Supplemental Material

Marker-free characterization of full-length transcriptomes of single live circulating tumor cells

Sarita Poonia[1], Anurag Goel[2,14], Smriti Chawla[1], Namrata Bhattacharya[2], Priyadarshini Rai[1], Yi Fang Lee[3,4], Yoon Sim Yap[13], Jay West[5,6], Ali Asgar Bhagat[3,7,8], Juhi Tayal[9], Anurag Mehta[10], Gaurav Ahuja[1], Angshul Majumdar[2,11,12*], Naveen Ramalingam[5*], Debarka Sengupta[1,2,11*]

1. Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
2. Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
3. Biolidics Limited, 81 Science Park Drive, 02-03 The Chadwick, Singapore 118257, Singapore.
4. Current address: Thermo Fisher Scientific, 33 Marsiling Industrial Estate Rd 3, #07-06, Singapore 739256.
5. Fluidigm Corporation, 2 Tower Place, Suite 2000, South San Francisco, CA 94080, USA.
6. Current address: BioSkryb Corporation, BioLabs, 701 W Main St, Suite 200, Durham, NC 27701, USA.
7. Current address: Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore, Engineering Drive 1, Singapore 117575, Singapore.
8. Current address: Institute for Health Innovation and Technology (iHealthtech), National University of Singapore, 14 Medical Drive, Singapore 117599, Singapore.
9. Department of Research, Rajiv Gandhi Cancer Institute and Research Centre-Delhi(RGCIRC-Delhi), D-18, Sec V Rohini. New Delhi 110085.
10. Department of Laboratory Services and Molecular Diagnostics, Rajiv Gandhi Cancer Institute and Research Centre-Delhi(RGCIRC-Delhi), D-18, Sec V Rohini. New Delhi 110085.
11. Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
12. Department of Electronics & Communications Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
13. National Cancer Centre Singapore, 11 Hospital Dr, Singapore 169610, Singapore
14. Department of Computer Science and Engineering, Delhi Technological University, New Delhi - 110042, India

*Corresponding authors: {debarka@iiitd.ac.in, naveen.ramalingam@fluidigm.com, angshul@iiitd.ac.in}.

Supplemental Note 1. Execution details of Seurat, fastMNN, and Harmony. We benchmarked unCTC against three well-reported integrative analysis methods — Seurat, fastMNN, and Harmony (**Supplemental Table S10**). We used Seurat version 4.1.1 and followed two alternative approaches for integrative analysis of single-cell studies (Butler et al. 2018). As part of the first approach, we employed Seurat's standard clustering pipeline, while combining individual scRNA-seq datasets (based on common genes) to construct the input gene expression matrix. This pipeline is referred to as Vanilla Seurat . We used NormalizeData(), FindVariableFeatures(), and ScaleData() for normalization, identification of highly variable genes, and scaling the data respectively, with default parameters. This is followed by the application of FindNeighbors() and FindClusters(), for clustering the cells. For Study 2, we skipped the normalization step since the matrices were already TPM normalized. However, we log-transformed the final matrix. We visualized the data using both UMAP **(Figure 3A-C, Figure 5A-C)** and PCA **(Supplemental Figure S1A-C, Supplemental Figure S2A-C)**. We also analyzed scRNA-seq count profiles corresponding to data used in Study 2. 2D scatter plots of cells were generated using UMAP **(Supplemental Figure S10A-C)** and PCA **(Supplemental figure S11A-C).** In the second approach, we used Seurat's integrative analysis pipeline comprising Canonical Correlation Analysis (CCA)/Reciprocal Principal Component Analysis (RPCA), referred to as Integrative Seurat. We used NormalizeData() for the normalization of Study 1 data. We log-transformed Study 2 data and skipped the normalization step. For the integrative analyses of CTC and WBC datasets from diverse studies, we used SelectIntegrationFeatures() to select the most informative features (i.e., genes). The anchors are then computed using the FindIntegrationAnchors() function. We used Canonical-Correlation Analysis (CCA) and Reciprocal PCA (RPCA) as a reduction method to find integration anchors with the following parameters: k.anchor = 3, dims = 1:5, k.score = 5 for Study 1 to avoid errors caused by a small number of samples in either dataset. These anchors were integrated using IntegrateData() function with the k.weight and dims parameters adjusted based on the sample size of the smallest dataset. In addition, clustering was accomplished using FindNeighbors() and FindClusters() with their default parameter settings. As a result, a batch-corrected matrix was generated with CCA **(Supplemental Figure S8)**, and RPCA **(Supplemental Figure S9)**.

We compared unCTC with fastMNN, which demonstrates the effectiveness of employing MNNs to correct data batches (Haghverdi et al. 2018). In the case of fastMNN, for both studies, we incorporated all WBCs and CTCs into a single scRNA-seq matrix based on common genes. To begin, we used the Seurat preprocessing workflow to filter, normalize, and scale the data. For the second study, we log-transformed the combined matrix with the addition of 1 pseudo count and skipped the normalization step. Finally, the RunFastMNN() function was used on the output from the previous step. We split combined Seurat objects into a list of multiple Seurat objects on the basis of the data sources. Additionally, the FindNeighbours() and FindClusters() functions are

used to cluster cells. Visualization and evaluation were performed using the batch-corrected outputs in the UMAP **(Figure 3D-F, Figure 5D-F)** and mnn **(Supplemental Figure S1D-F, Supplemental Figure S2D-F)** spaces. We also analyzed scRNA-seq count profiles corresponding to data used in Study 2 (**Supplemental Figure S10D-F, S11D-F**).

The third method, Harmony, aligns cells from distinct batches using an iterative clustering approach (Korsunsky et al. 2018). All WBCs and CTCs were merged into a single scRNA-seq matrix based on common genes for both studies. We ran Harmony within the Seurat preprocessing steps to filter, normalize, and scale the data. Finally, the RunHarmony() function was used on the output from the previous step, with parameter group.by.vars = data source. Additionally, the FindNeighbours() and FindClusters() functions were used to cluster cells. Visualization and evaluations were performed using the batch-corrected outputs in the UMAP **(Figure 3G-I, Figure 5G-I)** and harmony **(Supplemental Figure S1G-I, Supplemental Figure S2G-I)** spaces. We also analyzed scRNA-seq count profiles corresponding to data used in Study 2 **(Supplemental Figure S10G-I, S11G-I)**.

Supplemental Note 2. Copy number variation analysis at chromosomal p or q arm. For single-cell expression-based inference of the copy number variation (CNV) landscapes in single cancer cells, we used the InferCNV R package (Tickle et al. 2019). InferCNV does not implement any batch correction. As such, we performed separate analyses of Poonia and Ebright CTCs. Internal CD45 cells from a healthy individual **(Supplemental Note 3)** were used as a reference for Poonia CTCs. For Ebright et al. CTCs, we considered 1000 randomly sampled cells out of 7121 peripheral blood mononuclear cells (PBMCs), isolated by density-gradient centrifugation as a control. This dataset is referred to as the Xu et al. dataset (Xu and Jia 2021). For inferCNV, all data were used in the raw count form. For both datasets, the InferCNV plots show significant CNVs across CTCs as compared to the reference scRNA-Seq profiles. CTCs and single cell type for Ebright et al. CTCs. To infer the location of a particular gain or loss on a chromosome at the p or q-arm level, we used the cytoband information based on GRCh37 (Barrios and Prieto 2017).

Supplemental Note 3. Processing of CD45+ single cells from a healthy individual. 7.5 ml of blood (EDTA) was lysed for RBC. Following this, the CD45 cells were stained with CD45 antibody conjugated to Alexa647 dye (Biolegend, 304020). The CD45 cells were then sorted using Sony SH800 Cell Sorter. Subsequently, live cells were again sorted using SH800 by negatively depleting dead cells that were stained with Zombie Yellow™ Fixable dye (Biolegend, 423103). CD45 live cells were then captured using the C1 single-cell system (Fluidigm) and processed for whole transcriptome library preparation using the same chemistry that was used to process the CTCs. The libraries were sequenced using the Illumina Miseq system using v3 chemistry for 75bp

paired-end reads. 71 out of the 96 sequenced cells, having at least 10,000 raw reads were qualified for their use as a reference for CNV inference on Poonia et al. CTCs.

Supplemental Note 4. unCTC-based analysis of HNSCC scRNA-seq data of 18 patients. We subjected single-cell RNA-seq profiles from 18 primary head and neck squamous cell carcinoma (HNSCC) patients from a previously published study by Puram, Sidharth V., et al. to the unCTC pipeline (Puram et al. 2017). For our analysis, we discarded 324 cells with unknown annotations. Among the remaining 5578 cells, 2215 were marked annotated by the authors as malignant and 3363 as non-malignant. unCTC-based analysis of the scRNA-seq data offered clear segregation of the malignant and non-malignant categories **(Figure S7A)**. unCTC-based clustering revealed four distinct clusters **(Figure 7B)**. **Figure S7C** depicts the distribution of the number of malignant and non-malignant across four clusters obtained from unCTC. We assessed the clustering performance of unCTC, as compared to Vanilla Seurat. unCTC segregation of the malignant vs. non-malignant subtypes yielded higher values of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) **(Figure S7D).** As indicated in the figure, the cluster purity returned by unCTC was slightly lower as compared to Vanilla Seurat. On visualizing Stouffer's scores of combined expression of cell type-specific markers, including epithelial, T, and B lymphocytes, we observed scores of epithelial markers to be enriched in cluster 0 and cluster 2. This finding corroborates the original malignant annotations of the cells present in these two clusters **(Figure S7E)**.

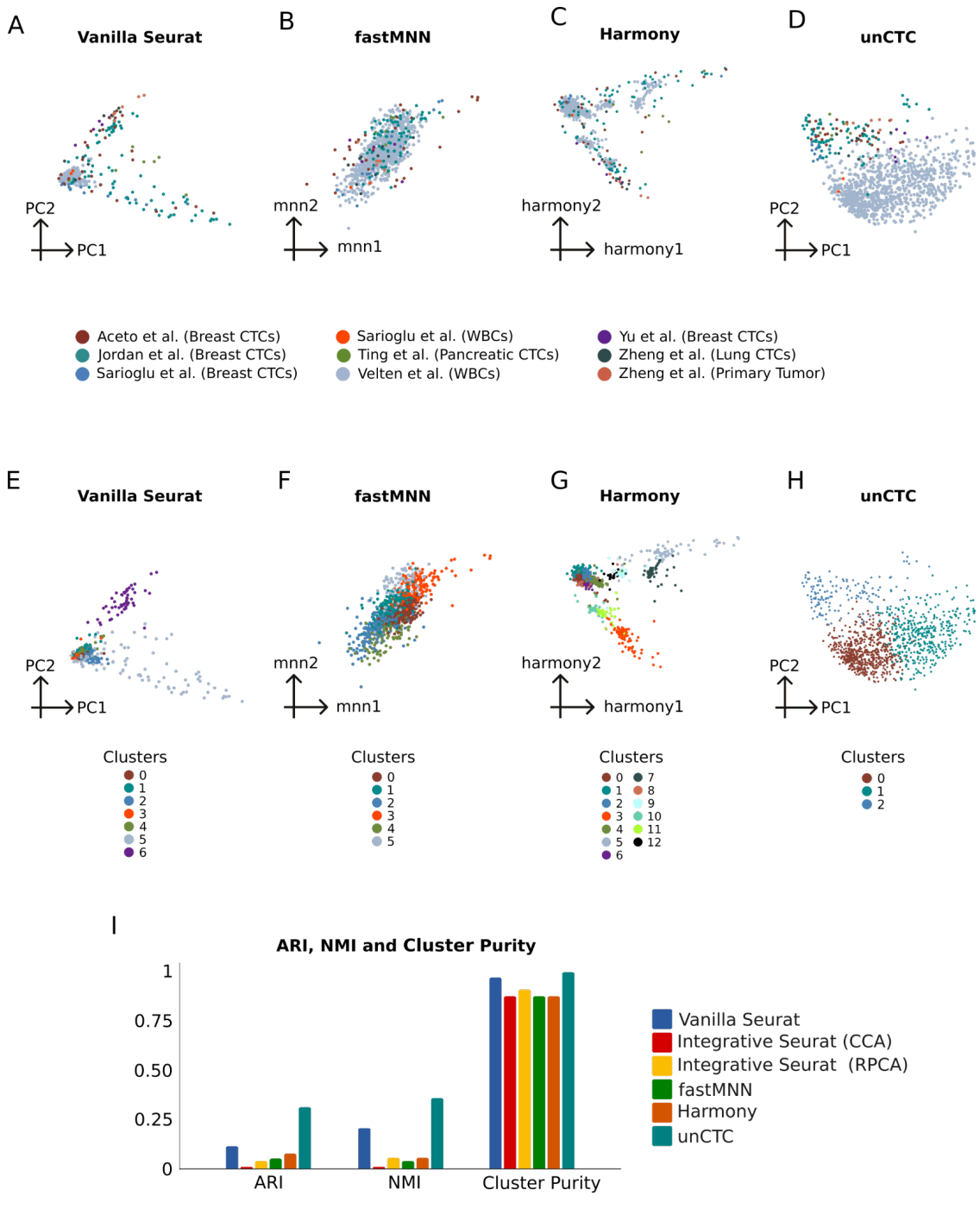
Supplemental Note 5. mRNA-seq library preparation and sequencing. SMARTer® Ultra® Low RNA Kit for Illumina® Sequencing (Clontech®, 634936) was employed to generate pre-amplified cDNA. The Polaris cell lysis mixture was used to lyse the selected and sequestered single cells. The 28 µL cell lysis mix is composed of 8.0 µL of Polaris Lysis Reagent (Fluidigm, 101-1637), 9.6 µL of Polaris Lysis Plus Reagent (Fluidigm, 101-1635), 9.0 µL of 3 SMART CDS Primer II A (12 M, Clontech, 634936), and 1.4 µL of Loading Reagent (20X, Fluidigm, 101-1004). The thermal profile for single-cell lysis is 37 °C for 5 min, 72 °C for 3 min, 25 °C for 1 min, and hold at 4 °C. The 48 µL preparation volume for reverse transcription (RT) contains 1X SMARTer Kit 5X First-Strand Buffer (5X; Clontech, 634936), 2.5-mM SMARTer Kit Dithiothreitol (100 mM; Clontech, 634936), 1 mM SMARTer Kit dNTP Mix (10 mM each; Clontech, 634936), 1.2 µM SMARTer Kit SMARTer II A Oligonucleotide (12 µM; Clontech, 634936), 1 U/µL SMARTer Kit RNase Inhibitor (40 U/µL; Clontech, 634936), 10 U/µL SMARTScribe™ Reverse Transcriptase (100 U/µL; Clontech, 634936), and 3.2 µL of Polaris RT Plus Reagent (Fluidigm, 101-1366). All the concentrations correspond to those found in the RT chambers inside the Polaris IFC. The thermal protocol for RT is 42 °C for 90 min (RT), 70 °C for 10 min (enzyme inactivation), and a final hold at 4 °C.

The 90 µL preparation volume for PCR contains 1X Advantage® 2 PCR Buffer [not short amplicon (SA)](10X, Clontech, 639206, Advantage 2 PCR Kit), 0.4-mM dNTP Mix (50X/10 mM, Clontech,

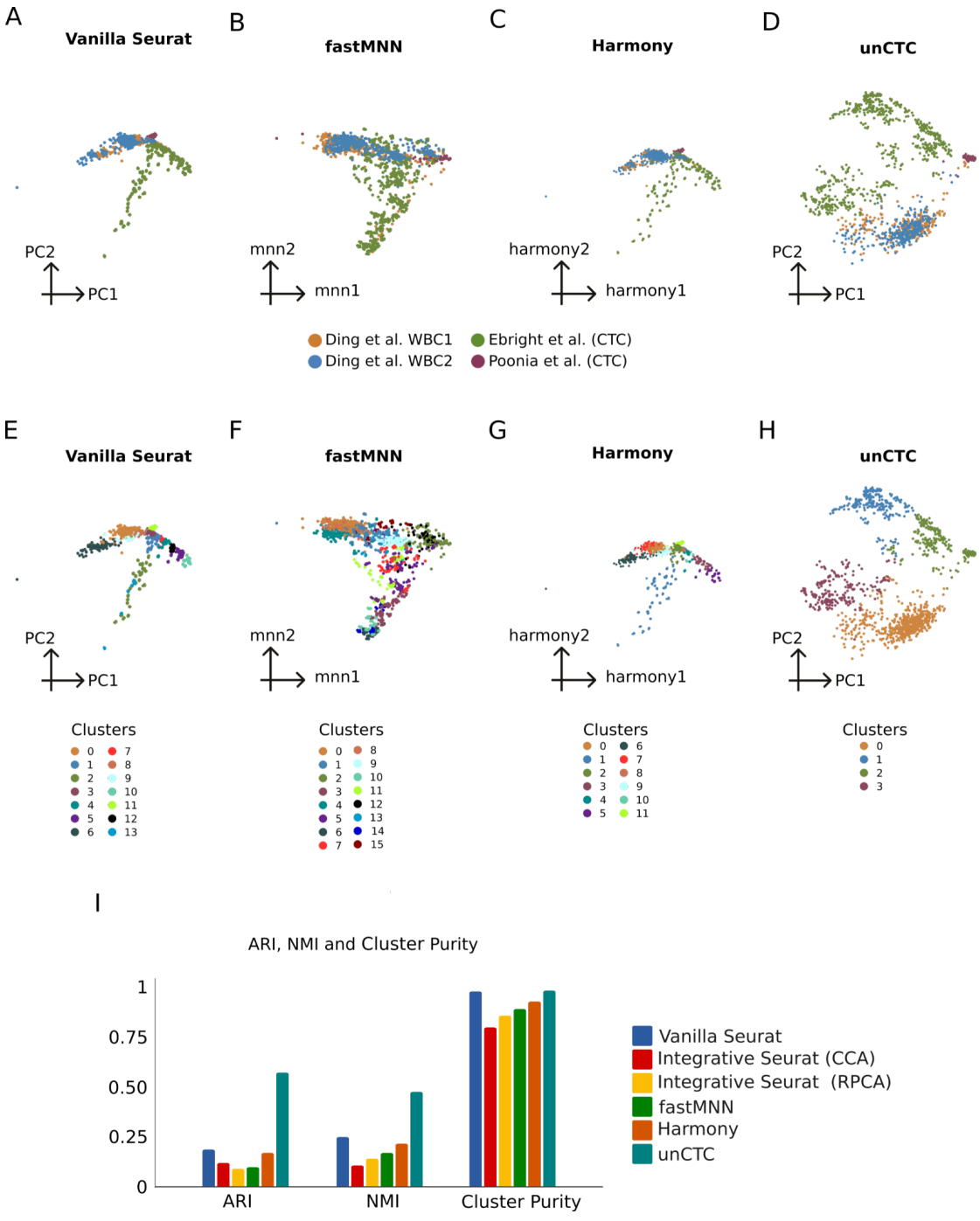
639206), 0.48-µM IS PCR Primer (12 µM, Clontech, 639206), 2X Advantage 2 Polymerase Mix (50X, Clontech, 639206), and 1X Loading Reagent (20X, Fluidigm, 101-1004). All the concentrations correspond to those found in the PCR chambers inside the Polaris IFC. The thermal protocol for preamplification consists of 95 °C for 1 min (enzyme activation), five cycles (95 °C for 20 s, 58 °C for 4 min, and 68 °C for 6 min), nine cycles (95 °C for 20 s, 64 °C for 30 s, and 68 °C for 6 min), seven cycles (95 °C for 30 s, 64 °C for 30 s, and 68 °C for 7 min), and final extension at 72 °C for 10 min. The preamplified cDNAs are harvested into 48 separate outlets on the Polaris IFC carrier. The cDNA reaction products were then converted into mRNA-seq libraries using the Nextera® XT DNA Sample Preparation Kit (Illumina, FC-131-1096 and FC-131-2001, FC-131-2002, FC-131-2003, and FC-131-2004) following the manufacturer's instructions with minor modifications. Specifically, reactions were run at one-quarter of the recommended volume, the tagmentation step was extended to 10 min, and the extension time during the PCR step was increased from 30 to 60 s. After the PCR step, samples were pooled, cleaned twice with 0.9× Agencourt® AMPure® XP SPRI beads (Beckman Coulter), eluted in Tris + EDTA buffer, and quantified using a high-sensitivity DNA chip (Agilent). The pooled library was sequenced on Illumina NextSeq® using reagent kit v3 (2 × 74 bp paired-end read).

Supplemental Note 6. Preprocessing of scRNA-seq datasets. The unCTC R package accepts scRNA-seq data in two forms: transcripts per million (TPM) and raw count data. In the first study, we downloaded scRNA-seq count data from all respective sources. The second study includes three different datasets, including our own. Ding et al. dataset (TPM) was downloaded from the Broad Institute's single-cell gateway. Ebright et al. and Poonia et al. datasets were obtained by processing the associated FASTQ files. We used the FastQC (Andrews 2010) tool to perform quality checks on both datasets for average percent GC content, mean quality score, and per-sequence quality score (Andrews 2010). For alignment purposes of the Ebright data set, we used the hg19 reference genome and hg19 GTF file from Ensembl (release 75) (Howe et al. 2021). To estimate the expression levels of genes, we used RNA-seq by Expectation-Maximization. v.1.3.1 (RSEM) (Li and Dewey 2011) with two scripts: *rsem-prepare-reference* and *rsem-calculate-expression*. Finally, length-normalized TPM datasets (reporting expression of 57773 transcripts) were obtained for both studies. **Supplemental Figure S13** shows the steps used in preprocessing the single-cell RNA-seq datasets. For alignment purposes of the Poonia data set, An index for RNA-seq by expectation maximization (RSEM) was generated based on the hg19/GRCh37. Compared to GRCh37, GRCh38 altered 8000 nucleotides, negligibly affecting sequences of functional genes. Our primary claims and conclusions will remain unaltered while using GRCh38 (Guo et al. 2017). RefSeq transcriptome was downloaded from the UCSC Genome Browser database (Karolchik et al. 2003). Raw read data were aligned directly to this index using RSEM/Bowtie (Li and Dewey 2011; Langmead 2010). Quantification of gene expression levels in counts for all genes in all samples was performed using RSEM v1.2.4. Genomic mappings were

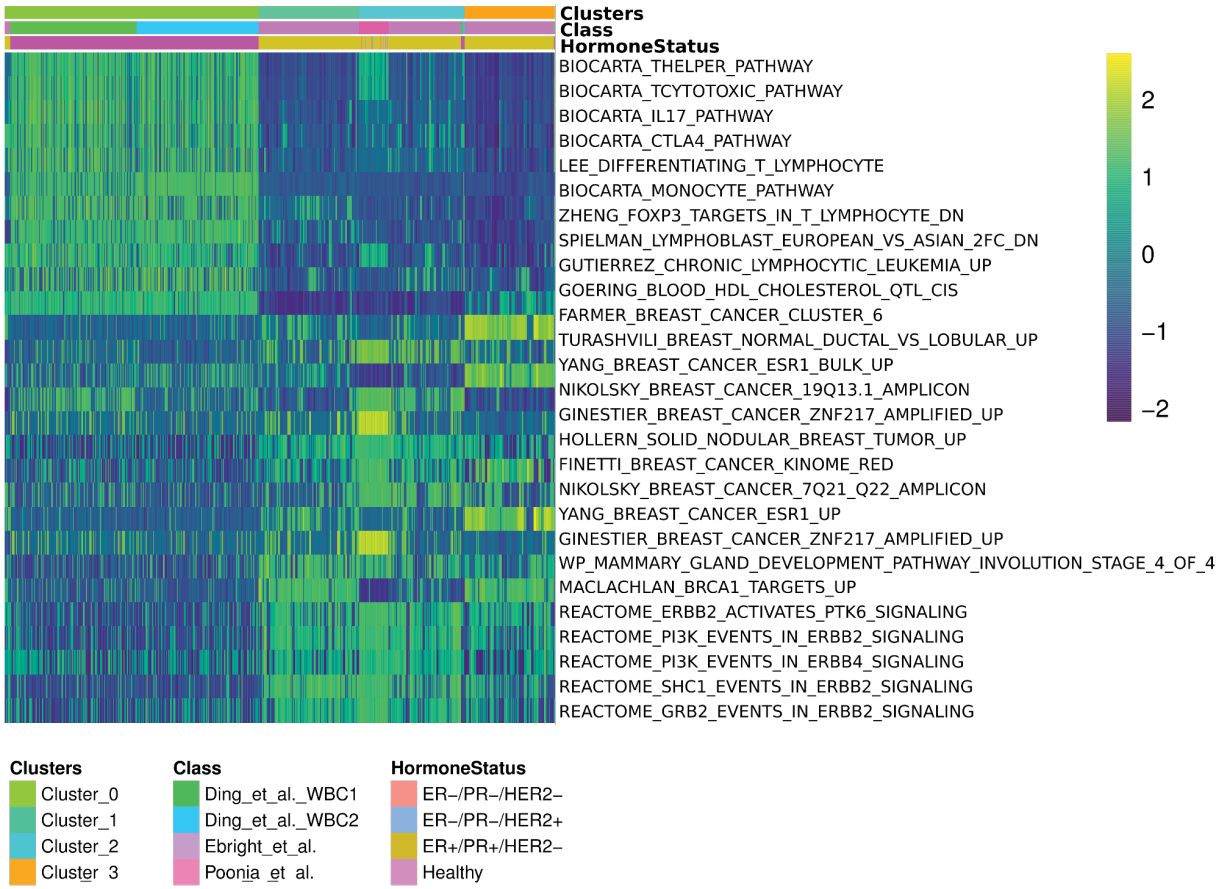
performed with TopHat2 v2.0.13 (Kim et al. 2013), and the resulting alignments were used to calculate genomic mapping percentages. Raw sequencing read data were aligned directly to the human rRNA sequences NR_003287.1 (28s), NR_003286.1 (18S) and NR_003285.2 (5.8S) using Bowtie 2 v2.2.4 (Langmead and Salzberg 2012).



Supplemental Figure S1. Visualization of WBCs and CTCs from Study 1. (A) PCA based visualization of cells, colored by study using Vanilla Seurat , (B,C) Cells, colored by studies are visualized using native embedding methods of fastMNN, and Harmony respectively. (D) Equivalent PCA based visualization of unCTC embeddings. (E-H) Equivalent clusters with all four methods (I) Summarizing the performance of different methods by comparing clusters with WBC/CTC annotations.

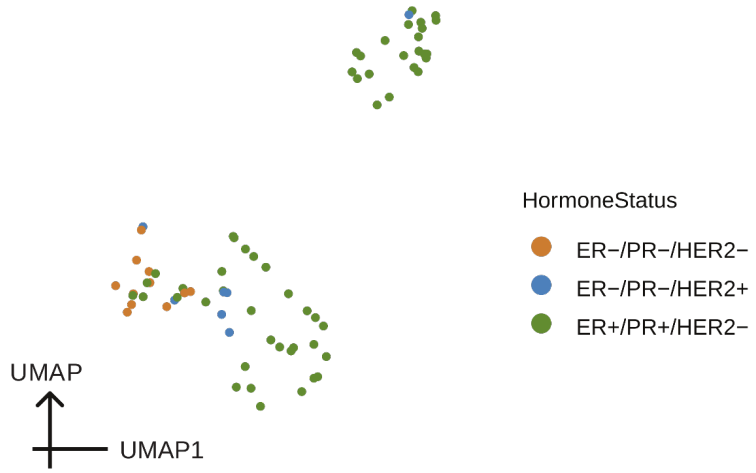


Supplemental Figure S2. Clustering of CTCs obtained from the ClearCell® FX - Polaris™ system with Study 2 scRNA-seq profiles. (A) PCA based visualization of cells, colored by study using Vanilla Seurat , (B,C) Cells, colored by studies are visualized using native embedding methods of fastMNN, and Harmony respectively. (D) Equivalent PCA based visualization of unCTC embeddings. (E-H) Equivalent clusters with all four methods (I) Summarizing performance of different methods by comparing clusters with WBC/CTC annotations.

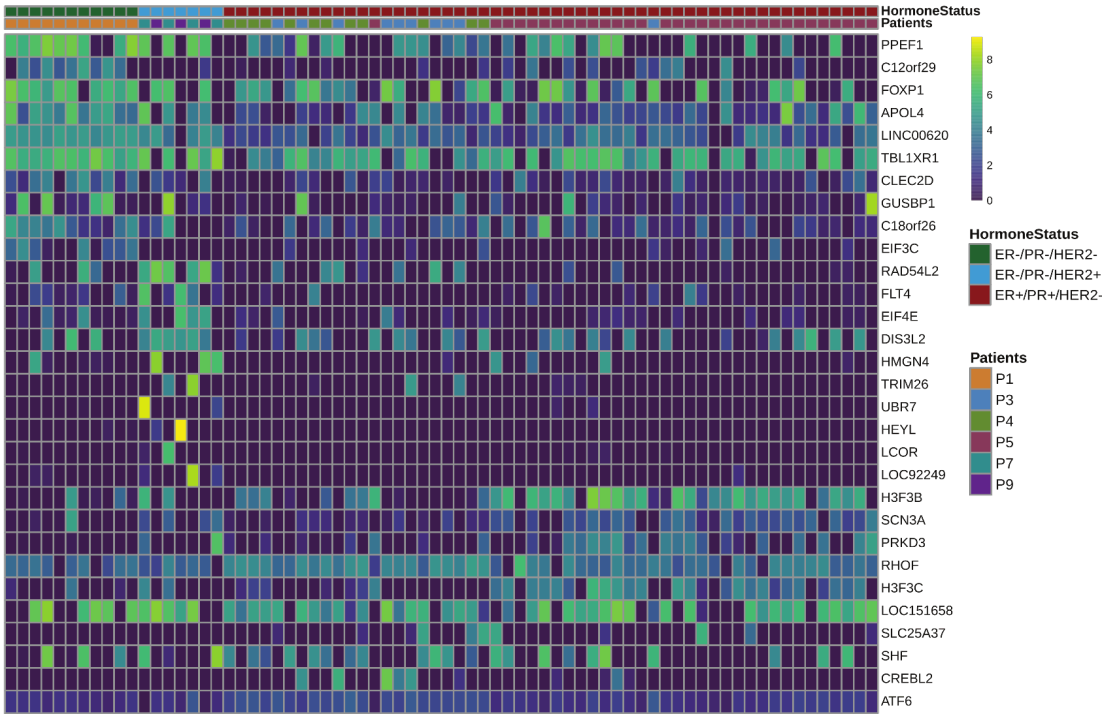


Supplemental Figure S3. Heatmap of differential pathway enrichment scores across four clusters detected by unCTC. Color bars indicate cluster identity, source data information, as well as molecular subcategories.

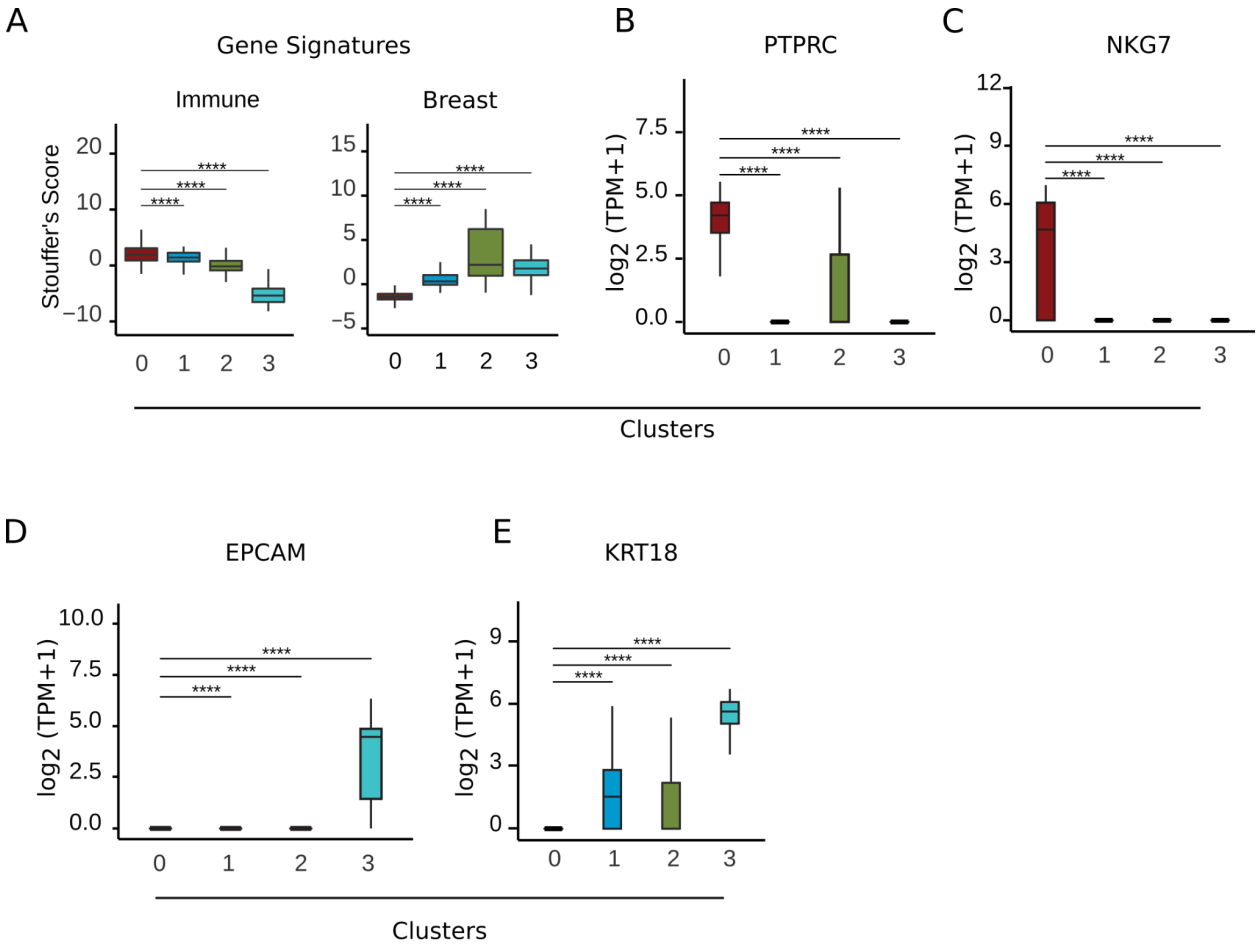
A



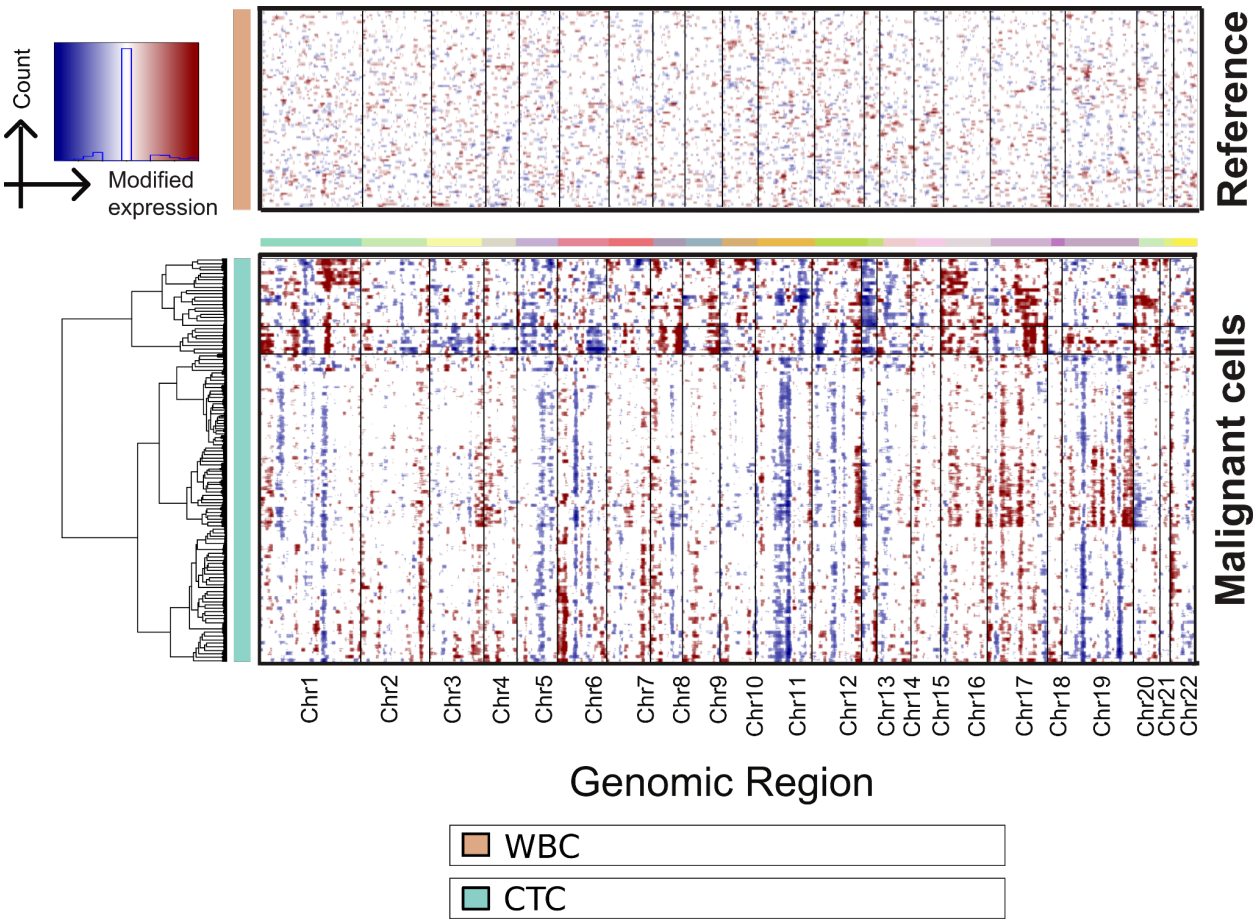
B



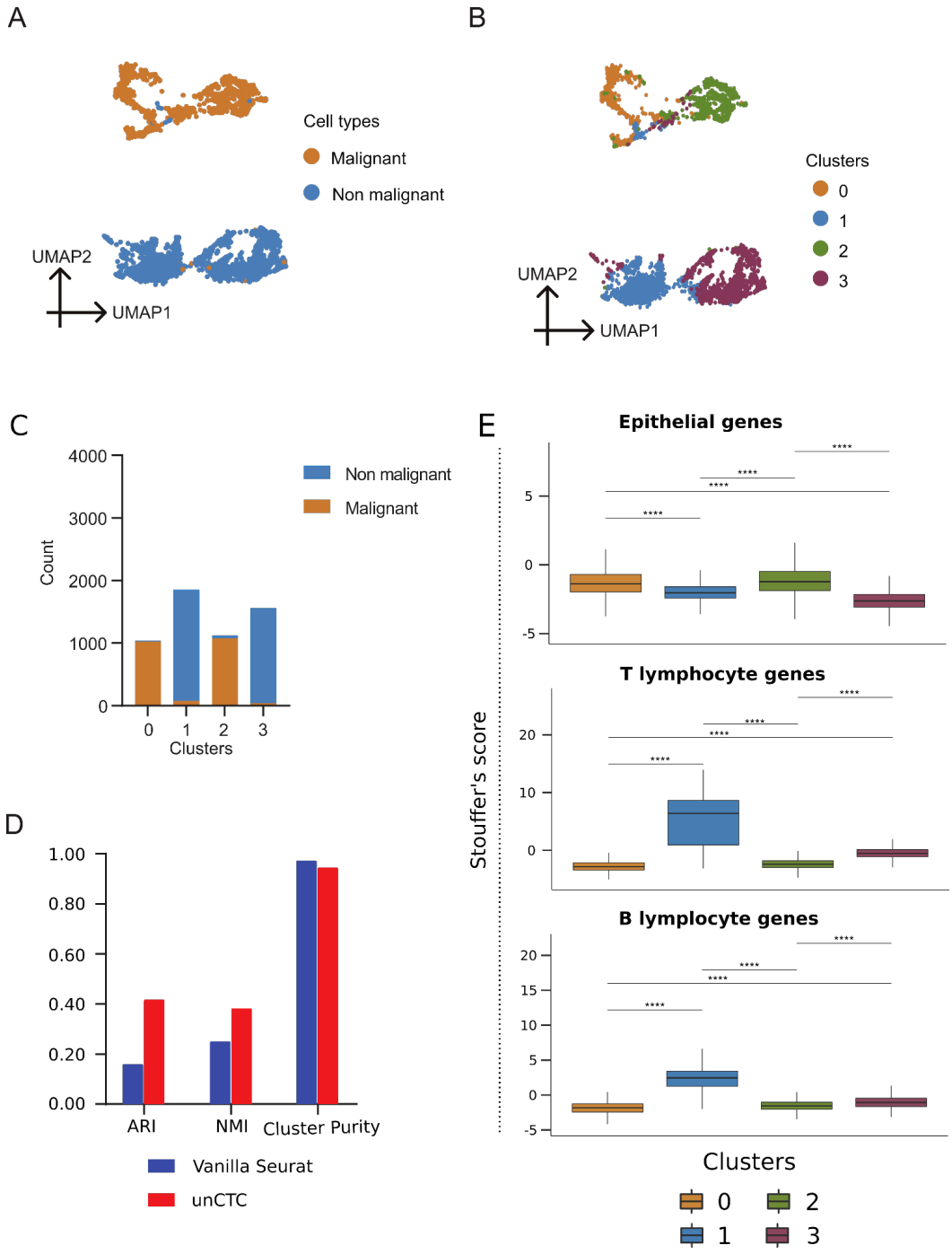
Supplemental Figure S4. Analysis of Poonia et al. RNA-seq count data. (A) UMAP projection of Poonia et al. CTCs based on pathway scores. We observe two apparent ER+/PR+/HER2- subpopulations and spatial segregation of TNBCs. **(B)** Heatmap depicting differential gene expression for ClearCell® FX and Polaris™ selected CTCs across three molecular subtypes. We used the Limma (Ritchie et al. 2015) package with its voom (Law et al. 2014) method for differential gene expression analysis with default parameter settings.



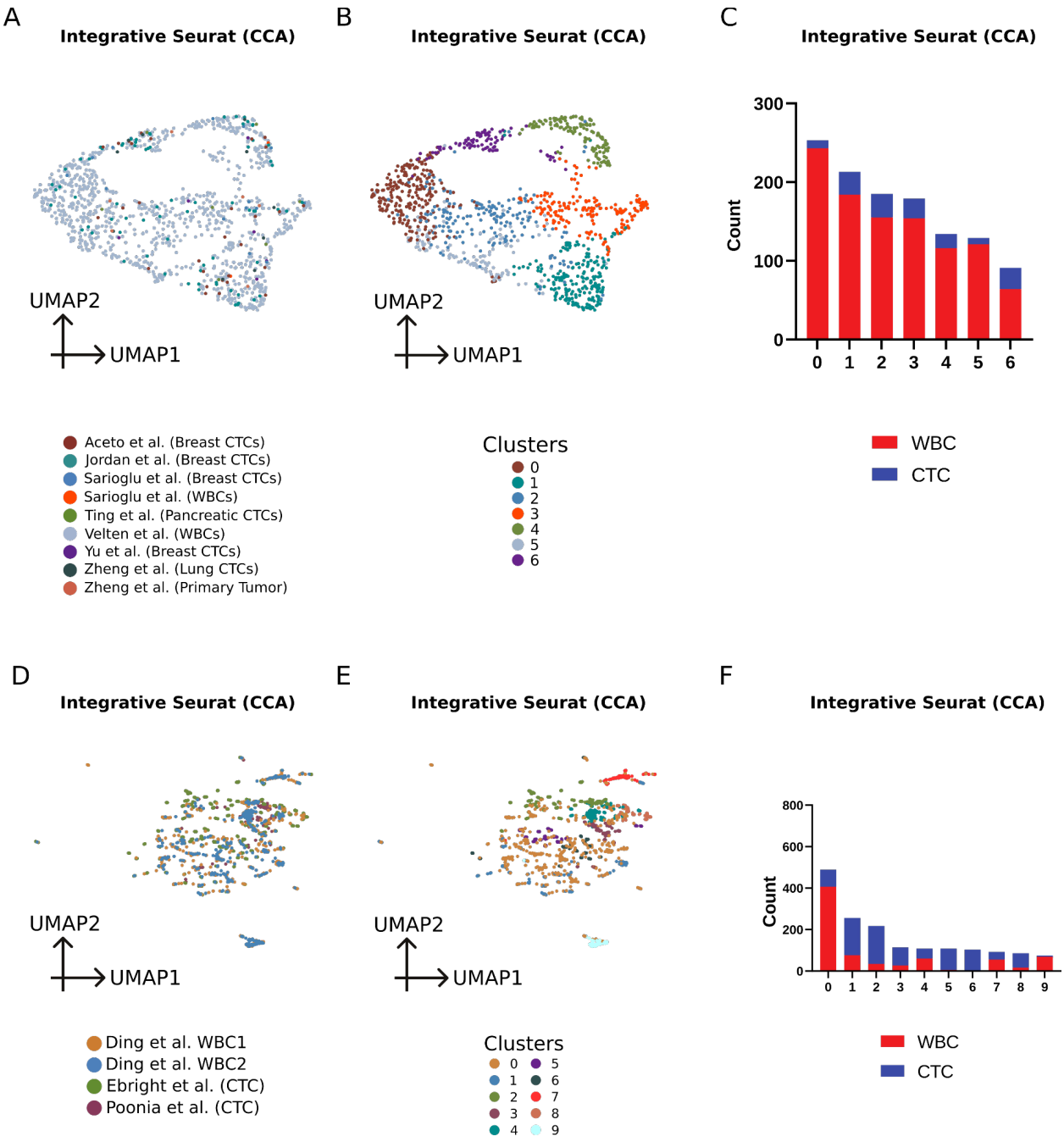
Supplemental Figure S5. (A) Box plot depicting the distribution of Stouffer's scores (Stouffer et al. 1949) associated with genesets specific to immune cells and breast epithelia **(Supplemental Table S4)**. Cluster 0 shows enrichment of immune cell-specific markers, whereas the remaining clusters show enrichment of markers specific to breast epithelia. (B, C) Box plots depicting differential enrichment of select immune cell markers i.e. *PTPRC* and *NKG7*, respectively. (D, E) Box plots depicting differential enrichment of select epithelial markers i.e. *EPCAM* and *KRT18*, respectively.



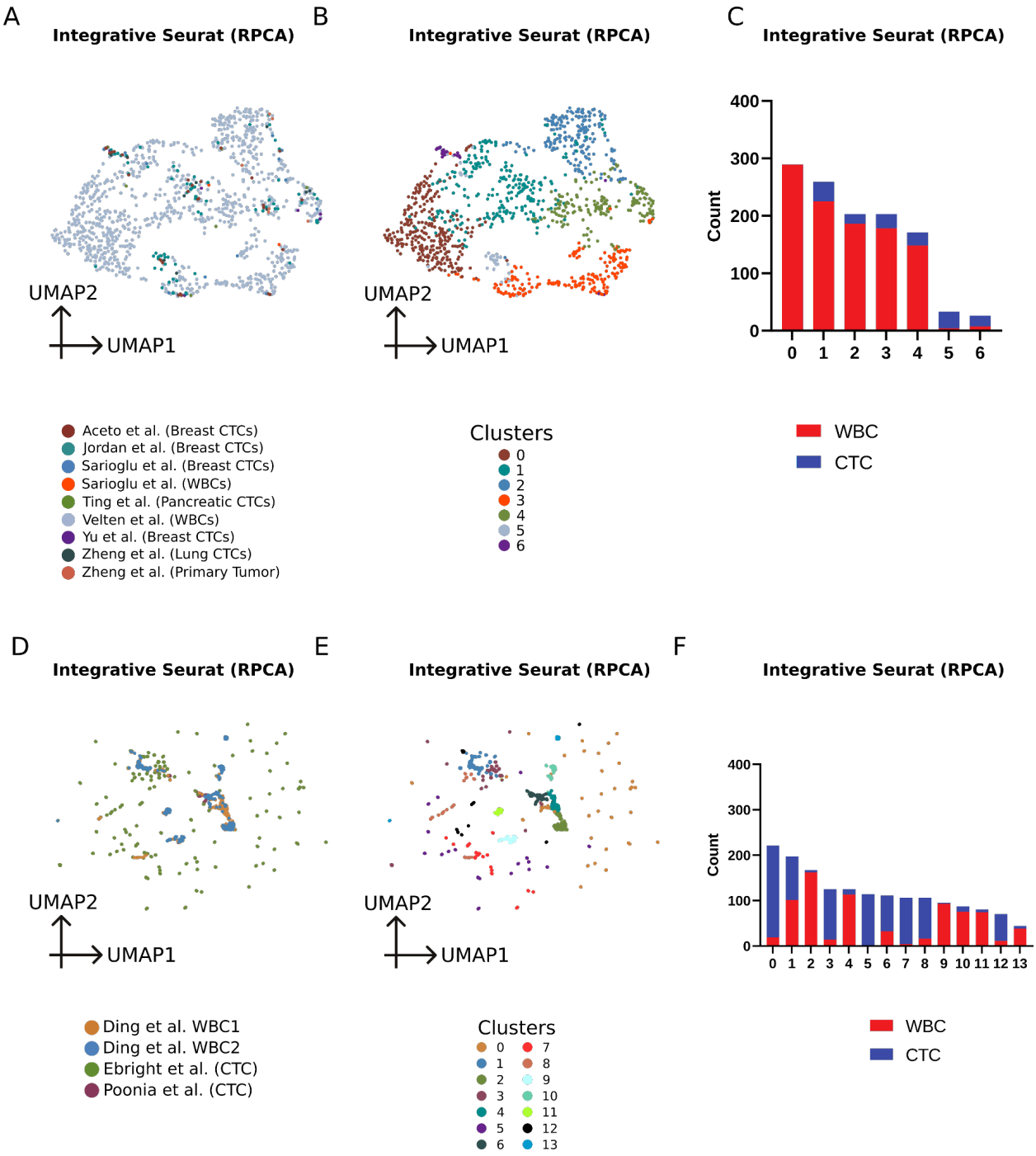
Supplemental Figure S6. Heatmap of Copy number variation obtained from inferCNV ((Tickle et al. 2019) tool in malignant cell (CTCs) from Ebright et al. dataset. Xu et al. PBMC data was used as a reference (WBC) dataset.



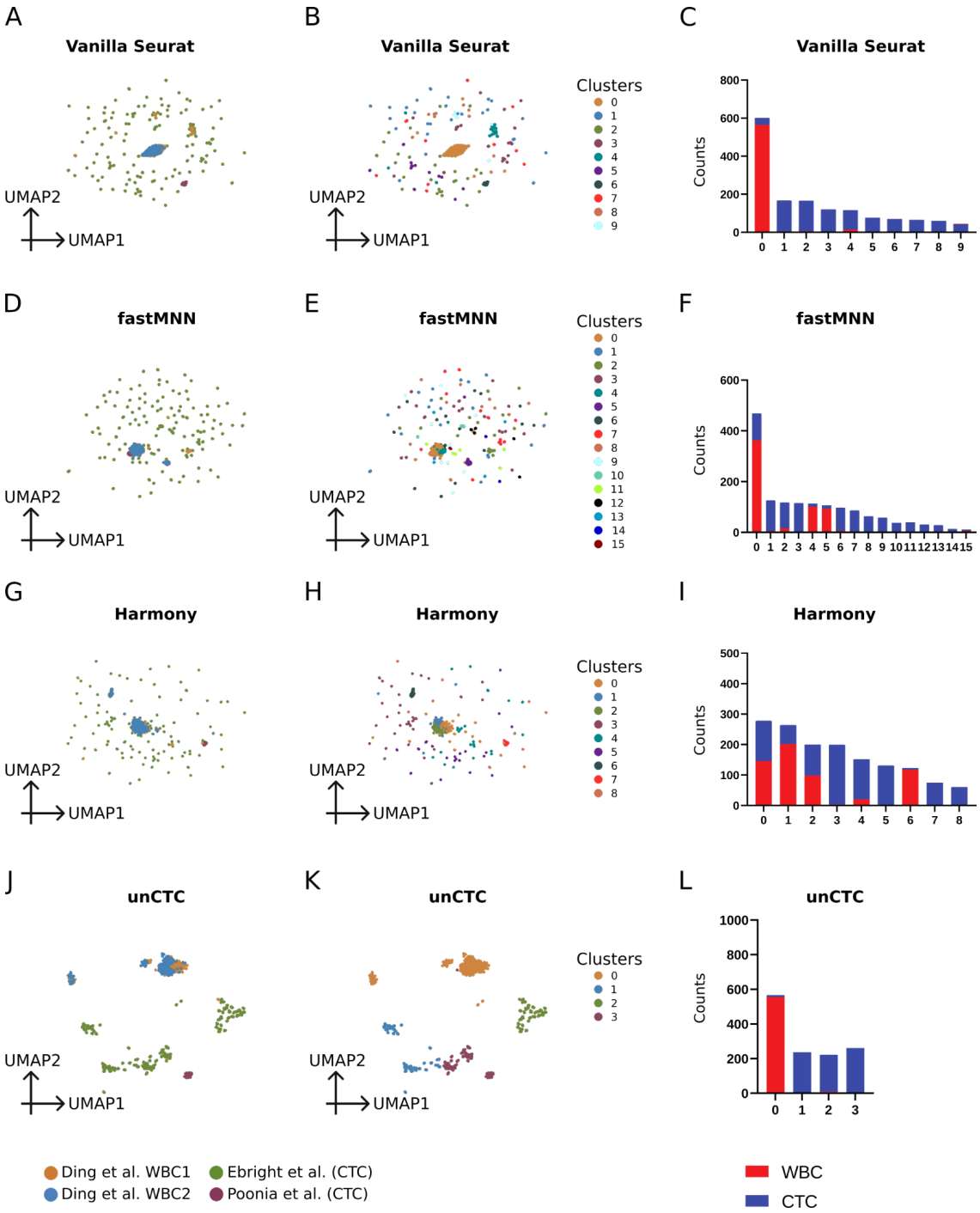
Supplemental Figure S7. (A) UMAP-based visualization of HNSCC data obtained through the unCTC pipeline. (B) Visualization of clusters obtained through unCTC. (C) Barplot shows the distribution of malignant and non-malignant across clusters obtained from unCTC. (D) Barplot showing ARI, NMI, and cluster purity for unCTC and Vanilla Seurat. (E) Boxplots depicting the distribution of Stouffer's scores computed based on known B cell, T cell, and epithelial cell markers, respectively, for cells in each of the unCTC-identified clusters.



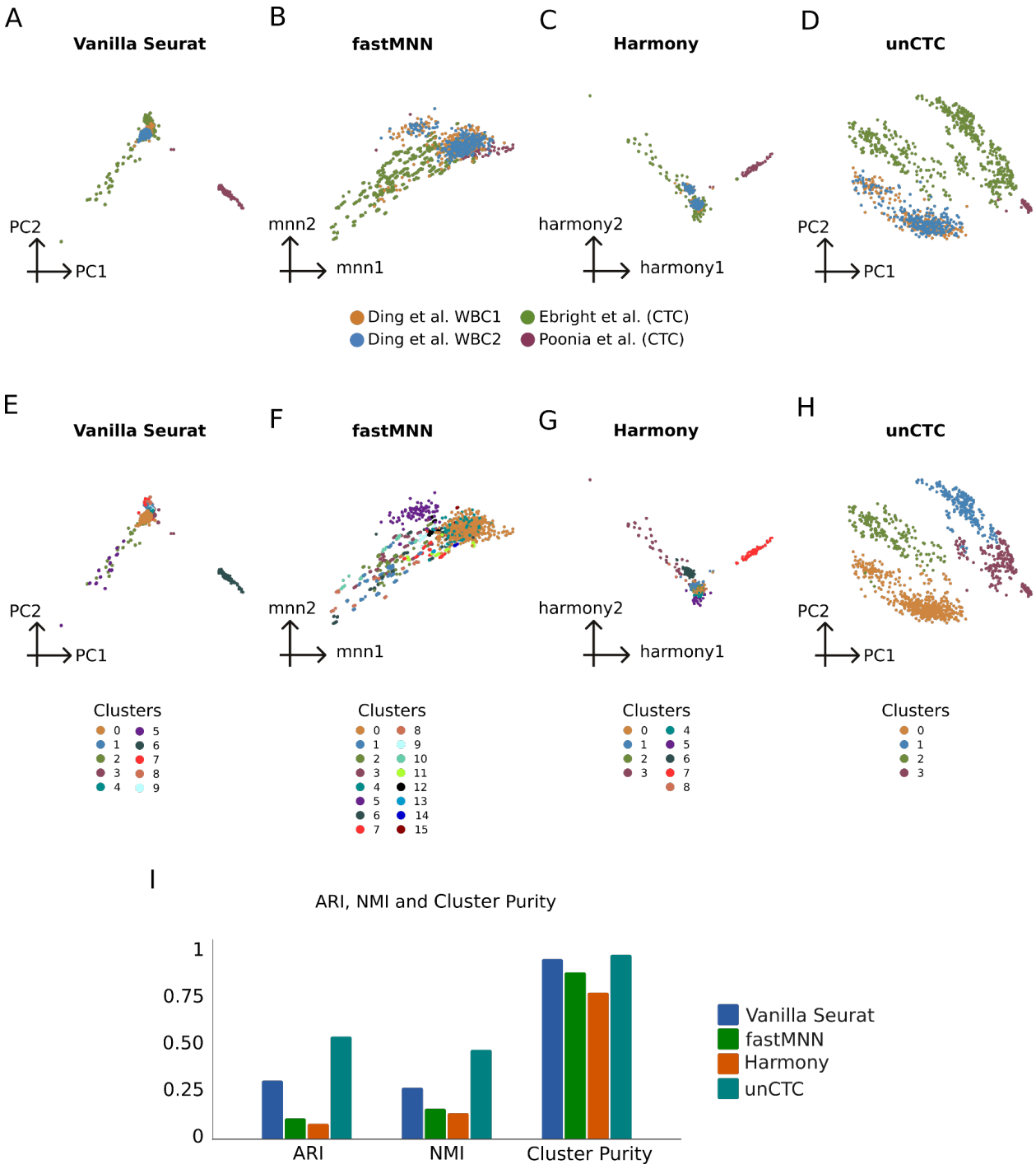
Supplemental Figure S8. (A-C) Seurat's canonical correlation (cca) based integrative analysis of data from Study 1. Subfigure A shows cells, coloured by studies, B shows cells by clusters, and C shows the cluster-wise distribution of CTCs and WBCs. (D-F) Similar figures for Study 2.



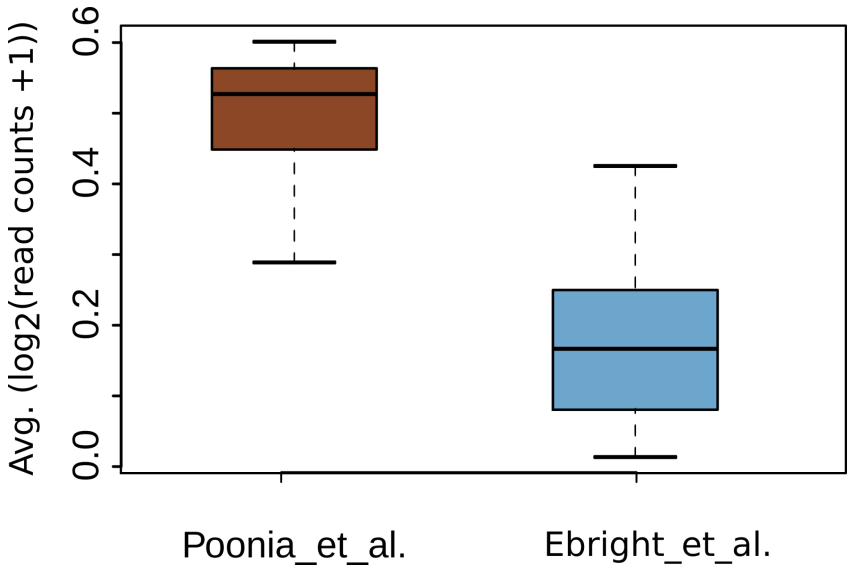
Supplemental Figure S9. (A-C) Seurat's reciprocal PCA (RPCA) based integrative analysis of data from Study 1. Subfigure A shows cells, coloured by studies, B shows cells by clusters, and C shows the cluster-wise distribution of CTCs and WBCs. (D-F) Similar figures for Study 2.



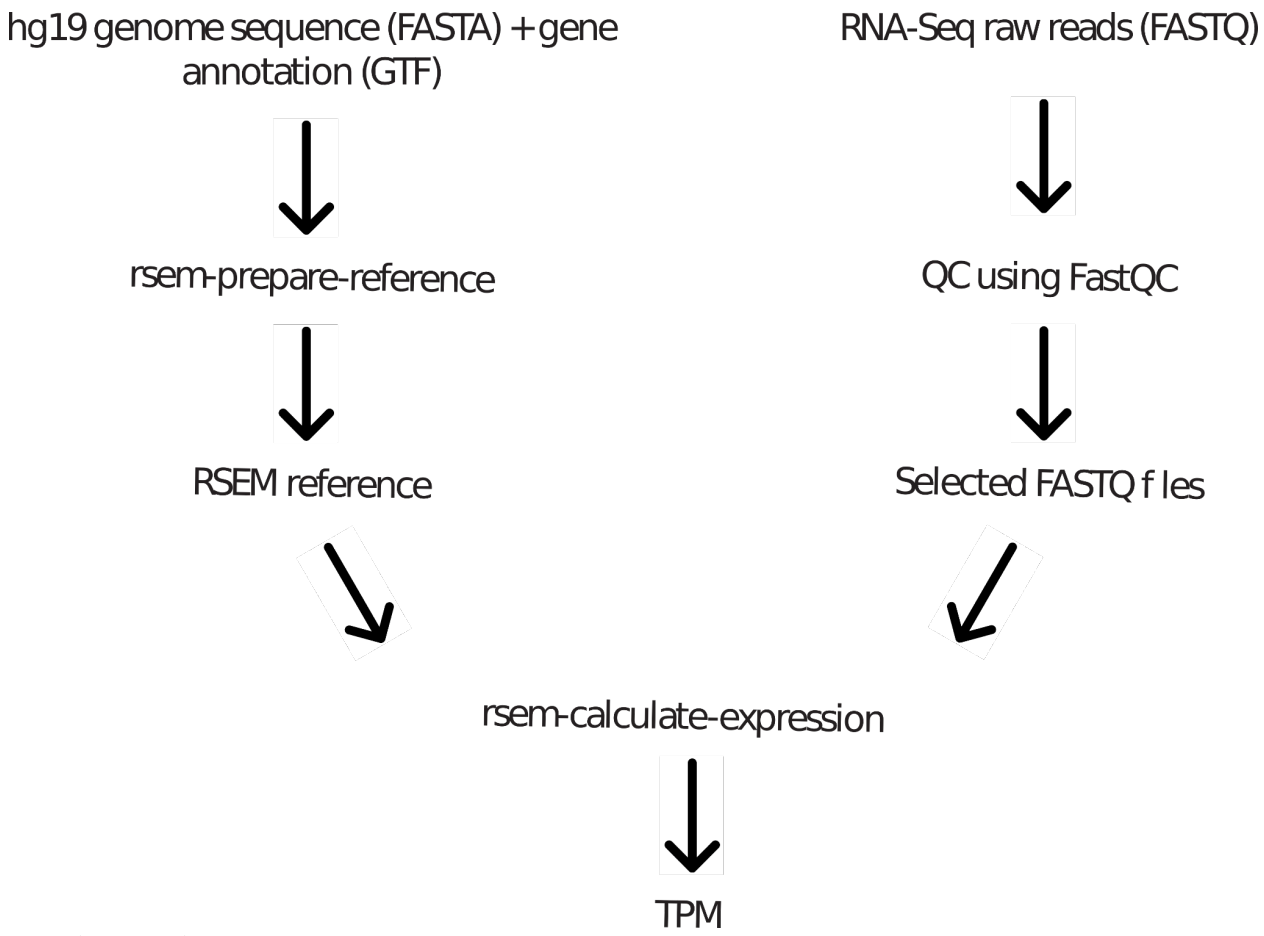
Supplemental Figure S10. Clustering and UMAP-based Visualization of scRNA-seq expression count profile of Study 2. (A-I) Visualization, Clustering, and distribution of CTCs' and WBCs' scRNA-seq profiles, using Vanilla Seurat, fastMNN and Harmony respectively. (J-L) Equivalent figures with unCTC.



Supplemental Figure S11. (A) PCA based visualization of cells, colored by study using Vanilla Seurat , (B,C) Cells, colored by studies are visualized using native embedding methods of fastMNN, and Harmony respectively. (D) Equivalent PCA based visualization of unCTC embeddings. (E-H) Equivalent clusters with all four methods (I) Summarizing the performance of different methods by comparing clusters with WBC/CTC annotations.



Supplemental Figure S12. Average of gene-specific read counts across Poonia and Ebright CTCs.



Supplemental Figure S13. Computational workflow depicting key steps involved in the generation of TPM matrix using the scRNA-Seq FASTQ files. Compared to hg19 (or GRCh37), GRCh38 altered only 8000 nucleotides, negligibly affecting sequences of functional genes. Our primary claims and conclusions will remain unaltered while using GRCh38 (Guo et al. 2017).

References

Andrews S. 2010. Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. *URL: https://www bioinformatics babraham ac uk/projects/fastqc*.

Barrios D, Prieto C. 2017. D3GB: An Interactive Genome Browser for R, Python, and WordPress. *J Comput Biol* **24**: 447–449.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**: 411–420.

Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. 2017. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**: 83–90.

Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**: 421–427.

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. 2021. Ensembl 2021. *Nucleic Acids Res* **49**: D884–D891.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**: 51–54.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Korsunsky F, Fan J, Slowikowski K, Zhang F, Wei K. 2018. Fast, sensitive, and flexible integration of single cell data with Harmony. bioRxiv, 461954.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**: Unit 11.7.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**: R29.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or

without a reference genome. *BMC Bioinformatics* **12**: 323.

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. 2017. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**: 1611–1624.e24.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**: e47.

Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM Jr. 1949. The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1. **1**: 599.

Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. 2019. inferCNV of the Trinity CTAT Project. *Klarman Cell Observatory, Broad Institute of MIT and Harvard*.