Heliyon 10 (2024) e28071

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Single cell analyses of cancer cells identified two regulatorily and functionally distinct categories in differentially expressed genes among tumor subclones

Wei Cao [a,b], Xuefei Wang [a], Kaiwen Luo [a], Yang Li [c], Jiahong Sun [a], Ruqing Fu [a], Qi Zhang [a], Ni Hong [a,*], Edwin Cheung [b,**], Wenfei Jin [d,***]

[a] *School of Life Sciences, Southern University of Science and Technology, Shenzhen, China*
[b] *Cancer Centre, Faculty of Health Sciences, University of Macau, Taipa, Macau SAR*
[c] *Shenzhen People's Hospital, The First Affiliated Hospital, Southern University of Science and Technology, Shenzhen, China*
[d] *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China*

ARTICLE INFO

Keywords:
Colorectal cancer
Single cell RNA sequencing
Single cell whole genome sequencing
Copy number variations
Differentially expressed gene
Cell-cell communication

ABSTRACT

To explore the feature of cancer cells and tumor subclones, we analyzed 101,065 single-cell transcriptomes from 12 colorectal cancer (CRC) patients and 92 single cell genomes from one of these patients. We found cancer cells, endothelial cells and stromal cells in tumor tissue expressed much more genes and had stronger cell-cell interactions than their counterparts in normal tissue. We identified copy number variations (CNVs) in each cancer cell and found correlation between gene copy number and expression level in cancer cells at single cell resolution. Analysis of tumor subclones inferred by CNVs showed accumulation of mutations in each tumor subclone along lineage trajectories. We found differentially expressed genes (DEGs) between tumor subclones had two populations: DEG_{CNV} and DEG_{reg}. DEG_{CNV}, showing high CNV-expression correlation and whose expression differences depend on the differences of CNV level, enriched in housekeeping genes and cell adhesion associated genes. DEG_{reg}, showing low CNV-expression correlation and mainly in low CNV variation regions and regions without CNVs, enriched in cytokine signaling genes. Furthermore, cell-cell communication analyses showed that DEG_{CNV} tends to involve in cell-cell contact while DEG_{reg} tends to involve in secreted signaling, which further support that DEG_{CNV} and DEG_{reg} are two regulatorily and functionally distinct categories.

1. Introduction

Single cell RNA-seq (scRNA-seq) is one of the most powerful technologies to study tumor microenvironment and profiles all cell types in tumor tissue, has significantly increased our knowledge about immune cell infiltration, tumor progression and facilitate the development of immunotherapy [1–8]. Meanwhile, intratumoral heterogeneity is continuously posing a challenge to diagnosis and

* Corresponding author.
** Corresponding author.
*** Corresponding author.

E-mail addresses: hongn@mail.sustech.edu.cn (N. Hong), echeung@um.edu.mo (E. Cheung), jinwf@sinh.ac.cn (W. Jin).

https://doi.org/10.1016/j.heliyon.2024.e28071
Received 12 October 2023; Received in revised form 10 March 2024; Accepted 11 March 2024
Available online 12 March 2024

treatment of cancer, leading to drug resistance and disease recurrence [9–11]. There are successful stories on distinguish the cancer cells from normal cells using gene expression matrix of scRNA-seq data [4,12], However, it is difficult to accurately infer tumor subclones using gene expression matrix of scRNA-seq data [4,13]. On the other hand, single cell whole genome sequencing (scWGS) is particularly useful in inferring intratumoral heterogeneity and tumor subclones [9,14,15]. However, scWGS could not distinguish

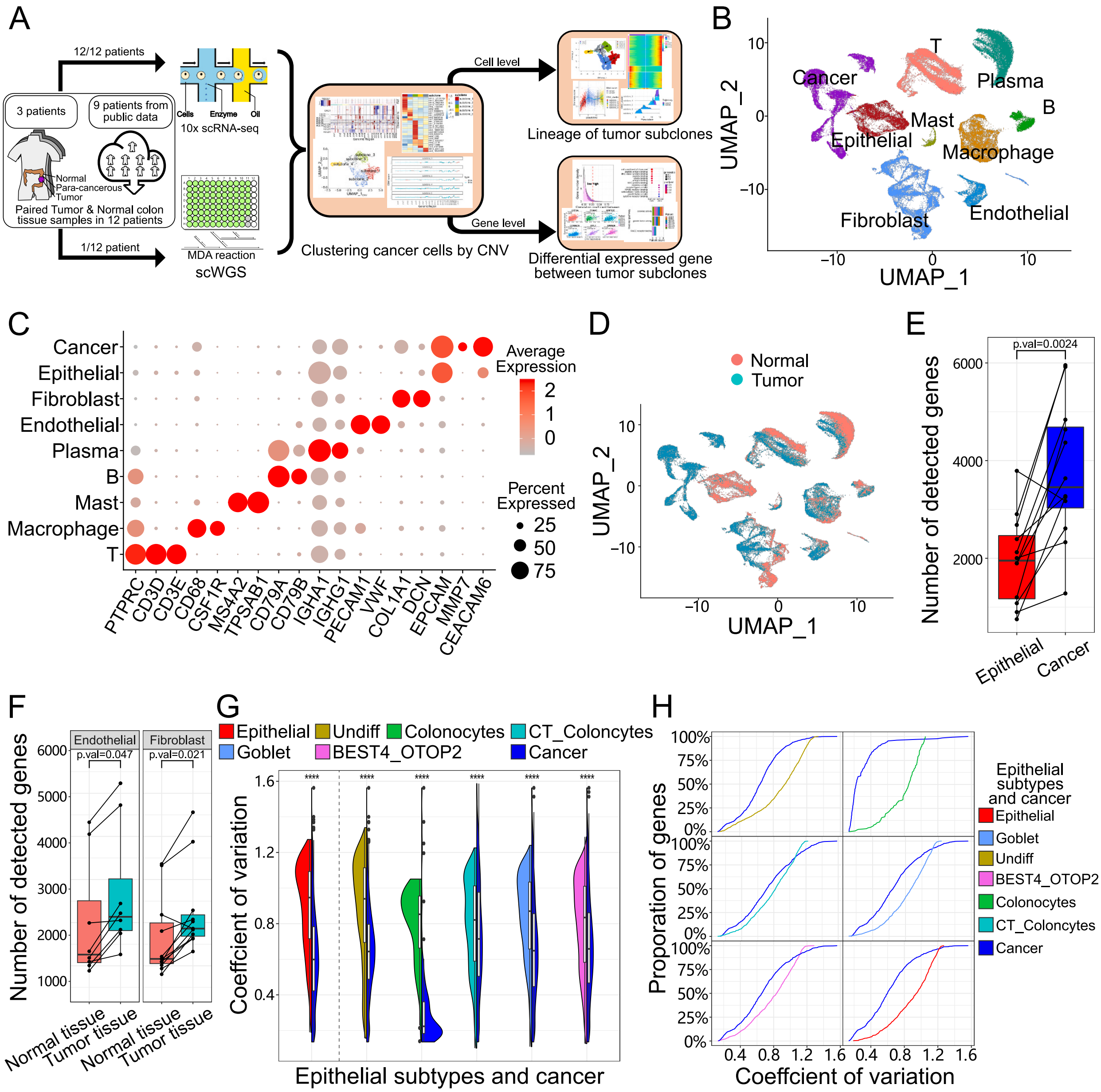

Fig. 1. Cell atlas of CRC and increased cell activities in tumor tissues. (A) The scheme of this study. (B) UMAP visualization of 101,065 cells from tumor tissue and normal colorectal tissue in 12 patients, colored by cell type. (C) Dot plot of normalized expression level and expression percentage of cell-type-specific genes in 9 cell types. (D) UMAP visualization of 101,065 cells from tumor tissue and normal colorectal tissue, colored by tissue. (E) Box plot of average number of detected genes in each patient for cancer cells and epithelial cells. Wilcoxon signed rank test were performed to examine whether the number of detected genes is significantly different between cancer cells and epithelial cells. (F) Box plot of average number of detected genes in each patient in normal tissue and tumor tissue for endothelial cells and stromal cells. Wilcoxon signed rank test were performed to examine whether the number of detected genes were significantly different between normal tissue and tumor tissue for endothelial cells and stromal cells. (G) Splited violin plot of coefficient of variation (CV) of gene expression level of detectable genes among cells in epithelial cells and epithelial subtypes compared with the same genes in cancer cells in patients. Paired Wilcoxon signed rank test was used to examine whether the differences were significant. (H) Accumulation curve of CV of gene expression level in epithelial cells and epithelial subtypes compared with same genes in cancer cells in all patients.

different cell types with normal genomes and is high cost, which hinders its wide application.

An alternative approach for inferring intratumoral heterogeneity is using copy number variations (CNVs) inferred from scRNA-seq data [2,12,16–19]. E.g., Patel et al. inferred the CNVs using scRNA-seq data and identified intratumoral heterogeneity in glioblastoma [19]. Puram et al. used the CNVs inferred from scRNA-seq to identify the intratumoral and intertumoral heterogeneity in head and neck cancer, based on which they found the partial epithelial-to-mesenchymal transition [17]. These success stories of identification of cancer cells and tumor subclones using scRNA-seq inferred-CNV are promising [20]. However, most studies focused on how tumor subclones involved in complex tumor ecosystem [21–24]. The features of cancer cells and tumor subclones is still not well investigation.

In this study, we characterized the features of cancer cells by comparing them with their epithelial cell counterparts using scRNA-seq data. We analyzed the relationship between gene copy number and expression level at single cell resolution. We used CNVs in each cell to infer the tumor subclones and their lineage trajectories. Most importantly, we found DEGs between tumor subclones had two populations, namely $\mathrm{DEG}_{\mathrm{CNV}}$ and $\mathrm{DEG}_{\mathrm{reg}}$. $\mathrm{DEG}_{\mathrm{reg}}$ mainly locates in regions without CNVs and low CNV variation regions, and enriched in cytokine signaling genes. While $\mathrm{DEG}_{\mathrm{CNV}}$ locates in high CNV variation regions and enriched in housekeeping genes and cell adhesion associated genes.

2. Result

2.1. Study design and cell atlas of human colorectal cancers

We designed a workflow to characterize the features of cancer cells and tumor subclones by integrating scRNA-seq and scWGS (Fig. 1A). In brief, we generated scRNA-seq of tumor tissues, para-cancerous tissues and normal tissues from three colorectal cancer (CRC) patients diagnosed with proficient mismatch repair (pMMR) adenocarcinoma. We analyzed the scRNA-seq data of total 36 samples from 12 CRC patients after integrating data from Lee et al. [25]. We further conducted scWGS on patient CRC#1, a patient with high quality scRNA-seq data. We finally integrated the scRNA-seq data and scWGS data to explore the feature of cancer cell and tumor subclones, thus potentially provide biological insight into the gene regulation in cancer.

After quality control, we obtained a total of 101,065 cells, among which 52,063 cells in-house data and 49,002 cells are from public data (Fig. S1A). Cells from para-cancerous tissues almost completely overlap with cells from normal tissues in UMAP plot, as well as cells from tumor border tissue almost completely overlap with cells from tumor tissue (Fig. S1B), indicating these cells from different samples are similar to each other for each patient and batch effects among these samples could be ignored. Therefore, we merged cells in para-cancerous tissues to cells in normal tissues in each patient, and merged cells in tumor border tissues to cells in tumor tissues in each patient as well. The 101,065 filtered cells were clustered into nine cell types, namely cancer cells, T cells, macrophages, mast cells, B cells, plasma cells, endothelial cells, stromal cells and epithelial cells (Fig. 1B and C; Fig. S1C). For each cell type except cancer cells, cells in tumor tissues and their counterparts in normal tissue partially overlap on UMAP. Cancer cells did not overlap with their normal counterparts (epithelial cells) on UMAP plot, indicating the differences between them are pronounced (Fig. 1D).

3. Comparison of cells from tumor tissues with their counterparts in normal tissues

We found cancer cells had much more detected UMIs and more detected genes than their epithelial cell counterparts (Fig. 1E; Fig. S1D), potentially indicating cancer cells have increased cell size and expressed more genes after epithelial cells cancerization. Endothelial cells and stromal cells in tumor tissues also have significantly more detected UMIs and more detected genes than their counterparts in normal tissues (Fig. 1F; Fig. S1E), which may be caused by the different microenvironment between normal tissues and tumor tissues. While detected UMIs and detected genes of other cell types in tumor tissues were not significantly different from their counterparts in normal tissues (Figs. S1F and G).

We identified total 600 DEGs between cancer cells and epithelial cells, of which 357 genes were cancer cell specific and 243 genes were epithelial cell specific (Fig. S2A). The cancer cells specific genes were enriched in ribonucleoprotein complex biogenesis, protein folding, neutrophil degranulation, regulation of apoptotic signaling pathway, cytokine signaling, VEGF pathway and MYC pathway (Figs. S2A and B), potentially indicating cancer cells have increased cell activity, and promotion of immune cell infiltration/maturation. The epithelial cell specific genes were significantly enriched in response to bacterium, metallothioneins bind metals, transport of small molecules, epithelial cell differentiation, response to hormone, primary alcohol metabolic process, regulation of defense response and response to nutrient (Figs. S2A and C), potentially indicating epithelial cells are more sensitive to external environments than their cancer cell counterparts.

We identified total 204 DEGs between tumor associated endothelial cells and normal endothelial cells, of which 100 genes were tumor associated endothelial cell specific and 104 genes were normal endothelial cell specific (Fig. S3A). The tumor associated endothelial cell specific genes were enriched in extracellular matrix organization, blood vessel development, regulation of angiogenesis and regulation of cell-substrate adhesion (Figs. S3A and B), indicating endothelial cells in tumor played important role in promotion of angiogenesis. The normal endothelial cell specific genes were enriched in positive regulation of cell migration, negative regulation of cell adhesion, negative regulation of cell population proliferation and negative regulation of catalytic activity (Figs. S3A and C), indicating normal endothelial cell have weaker cell adhesion and proliferation than their counterparts in tumor tissue. We further identified total 310 DEGs between tumor associated stromal cells and normal stromal cells, of which 164 genes were tumor associated stromal cell specific and 146 genes were normal stromal cell specific (Fig. S3D). The tumor associated stromal cell specific genes were enriched in extracellular matrix organization, vasculature development and response to wounding (Figs. S3D and E), while

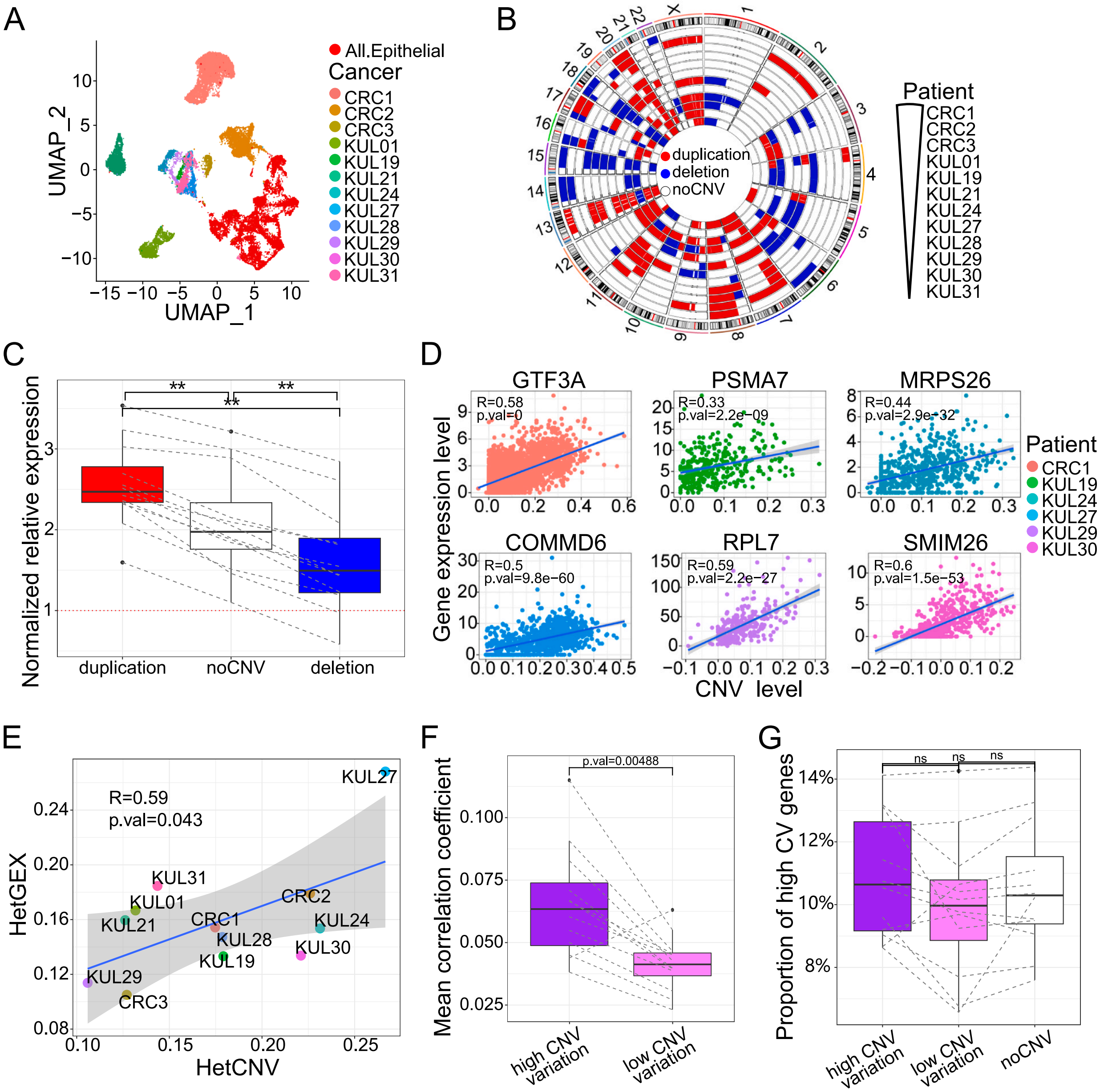

Fig. 2. Cancer cells have a specific regulatory mechanism that gene expression levels depend on its DNA copy number. (A) UMAP visualization of epithelial cells (red) and cancer cells, with cancer cells colored by patient. (B) Circos heatmap showing duplication, deletion, and noCNV genomic regions in 12 patients. (C) Box plot of gene expression in cancer cells normalized to their expression in normal counterpart in duplication, deletion, and noCNV genomic regions in each patient. Paired Wilcoxon signed rank test was used to examine whether the differences were significant. (D) Scatter plots showed the correlation between expression level and CNV level in CNV regions. Shaded areas corresponded to 0.95 confidence interval. (E) Scatter plot of the correlation between Intratumoral heterogeneity scores based on gene expressions (HetGEX) and Intratumoral heterogeneity scores based on CNVs (HetCNV) for the 12 patients. Shaded areas corresponded to 0.95 confidence interval. (F) Box plot of correlation coefficient between gene expression level and CNV level in CNV variation regions and low CNV variation regions. Paired Wilcoxon signed rank test was used to examine whether the differences were significant. (G) Box plot of the fraction of high CV genes in each chromosomal region. Chromosomal regions were classified into high CNV variation, low CNV variation region and no CNV region in each patient. Paired Wilcoxon signed rank test was used to test whether differences were significant. High CV genes are defined as the 2000 genes with the highest CV of gene expression in cancer cells of each patient.

normal stromal cell specific genes were enriched in negative regulation of cell population proliferation, negative regulation of cell adhesion and response to hypoxia (Figs. S3D and F).

3.1. Cancer cells have lower cell-cell variation than their epithelial cells

Although the high cellular heterogeneity of tumor tissue is well known, we found the cell-cell variations of cancer cells are significantly lower than that of epithelial cells (Fig. 1G). To further explored cell-cell variation, we further cluster epithelial cells into several subsets (Figs. S4A and B) using previously reported marker genes [26]. The cell-cell variations of cancer cells are significantly lower than that of each epithelial cell subset in pooled cells or in each CRC patient (Fig. 1G; Fig. S4C). The low cell-cell variation of

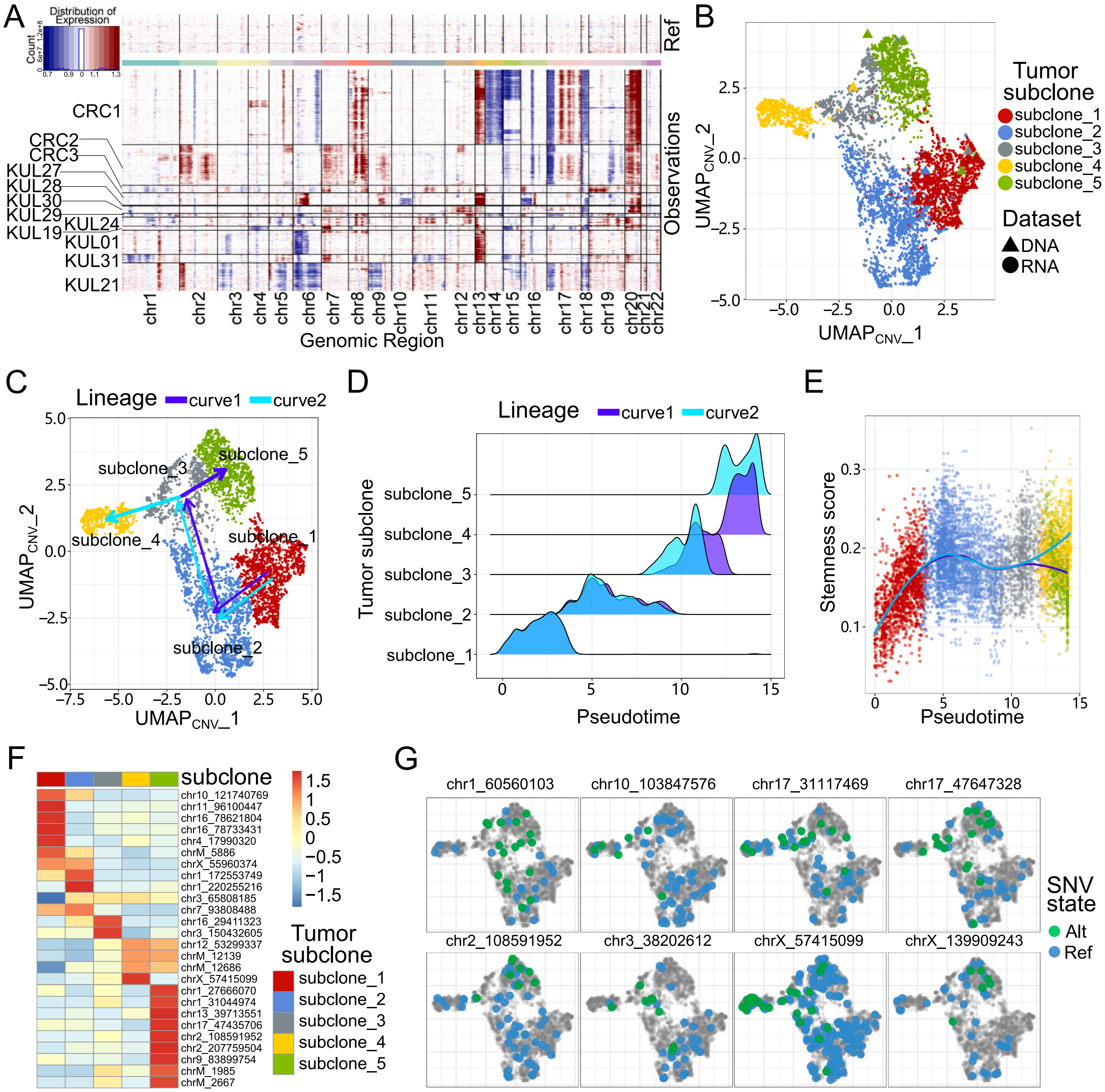

Fig. 3. Tumor subclones and its pseudotime trajectory revealed by single cell CNV analyses. (A) Heatmap of CNV level in reference normal cells and cancer cells, with rows and columns representing cells and genomic loci, respectively. (B) UMAP visualization of cancer cells from both scWGS and scRNA data in patient CRC#1, Cells colored by tumor subclone inferred by CNVs and shaped by datasets. (C) Lineage trajectories of cancer cells in CRC#1 on UMAP. (D) Density plot of the positioning of cells from each tumor subclone on pseudotime trajectories. (E) Scatter plot of the stemness score along pseudotime trajectories showed the stemness increase quickly at the early stage. (F) Heatmap of SNV alternative ratios' z-score by row in each tumor subclone. (G) UMAP visualization of cancer cells in CRC#1, colored by the SNV alternative ratio of each SNV.

cancer cells may be due to cancer cells were derived from the clone expansion of one or a few mutated cells, thus the cancer cells had a few states. To understand the cell-cell variation at gene level, we compared the cell-cell variation of each gene between epithelial cells and cancer cells. We found cell-cell variations of most genes in epithelial cells were much higher than that in cancer cells (Fig. 1H; Fig. S4D), consist with aforementioned observations that cell-cell variations of epithelial cells are higher than that of cancer cells.

3.2. Heterogeneity of cancer cells and inference of CNVs

To further understand the feature of cancer cells, we conducted UMAP analysis on epithelial-like cells including cancer cells and epithelial cells. We found epithelial cells from normal tissue in all patient samples mixed in a single cluster, while the cancer cells from each patient have a patient-specific cluster (Fig. 2A). The results indicate cancer cells are quite different from patient to patient, revealing high inter-patient diversity or inter-patient heterogeneity. Although the high inter-patient heterogeneity of cancer cells, the entropy of cancer cells from most patient is significantly higher than their epithelial cell counterparts (Fig. S5A), potentially indicating de-differentiation or increased stemness of cancer cells after epithelial cell carcinogenesis. Indeed, the cancer cells have much higher stemness scores than their epithelial cell counterparts in most patients (Fig. S5B). Since the inter-patient differences among cancer cells are dominant, we inferred copy number variations (CNVs) in each patient using scRNA-seq data. In order to better present CNVs, we separated the genomic regions into duplication regions, deletion regions and regions without CNVs (noCNV) in each patient according to CNV level of the genomic regions in cancer cells (Fig. 2B).

3.3. Gene expression level correlated with its DNA copy number in cancer cells

The expression level of each gene in cancer cells was normalized using its expression level in epithelial cell counterparts. The genes located in duplication regions express significantly higher than the genes in noCNV regions ($P = 4.9 \times 10^{-4}$) and deletion regions ($P = 4.9 \times 10^{-4}$), with genes in deletion regions have the lowest expression levels (Fig. 2C). These results indicate that gene expression in cancer cells are strongly associated with its DNA copy number. Indeed, there are many genes showing significant correlation between expression level and CNV level at single cell resolution (Fig. 2D). We further calculated intra-tumoral heterogeneity scores based on CNV level (HetCNV) and intra-tumoral heterogeneity scores based on gene expression (HetGEX). We found HetGEX highly correlated with HetCNV (Fig. 2E), potentially indicating the heterogeneity of CNV impacts the heterogeneity of gene expression. The copy number of CNV and heterogeneity of CNV influenced the gene expression should be a unique features of cancer cells since normal cells are diploid and gene expression were precisely regulated.

Gene expression levels are more likely to be determined by its DNA copy number in high CNV variation regions.

To further investigate the relationship between gene expression level and CNV level, CNV regions were further classified into high CNV variation regions and low CNV variation regions in each patient. We found genes in high CNV variation regions have significant higher correlation coefficient between gene expression level and CNV level than low CNV variation regions (Fig. 2F). Meanwhile, the fractions of genes showing high expression variation in high CNV variation regions, low CNV variation regions and noCNV regions are not significantly different (Fig. 2G), indicating genes with high expression variation have no bias in different CNV region categories.

3.4. Inference of tumor subclones using inferred CNVs

In order to better understand the heterogeneity of cancer cells, we inferred copy number variations (CNVs) in each epithelial-like cell using scRNA-seq data. Indeed, the epithelial cells from normal tissues did not have pronounced the chromosomal deletions or duplications; while the cancer cells from different patients have many patient-specific genomic alterations over large genomic regions (Fig. 3A), which could explain why cancer cells showed patient specific cluster on UMAP plot. Although the differences of CNVs among patients are major, we also found the differences of CNVs among cancer cells from the same patient (Fig. 3A). To better illustrate the heterogeneity of cancer cells in each patient, the inferred CNVs was used to cluster cancer cells into tumor subclones in each patient (Fig. S5C).

We also inferred CNVs and tumor subclones using scWGS data of patient CRC#1. We found that the CNVs inferred by scRNA-seq in each tumor subclones were essentially consistent with that inferred by scWGS (Figs. S6A–C), consistent with previous studies [19,20,27,28]. CNVs inferred in scWGS were more accurate and higher resolution than that inferred by scRNA-seq. Furthermore, cell abundance of non-cancer cells and each tumor subclone inferred based on scRNA-seq data is very similar to its tumor subclone counterpart inferred by scWGS data (Fig. 3B; Fig. S6D).

3.5. Tumor lineages and accumulation of mutations in each tumor subclone

We inferred pseudotime of cancer cells to better understand the relationship among these tumor subclones. We analyzed the CNV data using slingshot and identified two lineage trajectories of tumor subclones: (1) subclone 1 → subclone 2 → subclone 3 → subclone 4; (2) subclone 1 → subclone 2 → subclone 3 → subclone 5 (Fig. 3C and D). Furthermore, the two main lineages inferred by monocle 2 were essentially consistent with that inferred by slingshot (Figs. S6F–H). For the inferred lineages, the stemness scores of cancer cells increase rapidly at the beginning of the pseudotime. After separation of the two lineages, the scores continuously increase along lineage #2 while scores along lineage #1 are stable (Fig. 3E). Furthermore, the inferred CNV scores of each tumor subclone increased along with lineage trajectories (Fig. S6E), indicating the accumulation of CNV along lineages.

We further called single nucleotide variants (SNVs) in scRNA-seq data using GATK. We clustered alternative ratio of SNVs in tumor subclones and identify many tumor subclones specific SNVs (Fig. 3F). The distribution of alternative alleles on UMAP showed tumor subclones specific (Fig. 3G), which support the accumulation of mutations in each tumor subclone.

3.6. The DEGs between tumor subclones have two populations

We identified 288 DEGs and 85 DEGs between tumor subclones by setting average expression level fold change >1.5 and > 1.8, respectively (Fig. 4A). The DEGs are significantly enriched in interferon signaling, neutrophil degranulation, cellular response to cytokine stimulus and regulation of cell-cell adhesion (Figs. S7A and B). The genes showing the most significant difference include *PLCG2* and *CXCL14*, which have been reported as tumor-associated marker genes [29,30]. The distribution of correlation coefficient between gene expression level and CNV level revealed the existence of two populations of DEGs: DEGs with high CNV-expression correlation and DEGs with no CNV-expression correlation (Fig. 4B). The natural separation between the two populations of DEGs may indicate they have different features.

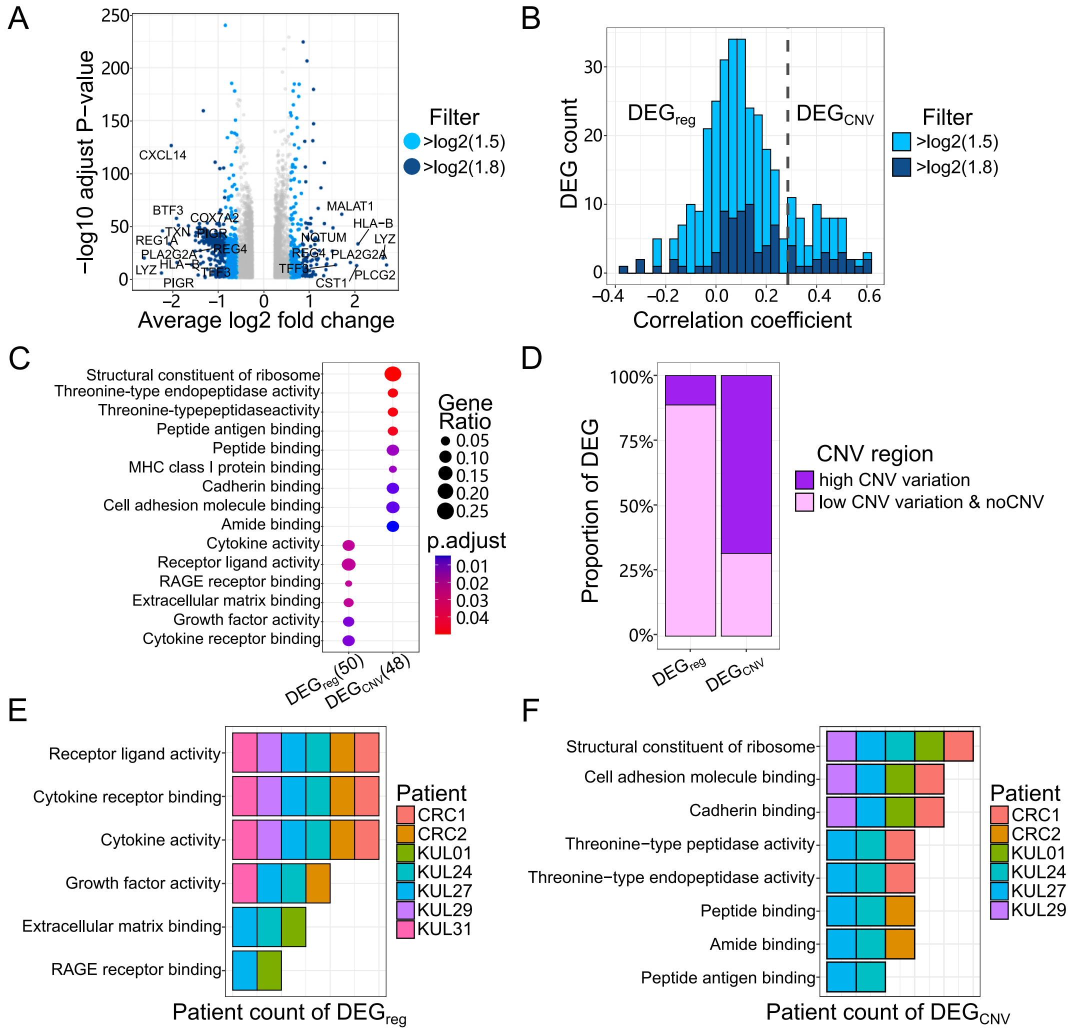

Fig. 4. DEGs between tumor subclones showed two regulatorily and functionally distinct populations. (A) Volcano plot of DEGs between tumor subclones in each patient. Light blue and blue represent fold change >1.5 and fold change >1.8, respectively. (B) Histogram plot of correlation coefficient between gene expression level and CNV level. The DEGs sets were set at fold change >1.5 and fold change >1.8, respectively. (C) GO enrichment analysis of DEGs with high CNV-expression correlation (DEG$_{CNV}$) and DEGs with low CNV-expression correlation (DEG$_{reg}$). Dot size and color represents gene ratio and adjusted p-value, respectively. (D) Percentage of DEGs located in high CNV variation region and NoCNV/low CNV variation regions. Purple and pink represents high CNV variation regions and noCNV/low CNV variation regions. (E) Bar plot of patient count representing whether the patient's DEG$_{reg}$ is enriched in the GO term. (F) Bar plot of patient count representing whether the patient's DEG$_{CNV}$ is enriched in the GO term.

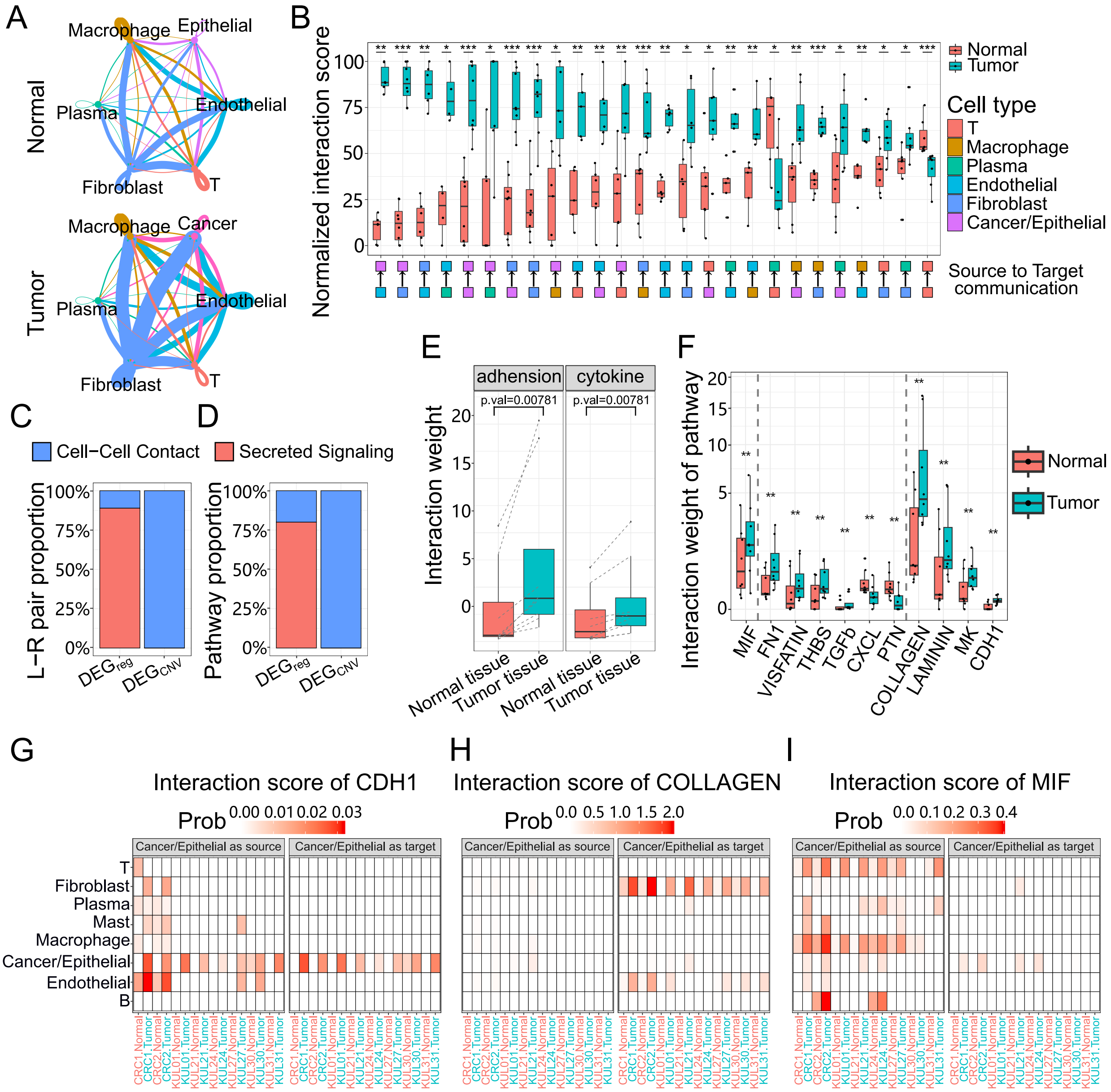

Fig. 5. The change of cell-cell communication between tumor tissues and normal tissues. (A) Cell-cell communication networks of ligand-receptor pairs in normal tissue (upper panel) and tumor tissue (lower panel) in all patients. Edge colors are consistent with the signaling senders. Edge weights are proportional to the interaction weight, and thicker edge line indicates a stronger signal. (B) Box plot of normalized interaction score of cell type pairs in all patients. Pairs are ranked by the interaction difference between normal tissue and tumor tissue. Paired Wilcoxon signed rank test was used to examine whether the differences between tumor tissue and normal tissue were significant in each source to target communication pairs, p values were adjusted by BH. (C) Bar plot of percentage of genes involved in cell-cell contact and secreted signaling in $\mathrm{DEG}_{\mathrm{CNV}}$ and $\mathrm{DEG}_{\mathrm{reg}}$. (D) Bar plot of percentage of pathways involved in cell-cell contact and secreted signaling in $\mathrm{DEG}_{\mathrm{CNV}}$ and $\mathrm{DEG}_{\mathrm{reg}}$. (E) Box plot of interaction weight of cytokine and adhesion signaling pairs in tumor tissue and normal tissue. Paired Wilcoxon signed rank test was used for examining whether differences were significant. (F) Box plot of interaction weight of 6 cytokine and adhesion shared pathways (FN1, VISFATIN, THBS and TGFb; PTN and CXCL), one cytokine specific pathway (MIF), four adhesion specific pathways (COLLAGEN, LAMININ, MK and CDH1) that in tumor and normal tissue. Wilcoxon signed rank test was used for examining whether differences were significant between tumor and normal tissue in each pathway, p values were adjusted by BH. (G–I) Heatmap of interaction score of cancer/epithelial cell to other cell types (left) and other cell types to cancer/epithelial cell (right) of CDH1 (G), COLLAGEN (H) and MIF (I) pathways in each patient's normal and tumor tissue.

3.7. DEGs with high or low CNV-expression correlation enrich different GO terms

GO analysis showed DEGs with high CNV-expression correlation were enriched in housekeeping genes and receptor-ligand binding genes such as peptide antigen binding, ($P = 3.0 \times 10^{-3}$), cadherin binding ($P = 0.041$), structural constituent of ribosome ($P = 2.6 \times 10^{-11}$) and MHC class I protein binding ($P = 0.033$) (Fig. 4C), which indicate differential expression of these housekeeping genes between tumor subclones depend on the change of their DNA copy number. While DEGs with low CNV-expression correlation were enriched in secretory signaling such as cytokine activity ($P = 0.026$), RAGE receptor binding ($P = 0.026$), cytokine receptor binding ($P = 0.038$) and growth factor activity ($P = 0.038$) (Fig. 4C), indicating differential expression of these cytokine signaling genes do not depend on the change of their DNA copy number.

Indeed, DEGs with high CNV-expression correlation are mainly in high CNV variation regions (Fig. 4D), whose expression level are strongly depended on CNV level thus called DEG_{CNV}. While DEGs with low CNV-expression correlation are mainly in noCNV regions and low CNV variation regions (Fig. 4D), which may be caused by differences of gene regulation thus called DEG_{reg}. Indeed, majority of the DEG_{reg} did not show correlation with CNV level (Fig. 4B), consistent with the definition of DEG_{reg}. There are only 3 DEGs, namely, VEGFA, ARGLU1 and N4BP2L2, showing negative correlation between gene expression level and CNV level. Ribosome genes, belonging to housekeeping genes, were pronounced enriched in DEG_{CNV} comparing with that in DEG_{reg} (Fig. S7C). Furthermore, the genes in DEG_{CNV} were expressed higher and were detected in much higher fraction of cells (Figs. S7D–F), which is consistent with previous report that housekeeping genes expressed much higher than other genes [31].

Analyses of signaling enriched in DEG_{reg} showed that cytokine activity was detected in most patients (Fig. 4E), indicating the change of expression of cytokine signaling genes, without gene copy change, played important roles in shaping tumor subclones or tumor heterogeneity in many patients. On the other hand, analyses of signaling enriched in DEG_{CNV} showed that structural constituent of ribosome, cell adhesion molecule binding and cadherin binding in several patients (Fig. 4F), indicating ribosome and cell adhesion molecule binding genes played important roles in shaping tumor subclones or tumor heterogeneity in many patients. The differences of molecule binding and the cytokine signal pathways between tumor subclones should be two quite different mechanisms for shaping tumor subclones differences.

3.8. DEG_{CNV} and DEG_{reg} have different cell-cell communication pattern

Cell-cell communications between most cell types in tumor tissue are significantly higher than that in normal tissue (Fig. 5A and B; Fig. S8A). Among all cell-cell communication pairs, the 3 top increased communication pairs are endothelial-cancer/epithelial, stromal-cancer/epithelial, endothelial-stromal (Fig. 5B), suggesting cell-cell communication among the non-immune cell pairs increased the most. Further analyses showed that cell-cell communications between immune cell and non-immune cell increased the second highest, while cell-cell communications between immune cell and immune cell increased but the least (Fig. 5B; Fig. S8B). We further compared the cell-cell communication between DEG_{CNV} and DEG_{reg}. The cell-cell communication of DEG_{CNV} and DEG_{reg} are mainly cell-cell contact and secreted signaling, respectively (Fig. 5C and D), which is consistent with our aforementioned enrichment analysis that DEG_{CNV} and DEG_{reg} enriches in cell adhesion and cytokine signaling genes, respectively. We found both the cell-cell interaction weight of cytokine signaling and cell-cell interaction weight of adhesion signaling in tumor tissue are significantly higher than that in normal tissue (Fig. 5E). There are total 11 pathways significant changed between tumor tissue and normal tissue. Among those pathways, six of them shared between cytokine signaling and adhesion signaling, four of which are significant increased (FN1, VISFATIN, THBS and TGFb) and two of which are significant decreased in tumor tissue (PTN and CXCL) (Fig. 5F). Other pathways are all significantly increased in tumor tissue. MIF is the only cytokine signaling specific pathway that significant increase the interaction between immune cell with cancer/epithelial cell. Among the adhesion signaling specific pathway, COLLAGEN, LAMININ, MK and CDH1 among most cell type pairs are significantly higher in tumor tissue than in normal tissue (Fig. 5F; Fig. S8C). Furthermore, CDH1 is mainly released and received by cancer/epithelial cells, MIF is mainly received by macrophage and T cells, and COLLAGEN, LAMININ, MK mainly released by stromal and endothelial cell (Fig. 5G–I; Figs. S8D and E), which potentially indicate DEG_{reg} mainly affect immune system and DEG_{CNV} mainly communicate with microenvironment and cancer cells themselves.

4. Discussion

It is well known that cancer cells are quite different from their normal counterparts, but the *in vivo* differences between cancer cells and their normal counterparts have not been systematically studied at single cell resolution. In this study, we found the cancer cells expressed much more genes than their normal epithelial cell counterparts. We found cancer cells highly expressed genes associated with protein folding, neutrophil degranulation, cellular responses to stress, cytokine signaling, VEGF pathway and MYC pathway, indicating cancer cells have increased cell activity and enhanced of immune cell infiltration/maturation compared with their normal counterpart. The results also indicated that cancer cells expressed much more genes, and stronger cell-cell communication with other cell types than their normal counterparts. Furthermore, the stromal cells and endothelial cells in tumor tissue also expressed much more genes and had much stronger cell-cell communication with other cell types than their counterparts in normal tissue. Stromal cells and endothelial cells in tumor tissue highly expressed genes associated with extracellular matrix organization, response to wounding and angiogenesis. Therefore, cancer cells and stromal cells played quite different role in shaping the tumor microenvironment.

The dramatic differences between the cells in tumor tissue and their counterpart in normal tissue could be attributed to their different tissue microenvironments. Interestingly, we found the significant correlation between gene expression and its DNA copy number in cancer cells, indicating gene expression in cancer cells are more likely to depend on its DNA copy number. This observation is consistent with the recent studies reported the highly expressed genes in tumor usually have a lot DNA copies in the form of extrachromosomal circular DNA (eccDNA) [32,33]. Considering that most genes have two copies in mammalian cells while their expression levels vary greatly, thus the copy number of DNA should not be a strong factor affecting the difference of gene expression in normal cells. In short, the CNV-expression correlation should be a very specific and unique features of gene regulation in cancer cells.

Cancer cells usually were classified into different cell subsets based on gene expression matrix of scRNA-seq [8,23,34–38]. In this study, we showed CNVs inferred by scRNA-seq data were essentially consistent with that inferred by scWGS data, consistent with previous studies [19,27,28]. We inferred the tumor subclones using the CNVs inferred by scRNA-seq and found these tumor subclones have subclone specific SNVs, which further support the reliability of the inferred tumor subclones. We further identified the tumor subclones and inferred the lineage relationship of these tumor subclones in CRC#1, which support the accumulation of mutations in each tumor subclone. We further identified the DEGs among tumor subclones and found these DEGs had two populations according to CNV-expression correlation. Interestingly, DEG_{CNV}, mainly located in high CNV variation regions, has high CNV-expression correlation and enriches in housekeeping genes and adhesion signaling genes. While DEG_{reg}, mainly located in noCNV regions and low CNV variation regions, has no/low CNV-expression correlation and enriched in secretory signaling particularly cytokine signaling, which changed cell communication between cancer cell with immune cells. These results indicate that cytokine signaling genes showed tumor subclones difference but these differences were not determined by its DNA copy but by gene regulation. While housekeeping genes and adhesion associated genes that showed tumor subclones difference between tumor subclones are more likely to depend on change at its DNA level.

In conclusion, we characterized the features of cancer cells in tumor tissues by comparing with their counterparts in normal tissues at single cell resolution. We found that cancer cells expressed much more genes, had stronger metabolic activities and stronger cell-cell interaction with other cell types than their normal counterparts. In particular, we found the strong correlation between gene expression and its DNA copy number in cancer cells, potential indicating CNV-expression correlation is a very specific and unique features of gene regulation in cancer cells. We identified the tumor subclones using CNVs and inferred the lineage relationship of these tumor subclones in CRC#1, which support the accumulation of mutations in each tumor subclone. We found two DEGs populations among tumor subclones according to CNV-expression correlation, DEG_{CNV} and DEG_{reg}. DEG_{reg} has low CNV-expression correlation and enriched in cytokine signaling genes, while DEG_{CNV} has high CNV-expression correlation enriched in housekeeping genes and adhesion signaling genes.

5. Materials and methods

5.1. Collection of colorectal cancer samples

This study was approved by IRBs at Southern University of Science and Technology (SUSTech). Three patients diagnosed with colorectal cancer were enrolled in the First Affiliated Hospital, SUSTech. All patients signed informed consent form approved by the IRB of SUSTech. None of them received chemotherapy, radiation or immunotherapy prior to tumor resection.

5.2. Tissue dissociation and cell suspension

Tissue obtained immediately after tumor resection was used to prepare cell suspension. In brief, freshly resected tissue was washed with RPMI 1640 (Thermo Fisher Scientific) and minced with scissors. The tissues were digested with Human Tumor Dissociation Kit (Miltenyi Biotec) according to manufacturer's manual. Dissociated cells were passed through a 70 μm cell sieve and dead cells were removed using Dead Cell Removal Kit (Miltenyi, 130-090-101). After centrifugation with 300g at 4 °C for 5 min, the supernatant was discarded and the cells were washed with PBS. Then, 1 ml cold PBS containing 2% FBS was added into the tube to suspend the cells. The cell suspension was used for further analysis.

5.2.1. scRNA-seq

scRNA-seq libraries were prepared using Chromium Single Cell 3' v3 Reagent Kits (10x Genomics) following manufacturer's protocols. In brief, approximately 1.6×10^4 cells were loaded into each lane to produce the Gel Beads-in-Emulsions (GEMs), followed by reverse transcription, cDNA amplification and library construction. Qubit and Qsep100 were used to measure DNA concentration and fragment length of each library, respectively. The scRNA-seq libraries were sequenced on Illumina NovaSeq6000, instrument with pair-end sequencing and dual indexing according to recommended Chromium platform protocol.

5.3. Single-cell DNA sequencing (scWGS) and bulk DNA-seq

scWGS library were prepared using QIAseq FX Single Cell DNA Library Kit. Cell suspension of tumor tissue and normal tissue from patient CRC#1 were used for library preparation. In brief, 4ul PBS was added to each well of a 96-well plate. Single cells were sorted into individual well of the 96-well plate containing PBS and 3 μl Buffer D2. Cells were incubated for 10 min at 65 °C and the reaction was stopped by adding 3 μl of Stop Solution. For each amplification reaction, 40 μl master mix was added to the 10 μl reaction to denature the DNA, followed by incubation at 30 °C for 2 h and stopping by incubating at 65 °C for 3 min. In this way, we completed genomic DNA amplification from single cells. PCR-free library construction from the amplified gDNA proceeded following manufacturer's protocols. A total of about 120 cells from tumor tissue and 4 cells from normal tissue were used for construction of scWGS libraries. The bulk DNA-seq libraries (containing about 1000 cells) were prepared exactly following the scWGS library preparation protocol. We prepared bulk DNA-seq libraries with 2 tumor samples and 1 normal sample. The scWGS libraries and bulk DNA-seq libraries were sequenced on Illumina NovaSeq6000 system, setting with paired-end 150bp.

6. Data and samples

We generated scRNA-seq data of 9 samples from 3 colorectal cancer patients. We also generated scWGS data of 187 cells from tumor tissue and 5 cells from normal tissue from patient CRC#1, as well as bulk DNA-seq data of 2 tumor samples and bulk DNA-seq data from 2 normal samples from patient CRC#1. By integrating the 10x Genomics scRNA-seq data from Lee et al. [25], we obtained scRNA-seq data of 36 samples from 12 patients. Therefore, all scRNA-seq data were prepared using 10X Genomics Chromium, with Single Cell 3' Reagent Kits v2 for data in Lee, H.O et al. [25]. and Single Cell 3' Reagent Kits v3 for in-house data. The samples include 12 tumor tissues, 12 normal tissues, 9 tumor border tissues and 3 para-cancerous tissues.

6.1. scRNA-seq data pre-processing

Cell Ranger v3.1.0 was used to perform reads alignment (human reference genome GRCh38), barcode demultiplexing, transcripts assemblies and expression counting, similar to our previous studies [4,39,40]. The raw count matrices outputted by Cell Ranger were further processed by Seurat v3.2.1 package [41]. We filtered out the low-quality cells using the following criteria: 1) cell with <500 genes; 2) cell with <1000 UMIs; 3) cell with >50% mitochondrial genes; 4) potential doublets. We further removed plasma & red blood cell marker genes potentially caused by soup pollution.

6.2. scWGS data pre-processing

The reads of scWGS were mapped to human reference genome GRCh38 using BWA VN:0.7.17 [42]. CNVs in each cell were identified by CNVkit-0.9.6 [43]. In brief, the scWGS library with extremely duplicated probes were considered as low quality, thus cells will be filtered out if the average depth of the 10 deepest contigs is greater than $10^{2.5}$ fold the median depth of the genome. We obtained the normalized CNV after the probes (log2 ratio) in each cell were segmented and smoothed by winsorize and multipcf in copynumber v1.15.0 R package [44]. SNVs were identified by Mutect 2 in GATK 4.0 [45], with bulk DNA-seq data of normal tissue as control (Panel of Normal). SNVs that detected in <3 cells will be filtered out.

6.3. Dimension reduction and single-cell clustering

scRNA-seq data was normalized using NormalizeData function in Seurat [41] with default parameter. High-variable genes in scRNA-seq data were selected by FindVariableFeatures function in Seurat. We selected 2000 high-variable genes for dimension reduction and single-cell clustering. To reduce the impact of soup pollution, plasma and red blood cell marker genes (plasma cell marker genes include *IGHA1*, *IGHA2*, *IGHG1*, *IGHG2*, *IGHG3*, *IGHG4*, *IGHD*, *IGHE*, *IGHM*, *IGLC1*, *IGLC2*, *IGLC3*, *IGLC4*, *IGLC5*, *IGLC6*, *IGLC7*, *IGKC* and *JCHAIN*, red blood cell marker genes include *HBA1*, *HBA2*, *HBB*, *HBD*, *HBE1*, *HBG1*, *HBG2*, *HBM*, *HBQ1* and *HBZ*) were used to perform regression to get an unbiased z-score expression matrix by ScaleData function. Principal component analysis (PCA) was conducted on the unbiased z-score matrix. The top 20 principal components (PCs) were used to conduct Uniform Manifold Approximation and Projection (UMAP). Based on the 20 PCs, we clustered the cells using FindNeighbors and FindClusters with default parameter.

7. Annotation of cell clusters

Cell clusters identified based on scRNA-seq data were annotated by cell type specific markers: immune cell (*PTPRC*), T cell (*CD3D*, *CD3E*), macrophage (*CD68*, *CSF1R*), mast cell (*MS4A2*, *TPSAB1*), B cell (*CD79A*, *CD79B*), plasma cell (*IGHA1*, *IGHG1*), endothelial cell (*PECAM1*, *VWF*), stromal cell (*COL1A1*, *DCN*), epithelial cell (*EPCAM*), and cancer cell (*MMP7*, *CEACAM6*). Those marker genes were essentially consistent with parikh et al. [26].

8. Inference of CNVs to identify cancer cells using scRNA-seq data

Normalized expression matrix was used to infer CNV in each cell. In detail, the genes with mean normalized expression level >0.1 across all cells were sorted by their genomic position. The genes were grouped into bins by applying a sliding window of 51 genes with a 10-genes step in each chromosome, in which each bin was called a probe. The mean of all genes' z-score of normalized expression level in a probe was called the inferred raw CNV level. The raw inferred CNV level of a probe minus the average raw inferred CNV level in reference cells is the inferred CNV level of a probe.

By using stromal cells and endothelial cells as reference, the inferred CNV level matrix of all epithelial-like cells was obtained from normalized expression matrix in each patient. Then we define each cell's inferred CNV score as the sum of the squared inferred CNV level of each probe in each cell. By comparing the mean value of all cells' inferred CNV scores between cell clusters, we separated malignant cells and normal epithelial cells in each patient. Infercnv v1.12 [19] from Board Institute was also used to display patients' CNV level.

8.1. Identification of tumor subclone using scRNA-seq

To measure cell similarity at CNV level, the inferred CNV level matrix was processed for chromosome arm weighted PCA and unsupervised clustering. We apply hclust function with ward.D2 method on the top 10 PCs of inferred CNV level to perform cell clustering, and use cutree function to classify cells into tumor subclones. We set the number of tumor subclones (k) from 6 to 2 and chose the largest k as the number of tumor subclones when each tumor subclone has unique CNV pattern in each patient.

9. Comparison of CNVs inferred by scWGS and CNVs inferred by scRNA-seq

Because we have both scWGS data and scRNA-seq data from patient CRC#1, we can compare whether the CNVs inferred by scRNA-seq are consistent with that inferred by scWGS. We mapped the high-resolution CNV regions inferred by scWGS to CNV regions inferred by scRNA-seq data for integration analysis. The integrated CNV level is the average CNV level weighted by its original CNV region percent. In order to assign the cells from scWGS data to the tumor subclones inferred by scRNA-seq data, we calculated the Elucidate distance of the cell to the centroid of each tumor subclone based on CNV level. The cell was assigned to the tumor subclone with the shortest distance to the cell.

10. Pseudotime inferences of tumor subclones

We used UMAP_{CNV} in slingshot [46] to calculate pseudotime of tumor subclones in CRC#1. We use reduceDimension with DDRtree method in monocle [47] to calculate pseudotime and embedded trajectory.

11. Identification of SNVs in scWGS and scRNA-seq

We identified the SNVs in each cell from scWGS data following aforementioned method [45] (χ^2 test; BH adjusted p value < 0.05). We identified total 1403 SNVs on autosome and X chromosome. For scRNA-seq data, we extracted cancer cells and epithelial cells from raw bam files that output from Cell Ranger. We called the SNVs in scRNA-seq data following our previous studies [4,48]. We finally obtained 150 SNVs after removing recorded RNA editing sites in database RADAR [49] and DARNED [50]. If the frequency of alterative allele is $\geq$ 1%, or if there is one alterative allele when total number reads is $\leq$ 100, the cell will be marked as an alternative cell at a this SNV. A SNV detected in $>$10 cells and had at least 5% alternative cells will be regarded as a validated SNV. A SNV with significantly different alternative cell ratio between tumor subclones (Wilcoxon test) will be regarded as tumor subclone specific SNV.

12. Circos plot of CNV regions

CNV region was defined as a duplication region or a deletion region when the mean CNV level of this region across all cells is larger than 0.01 or less than -0.01 in a patient. The chromosomal regions without detected CNVs were called noCNV regions. All patients' ploidy CNV regions are shown on a circos plot [51].

12.1. Entropy, stemness score and inferred CNV score of each tumor subclone

We used LandSCENT (version 0.99.3) [52] to calculate each cell's entropy. And we used the mean expression level of genes in a stem associated gene set to calculate stemness score in each cell [53]. The entropy and inferred CNV score of each tumor subclone are calculated by averaging the entropies and inferred CNV scores of all the cells in the subclone, respectively.

12.2. Coefficient of variation (cell-cell variation) of gene expression

The coefficient of variation of gene expression was calculated for genes that expressed in more than half cells in cancer cells, epithelial cells, and each epithelial cell subtype which contain more than 100 cells and described in Ref. [26]. Undiff, Colonocytes, CT_colonocytes, Goblet, EECs and BEST4_OTOP2 among epithelial cell subtypes were defined by their marker genes, others were defined by high expressed gene in that cluster. There are 1440 genes expressed in more than 50% of cancer cells and at least 1 normal epithelial cell subsets. Wilcoxon Rank-Sum Test (paired samples) was used to test whether the coefficient of variations between two cell types/subtypes were significantly different.

13. Relationship between gene expression and its DNA copy number

The expression level of each gene in a cell was calculated by dividing the UMI count of this gene by the total UMI count in this cell. The relative expression level of each gene in cancer cells was normalized by its expression level in epithelial cells. In brief, we calculated the fold change of gene expression in cancer cells relative to epithelial cells and clamped it between 0.2 and 5 to avoid the strong influence of outliers. Then, we divided genes that expressed in $\geq$5% cancer cells or normal epithelial cells into noCNV, deletion and duplication regions in each patient, and calculated average relative expression level of those genes in each category.

Correlation coefficient between gene expression level and its CNV level in cancer cells were calculated. Variations of CNV level in each CNV region across all cells were calculated in each patient. CNV regions were classified into 2 groups according to whether the variation >0.001 or not. Pearson's correlation coefficient between the expression level of a gene and the CNV level of the gene located region cross all cells was calculated.

14. Calculation of intra-tumoral heterogeneity scores

The intra-tumoral heterogeneity scores were calculated following Wu et al., [24]. To calculate the intra-tumoral heterogeneity scores based on CNV (HetCNV), cell-cell similarities were calculated by Pearson's correlation coefficients between the inferred CNV scores in each patient. Interquartile range (IQR) of the distribution of cancer cell pairs' similarities was HetCNV. The intra-tumoral heterogeneity scores based on gene expression (HetGEX) were calculated similar to HetGEX but using normalized gene expression matrix.

15. Identification of DEGs between cell types or cell subpopulations

DEGs between cell types or cell subpopulations were identified by FindMarkers function in Seurat. We set logfc.threshold = log2(1.5) to identify the DEGs between cancer cells and epithelial cells, or DEGs between tumor associated endothelial cells (or stromal cells) and their normal counterparts. The subclone specific genes for each tumor subclone in each patient were identified using FindAllMarkers with parameter logfc.threshold = log2(1.5) and logfc.threshold = log2(1.8). Enrichment analysis of each gene set was conducted using Metascape [54].

16. Separating DEGs between tumor subclones into two populations

The distribution of correlation coefficient between gene expression level and CNV level revealed the existence of two populations of DEGs. DEGs among tumor subclones were separated into two population at the lowest point of the distribution ($r = 0.29$). To keep the number of genes with low CNV-expression correlation similar to that with high CNV-expression correlation, we set logfc.threshold = log2(1.5) for the genes with high CNV-expression correlation and logfc.threshold = log2(1.8) for the genes with low CNV-expression correlation. GO term enrichment in molecular function by enrichGO in clusterprofile [55].

17. Analyses of cell-cell communication

We inferred cell-cell communication signatures and communication strengths by using cellchat v1.4.0 [56]. There're 8 patients among all 12 patients used in cell-cell communication analysis, other 4 patients are filtered due to few cells in many cell types or cells concentrated in few cell types. There are 78 cytokine signaling related pathways and 73 adhesion signaling related pathways for cytokine signaling analysis and adhesion signaling analysis, respectively. We calculated the significantly changed pathways between two cell types by rankNet function. We calculated the communication distance of a pathway by rankSimilarity function between normal and tumor sample in each patient. We plotted interaction weight between cell types by netVisual_circle function.

17.1. Ethics statement

This study was approved by internal review board (IRB) at Southern University of Science and Technology (APPROVAL NUMBER: SUSTech-H-2018003).

18. Data availability statement

The sequence data have been deposited in the Genome Sequence Archive in BIG Data Center under accession number HRA004167 and HRA000086. Code used for analysis is available https://github.com/CaoWei-UM/Single-cell-CRC-pipeline-2023.

CRediT authorship contribution statement

Wei Cao: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Xuefei Wang:** Project administration, Methodology. **Kaiwen Luo:** Data curation. **Yang Li:** Data curation. **Jiahong Sun:** Project administration, Methodology. **Ruqing Fu:** Project administration, Methodology. **Qi Zhang:** Data curation. **Ni Hong:** Writing – review & editing, Supervision. **Edwin Cheung:** Writing – review & editing, Supervision. **Wenfei Jin:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by National Key R&D Program of China (2021YFF1200900), National Natural Science Foundation of China (32170646, 32370688), Key-Area Research and Development Program of Guangdong Province (2023B1111020006), Guangdong Basic and Applied Basic Research Foundation to N. H. (2023A1515011908), Shenzhen Innovation Committee of Science and Technology (JCYJ20220818100401003) and Shenzhen Science and Technology Program (KQTD20180411143432337). We thank all members from the Jin lab for the helpful discussion. The computational work was supported by Center for Computational Science and Engineering at SUSTech. We acknowledge the assistance of Core Facilities of SUSTech.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e28071.

References

[1] X. Fan, A.Y. Rudensky, Hallmarks of tissue-resident lymphocytes, Cell 164 (6) (2016) 1198–1211.

[2] E. Azizi, et al., Single-cell map of diverse immune phenotypes in the breast tumor microenvironment, Cell 174 (5) (2018) 1293–1308 e36.

[3] L. Zheng, et al., Pan-cancer single-cell landscape of tumor-infiltrating T cells, Science 374 (6574) (2021) abe6474.

[4] P. Qin, et al., Integrated decoding hematopoiesis and leukemogenesis using single-cell sequencing and its medical implication, Cell Discov 7 (1) (2021) 2.

[5] L. Zhang, et al., Lineage tracking reveals dynamic relationships of T cells in colorectal cancer, Nature 564 (7735) (2018) 268–272.

[6] J. Qi, et al., Single-cell and spatial analysis reveal interaction of FAP(+) fibroblasts and SPP1(+) macrophages in colorectal cancer, Nat. Commun. 13 (1) (2022) 1742.

[7] L. Zhang, et al., Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer, Cell 181 (2) (2020) 442–459 e29.

[8] H. Li, et al., Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors, Nat. Genet. 49 (5) (2017) 708–718.

[9] P.L. Bedard, et al., Tumour heterogeneity in the clinic, Nature 501 (7467) (2013) 355–364.

[10] I. Dagogo-Jack, A.T. Shaw, Tumour heterogeneity and resistance to cancer therapies, Nat. Rev. Clin. Oncol. 15 (2) (2018) 81–94.

[11] M.R. Junttila, F.J. de Sauvage, Influence of tumour micro-environment heterogeneity on therapeutic response, Nature 501 (7467) (2013) 346–354.

[12] I. Tirosh, et al., Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma, Nature 539 (7628) (2016) 309–313.

[13] W.R. Becker, 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. 54 (7) (2022) 985–995.

[14] N. Navin, et al., Tumour evolution inferred by single-cell sequencing, Nature 472 (7341) (2011) 90–94.

[15] C. Kim, et al., Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing, Cell 173 (4) (2018) 879–893 e13.

[16] A. Maynard, et al., Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing, Cell 182 (5) (2020) 1232–1251 e22.

[17] S.V. Puram, et al., Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer, Cell 171 (7) (2017) 1611–1624 e24.

[18] A.S. Venteicher, et al., Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq, Science 355 (6332) (2017).

[19] A.P. Patel, et al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science 344 (6190) (2014) 1396–1401.

[20] W.E. Hu, et al., HeLa-CCL2 cell heterogeneity studied by single-cell DNA and RNA sequencing, PLoS One 14 (12) (2019) e0225466.

[21] L.M. Richards, et al., Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity, Nat Cancer 2 (2) (2021) 157–173.

[22] K. Xu, et al., Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis, Oncogenesis 10 (10) (2021) 66.

[23] R. Xiang, et al., Identification of subtypes and a prognostic gene signature in colon cancer using cell differentiation trajectories, Front. Cell Dev. Biol. 9 (2021) 705537.

[24] F. Wu, et al., Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer, Nat. Commun. 12 (1) (2021) 2540.

[25] H.O. Lee, et al., Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer, Nat. Genet. 52 (6) (2020) 594–603.

[26] K. Parikh, et al., Colonic epithelial cell diversity in health and inflammatory bowel disease, Nature 567 (7746) (2019) 49–55.

[27] R. Gao, et al., Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes, Nat. Biotechnol. 39 (5) (2021) 599–608.

[28] J. Fan, et al., Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data, Genome Res. 28 (8) (2018) 1217–1227.

[29] E. Sjoberg, et al., Expression of the chemokine CXCL14 in the tumour stroma is an independent marker of survival in breast cancer, Br. J. Cancer 114 (10) (2016) 1117–1124.

[30] Z. Li, et al., PLCG2 as a potential indicator of tumor microenvironment remodeling in soft tissue sarcoma, Medicine (Baltim.) 100 (11) (2021) e25008.

[31] W. Jin, et al., A systematic characterization of genes underlying both complex and Mendelian diseases, Hum. Mol. Genet. 21 (7) (2012) 1611–1624.

[32] K.M. Turner, et al., Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity, Nature 543 (7643) (2017) 122–125.

[33] S. Wu, et al., Circular ecDNA promotes accessible chromatin and high oncogene expression, Nature 575 (7784) (2019) 699–703.

[34] X. Sun, et al., Colon cancer-related genes identification and function study based on single-cell multi-omics integration, Front. Cell Dev. Biol. 9 (2021) 789587.

[35] M.K. Zowada, et al., Functional states in tumor-initiating cell differentiation in human colorectal cancer, Cancers 13 (5) (2021).

[36] S. Chowdhury, et al., Implications of intratumor heterogeneity on consensus molecular subtype (CMS) in colorectal cancer, Cancers 13 (19) (2021).

[37] R.Q. Wang, et al., Single-cell RNA sequencing analysis of the heterogeneity in gene regulatory networks in colorectal cancer, Front. Cell Dev. Biol. 9 (2021) 765578.

[38] A. Sacchetti, et al., Phenotypic plasticity underlies local invasion and distant metastasis in colon cancer, Elife 10 (2021).

[39] B. Zhou, W. Jin, Visualization of single cell RNA-seq data using t-SNE in R, Methods Mol. Biol. 2117 (2020) 159–167.

[40] X. Wang, et al., Reinvestigation of classic T cell subsets and identification of novel cell subpopulations by single-cell RNA sequencing, J. Immunol. 208 (2) (2022) 396–406.

[41] T. Stuart, et al., Comprehensive integration of single-cell data, Cell 177 (7) (2019) 1888–1902 e21.

[42] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (5) (2010) 589–595.

[43] E. Talevich, et al., CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing, PLoS Comput. Biol. 12 (4) (2016) e1004873.

[44] G. Nilsen, et al., Copynumber: efficient algorithms for single- and multi-track copy number segmentation, BMC Genom. 13 (2012) 591.

[45] BD, V.d.A.G.O.C., Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, first ed., O'Reilly Media, 2020.

[46] K. Street, et al., Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics, BMC Genom. 19 (1) (2018) 477.

[47] X. Qiu, et al., Reversed graph embedding resolves complex single-cell trajectories, Nat. Methods 14 (10) (2017) 979–982.

[48] R. Fu, et al., A comprehensive characterization of monoallelic expression during hematopoiesis and leukemogenesis via single-cell RNA-sequencing, Front. Cell Dev. Biol. 9 (2021) 702897.

[49] G. Ramaswami, J.B. Li, RADAR: a rigorously annotated database of A-to-I RNA editing, Nucleic Acids Res. 42 (2014) D109–D113 (Database issue.

[50] A. Kiran, P.V. Baranov, DARNED: a DAtabase of RNa EDiting in humans, Bioinformatics 26 (14) (2010) 1772–1776.

[51] M. Krzywinski, et al., Circos: an information aesthetic for comparative genomics, Genome Res. 19 (9) (2009) 1639–1645.

[52] A.E. Teschendorff, T. Enver, Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome, Nat. Commun. 8 (2017) 15599.

[53] A. Miranda, et al., Cancer stemness, intratumoral heterogeneity, and immune response across cancers, Proc Natl Acad Sci U S A 116 (18) (2019) 9020–9029.

[54] Y. Zhou, et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets, Nat. Commun. 10 (1) (2019) 1523.

[55] T. Wu, et al., clusterProfiler 4.0: a universal enrichment tool for interpreting omics data, Innovation 2 (3) (2021) 100141.

[56] S. Jin, et al., Inference and analysis of cell-cell communication using CellChat, Nat. Commun. 12 (1) (2021) 1088.