclear" and removed from the downstream analysis.

For detailed epithelial cell type annotation, we used scVI and scANVI to integrate datasets and learn cell type labels from Uhlitz et al.[16] The scVI models were trained on the raw count matrix (adata.layer["count"]) of 2000 highly variable genes using scvi-tools v0.19.0 with patient and percent_ribo as covariates. These models were used by scANVI as input to predict cell type labels of newly included cells based on the annotation of previously annotated cells.

The linear mixed model for cell type composition was composed using the "glmer" function with binomial distribution from the lmer package.[30] For each cell type, we tested if there is a difference between genomically normal cells and healthy cells from normal samples, where patient was included as a random effect variable.

To enhance concrete transcriptomic contrasts between cancer and normal cells, 1498 cells from normal samples that were assigned as tumor-centric cell type, namely TC1-4, were removed from the downstream analysis. The epithelial cell type of genomically cancer cells was then assigned as "cancer-like" in transcriptomic analysis. Diffusion maps were calculated with 15 neighbors and CytoTrace pseudotime as implemented in CellRank 1.5.2. dev236 + gab03900.[31]

2.6 | Methodology for scoring CRC signaling pathways and inferring paracrine interactions

We curated a list of known ligands and receptors of key signaling pathways in CRC and a list of CRC signature genes for specific phenotypes from literature (Table S2). The expression levels of CRC signatures were calculated using "score_gene" function in Scanpy. The paracrine interactions within normal and tumor samples were inferred by CellChat v1.6.1.

2.7 | Re-analyzing lung adenocarcinoma data

For the lung adenocarcinoma data,[32] we performed whole-genome sequencing followed by calling of somatic mutations as described above. We then used scRNAseq bam files for these samples to run cellsnp-lite and CCISM. We finally merged CCISM cancer cell calls into an R object with cell type annotation and inferCNV results provided by the authors.

3 | RESULTS

3.1 | Transcriptome information is insufficient for cancer cell calling in CRC

To reliably distinguish cancer from normal cells in single-cell RNAseq data, we complemented single-cell data of 10 treatment-naive CRC patients of a previous study[16] with whole-genome sequencing data of cancer and normal samples. Clinical and pathology assessment of the cohort shows a broad distribution along the longitudinal axis of the colon, and driver mutations in APC, BRAF, P53, beta-Catenin, and KRAS in subsets of the cancers (Figure 1A). Using updated bioinformatic pipelines, 73,294 cells passed quality controls after ambient RNA and doublet removal. Of these, 43,110 transcriptomes were from cancer tissue and 30,184 were from normal tissue samples adjacent to tumor. Across all samples, 39,168 cells were annotated as epithelial, 31,663 as immune and 2463 as stromal cells(Figure 1B).

We first sought to distinguish cancer from normal epithelial cells in the cancer samples using transcriptome information. In a UMAP representation of all epithelial cell transcriptomes, a fraction of the 17,623 transcriptomes derived from cancer samples clustered as a separate "community" while another fraction interspersed with the normal tissue-derived epithelial cells (Figure 1C). We used probabilistic label transfer from published gene expression data[10] to assign cancer sample epithelial cells to the cancerous iCMS2 or iCMS3 epithelial cell states, or a normal cell state (Figure 1D,E). In total, 10,589 cells were classified as iCMS2 or iCMS3 and therefore were assigned as cancer cells by this method. Cancer cells from P09, P13, P16, and P21 were predominantly called as iCMS2, whereas P07, P08, P14, P20, P26, and P35 were mostly iCMS3. In line with previous observations, we find that cells from each sample were mostly from a single dominant iCMS type, with only a small minority of cells assigned to the

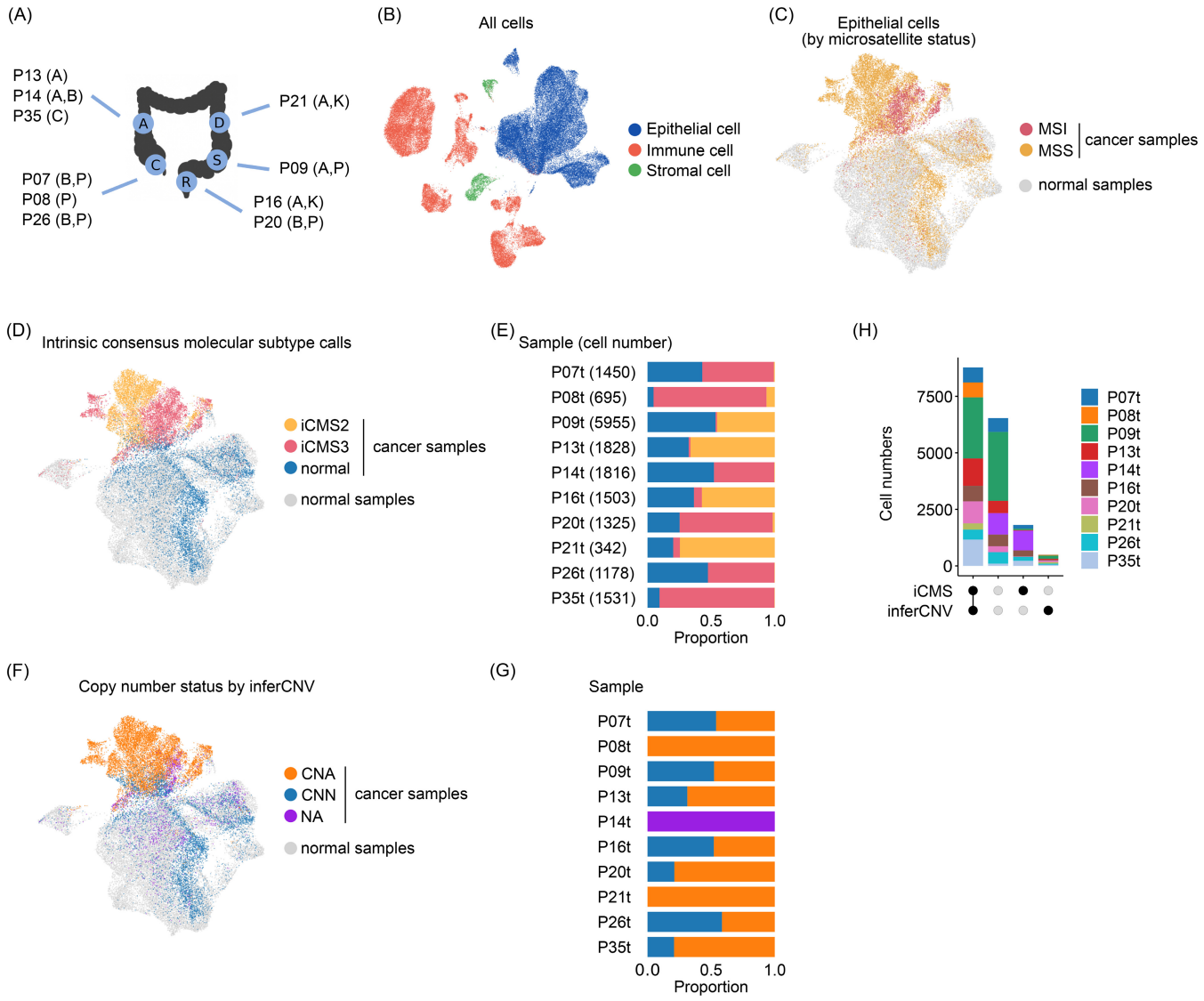

FIGURE 1 Cancer cell calling based on transcriptome information. (A) Anatomical locations and mutational patterns of the samples. C, cecum; A, ascending colon; D, descending colon; S, sigmoid; R, rectum. Mutations (in brackets) A: *APC*, B: *BRAF*, C: *CTNNB1*, K: *KRAS*, P: *TP53*. (B) UMAP of all 73,294 cells, colored by three major cell type compartments: Epithelial (blue), immune (orange), and stromal cells (green). (C, D, F) UMAPs of epithelial cells only. (C) Color code by the sample origin and the microsatellite status. Cancer sample (MSI), red; cancer sample (MSS), yellow; normal sample, gray. (D) Color code for cancer sample cells by iCMS assignment; iCMS2 (yellow), iCMS3 (pink), or normal (blue), normal samples (not scored, gray). (F) Color code of cancer sample cells by inferCNV. Copy number status aberrant (CNA; orange), normal (CNN; blue), or not applicable (NA; purple) when the clones in the sample are not differentiable, normal samples (not scored, gray). (E,G) Stacked bar plots summarizing iCMS and inferCNV information, respectively, by cancer sample. (H) Quantification of the agreement between iCMS and inferCNV calls as an upset plot, color-coded by patient, as indicated.

other type[10] (Figure 1E), and the results also confirm that MSI cancers are usually iCMS3. Almost all the cells receiving iCMS2 or iCMS3 calls were located on the cancer cell community of the UMAP, in contrast to the 7034 cancer tissue-derived epithelial cells receiving the "normal" label that were mostly scattered among cells derived from normal tissue samples.

We next inferred cancer cell identity by expression-derived copy number calls, using inferCNV[28] (Figure 1F,G). Using hierarchical clustering based on copy number-driven genome-averaged expression patterns (Figure S1), we assigned cell clusters as cancer when their averaged expression pattern deviated more than three standard deviations from epithelial cells in the normal tissue samples. This method did not yield results for the MSS cancer P14, which did not exhibit detectable alterations in the averaged expression patterns. For the remaining cancer samples, inferCNV identified a total of 10,509 abnormal transcriptomes, whereas 7114 transcriptomes were assigned as derived from normal epithelium.

Taken together, the transcriptome-based analyses showed a large overlap for calling cancer versus genomically normal cells (Figure 1H). However, 1441 cells received conflicting calls, and cells from P14

could not be properly assigned. Thus, these methods are not suitable to generally define genomically-normal versus cancer epithelial cells in CRC samples with high accuracy.

3.2 | Exploiting cancer-specific SNV information for cancer cell calling with CCISM

Given that transcriptome analyses can potentially be confounded by expression similarities between cancer and normal epithelial cell states, we hypothesized that independently derived somatic variants that are observed in single-cell sequencing reads constitute the most unambiguous evidence that a cell originated from a cancer lineage. We therefore utilized cancer-specific somatic variants derived from bulk whole-genome sequencing data of matched samples to interrogate the associated single-cell transcriptomes.

Comparison of normal and cancer genomes yielded 2–12 cancer-specific somatic single nucleotide variants (SNVs) per million bases of genome sequence (MB) in most CRCs, except for the MSI CRCs P26 and P35 which had up to 50 SNVs/MB (Figure 2A). The mean number of expressed SNVs per cell in the single-cell transcriptomes correlated with the SNV frequency in the whole-genome sequencing data and

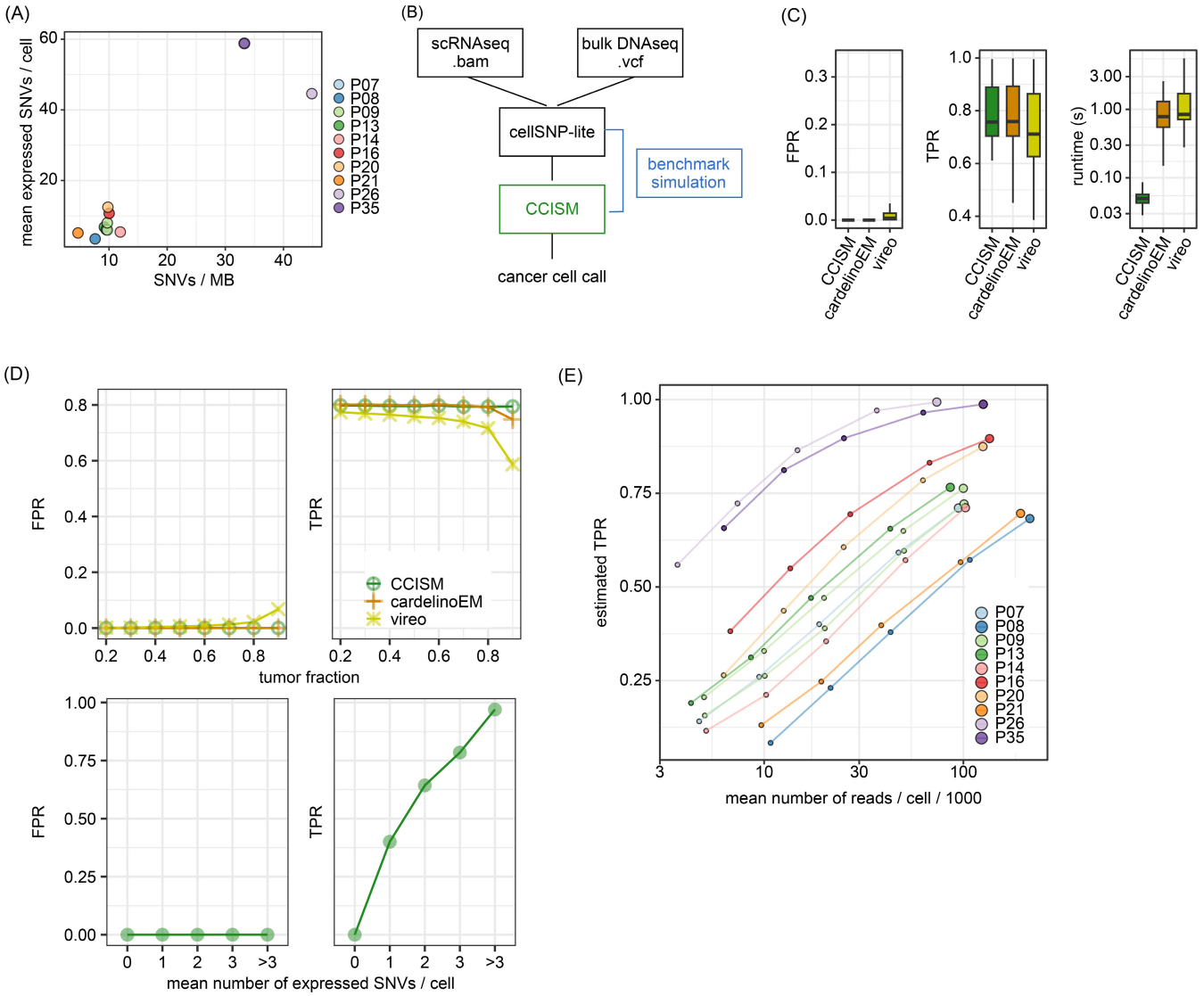

FIGURE 2 CCISM identifies cancer cells with somatic single nucleotide variants. (A) Scatterplot of the number of SNVs in whole genome sequencing data and the average number of expressed SNVs per cell in single-cell RNA sequencing data colored by patient. (B) CCISM's workflow diagram from input data (scRNAseq and bulk DNAseq data), allele count calculation by cellSNP-lite to CCISM modelling. Benchmark simulations can be generated from input counts (blue). (C) Boxplots of tool performances in simulation data regarding runtime in seconds (right), false positive rate (FPR, left), and true positive rate (TPR, mi) between CCISM (green), cardelinoEM (orange), and vireo (pear). (D) Line plots comparing model performances (CCISM, green circle; cardelinoEM, orange cross; vireo, pear star) as function of tumor fraction (upper) and mean number of expressed SNVs per cell (lower). (E) Line plot of CCISM's performance (TPR) in single-cell transcriptomes subsampled to five different mean numbers of reads per cell, color-coded by patient.

was for many CRCs less than 10 SNVs per cell, but up to 60 SNVs/cell for the MSI CRC P35.

To make use of SNV patterns for the classification of single-cell data, we developed CCISM (for Cancer Cell Identification by Somatic Mutations). Input data are the UMI-collapsed read counts for reference and alternative allele observed per cell and variant, which are obtained from the single-cell sequencing reads as well as a list of high-quality somatic variants derived from bulk whole-genome or whole-exome data. Based on this input, CCISM computes for each cell a posterior cancer cell assignment by expectation maximization. Importantly, these are cell-specific values and not derived from clustering. At the same time, benchmark simulations can be used to estimate expected sensitivity and specificity values for the dataset at hand (Figure 2B).

We first used simulations based on the total allele count matrices from our single-cell RNAseq datasets to benchmark CCISM against cardelinoEM[26] and vireo.[33] Compared to these existing tools with related functionality, CCISM has similar specificity but superior computational efficiency (Figure 2C). We also obtained better sensitivity especially at high tumor content, mainly because we employed a fixed parameter for the probability of observing variant alleles in normal cells instead of estimates. It is of note that sensitivity depends on the number of expressed SNVs per cell and reaches optimal values at three or more expressed SNVs per cell (Figure 2D). Across the datasets used to initiate the simulations, we found sensitivity strongly associated with mutational burden and therefore highly correlated to the average number of expressed SNVs per cell (Figure S2). A subsampling analysis revealed that most datasets were not saturated for SNV coverage despite being sequenced to depths of more than 90,000 autosomal reads per cell on average (Figure 2E).

3.3 | CCISM and Numbat can be used cooperatively to define consensus normal and cancer cell lineage populations

We applied CCISM to our CRC single-cell RNA dataset resulting in 9738 cancer cell calls (Figure 3A). The predicted cancer cells show a widely overlapping localization with cells previously classified using expression-based copy-number variation inference with inferCNV or iCMS2/iCMS3 gene expression (Figure 1D,F). However, CCISM generated more cancer cell calls in UMAP neighborhoods identified mainly as normal by iCMS or inferCNV (Figure 3A, see rectangular insets), suggesting that the use of cancer-specific variant information retrieves cells of cancer lineages that are transcriptomically less divergent from genomically normal epithelial cells.

For comparison, we employed Numbat,[29] a recently developed tool using allele frequency shifts of common germline variants to facilitate cancer cell calling via the detection of copy number changes. In our single-cell dataset, Numbat identified 11,008 cells as of cancerous origin, again showing an incomplete overlap with cancer cells identified by the other methods (Figure 3B,C).

Initially, 2562 cells received conflicting assignments by CCISM and Numbat (Figure 3D). Therefore, we studied strengths and weaknesses of both tools, considering individual tumor characteristics (Figures 3D,E and S3). On the one hand, we found that in MSS CRCs, most cells with a conflicting assignment were earmarked as cancer cells by Numbat; however, these cells did not receive a high-confidence cancer cell score by CCISM, as they contained only a median of one SNV, with 707 cells expressing no SNV at all (Figure S3A). On the other hand, cells with conflicting assignments in MSI CRC samples mostly (272/314) received a high-confidence cancer cell score by CCISM, and these contained a median of 16 SNVs, while cancer cell scores computed by Numbat were generally low (Figure S3A). Therefore, we developed a set of rules to arrive at a cancer cell consensus based on genomic information (Figure 3F): epithelial cells of cancer samples receiving high scores (>0.5) by Numbat were assigned as cancer cells, except for cells of MSI cancers that were assigned as normal by CCISM (<0.5), which then received a normal call. Epithelial cells of cancer samples receiving high scores by CCISM (>0.5) were also assigned cancer cells, except when this call of MSS CRC cells conflicted with a low score by Numbat (<0.5), in which case the cell was called "unclear." Using these consensus call rules, we were able to assign 11,238 cells as cancer cells (Figure 3G,H). Exactly 570 of these were not recognized as cancer cells by iCMS transcriptional signatures or by inferCNV. A total of 5969 cells were assigned as derived from normal epithelial lineages, using SNV or haplotype information. A remaining set of only 416 cells was assigned as "unclear" and removed from further analysis, as they contained no reliable SNV or haplotype information.

3.4 | Consensus cell identity leads to higher homogeneity of transcriptome clustering and enables phenotypic comparison

The final cell assignment to cancer or normal lineages resulted in a substantial separation of the populations when visualized on the UMAP (Figure 3H). We additionally performed a transcriptome-based Louvain clustering of cells (Figure 4A; see Table S4 for associated marker genes) and quantified how cancer and normal calls are distributed across these clusters (Figure 4B). We found that normal and cancer cell communities were best separated when using the consensus call, compared to relying on the different methods that use transcriptome or genomic information individually (Figures 4C and S4A–D). Using the consensus annotation, cancer cells were distributed in a highly patient-specific manner, but genomically normal epithelial cells intermingled as well as epithelial cells derived from normal tissue samples (Figure S4E). While the consensus call requires additional genomic data, the correspondence to the Louvain cluster structure also implies that transcriptomes alone may contain sufficient information for the disambiguation of cancer and normal lineage epithelial cells, at least in our CRC single-cell data set.

We found that the genomically normal epithelial cells from cancer samples showed distinct cluster distributions when compared to the

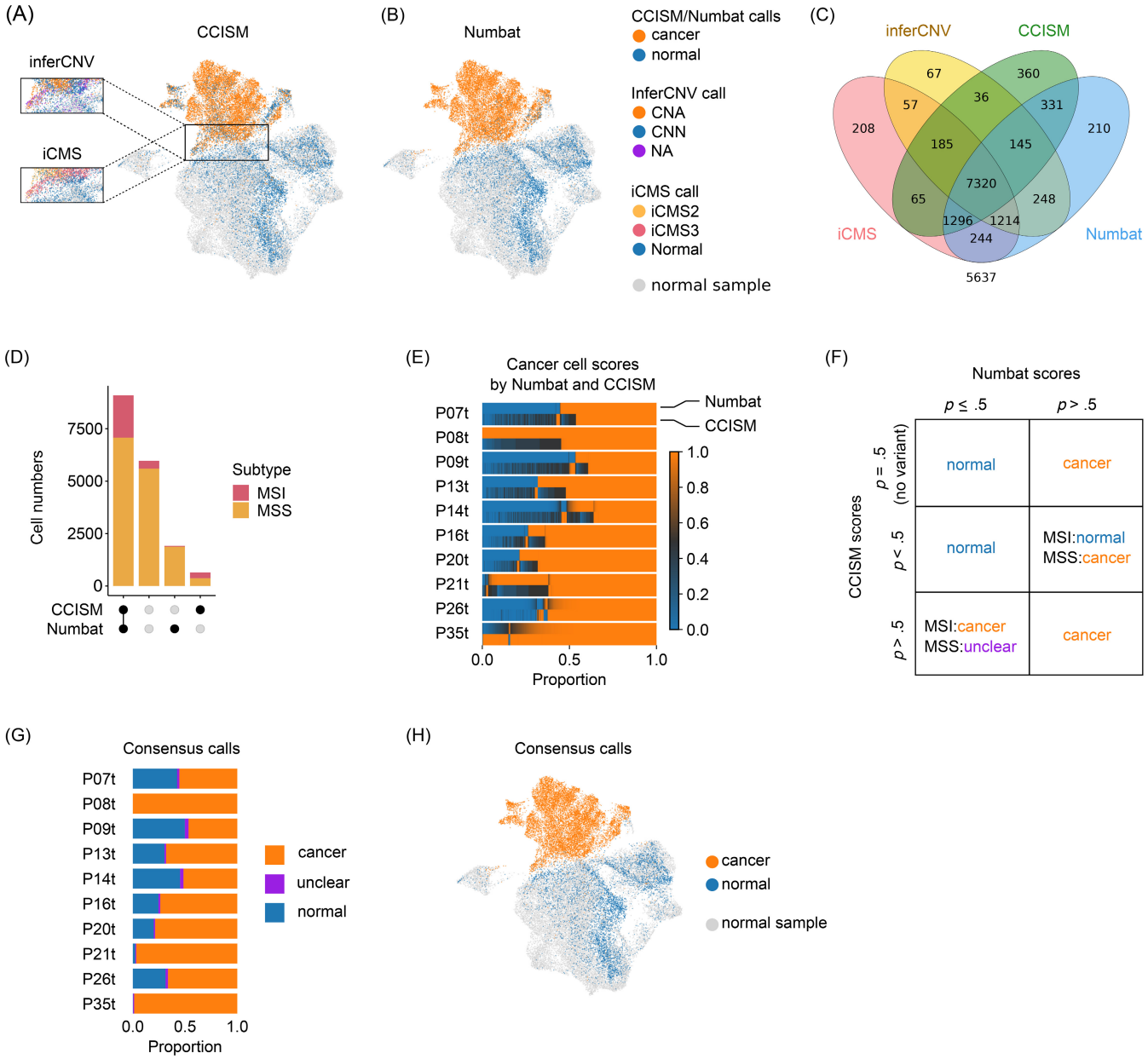

FIGURE 3 Cancer cell calling based on genomic information. (A,B) UMAPs of epithelial cells. (A) Color-code by CCISM calls (cancer cell, orange; normal cell, blue). Insets given for inferCNV and iCMS calls. Cells from normal samples are given in gray. (B) Color-code by Numbat call (cancer cell, orange; normal cell, blue). Cells from normal samples are given in gray. (C) Venn diagram of the intersections of cancer cell calls from iCMS (pink), inferCNV (yellow), CCISM (green), and Numbat (blue). 5,637 cells are called as normal by all four tools. (D) Intersections of cancer cell calls from CCISM and Numbat colored by microsatellite status of the sample (MSI, red; MSS, yellow), given as an upset plot. (E) Heatmaps of the cancer cell scores (0.0, blue; 0.5, dark gray; 1.0, orange) from Numbat (upper) and CCISM (lower) across cancer samples. (F) Decision matrix for consensus cancer cell calls, based on CCISM, Numbat and microsatellite status. (G) Stacked barplot of the consensus derived from CCISM and Numbat (cancer cell, orange; normal cell, blue; undefined, purple). (H) UMAP of the consensus calls, color code as in G, excluding cells with an "unclear" call.

normal tissue epithelial cells (Figure 4B; see Table S5 for marker genes associated with these groups). In particular, Louvain cluster 9 was almost exclusively composed of genomically-normal epithelial cells of cancer samples, and these were derived predominantly from tissue samples of patients P09, P16, P20, and P21 with a left-sided (sigmoid colon or rectum) origin (Figures 1A and 4D). We further explored the identities of epithelial cells using label transfer.[16] Cluster 9 contained mainly stem cells, transiently amplifying cells, or enterocyte precursors (Figure 4E), and their assignment to a distinct Louvain cluster suggested that these cells adopted a cell state that was induced by the cancer microenvironment and therefore not found in normal colon. When we analyzed cluster 9-specific expression patterns, the most strongly defining gene for cluster 9 epithelial cells was *PLA2G2A*, encoding a secreted phospho-lipase (Figure 4F,G). Similarly, when we

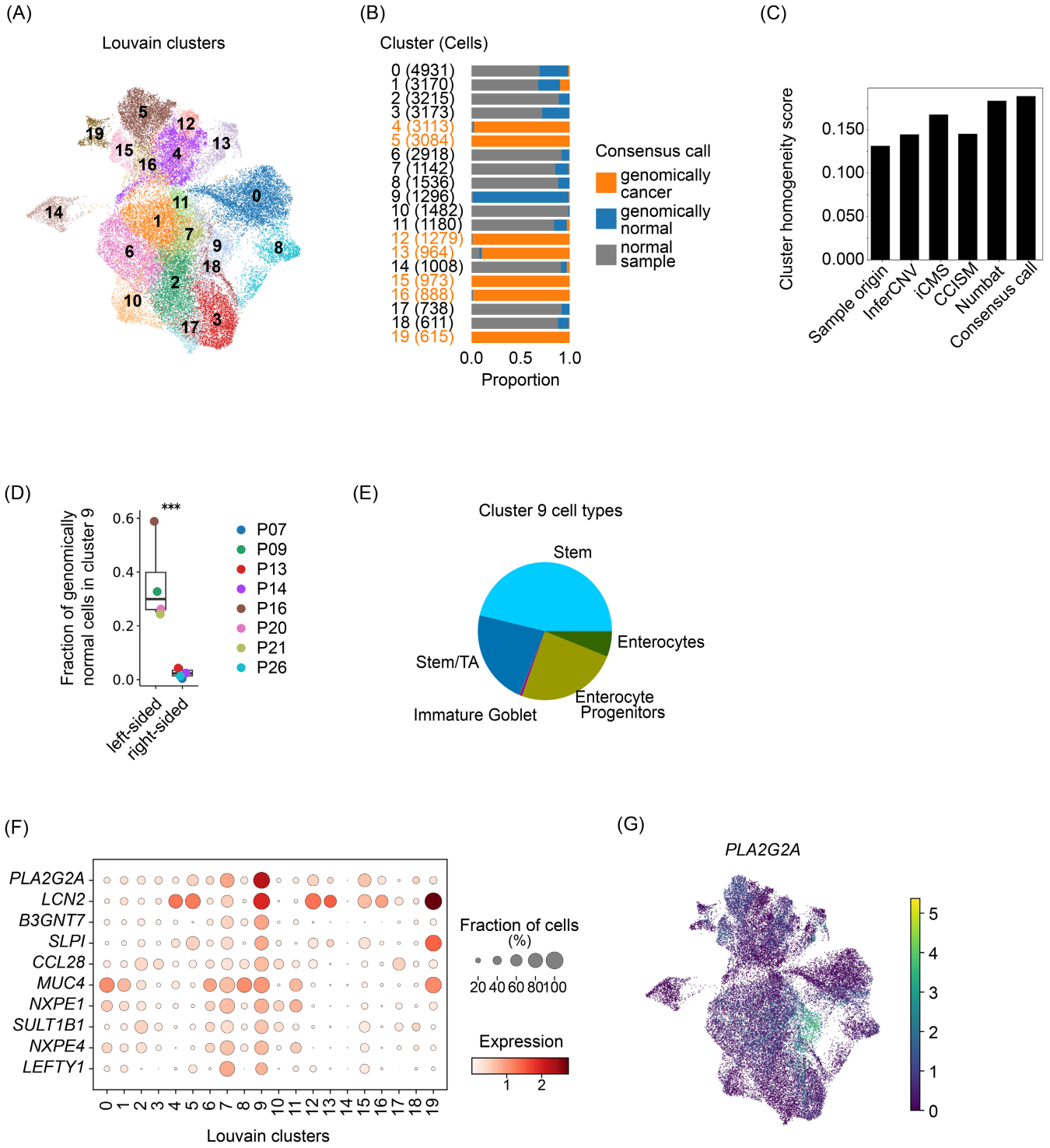

FIGURE 4 Consensus calls identify a cluster of genomically normal cells unique to left-sided cancer samples. (A) UMAP of epithelial cells, colored by louvain clustering. (B) Stacked bar plot of consensus calls across 20 louvain clusters (cancer sample and genomically cancer, orange; cancer sample and genomically normal, blue, normal sample, grey). (C) Bar plot of cluster homogeneity scores for cancer cell calls by different methods as indicated. (D) Relative fractions of genomically normal cells in cluster 9, by cancer location (see Figure 1A). *P*-value from mixed-effects binomial model, *** $P < .001$. (E) Pie chart of the epithelial cell types in louvain cluster 9, as indicated. Color code: Enterocyte (dark green), Enterocyte progenitor (light green), Immature Goblet (light purple), Stem/TA (dark blue), and Stem (light blue). (F) Dot plot of top 10 marker genes for louvain cluster 9. Color of dot represents the mean normalized expression of the gene, and the size of the dot shows the fraction cells expressing the gene. (G) UMAP colored by *PLA2G2A* expression, which is the top gene marker specific to louvain cluster 9.

compared genomically normal stem cells from cancer samples to stem cells in normal tissue, *PLA2G2A* and other markers for cluster 9 were among the top differential genes (Figure S5A).

Mapping of well-established colon and CRC cell-type signatures (Table S2) onto the epithelial single-cell transcriptomes derived from cancer and normal samples unveiled further differences in differentiation programs in the cancer's vicinity, as Goblet cell transcriptomes derived from cancer samples were enriched for a Paneth cell signature, indicating that the cancer microenvironment perturbs secretory lineage fate decisions (Figure S5B). Indeed, the occurrence of metaplastic Paneth cells has been widely documented in inflammation and also in cancer of the colon.[34,35]

3.5 | The CRC microenvironment modulates epithelial cell states and developmental trajectories

We next assessed cell type frequencies among the genomically normal epithelial cells from cancer samples and compared them to normal tissue sample epithelium, excluding patients P08, P21, P26, and P35 which either had no matched normal sample or very few genomically normal cells (Figure 5A). We found that the cancer-adjacent epithelial cells were significantly enriched for stem cells, immature goblet cells, and enterocyte progenitors, while they contained lower proportions of terminally differentiated cell types, such as differentiated enterocytes, goblet cells and tuft cells (Figure S5C).

We then wanted to infer cell developmental trajectories. For this, we first embedded epithelial cells from normal and cancer samples into a common diffusion map, thereby emphasizing continuous cell distributions (Figure 5B). In this embedding, diffusion component (DC) 1 was largely correlated to tuft cell identity, whereas DC2 distributed all other cell types along an apparent differentiation axis, with genomically cancer cells occupying one end. Binning the non-cancer cell types along the DC2 axis (Figure 5C), we observed that genomically normal stem cells from cancer samples occupied a larger range on the DC2 axis compared to stem cells from normal tissue samples. In contrast, while immature goblet cells and enterocyte progenitors were also more frequent among the cancer-adjacent normal epithelium, they were confined to a similar range on the DC2 diffusion axis compared to normal tissue samples. These results were corroborated by ordering the cell lineages along a pseudo-time axis using CytoTrace[36] (Figure 5D,E). Here, stem cells had a wider distribution in the cancer microenvironment samples, whereas all other cell types were distributed in a fashion comparable to normal tissue. The cancer sample-specific stem cell zone extending into the developmental trajectory is composed mainly of cluster 9 stem cells (Figure 5F), derived from CRCs in the left colon. Together, these analyses suggest that the cancer microenvironment affects differentiation trajectories of normal colonic epithelial cells in their vicinity. The primary difference appears to be the stabilization of the stem cell transcriptional state, which in a left-sided CRC microenvironment extends further along the developmental trajectory. In addition, proportions of immature to terminally differentiated cell states are shifted toward the immature cell states in vicinity of CRC.

3.6 | The CRC tumor microenvironment is enriched for morphogenetic signal interactions

We next analyzed potential paracrine interactions that could underlie the observed differences in cell type frequencies and developmental trajectories between the CRC microenvironment and the normal colon. Our dataset contains a high proportion of immune cells and a lower proportion of stromal cells (Figures 1B and S6A). Specifically, among the 31,663 immune cells, 23,433 were derived from cancer, as were 2054 of the 2463 stromal cells. We annotated stromal and immune cell types at a medium granularity using established signatures (Figures 6A and S6B), to strike a balance between accuracy and cluster size. We found that among immune cells, monocytes, macrophages, and regulatory T cells were most enriched in the cancer samples, while among stromal cells, fibroblasts were overrepresented in the cancer microenvironment.

We then used CellChat[37] to infer interactions in the normal and the cancer samples on a comprehensive basis (Figure S6C for all interactions). Quantitative analysis revealed that fibroblasts had the most extensive network of outgoing signaling interactions (Figure 6B) and this network was even larger in cancer samples (Figure 6C). Endothelial cells and pericytes were rich sources of outgoing signaling interactions in cancer compared to normal. In contrast, endothelial cells, macrophages and pericytes were prominent signal receivers particularly in the cancer microenvironment, whereas CD8+ T cells received the most signals in both, normal and cancer samples (Figure 6B). Normal epithelial cells emitted and received relatively few signals. Therefore, we analyzed key morphogenetic signaling pathway interactions, WNT, BMP and FGF, known to pattern the epithelium in more detail (Figure 6D). We found that fibroblasts were rich sources of FGF signals potentially received by goblet cells, and of Wnt signals received, for example, by stem cells, and these interactions were seen in both tumor and normal tissue. In addition, BMP interactions known to abrogate the stem cell state[38] were diminished in the cancer microenvironment, in particular due to lower BMP expression from fibroblasts. Thus, our data predicts that differences in fibroblast signaling could underlie the changes in normal epithelial cell developmental trajectories that were mainly detected in stem and immature cell population. Indeed, cross-referencing the interactions predicted by CellChat with a curated list of signaling pathway ligands and receptors (Figures 6E and S6D,E; Table S3), we found that WNT2 and the TGF-beta ligand INHBA were most strongly overexpressed by cancer-associated fibroblasts compared to normal fibroblasts, while BMP4 and the WNT co-ligand RSPO3 were expressed at lower levels compared to normal tissue samples.

4 | DISCUSSION

Single-cell data of cancer tissue often contain transcriptomes of both cancerous and normal epithelial cells. In this study, we compared different strategies that exploit transcriptome and genome sequence information to trace back the origins of epithelial cell transcriptomes.

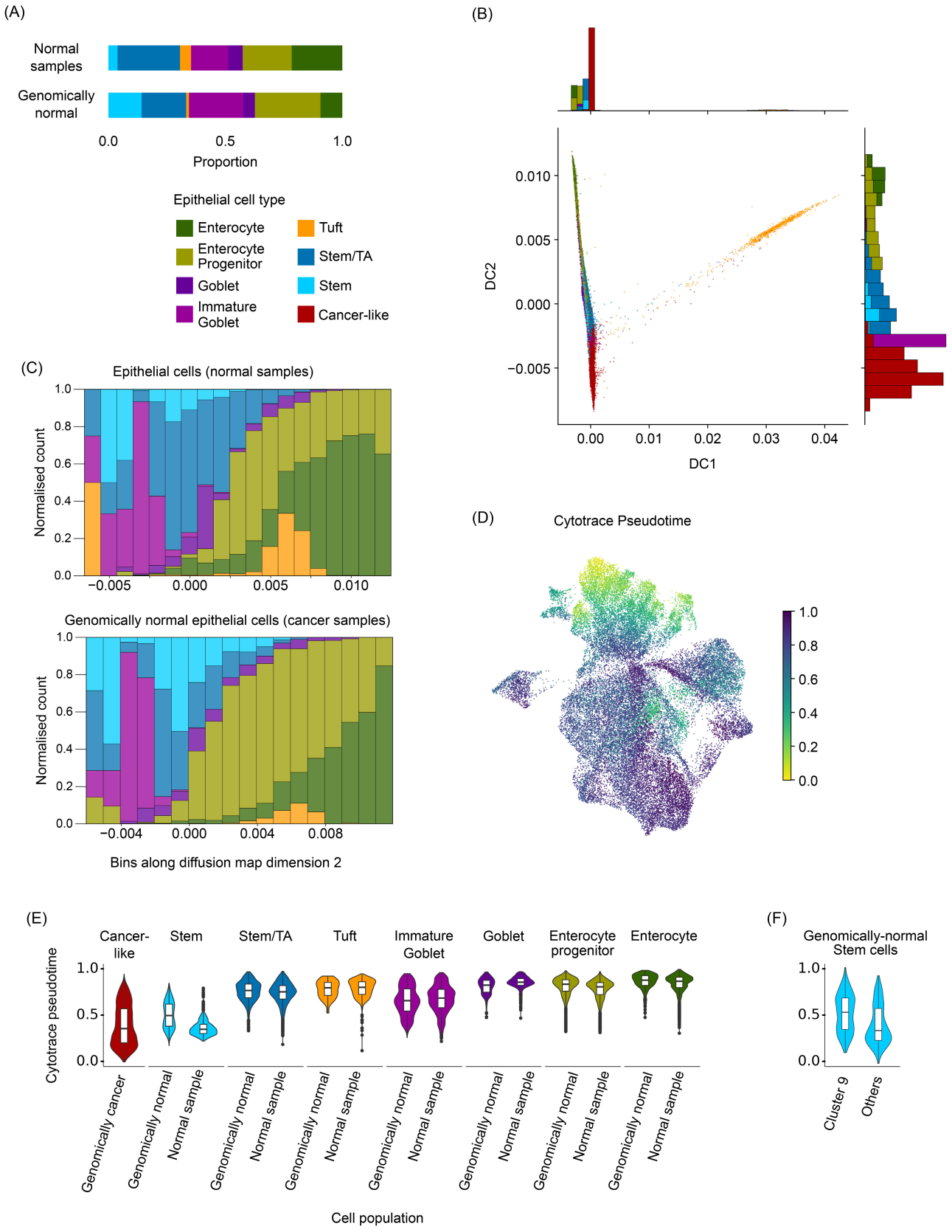

FIGURE 5 Legend on next page.

Across a cohort of CRCs of stages T1–T4 and with different molecular characteristics, a combination of haplotype-aware copy number inference and our new method based on somatic SNVs allowed us to differentiate with high accuracy between cancerous cells and those that are found within cancer tissue but are genomically normal. Using consensus sets of normal and cancer cells, we identified one cluster of genomically normal epithelial cells that were derived from cancer tissue samples exclusively, implying that the cancer microenvironment can result in the adoption of non-standard epithelial cell states in the colon.

Our new tool CCISM makes use of somatic SNVs observed in single-cell sequencing reads for cancer cell identification. Making use of the most unambiguous evidence that a cell originated from a cancer lineage, this approach currently requires somatic SNVs independently obtained from matched tumor-normal whole-genome or whole-exome sequencing of the same cancers. To benchmark CCISM, we used Numbat, which estimates copy-number variation from shifts in haplotype frequencies over common genetic variants to identify cancer cells, as well as two additional methods that use transcriptome information exclusively. Although cancer cell calls from the different approaches show substantial overlap, our analysis of CRC reveals distinct strengths and limitations contingent on the underlying biology of each individual cancer. Consequently, workflows for cancer cell identification should be specifically tailored for the data under analysis. Notably, the mutational load of cancer types differs by several orders of magnitude[39]; our benchmarking of CCISM suggests a lower threshold of 2–3 high-quality SNVs per transcriptome to achieve ~75% sensitivity in the disambiguation of single-cell transcriptomes, which translates roughly to a mutational load above 10 SNVs/MB.

In the final cell annotation of our CRC dataset, cancer and genomically normal cells were largely separated in the underlying Louvain cluster structure, implying that cancer and normal epithelial cells do not share common cell states during their developmental trajectories. Nevertheless, we observed altered cancer sample-specific cell states in genomically normal epithelial cells, which could easily be mistaken for genuine cancer cells. It is important to note that our result of largely non-overlapping cell states between normal and cancer may not transfer to cohorts of other stages or types of cancer.

To demonstrate that CCISM and our consensus call strategy can be applied to other cancer types, we analyzed a scRNAseq dataset of six lung adenocarcinomas[32] with matched tumor-normal whole genome sequencing data and used CCISM to call cancer cells. More pronounced than in the CRC cohort, normal epithelial cells clustered by cell type, while cancer epithelial cells clustered by patient (Figure S7A–C). In contrast to the CRC data, the lung cancer samples contained very few non-cancer epithelial cells. Cancer cell calls by CCISM overlapped with calls from inferCNV, with patient-dependent accuracy in-line with the mutational burden and the estimated sensitivity from benchmark simulations (Figure S7D–G). Generally, we expect that CCISM works best for cancer entities with high mutational load such as melanoma or lung cancers,[39] while tools based on copy-number variation will work better for cancer entities with frequent structural aberrations such as neuroblastoma[40] or high-grade breast cancer.[41] Our consensus strategy will strike a balance between both types of genomic aberrations that exist side-by-side in many cancer entities.

Using the consensus sets of genomically normal and cancer cells defined here for colorectal cancers, we identified genomically normal *PLA2G2A*-positive stem-like cells arising specifically in the cancer context in the left colon (sigmoid and rectum). PLA2G2A is the human homologue of the gene underlying the mouse Mom-1 locus,[42] a genetic modifier of familial cancer susceptibility shown to confer cancer resistance in mouse models.[43] The functional relevance of these stem-like cells remains elusive. On the one hand, the extension of stem-like and immature cell states along the differentiation trajectory could represent a misguided regenerative process hijacked by paracrine signals of the cancer microenvironment.[44,45] Indeed, we identify novel paracrine interactions in the CRC microenvironment that were dominated by fibroblasts, as recently also found for breast cancer.[46] These signals could guide tissue remodeling in the proximity of cancer, which is commonly accompanied by inflammation.[47] On the other hand, the induction of *PLA2G2A*, which we identified as the most specific marker gene of the novel stem-like cells arising near the cancer, could be part of a feedback mechanism to protect the organ from cancer under inflammatory conditions. In agreement with such a function, PLA2G2A is a secreted phospho-lipase that controls tissue homeostasis via modulation of inflammatory responses and is a key player in reducing cancer susceptibility.[48] The exclusive occurrence of the cancer-induced *PLA2G2A*-positive cells in the left colon suggests regional specificity of the underlying mechanisms along the longitudinal axis of the colon. Supporting region-specific models of cell differentiation, different cell compositions and interactions have been identified in the left-sided/sigmoid colon, such as increased plasma cell interactions.[49]

Cancer tissue has been shown to extend its influence far beyond its perimeter. Several potential mechanisms with different ranges exist: tumors expressing hormones will affect the complete patient's body regardless of localization,[50] while inflammatory responses and

FIGURE 5 Cell states and developmental trajectories are altered in genomically normal cells of cancer samples compared to normal colon epithelium. (A) Stacked bar plots of epithelial cell types in normal samples (upper) and genomically normal cell populations (lower), including Enterocyte (dark green), Enterocyte progenitor (light green), Goblet (dark purple), Immature Goblet (light purple), Tuft (yellow), Stem/TA (dark blue), and Stem (light blue). (B) Diffusion map with additional histograms of first and second dimensions/axes colored by epithelial cell types. Color code as in A, with the addition of genomically cancer cells (red). (C) Stacked bar plots of the epithelial cell type compositions across binned diffusion map dimension 2 in normal sample and genomically normal cells, as indicated. (D) UMAP colored by Cytotrace developmental pseudotime, from early (0, yellow) to late (1, dark purple) in pseudotime space. (E,F) Violin plots of Cytotrace pseudotime across epithelial cell types and consensus call groups, as indicated.

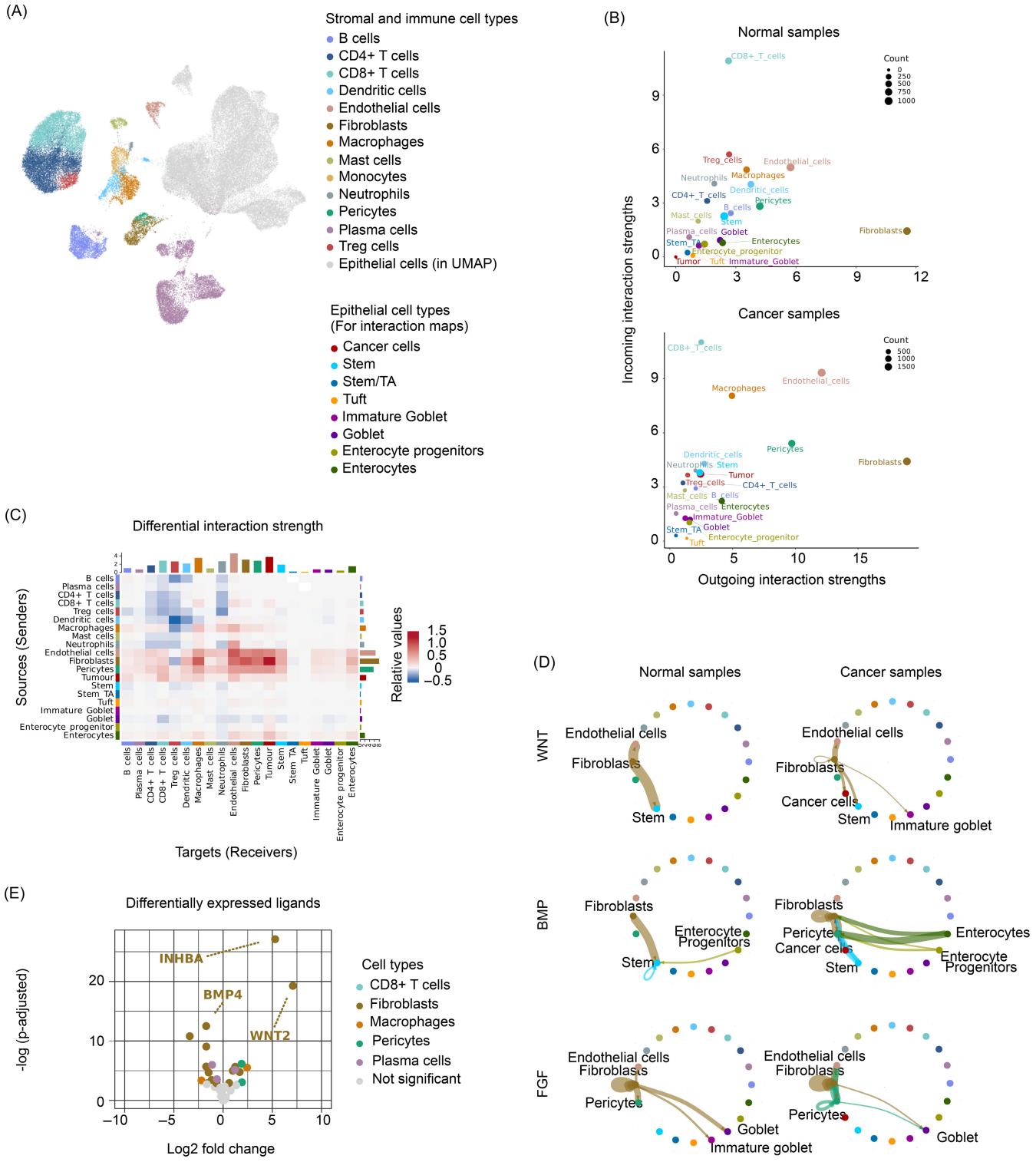

FIGURE 6 Signaling networks of normal epithelial and genomically normal cells with their respective microenvironments. (A) UMAP of all the cells under analysis, colored by detailed immune and stromal cell types. Epithelial cells given in gray. (B–D) Analyses by CellChat (B) Scatterplots of incoming and outgoing signals in normal and cancer samples, as indicated. (C) Heatmap of differential cell–cell communications of cancer samples in contrast to normal samples. (D) Aggregated network graphs of WNT, BMP, and FGF pathways in normal samples versus cancer samples, as indicated. (E) Volcano plot of differentially expressed ligand genes in immune and stromal cell types, as indicated. For a complete list of scored ligands, see Table S3.

other differences in cell composition can have long-range, yet local, effects.[51] A recent study found a prognostic value of gene expression signatures derived from normal-adjacent to CRC issue harvested at a distance of ~10 cm from the cancer,[52] suggesting the existence of long-range interactions between the CRC and surrounding tissues. Thus, gene expression patterns of our normal controls harvested

~10–30 cm from the cancer, may not represent a true normal state, and in extension, our study may underestimate the influence of cancer cells and the cancer microenvironment on adjacent genomically normal colon cells.

New technological developments constantly change single-cell methodology. Employing advances in sequencing depth and transcriptome coverage, for example, by long-read sequencing or specific protocols,[53] a more comprehensive readout of somatic SNVs could be achieved. This would help improve cell lineage determination, for example, for cancers with few genomic aberrations, such as childhood cancers. With increased coverage, robust de novo calling of somatic SNVs could even be feasible directly from single-cell data.[54–56] In summary, our study provides general rules for distinguishing between cancer and non-cancer single-cell transcriptomes and provides recommendations on how to account for the biology and genetic characteristics of CRC. The rules can easily be adapted for cancers of different origins.

AUTHOR CONTRIBUTIONS

Tzu-Ting Wei: Data curation; formal analysis; investigation; methodology; visualization. **Eric Blanc:** Data curation; formal analysis. **Stefan Peidli:** Methodology; validation. **Philip Bischoff:** Investigation; methodology; resources. **Alexandra Trinks:** Investigation; methodology; resources. **David Horst:** Resources; supervision. **Christine Sers:** Resources; supervision. **Nils Blüthgen:** Conceptualization; resources; supervision. **Dieter Beule:** Conceptualization; project administration; resources; supervision. **Markus Morkel:** Conceptualization; project administration; supervision; writing – original draft; writing – review and editing. **Benedikt Obermayer:** Conceptualization; formal analysis; project administration; software; supervision; visualization; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

We thank Edda von der Wall and Hedwig Lammert (Charité, Institute of Pathology) for excellent technical assistance. The work was in part funded by Deutsche Forschungsgemeinschaft (RTG CompCancer GRK2424/1) and by the BIH-funded PeDiOn and Clinical Scientist programs. We acknowledge excellent services of the BIH Sequencing core facility. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interests related to this publication.

DATA AVAILABILITY STATEMENT

Processed WGS and single-cell data generated in this study are available on Zenodo via zenodo.org/records/10692019. CCISM is available from github.com/bihealth/CCISM. Analysis code is available from github.com/bihealth/Wei_et_al_2024. Raw and other data that support the findings of this study are available from the corresponding author upon request, in accordance with national legislation.

ETHICS STATEMENT

All patients were aware of the planned research and agreed to the use of tissue. Research was approved by vote EA4/164/19 of the ethics commission of Charité – Universitätsmedizin Berlin.

ORCID

Tzu-Ting Wei https://orcid.org/0000-0001-5719-721X
Eric Blanc https://orcid.org/0000-0002-4369-0254
Stefan Peidli https://orcid.org/0000-0002-4257-8690
Philip Bischoff https://orcid.org/0000-0002-4442-7116
Alexandra Trinks https://orcid.org/0000-0001-9983-1506
David Horst https://orcid.org/0000-0003-4755-5743
Christine Sers https://orcid.org/0000-0002-6219-1514
Nils Blüthgen https://orcid.org/0000-0002-0171-7447
Dieter Beule https://orcid.org/0000-0002-3284-0632
Markus Morkel https://orcid.org/0000-0002-2553-9999
Benedikt Obermayer https://orcid.org/0000-0002-9116-630X

REFERENCES

1. Baghban R, Roshangar L, Jahanban-Esfahlan R, et al. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun Signal*. 2020;18:59.
2. Ungefroren H, Sebens S, Seidl D, Lehnert H, Hass R. Interaction of tumor cells with the microenvironment. *Cell Commun Signal*. 2011;9:18.
3. Lugli A, Zlobec I, Berger MD, Kirsch R, Nagtegaal ID. Tumour budding in solid cancers. *Nat Rev Clin Oncol*. 2021;18:101-115.
4. Koelzer VH, Dawson H, Andersson E, et al. Active immunosurveillance in the tumor microenvironment of colorectal cancer is associated with low frequency tumor budding and improved outcome. *Transl Res*. 2015;166:207-217.
5. De Wever O, Mareel M. Role of tissue stroma in cancer cell invasion. *J Pathol*. 2003;200:429-447.
6. Lochhead P, Chan AT, Nishihara R, et al. Etiologic field effect: reappraisal of the field effect concept in cancer predisposition and progression. *Mod Pathol*. 2015;28:14-29.
7. Schürch CM, Bhate SS, Barlow GL, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*. 2020;182:1341-1359.
8. William Zhao A, Kepecs B, Mahadevan NR, et al. A cellular and spatial atlas of TP53-associated tissue remodeling in lung adenocarcinoma. *bioRxiv*. 2024;2023.06.28.546977. doi:10.1101/2023.06.28.546977
9. Aran D, Camarda R, Odegaard J, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*. 2017;8:1077.
10. Joanito I, Wirapati P, Zhao N, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat Genet*. 2022;54:963-975.
11. Worthley DL, Leggett BA. Colorectal cancer: molecular features and clinical opportunities. *Clin Biochem Rev*. 2010;31:31-38.
12. Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol*. 2011;6:479-507.
13. Ijspeert JEG, Vermeulen L, Meijer GA, Dekker E. Serrated neoplasia-role in colorectal carcinogenesis and clinical implications. *Nat Rev Gastroenterol Hepatol*. 2015;12:401-409.
14. Weisenberger DJ, D Siegmund K, Campan M, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet*. 2006;38:787-793.
15. Lee HO, Hong Y, Etlioglu HE, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet*. 2020;52:594-603.
16. Uhlitz F, Bischoff P, Peidli S, et al. Mitogen-activated protein kinase activity drives cell trajectories in colorectal cancer. *EMBO Mol Med*. 2021;13:e14123.
17. Becker WR, Nevins SA, Chen DC, et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant

transformation of polyps to colorectal cancer. *Nat Genet*. 2022;54:985-995.
18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 2013;1303.3997.
19. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:1.
20. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19:15.
21. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019;20:296.
22. Fleming SJ, Chaffin MD, Arduini A, et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat Methods*. 2023;20:1323-1335.
23. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*. 2019;8:281.e9-291.e9.
24. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling somatic SNVs and Indels with Mutect2. *bioRxiv*. 2019;861054.
25. Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*. 2021;37:4569-4571.
26. McCarthy DJ, Rostom R, Huang Y, et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nat Methods*. 2020;17:414-421.
27. Smillie CS, Biton M, Ordovas-Montanes J, et al. Intra- and intercellular rewiring of the human colon during ulcerative colitis. *Cell*. 2019;178:714-730.
28. Tickle T, Tirosh I, Georgescu C, Brown MHB. *inferCNV of the Trinity CTAT Project*. Klarman Cell Observatory, Broad Institute of MIT and Harvard; 2019.
29. Gao T, Soldatov R, Sarkar H, et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat Biotechnol*. 2023;41:417-426.
30. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67:1-48. doi:10.18637/jss.v067.i01
31. Lange M, Bergen V, Klein M, et al. CellRank for directed single-cell fate mapping. *Nat Methods*. 2022;19:159-170.
32. Bischoff P, Trinks A, Obermayer B, et al. Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene*. 2021;40:6748-6758.
33. Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol*. 2019;20:273.
34. Elphick DA, Mahida YR. Paneth cells: their role in innate immunity and inflammatory disease. *Gut*. 2005;54:1802-1809.
35. López-Arribillaga E, Yan B, Lobo-Jarne T, et al. Accumulation of paneth cells in early colorectal adenomas is associated with beta-catenin signaling and poor patient prognosis. *Cells*. 2021;10:2928.
36. Gulati GS, Sikandar SS, Wesche DJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*. 2020;367:405-411.
37. Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun*. 2021;12:1088.
38. He XC, Zhang J, Tong W-G, et al. BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt–β-catenin signaling. *Nat Genet*. 2004;36:1117-1121.
39. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415-421.
40. Paolini L, Hussain S, Galardy PJ. Chromosome instability in neuroblastoma: a pathway to aggressive disease. *Front Oncol*. 2022;12:98897.
41. Kuzmin E, Baker TM, Lesluyes T, et al. Evolution of chromosome-arm aberrations in breast cancer through genetic network rewiring. *Cell Rep*. 2024;43:113988.
42. Dietrich WF, Lander ES, Smith JS, et al. Genetic identification of Mom-1, a major modifier locus affecting Min-induced intestinal neoplasia in the mouse. *Cell*. 1993;75:631-639.
43. Cormier RT, Hong KH, Halberg RB, et al. Secretory phospholipase Pla2g2a confers resistance to intestinal tumorigenesis. *Nat Genet*. 1997;17:88-91.
44. Liu Y, Chen YG. Intestinal epithelial plasticity and regeneration via cell dedifferentiation. *Cell Regen*. 2020;9:14.
45. Beumer J, Clevers H. Cell fate specification and differentiation in the adult mammalian intestine. *Nat Rev Mol Cell Biol*. 2021;22:39-53.
46. Mayer S, Milo T, Isaacson A, et al. The tumor microenvironment shows a hierarchy of cell-cell interactions dominated by fibroblasts. *Nat Commun*. 2023;14:5810.
47. Flier JS, Underhill LH, Dvorak HF. Tumors: wounds that do not heal. *N Engl J Med*. 1986;315:1650-1659.
48. Schewe M, Franken PF, Sacchetti A, et al. Secreted phospholipases A2 are intestinal stem cell niche factors with distinct roles in homeostasis, inflammation, and cancer. *Cell Stem Cell*. 2016;19:38-51.
49. Hickey JW, Becker WR, Nevins SA, et al. Organization of the human intestine at single-cell resolution. *Nature*. 2023;619:572-584.
50. Dimitriadis GK, Angelousi A, Weickert MO, Randeva HS, Kaltsas G, Grossman A. Paraneoplastic endocrine syndromes. *Endocr Relat Cancer*. 2017;24:R173-R190.
51. Zhao H, Wu L, Yan G, et al. Inflammation and tumor progression: signaling pathways and targeted intervention. *Signal Transduct Target Ther*. 2021;6:263.
52. Kim J, Kim H, Lee MS, et al. Transcriptomes of the tumor-adjacent normal tissues are more informative than tumors in predicting recurrence in colorectal cancer patients. *J Transl Med*. 2023;21:209.
53. Salmen F, De Jonghe J, Kaminski TS, et al. High-throughput total RNA sequencing in single cells using VASA-seq. *Nat Biotechnol*. 2022;40:1780-1793.
54. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med*. 2020;52:1452-1465.
55. Muyas F, Sauer CM, Valle-Inclán JE, et al. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. *Nat Biotechnol*. 2023;42:758-767.
56. Dou J, Tan Y, Kock KH, et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. *Nat Biotechnol*. 2023;42:803-812.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wei T-T, Blanc E, Peidli S, et al. High-confidence calling of normal epithelial cells allows identification of a novel stem-like cell state in the colorectal cancer microenvironment. *Int J Cancer*. 2024;155(9):1655-1669. doi:10.1002/ijc.35079