Single-cell DNA and RNA sequencing reveals the dynamics of intra-tumor heterogeneity in a colorectal cancer model

Short title: ITH dynamics by single-cell sequencing

AUTHORS

Hanako Ono[1], Yasuhito Arai[2], Eisaku Furukawa[1], Daichi Narushima[1], Tetsuya Matsuura[3], Hiromi Nakamura[2], Daisuke Shiokawa[4], Momoko Nagai[1], Toshio Imai[3], Koshi Mimori[5], Koji Okamoto[6], Yoshitaka Hippo[3,7], Tatsuhiro Shibata[2,8], and Mamoru Kato[1,*]

AFFILIATIONS

[1] Department of Bioinformatics, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[2] Division of Cancer Genomics, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[3] Department of Animal Experimentation, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[4] Division of Cancer Differentiation, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[5] Department of Surgery, Kyushu University Beppu Hospital, 101 Hasamamachiidaigaoka, Yufu, Oita 879-5593, Japan

[6] Division of Cancer Differentiation, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[7] Division of Biochemistry and Molecular Carcinogenesis, Chiba Cancer Center Research Institute, 666-2 Nitona-cho, Chiba Chuo-ku, Chiba 260-8717, Japan

[8] Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shiroganedai, Minato-ku, Tokyo 108-8639, Japan

* Corresponding author

Tel: +81-3-3542-2511

E-mail: mamkato@ncc.go.jp

Abstract

Intra-tumor heterogeneity (ITH) encompasses cellular differences in tumors and is related to clinical outcomes, such as drug resistance. However, little is known about the dynamics of ITH, owing to the lack of time-series analysis at the single-cell level. We performed single-cell exome and transcriptome sequencing of 200 cells and investigated how ITH is generated from one single cell in a mouse colorectal cancer model. The ITH of the transcriptome increased after transplantation from cultured organoids, while that of the exome decreased. The RNA ITH increase was due to the emergence of new transcriptional subpopulations. In contrast to the initial cells expressing mesenchymal-marker genes, new subpopulations repressed these genes at transplantation, suggesting that the birth of transcriptional subpopulations without substantial genetic changes is associated with mesenchymal-epithelial transformation at metastasis. Analyses of colorectal cancer data from The Cancer Genome Atlas, revealed a higher proportion of patients with metastatic tumor among human subjects with expression patterns similar to those of mouse cell subpopulation. This study revealed an evolutionary pattern of single-cell RNA and DNA changes in tumor progression, giving clinical insights into the mesenchymal-epithelial transformation of tumor cells and subclasses of colorectal cancer.

Author summary

“Intra-tumor heterogeneity (ITH)” is one of the root causes of cancer malignancy, including drug resistance; however, little is known about the time-dependence of ITH. To investigate how ITH is generated, we combined single cell DNA and RNA sequencing technologies with a mouse colorectal cancer model, ideal for time-series analysis. Our results suggested that mouse cancer cells, with sufficient mutations, adapted to the drastic environmental changes of allograft into a mouse. Transcriptional and genetic ITH increased and somewhat decreased, respectively. New transcriptional subpopulations emerged, showing mesenchymal-epithelial transformation. Using human colorectal cancer data, we found a remarkable trend of metastasis in a fraction of human patients whose expression patterns were similar to those of the mouse-cell subpopulations.

Introduction

It is well established that cancer is pathologically composed of different types of cells [1]; however, intra-tumor heterogeneity (ITH) has only been recently addressed at the genomic level [2]. ITH is clinically important. For example, elevated copy-number heterogeneity is related to an increased risk of recurrence or death in non-small-cell lung cancer [3]. High levels of ITH ultimately provide the seeds for the emergence of anti-cancer drug resistance [4]. High levels of genetically-characterized heterogeneity in Barrett's esophagus are associated with incidence of esophageal adenocarcinoma [5].

ITH essentially stands for the cellular differences in tumor tissue arising from genetic changes, called clonal evolution, or non-genetic changes, such as cancer stem cells and simple transcriptional responses to the environment. In clonal evolution, as in Darwinian evolution, cancer cells with advantageous genetic mutations evolve into different types of cancer cells [6]. In contrast, cancer stem cells, like normal stem cells, produce a variety of differentiated daughter cells that constitute phenotypically distinct cancer cells without genetic differences through epi-genetic and the resultant transcriptional mechanisms [7, 8].

A flood of studies have addressed ITH through the variant allele frequencies (VAFs) of tumor cells in bulk, which are calculated from sequence reads with variants identified through next-generation sequencing (reviewed in [2, 9]). In this bulk-cell sequencing approach, the presence of minor clones is often reflected on lower VAFs than the VAF of the major clone [10]. However, this bulk-cell DNA sequencing approach is limited in revealing genetic ITH because it only infers the presence of clones, not directly observing individual cells. In addition, the bulk-cell approach is generally not suitable to resolve transcriptomic ITH, where transcript mixtures from different cells are sequenced.

Single-cell sequencing is a powerful technology for investigating ITH by identifying genomic alterations and distinct transcriptomic states in single tumor cells [11-19]. For example, in clinical samples of glioblastoma, single-cell RNA sequencing showed that individual tumor cells vary in terms of their degree of stemness-related gene expression from extremely stem-like to differentiated states [13]. Additionally, the existence of cancer stem cells that continuously differentiate into astrocyte- and oligodendrocyte-like cells has been demonstrated in oligodendrogliomas by single-cell RNA sequencing

[14]. Single-cell DNA sequencing has also been applied to breast cancer samples to evaluate ITH originating in genomic DNA, leading to the suggestion of stepwise/sweepstake or gradual evolution of cancer cells from single nucleotide variation (SNV) data, respectively [11, 12, 20]. However, these types of ITH and their respective evolutionary mechanisms are based on snapshot data at one-time point. Furthermore, either RNA or DNA was solely examined. It is necessary to address both RNA and DNA over time for the full elucidation of tumor evolutionary dynamics associated with ITH.

Mouse models are more useful than human clinical samples for examining changes in genomic and transcriptomic states over time. In a breast tumor xenograft mouse model, single-cell DNA sequencing of serially passaged samples identified tumor cell subpopulations and suggested that tumor cells in the same initial state followed the same evolutionary trajectory [21]. In the present study, we employed a modified version of the mouse colorectal cancer model that we previously established [22] and sequenced both single-cell DNA and RNA. We thus investigated how ITH based on the exome and transcriptome changes over time at the single-cell level.

Results

Colorectal cancer mouse model

The colorectal cancer mouse model was established by knocking down *APC* expression in normal epithelial cells taken from mouse intestinal crypts using short hairpin RNA (shAPC; **Fig. 1A**) [22]. In the previous system, we used bulk cells from a tissue for culture; however, in this study, we cultured organoids from *one single cell* so that heterogeneity observed in these cultures could not be confused with heterogeneity originating from the knock-down efficiency or intestinal crypts [23].

We grew organoids for a period of five months so that a single cell having only artificial *APC* intervention could naturally obtain mutations to transform into tumor cells. Cultured cells were subcutaneously transplanted into a nude mouse. One month after transplantation, the mouse was sacrificed, and the tumor was harvested; half of the tumor tissue was re-cultured in our three-dimensional (3D) culture system for one month. Using half-samples preserved the same genetic lineage over time. The process was repeated once more. We sequenced single-cell RNA and DNA separately taken from the different single cells of multiple organoids, which descended from one single cell.

Hematoxylin-eosin (HE) staining revealed that subcutaneously transplanted organoids formed tumors consisting of both glandular and non-glandular structures (HE *a* and *b* in **Fig. 1A**). Glandular components in HE *a* were mainly lined with single-layered epithelia, while those in HE *b* were characterized by increased multi-layered regions, loss of cellular polarity, and nuclear enlargement. Non-glandular components had a stromal/medullary structure consisting of spindle-shaped or round to polygonal cells, were characteristically gelatinous/fibrous, and had an abundance of fibrous stroma.

The *APC* expression was decreased in the *APC* knockdown samples according to bulk-cell RNA sequencing (**Fig. 1B**). Out of the 31 significantly mutated genes (excluding *TTN*) defined by The Cancer Genome Atlas (TCGA) colorectal cancer study [24], we found two mutations in *KRAS* and *TP53* by bulk-cell DNA sequencing in our model (**Fig. 1C**) , though the *KRAS* mutation was located outside of, but close (9 bps) to, an exon and the position was evolutionary conserved as much as exons (S1 Appendix: **Figure S1**). The *KRAS* mutation occupied only a small fraction (2.5%) of the population at T0.5 but increased to 46.4% at T3. Additionally, we found nonsynonymous mutations in six, *CLTC*, *LRP1B*, *ALK*, *GRIN2A*,

MSH2, and *SALL4* out of the cancer-related genes in COSMIC Gene Census [25] (**Fig. 1C**). It seems that the substitution of clones occurred between T0.5 and T1.

Single-cell transcriptome analysis

We checked various indices of single-cell transcriptome data to filter 42, 42, and 51 cells out of the 50 T1, 43 T2, and 52 T3 cells, respectively (S1 Appendix: **Figure S2**). The median (± inter-quartile range) number of mapped reads, mapping rate, and number of expressed genes across selected cells were 6.2×10^6 ($\pm 2.0 \times 10^6$), 61.9% (± 5.39%), and 3814 (± 889.5), respectively. There was a strong correlation between gene expression levels in the bulk sequencing data and average expression levels across single cells (S1 Appendix: **Figure S2**; $R^2 = 0.9$).

A principal component analysis (PCA) plot of cells based on expression levels revealed increased heterogeneity from T1 to T2 (**Fig. 2A**). This was quantitatively confirmed by the diversity index (distance from the centroid in the PCA space) (**Fig. 2B**). In the plot, T2 and T3 cells partly overlapped but were separate from T1 cells. We identified genes whose expression levels varied greatly across cells at each time point; that is, these genes had high corrected coefficient of variation (c*CV*) values (S1 Appendix: **Figure S3**), and were thus referred to as highly variable genes. There were eight, 14, and 16 highly variable genes at T1, T2, and T3, respectively, reflecting an increase in variability from T1 to T2.

A cluster analysis of highly variable genes identified three gene groups (S1 Appendix: **Figure S4**); expression levels were correlated within two of the groups, but not within the third group. Gene set enrichment analysis showed that one of the correlated groups was associated with negative regulation of keratinocyte differentiation (referred to as Anti-Epithelial genes) ($P = 3.80 \times 10^{-3}$), whereas the other was associated with positive regulation of cGMP and guanylate cyclase (GC) activity (referred to as cGMP/GC genes) ($P = 1.30 \times 10^{-3}$), which are known to be associated with negative regulation of β-catenin signaling and matrix metalloproteinase activity in colorectal cancer [26, 27].

A heatmap generated from the cluster analysis revealed that T1 cells were relatively homogenous and formed one group that highly expressed Anti-Epithelial genes but showed negligible expression of cGMP/GC genes (**Fig. 2C**). This group was therefore termed Anti-Epithelial. In addition to an Anti-

Epithelial cell group, two new groups appeared at T2: one showing the opposite pattern, repression of Anti-Epithelial and activation of cGMP/GC gene expression, referred to as the cGMP/GC cell group; the other showed repression of both Anti-Epithelial and cGMP/GC genes. Notably, as shown in the heat map, bulk-cell sequencing analysis alone could not have identified these cell groups, where their distinct expression patterns were offset by bulk-cell expression levels (labeled as T1, T2, and T3 bulk in **Fig. 2C**).

T3 cells showed similar grouping to T2 cells. In a PCA plot based on highly variable gene expression (**Fig. 2D**), cells of the Anti-Epithelial group seemed close together across all time points, but seemed to form two groups—i.e., T1 main (referred to as T1 main) and T1/T2/T3 mixture (T1+T2+T3). The cGMP/GC and other groups seemed close together and contained T2 and T3 only (T2+T3 only). This grouping based on PCA will be discussed later in association with exome analysis.

Marker gene expression

We examined the expression of several types of marker genes. We first looked at proliferation/cell cycle markers (S1 Appendix: **Figure S5**) and performed PCA to summarize the multiple expression levels (**Fig. 3A**). Remarkably, most cells in the Anti-Epithelial group at T1 expressed high levels of proliferation- and cell cycle-related genes according to the PCA loading plot. In contrast, nearly all cells in the unnamed group at T3 showed a downregulation of the marker genes, so we termed the cell group Dormant. At T2, about half of the cells showed a downregulation of the proliferation/cell-cycle genes.

We next examined epithelial and mesenchymal markers (S1 Appendix: **Figure S5**). A PCA plot of the markers showed that expression of mesenchymal cell-related genes decreased with time (T2 and T3), with cells forming two groups (**Fig. 3B**): one (upper left) overlapping with some T1 Anti-Epithelial cells with decreased mesenchymal N-cadherin (*CDH2*) and fibronectin (*FN1*) levels; the other (middle right) group was composed only of T3 cells with decreased mesenchymal vimentin (*VIM*) and increased epithelial E-cadherin (*CDH1*) levels. These results suggest a similarity between the processes occurring at T2 and T3 and mesenchymal-epithelial transition (MET).

Stem cell and differentiation markers showed that over time, cells generally expressed more differentiation than stem cell markers (**Fig. 3C**; S1 Appendix: **Figure S5**). Nevertheless, a remarkable

variation across individual cells was also observed; for example, many T3 cells tended to express more differentiation markers, while others tended to express more stem cell markers. Among the markers for crypt base stem cells, *SOX9* appeared to be the most influential; *LGR5*, *OLFM4*, and *MSI1* were not substantially expressed. It seems that with time, cells differentiated into those expressing a marker for absorption cells (*KRT20*) and those for secretion cells (*MUC2*) in the digestive tract.

There was no remarkable change in the expression of drug efflux genes [28, 29] at any time point (**Fig. 2C**), although *ABCB1* expression was slightly lower in the T3 Dormant group (S1 Appendix: **Figure S5**) and *ABCE1* was downregulated at T2 and T3. There was variable expression of glycolysis-related gene *PDK1* [29] across all cells, irrespective of groups (**Fig. 2C**; S1 Appendix: **Figure S5**).

Single-cell exome analysis

Based on several indices from single-cell exome sequencing (S1 Appendix: **Figure S6**), we selected 21, 23, and 23 cells out of the 23 T1, 24 T2, and 24 T3 cells for analysis. On average (expressed as the median [± inter quartile range] across selected cells), the number of mapped reads was 1.2×10^8 ($\pm 2.2 \times 10^7$), mapping rate was 76.6% (± 4.9%), coverage with > 0 depth regions was 76.9% (± 34.2%), average depth was 43 (± 34.5), Gini coefficient was 0.85 (± 0.15), allelic drop-out (ADO) rate was 47.0 (± 36.1), and number of called SNVs was 462 (± 313.5). The false positive rate in single-cell sequencing was estimated to be $0.1–1.1 \times 10^{-7}$ per chromosomal site, based on normal intestinal tract tissue samples from two mice and four single cells obtained from one of these samples (**S1 Appendix: Supplementary Results**). We compared the fractions of single cells with SNVs to the variant allele frequencies (VAFs) of the bulk-cell sequencing; in theory, the single-cell fractions should be equal to half of the VAFs. We confirmed a good concordance between these variables, although the cell fractions were slightly lower than those expected from bulk VAFs (S1 Appendix: **Figure S6**).

We first examined the bulk-cell sequence data. The T0.5 tissue had much fewer SNVs than the later stages (**Fig. 4A**), which suggests that DNA heterogeneity only weakly appeared soon (1.5 months) after culture initiation. The numbers of SNVs increased markedly from T0.5 to T1, a five-month period (**Fig. 4A**). Although these numbers decreased slightly at T2 before recovering at T3, they were all mostly

saturated at T1, T2, and T3. Thus, new SNVs were largely generated from T0.5 to T1, and most of these SNVs remained in the genome after T1 at the bulk-cell level (**Fig. 4B**).

We then used single-cell sequencing data to draw a multi-dimensional scaling (MDS) plot based on single-cell SNVs at polymorphic SNV sites (defined as SNVs with 10–35% bulk VAFs) (**Fig. 4C**). T1 cells showed the greatest genetic divergence, whereas T2 and T3 cells showed convergence. This decrease in diversity was confirmed by a statistical significance of the diversity index (average distance from the centroid), where the bias due to ADO rates was taken into account by a bootstrapping test (**Fig. 4D**). Interestingly, this diversity tendency was the complete opposite of the transcriptomic pattern (**Fig. 2A**, **B**). Although transitional, cells can be classified into three groups composed of T1 cells only (T1 main); T1, T2, and T3 cells (T1+T2+T3); and T2 and T3 cells (T2+T3 only) (**Fig. 4C**).

Association with human cancer

We used TCGA colorectal cancer data [24] to identify human samples with gene expression patterns similar to the groups of mouse single cells. The human sample clusters were separated from the mouse cell groups, but we found 94 (38.5%), 42 (17.2%), and 13 (5.3%) samples out of the 244 TCGA samples that were respectively close to the Anti-Epithelial, cGMP/GC, and Dormant mouse cell groups (**Fig. 5A**). TCGA Anti-Epithelial samples showed enhanced *REG* and repressed cGMP/GC gene expression; TCGA cGMP/GC samples showed the opposite pattern; and TCGA Dormant samples had both repressed *REG* and GC-related gene expression (**Fig. 5B**). TCGA cGMP/GC and TCGA Dormant samples tended to be more closely associated with metastasis than those with patterns similar to the Anti-Epithelial group (two-sided Fisher's exact test $P = 0.04$; **Fig. 5C**). We determined that our mouse tumor cells were molecularly similar to cells of human colon adenocarcinoma classified as high microsatellite instability type (S1 Appendix: **Figure S7**; S1 Appendix; **Figure S8**; S1 Appendix: **Supplementary Results**).

Discussion

Our results suggest a scenario in which, once cancer cells accumulate a sufficient number of genetic alterations (SNVs/indels), they can adapt to drastic environmental changes, such as the shift from a 3D culture to a live mouse. Only by altering their transcriptional profiles, cancer cells generate new subpopulations of cells with ever increasing transcriptional heterogeneity. In turn, genetic heterogeneity decreases, possibly as a result of microscale natural selection that occurs during the environmental transition. Though expected, it is nonetheless surprising to see that this diversity was indeed generated from *one single cell*.

Because T1 cells express mesenchymal genes, they are considered as a late stage of tumor development, when typically they move out from the niches or microenvironment of intestinal crypts [30]. For example, we did not observe the expression of *LGR5,* a stem cell marker and a tumor suppressor that delimits stem cell expansion in the niches; ablation of *LGR5* reduces cell-cell adhesion and induces invasion and metastasis [31-33]. Our observation that cells lose their mesenchymal-like phenotype and acquire epithelial-like characteristics after subcutaneous transplantation may be analogized to MET during metastasis.

There is a possibility that the observed subpopulations were derived from distinct genetic lineages. The RNA cell categories of T1 main (composed of the Anti-Epithelial cell group), T1+T2+T3 (Anti-Epithelial cell group), and T2+T3 only (cGMP/GC and Dormant cell groups) in **Fig. 2D** correspond to the DNA cell categories of T1 main, T1+T2+T3, and T2+T3 only, respectively, in **Fig. 4C**. This suggests that the two subpopulations (cGMP/GC and Dormant) that emerged after transplantation were genetically distinct from the initial Anti-Epithelial group and that transcriptional differences between the cGMP/GC or Dormant groups and the Anti-Epithelial group were due to their genetic differences, though simultaneous single-cell sequencing of both DNA and RNA from the same cells([34], [35]) is required for further clarification.

Classically, cells that generate a tumor by subcutaneous transplantation are called tumor-initiating cells or cancer stem cells (CSCs) [29]. It is thought that differentiated cells die while CSCs can survive at the start of subcutaneous transplantation and 3D culture; then, CSCs re-generate differentiated cells. In our

transplantation approach, at the cell-population level we observed decreased expression of stem-cell markers and increased expression of differentiation markers over time, with varying degrees of expression across single cells. This suggests that once cells experience subcutaneous environments, CSCs that more efficiently generate differentiated cells may survive and prevail. Alternatively, contrary to naïve expectations, tumor cells expressing differentiation markers may also survive at the start of the transplantation and 3D culturing, supporting the idea that differentiated cells that experience MET can colonize and are not necessarily generated from CSCs.

Clinically, the proportion of TCGA samples with metastasis in the Dormant and cGMP/GC groups was higher than in the TCGA Anti-Epithelial group (**Fig. 5C**). This is probably because the appearance of the former two subgroups, and resultant increased ITH, may be a sign for later stages of tumor progression. This is surprising because mouse expression patterns decomposed by single-cell sequencing may provide us with clinical significance, though further investigations including single-cell sequencing of the TCGA samples will be necessary to clarify a relationship between the tumor progression and metastasis.

Recently, more fine-scale single-cell sequencing technology, such as 10X/Drop-Seq, has emerged for RNA-seq, enabling researchers to capture tens of thousands of cells. Although the number of cells we addressed was relatively small compared to that technology, we believe that we successfully captured a major part of the heterogeneity constructed by cell clones, constituting as small as ~2% (an inverse number of 42, 42, and 51 cells at T1, T2, and T3) of the tumor cell population. Nevertheless, to investigate rarer cells, for example, CSCs related to the above issue, 10X/Drop-Seq will be needed.

We demonstrated that time-series ITH analysis by single-cell DNA and RNA sequencing for a mouse model is able to provide clinical insights, such as finding associations with MET and metastasis, and the birth of transcriptional subpopulations of cells with sufficient genetic alterations at a drastic micro-environmental change. It will be crucial to examine how such genetic changes accumulate in the earlier stages of tumorigenesis and how transcriptional subpopulations develop to increase malignancy in the further later stages of tumor progression.

Materials and Methods

Ethics approval and consent to practice

Animal studies were carried out according to the Guideline for Animal Experiments established by the Committee for Ethics in Animal Experimentation of the National Cancer Center (T10-033-M05), which meets the ethical standards required by law and guidelines for animal experimentation in Japan. All sacrificed mice were anesthetized by inhalation of isoflurane. And cervical dislocation was used as a euthanasia method.

Organoid culture of small intestinal cells and lentiviral transduction

C57BL/6J mice and BALB/cAnu/nu immune-deficient nude mice were purchased from CLEA Japan (Tokyo, Japan). The small intestine was harvested from wild-type male C57BL/6J mice at 3–5 weeks of age. Crypts were purified and dissociated into single cells, which were then put into serum-free Matrigel-based organoid culture as previously described [22, 36]. Transduced organoids were maintained in culture medium lacking R-spondin 1. Single cell-derived shAPC-transduced organoids were obtained by limiting dilution of dissociated organoids in a 96-well plate. Organoids composed of 5×10^5 cells were mixed with 200 µl of Matrigel and injected into one side of the dorsal skin of nude mice, while uninjected cells were maintained in 3D cultures for later use.

Analysis of subcutaneous tumors in nude mice

Palpable tumors from the injection sites were harvested for histological examination or cell culture. Half of the subcutaneous tumors were formalin-fixed, paraffin-embedded, and sectioned at a thickness of 5 µm, followed by HE staining to assess histological features. The other half of the tumors were minced and digested to recover cells as described previously (22), then seeded onto solidified Matrigel to resume organoid culture. We defined the time points as follows; before the first transplantation was time point T1, and two time points following the first and second transplantations were T2 and T3, respectively (**Fig. 1A**).

Single-cell transcriptome and exome sequencing

Cultured mouse organoids derived from a single cell were harvested and treated with Accumax (Innovative Cell Technologies, AM105) to generate a single-cell suspension. To capture cells and extract RNA or DNA from a single cell, the cell suspensions (4.4×10^5 cells/ml) were loaded on a C1 Single Cell Auto Prep System (Fluidigm, C1) with medium-sized (10–17 μm) microfluidic chips for 96 cells. Approximately 1300 cells were applied to each chip. Captured cells were imaged on a BZ-9000 digital microscope (Keyence, BZ-9000) and a visual inspection was performed to determine whether a single cell was captured in each well of the chip. Capture efficiency for a single cell was determined as 71–82%.

For single-cell whole transcriptome (RNA) sequencing, captured cells were lysed on the chip using a C1 Single-Cell Auto Prep Reagent Kit for mRNA Seq (Fluidigm, 100-6201). Full-length cDNA fragments were transcribed and amplified from poly-A RNA in each single cell using the SMARTer Ultra Low RNA kit (Takara Bio, 634832). Tagmentation of cDNA was performed and sequencing libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina, FC-131-1096) according to the manufacturer's protocol. Up to 52 independent single-cell RNA-seq libraries were prepared for sequencing.

For single-cell DNA sequencing, genomic DNA was prepared from single cells using the C1 Single-cell Auto Prep Reagent Kit for DNA Seq (Fluidigm, 100-7357) and whole genome amplification was achieved by multiple displacement amplification with Phi29 DNA polymerase and the Illustra GenomiPhi v.2 kit (GE Healthcare, 25660032). Amplified genomic DNA (70 ng) was used to generate exome sequence libraries using the Hyper Prep kit (Kapa Biosystems, KK8504), SureSelect Target Enrichment kit (Agilent Technologies, 931171), and SureSelect XT Mouse All Exon v.1 probe (Agilent Technologies, 5190-4642).

Bulk-cell transcriptome and exome sequencing

Among the cells that were not used for single-cell capture with the C1 system, suspensions of about 200 cells were subjected to whole transcriptome (RNA) sequencing for bulk-cell RNA-seq (T1, T2, and T3 samples). The sequencing libraries were prepared using the same reagents as the single cell RNA-seq. As

358 control bulk cells, normal intestinal crypt epithelial cells from two wild-type mice of the same strain were
359 grown in the 3D culture system for seven days, then harvested and lysed for total RNA preparation using
360 the miRNAeasy Mini kit (Qiagen, 217004). RNA-seq libraries for control bulk RNA were generated using
361 the SureSelect Strand-specific kit (Agilent Technologies, G9691B). Bulk DNA from over 1×10^5 cells
362 was obtained from the cell culture (T0.5 sample, 1.5 months after culture initiation) and the remaining
363 cells in single-cell capture (T1, T2, and T3 samples) using the QIAamp DNA Mini kit (Qiagen, 51304),
364 and 500 ng of DNA were used to construct exome sequencing libraries with the same reagents as the
365 single cell DNA-seq.
366
367 **Sequencing**
368 All the sequencing libraries were subjected to paired-end sequencing of 101-bp fragments on the
369 HiSeq2500 DNA sequencer (Illumina, SY–401–2501).
370
371 **Transcripts per kilobase million (TPM) calculation for single and bulk cells**
372 The procedure for calculating TPM is summarized in S1 Appendix: **Figure S9.** Specifically, sequence
373 reads were quality-filtered and trimmed using PrinSeq [37], and then used as input for quality-check by
374 FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We used the following parameters:
375 --min_len 30 (removing reads ≤30 bases); --min_qual_mean 20 (average read quality ≤ 20); --
376 trim_tail_right 5, --trim_tail_left 5 (trim bases if the 3' and 5' end poly A/Ts are ≥ five bases); and --
377 trim_qual_right 20, --trim_qual_left 20 (trim 3' or 5' end for read quality ≤ 20). Paired-end reads were
378 mapped to the University of California Santa Cruz mouse genome sequence (mm10) using Bowtie2 [38]
379 built in RSEM [39]. Expression levels (in TPM) were estimated by RSEM using the command rsem-
380 calculate-expression with the parameters --estimate-rspd, --paired-end, --bowtie2, -p 30, and --ci-memory
381 10192. We removed eight T1 cell samples due to an excessive number of genes (≥ 5,200) with TPM ≥ 10
382 or with too few unique mapping reads ($< 2.2 \times 10^6$). We also removed two samples with unique mapping
383 rates that were too low (< 20%) and discarded genes with low expression levels (≤ 10 TPM) across all cell
384 samples, leaving 14,258 out of 32,732 genes for analysis.

385

386 **Detection of highly variable genes**

387 To detect genes with variable expression levels across cells, we defined highly variable genes according to
388 the *CV*, corrected in the locally weighted scatterplot smoothing (LOWESS) method using the "lowess"
389 function in R. To fit a single LOWESS curve across all ranges, we divided average expression level data
390 into three ranges: < 4, 4–8.5, and > 8.5. c*CV* values were yielded by dividing *CV* values by the value of
391 the upper variability band (± 1.96 times the standard deviation) of smoothed curve estimated using
392 "loess.sd" in the "msir" package. Because of the large bias in original *CV* values against low average
393 expression levels, only those with c*CV* values > 1.3 and high average expression levels ($\log_2$ [TPM+1] ≥
394 4) were defined as highly variable genes.

395

396 **PCA of RNA data**

397 PCA was carried out for gene expression levels ($\log_2$ [TPM + 1]) without scaling. For the loading analysis
398 of marker genes, we used the following genes; *MKI67* and *PCNA* for positive markers and *CDKN1A* for a
399 negative marker for cell proliferation in colorectal cancer [40]. *CCND2* and *CCND3* for positive markers
400 for cell cycle in this cancer [41]. E-cadherin (*CDH1*) for an epithelial marker; N-cadherin (*CDH2*),
401 vimentin (*VIM*), and fibronectin (*FN1*) for mesenchymal markers [42]. *LGR5*, *ASCL2*, *OLFM4*, *MSI1*, and
402 *SOX9* for crypt base stem cell markers, *HOPX*, *BMI1*, and *LRIG1* for +4 (position from the crypt base)
403 stem cell markers, *AQP8*, *CAR1*, *CEACAM1*, *KRT20*, and *SLC26A3* for differentiation makers for
404 absorption cells, and *MUC2*, *SPINK1* for differentiation markers for secretion cells [43].

405

406 **Hierarchal clustering, correlation plot, and heatmap analysis**

407 For hierarchal clustering, we used the "hclust" function in the "stats" package of R software, where we
408 calculated the Euclidean distance of expression levels ($\log_2$ [TPM +1]) of all highly variable genes
409 between cells and used the Ward method for agglomeration. We generated correlation plots of highly
410 variable genes using the "corrplot" function in the R "corrplot" package, where we used the Ward method
411 for agglomeration. We divided genes into three clusters based on these hierarchical clustering results using

412 the "addrect = 3" option. A heatmap was generated using the "heatmap.2" function in the "ggplot2"
413 package. In the heatmap, cells were arranged according to their order in the dendrogram described above
414 and genes were arranged according to their order in the correlation plot of highly variable genes.
415
416 **Gene set enrichment analysis**
417 DAVID [44] was used to identify gene ontologies (biological processes) in which genes of an identified
418 group were enriched ($P < 0.01$).
419
420 **SNV detection for single and bulk cells**
421 For bulk-cell data, we used a previously described method for SNV/indel calling [45] by cisCall with cell-
422 line/frozen parameters [46], mapping reads to the mouse genome (mm9) by BWA [47]. We filtered out
423 PCR-duplicated reads as well as reads and bases with low mapping and base qualities. The remaining
424 variants were further filtered statistically using Fisher's exact test to compare fore- and background
425 samples, which came from two different individuals of the same pure C57BL/6J strain. We verified the
426 negligible effects of using a different individual for the background sample (**Supplementary Results**). A
427 series of filters was used to remove suspicious variant calls, such as those related to misalignments.
428 Variants for which allele frequencies were significantly greater than 1% in the binomial test were retained.
429 The procedure is summarized in S1 Appendix: **Figure S9**.
430 For single-cell sequencing data, we called SNVs only at SNV sites called in bulk-cell sequencing
431 data. Specifically, we counted nucleotide bases with high qualities (mapQ ≥ 20, BaseQ ≥ 10) in single-cell
432 sequencing data as well as in background data (same as those used in bulk-cell SNV calling) with the
433 Samtools mpileup function [48]. We then selected variants with the largest χ^2 test statistic (with Yates's
434 correction) among all possible variants at each position to identify those that were most likely to differ
435 between single-cell and background data. We required a variant count ≥ 2 and VAF ≥ 2% for such variants
436 in single-cell data. We then selected variants that overlapped with SNV sites called in bulk-cell data.
437
438 **Detecting mutation in cancer-related genes**

We investigated nonsynonymous mutations in cancer-related genes contained in Tier1 in COSMIC Cancer Gene Census [25].

MDS of DNA data and the diversity index

We performed MDS from the cell × site matrix composed of zero and one, which respectively represent the absence and presence of SNVs (both synonymous and non-synonymous SNVs) and NA, which represents an undetermined call due to shallow depth. We assigned zero to non-called sites as the true negative when those sites had depths ≥ 30 and assigned NA to non-called sites when the depth was < 30. We only used SNV sites where a variant was called in at least one cell and the VAFs at the same sites in bulk data ranged from 10–35% (polymorphic) for at least one time point. We removed cells and sites (two each) with too few or too many NAs, yielding 104 sites and 69 cells. Using this 0/1/NA matrix, we calculated the *p*-distance (proportion of different sites) used in molecular evolution without using NA, and then performed MDS.

The diversity index was calculated as the average Euclidian distance from the centroid over cells in the MDS space, where we used up to the sixth dimension because of a sudden drop in the eigenvalues over this dimension. To calculate the statistical significance of differences between cell groups, we used a bootstrapping approach in which we randomly re-sampled cells' sequences from the 0/1/NA matrix of each cell group 10,000 times and performed the same MDS as in the observed data for each replicate. We then calculated the diversity index for each replicate to determine the 95% confidence interval and standard deviation for each cell group.

Lorenz curve and Gini coefficients

A Lorenz curve was generated with read depth at each site using the "Lc" function in the "ineq" package of R software. To quantify uniformity, the Gini coefficient was calculated using the "Gini" function in the "ineq" package.

466 **ADO rate**

467 The ADO rate was defined as the rate of homozygous sites in single-cell samples where the bulk sample
468 was heterozygous (defined as sites where VAFs were 45–55%) at the same nucleotide site. We removed
469 outlier cells with high ADO rates at each time point (one cell each with an ADO rate > 80% at T2 and T3).

470

471 **Average copy number**

472 The average copy number (ACN) was calculated as follows:

473 $$\mathrm{ACN} = 2 \times \left\{ \left(2^{\frac{\sum(\log_2 R_i \times L_i)}{\sum L_i}} \right) \times \left(\frac{\sum L_i}{GL} \right) + \left(1 - \frac{\sum L_i}{GL} \right) \right\} , \quad (1),$$

474 where $\log_2 R_i$, L_i, and GL represent the log-ratio of CNA segment i, length of CNA segment i, and genome
475 length (50 Gb for mouse, 40 Gb for human), respectively. CNAs of mouse bulk data were detected as
476 previously described [45]. Briefly, segments were called for the same fore- and background BAM files as
477 those used in SNV with Exome CNV [49] and Varscan2 [50]. Overlapping segments called by both tools
478 were used as CNA segments.

479

480 **Random Forest**

481 Random Forest was used to generate the classifier for the histological type and MSI status of human
482 cancer. We used gene expression levels, number of SNVs in each gene, total mutation (SNV/indel)
483 number, and mutation density (total number of SNVs/indels divided by target region size) as explanatory
484 variables. Using TCGA data [24], we first filtered out unimportant explanatory variables using the two-
485 sided Kruskal-Wallis test with P values of 5.00×10^{-5} and 1.00×10^{-9} yielding 171 and 78 variables for
486 histological type and MSI status, respectively. These were used to train a Random Forest classifier with
487 the "randomForest" function in the "randomForest" package of R software, with the options ntree = 10000
488 (setting the number of trees to grow to 1000) and mtry = 5 (setting the number of variables randomly
489 sampled to five). Using the created classifier, the same explanatory variables for mouse data were used to
490 classify each feature in the mouse model.

MDS of mouse cell and TCGA samples

We first identified TCGA samples with gene expression patterns similar to the mouse single-cell groups. For that purpose, we calculated a normalized 1-*r* distance as follows:

$$d_{h,G} = \frac{1 - r_{h,m^G}}{\mathrm{MADN}(1 - r_{m_i^G, m^G})} \quad , \qquad (2),$$

where $r_{i,j}$ is a Pearson correlation coefficient between vectors *i* and *j* of expression levels in log across highly variable genes, *h* represents a human TCGA sample, *G* represents a mouse single-cell group, $\boldsymbol{m}_i^G$ represents mouse single cell *i* in group *G*, m^G represents the centroid of $\boldsymbol{m}_i^G$ that was calculated by the median, and MADN represents the median absolute deviation adjusted by a factor for asymptotically normal consistency. We calculated this distance from a TCGA sample to every mouse group and selected a TCGA sample for those whose minimum distance across the groups was less than 4.05 and the difference between the first and second minimum distances was larger than 0.31. For selected TCGA and all mouse single-cell samples, MDS was performed based on the distance of 1-*r*.

Availability of data and materials

Sequence data used in this study are available in the DDBJ Sequenced Read Archive under Accession Nos. DRX100507-DRX100729. [These data are held until the acceptance. URL will be added after the data release.]

Competing interests

None to be declared.

Authors' contributions

Y.A., Y.H., T.S., and M.K. designed the study. Y.A., T.M., T.I., and Y.H. performed the experiments. H.O., E.F., D.N., H.N., M.N., and M.K. analyzed the data. H.O., Y.A., E.F., T.I., Y.H., and M.K. wrote the manuscript. D. S., K.M., K.O., and T.S. reviewed the manuscript. M.K. led the project.

Acknowledgements

We thank Joe Miyamoto, Asmaa Elzawahry, Iku Orihashi, Masako Ochiai, and Wakako Mukai for technical assistance, and Ryuichi Sugino and Daniel A. Vasco for useful suggestions.

References

1. Almendro V, Marusyk A, Polyak K. Cellular heterogeneity and molecular evolution in cancer. Annu Rev Pathol. 2013;8:277-302. Epub 2012/10/25. doi: 10.1146/annurev-pathol-020712-163923. PubMed PMID: 23092187.

2. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-28. Epub 2017/02/12. doi: 10.1016/j.cell.2017.01.018. PubMed PMID: 28187284.

3. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med. 2017;376(22):2109-21. Epub 2017/04/27. doi: 10.1056/NEJMoa1616288. PubMed PMID: 28445112.

4. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018;15(2):81-94. Epub 2017/11/09. doi: 10.1038/nrclinonc.2017.166. PubMed PMID: 29115304.

5. Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet. 2006;38(4):468-73. Epub 2006/03/28. doi: 10.1038/ng1768. PubMed PMID: 16565718.

6. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8. Epub 1976/10/01. PubMed PMID: 959840.

7. Fulawka L, Donizy P, Halon A. Cancer stem cells--the current status of an old concept: literature review and clinical approaches. Biol Res. 2014;47:66. Epub 2015/02/28. doi: 10.1186/0717-6287-47-66. PubMed PMID: 25723910; PubMed Central PMCID: PMCPMC4335556.

8. Kreso A, Dick JE. Evolution of the cancer stem cell model. Cell Stem Cell. 2014;14(3):275-91. Epub 2014/03/13. doi: 10.1016/j.stem.2014.02.006. PubMed PMID: 24607403.

9. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochim Biophys Acta Rev Cancer. 2017;1867(2):151-61. Epub 2017/01/23. doi: 10.1016/j.bbcan.2017.01.003. PubMed PMID: 28110020; PubMed Central PMCID: PMCPMC5558210.

10. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nat Genet. 2018;50(6):895-903. Epub 2018/05/29. doi: 10.1038/s41588-018-0128-6. PubMed PMID: 29808029.

11. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90-4. Epub 2011/03/15. doi: 10.1038/nature09807. PubMed PMID: 21399628; PubMed Central PMCID: PMCPMC4504184.

12. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512(7513):155-60. Epub 2014/08/01. doi: 10.1038/nature13600. PubMed PMID: 25079324; PubMed Central PMCID: PMCPMC4158312.

13. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396-401. Epub 2014/06/14. doi: 10.1126/science.1254257. PubMed PMID: 24925914; PubMed Central PMCID: PMCPMC4123637.

14. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539(7628):309-13. Epub 2016/11/05. doi: 10.1038/nature20123. PubMed PMID: 27806376; PubMed Central PMCID: PMCPMC5465819.

15. Suzuki A, Matsushima K, Makinoshima H, Sugano S, Kohno T, Tsuchihara K, et al. Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol. 2015;16:66. Epub 2015/04/19. doi: 10.1186/s13059-015-0636-y. PubMed PMID: 25887790; PubMed Central PMCID: PMCPMC4450998.

16. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc Natl Acad Sci U S A. 2014;111(50):17947-52. Epub 2014/11/27. doi: 10.1073/pnas.1420822111. PubMed PMID: 25425670; PubMed Central PMCID: PMCPMC4273416.

17. Kim KT, Lee HW, Lee HO, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015;16:127. Epub 2015/06/19. doi: 10.1186/s13059-015-0692-3. PubMed PMID: 26084335; PubMed Central PMCID: PMCPMC4506401.

18. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148(5):873-85. Epub 2012/03/06. doi: 10.1016/j.cell.2012.02.028. PubMed PMID: 22385957.

19. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012;148(5):886-95. Epub 2012/03/06. doi: 10.1016/j.cell.2012.02.025. PubMed PMID: 22385958.

20. Kato M, Vasco DA, Sugino R, Narushima D, Krasnitz A. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. R Soc Open Sci. 2017;4(9):171060. doi: 10.1098/rsos.171060. PubMed PMID: 28989791; PubMed Central PMCID: PMCPMC5627131.

21. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature. 2015;518(7539):422-6. Epub 2014/12/04. doi: 10.1038/nature13952. PubMed PMID: 25470049; PubMed Central PMCID: PMCPMC4864027.

22. Onuma K, Ochiai M, Orihashi K, Takahashi M, Imai T, Nakagama H, et al. Genetic reconstitution of tumorigenesis in primary intestinal cells. Proc Natl Acad Sci U S A. 2013;110(27):11127-32. Epub 2013/06/19. doi: 10.1073/pnas.1221926110. PubMed PMID: 23776211; PubMed Central PMCID: PMCPMC3703980.

23. Itzkovitz S, Blat IC, Jacks T, Clevers H, van Oudenaarden A. Optimality in the development of intestinal crypts. Cell. 2012;148(3):608-19. Epub 2012/02/07. doi: 10.1016/j.cell.2011.12.025. PubMed PMID: 22304925; PubMed Central PMCID: PMC3696183.

24. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7. Epub 2012/07/20. doi: 10.1038/nature11252. PubMed

627 PMID: 22810696; PubMed Central PMCID: PMCPMC3401966.

628 25. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human
629 cancer genes. Nature Reviews Cancer. 2004;4:177. doi: 10.1038/nrc1299

630 26. Fajardo AM, Piazza GA, Tinsley HN. The role of cyclic nucleotide signaling pathways in cancer:
631 targets for prevention and treatment. Cancers (Basel). 2014;6(1):436-58. Epub 2014/03/01. doi:
632 10.3390/cancers6010436. PubMed PMID: 24577242; PubMed Central PMCID:
633 PMCPMC3980602.

634 27. Lubbe WJ, Zhou ZY, Fu W, Zuzga D, Schulz S, Fridman R, et al. Tumor epithelial cell matrix
635 metalloproteinase 9 is a target for antimetastatic therapy in colorectal cancer. Clin Cancer Res.
636 2006;12(6):1876-82. Epub 2006/03/23. doi: 10.1158/1078-0432.CCR-05-2686. PubMed PMID:
637 16551873.

638 28. Li CJ, Zhang X, Fan GW. Updates in colorectal cancer stem cell research. J Cancer Res Ther.
639 2014;10 Suppl:233-9. Epub 2015/02/20. doi: 10.4103/0973-1482.151449. PubMed PMID:
640 25693926.

641 29. Qureshi-Baig K, Ullmann P, Haan S, Letellier E. Tumor-Initiating Cells: a criTICal review of
642 isolation approaches and new challenges in targeting strategies. Mol Cancer. 2017;16(1):40. Epub
643 2017/02/18. doi: 10.1186/s12943-017-0602-2. PubMed PMID: 28209178; PubMed Central
644 PMCID: PMCPMC5314476.

645 30. Batlle E, Clevers H. Cancer stem cells revisited. Nat Med. 2017;23(10):1124-34. Epub
646 2017/10/07. doi: 10.1038/nm.4409. PubMed PMID: 28985214.

647 31. Walker F, Zhang HH, Odorizzi A, Burgess AW. LGR5 is a negative regulator of tumourigenicity,
648 antagonizes Wnt signalling and regulates cell adhesion in colorectal cancer cell lines. PLoS One.
649 2011;6(7):e22733. Epub 2011/08/11. doi: 10.1371/journal.pone.0022733. PubMed PMID:
650 21829496; PubMed Central PMCID: PMCPMC3145754.

651 32. Carmon KS, Gong X, Yi J, Wu L, Thomas A, Moore CM, et al. LGR5 receptor promotes cell-cell
652 adhesion in stem cells and colon cancer cells via the IQGAP1-Rac1 pathway. J Biol Chem.
653 2017;292(36):14989-5001. Epub 2017/07/26. doi: 10.1074/jbc.M117.786798. PubMed PMID:

28739799; PubMed Central PMCID: PMCPMC5592675.

33. Zhou X, Geng L, Wang D, Yi H, Talmon G, Wang J. R-Spondin1/LGR5 Activates TGFbeta Signaling and Suppresses Colon Cancer Metastasis. Cancer Res. 2017;77(23):6589-602. Epub 2017/09/25. doi: 10.1158/0008-5472.CAN-17-0219. PubMed PMID: 28939678.

34. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33(3):285-9. Epub 2015/01/20. doi: 10.1038/nbt.3129. PubMed PMID: 25599178; PubMed Central PMCID: PMCPMC4374170.

35. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12(6):519-22. Epub 2015/04/29. doi: 10.1038/nmeth.3370. PubMed PMID: 25915121.

36. Maru Y, Orihashi K, Hippo Y. Lentivirus-Based Stable Gene Delivery into Intestinal Organoids. Methods Mol Biol. 2016;1422:13-21. Epub 2016/06/02. doi: 10.1007/978-1-4939-3603-8_2. PubMed PMID: 27246018.

37. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863-4. Epub 2011/02/01. doi: 10.1093/bioinformatics/btr026. PubMed PMID: 21278185; PubMed Central PMCID: PMCPMC3051327.

38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.

39. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. Epub 2011/08/06. doi: 10.1186/1471-2105-12-323. PubMed PMID: 21816040; PubMed Central PMCID: PMCPMC3163565.

40. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. Nat Rev Cancer. 2006;6(2):99-106. Epub 2006/02/24. doi: 10.1038/nrc1802. PubMed PMID: 16491069.

41. Tominaga O, Nita ME, Nagawa H, Fujii S, Tsuruo T, Muto T. Expressions of cell cycle regulators in human colorectal cancer cell lines. Jpn J Cancer Res. 1997;88(9):855-60. Epub 1997/11/25. PubMed PMID: 9369933.

681 42. Weinberg R. The biology of cancer: Garland science; 2013.

682 43. Hong SN, Dunn JC, Stelzner M, Martin MG. Concise Review: The Potential Use of Intestinal
683 Stem Cells to Treat Patients with Intestinal Failure. Stem Cells Transl Med. 2017;6(2):666-76.
684 Epub 2017/02/14. doi: 10.5966/sctm.2016-0153. PubMed PMID: 28191783; PubMed Central
685 PMCID: PMC5442796.

686 44. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the
687 comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13. Epub
688 2008/11/27. doi: 10.1093/nar/gkn923. PubMed PMID: 19033363; PubMed Central PMCID:
689 PMCPMC2615629.

690 45. Takahashi T, Elzawahry A, Mimaki S, Furukawa E, Nakatsuka R, Nakamura H, et al. Genomic
691 and transcriptomic analysis of imatinib resistance in gastrointestinal stromal tumors. Genes
692 Chromosomes Cancer. 2017;56(4):303-13. Epub 2016/12/21. doi: 10.1002/gcc.22438. PubMed
693 PMID: 27997714; PubMed Central PMCID: PMCPMC5324566.

694 46. Kato M, Nakamura H, Nagai M, Kubo T, Elzawahry A, Totoki Y, et al. A computational tool to
695 detect DNA alterations tailored to formalin-fixed paraffin-embedded samples in cancer clinical
696 sequencing. Genome Med. 2018;10(1):44. Epub 2018/06/09. doi: 10.1186/s13073-018-0547-0.
697 PubMed PMID: 29880027; PubMed Central PMCID: PMCPMC5992758.

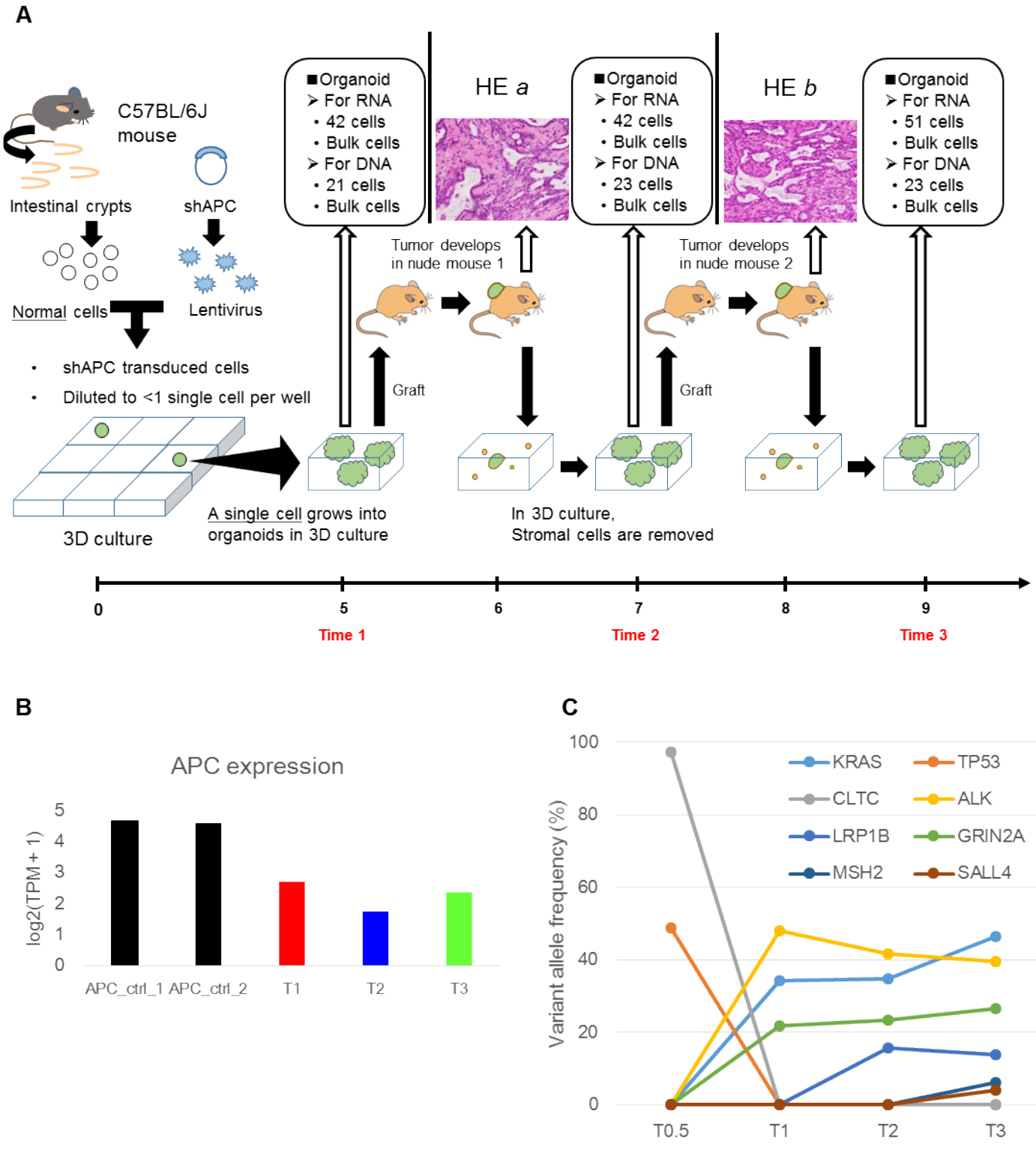
698 47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
699 Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324.
700 PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234.

701 48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
702 format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.
703 PubMed PMID: 19505943; PubMed Central PMCID: PMCPMC2723002.

704 49. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome
705 sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV.
706 Bioinformatics. 2011;27(19):2648-54. Epub 2011/08/11. doi: 10.1093/bioinformatics/btr462.
707 PubMed PMID: 21828086; PubMed Central PMCID: PMCPMC3179661.

50. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76. Epub 2012/02/04. doi: 10.1101/gr.129684.111. PubMed PMID: 22300766; PubMed Central PMCID: PMCPMC3290792.

735 **Figures**



736

737 **Fig. 1.** The mouse model. (A) The experimental procedure and HE staining of subcutaneously transplanted

738 tumors. *One single cell* was 3D-cultured in a compartment in a 96-well plate, and single cell -derived

739 organoids were taken to separate single cells. RNA and DNA were separately extracted from the different

740 single cells of multiple organoids and then sequenced. The numbers of cells for RNA and DNA sequencing
741 in boxes are those obtained after quality control of data. (B) The *APC* gene expression from bulk-cell RNA
742 sequencing. "APC_ctrl" indicates control samples that were cultured in our 3D culture system and derived
743 from normal cells without *APC* knockdown. (C) Variant allele frequencies of mutations found in the
744 significantly mutated genes of colorectal cancer by bulk-cell DNA sequencing. See S1 Appendix: **Figure**
745 **S1** for the annotations of the mutations.
746

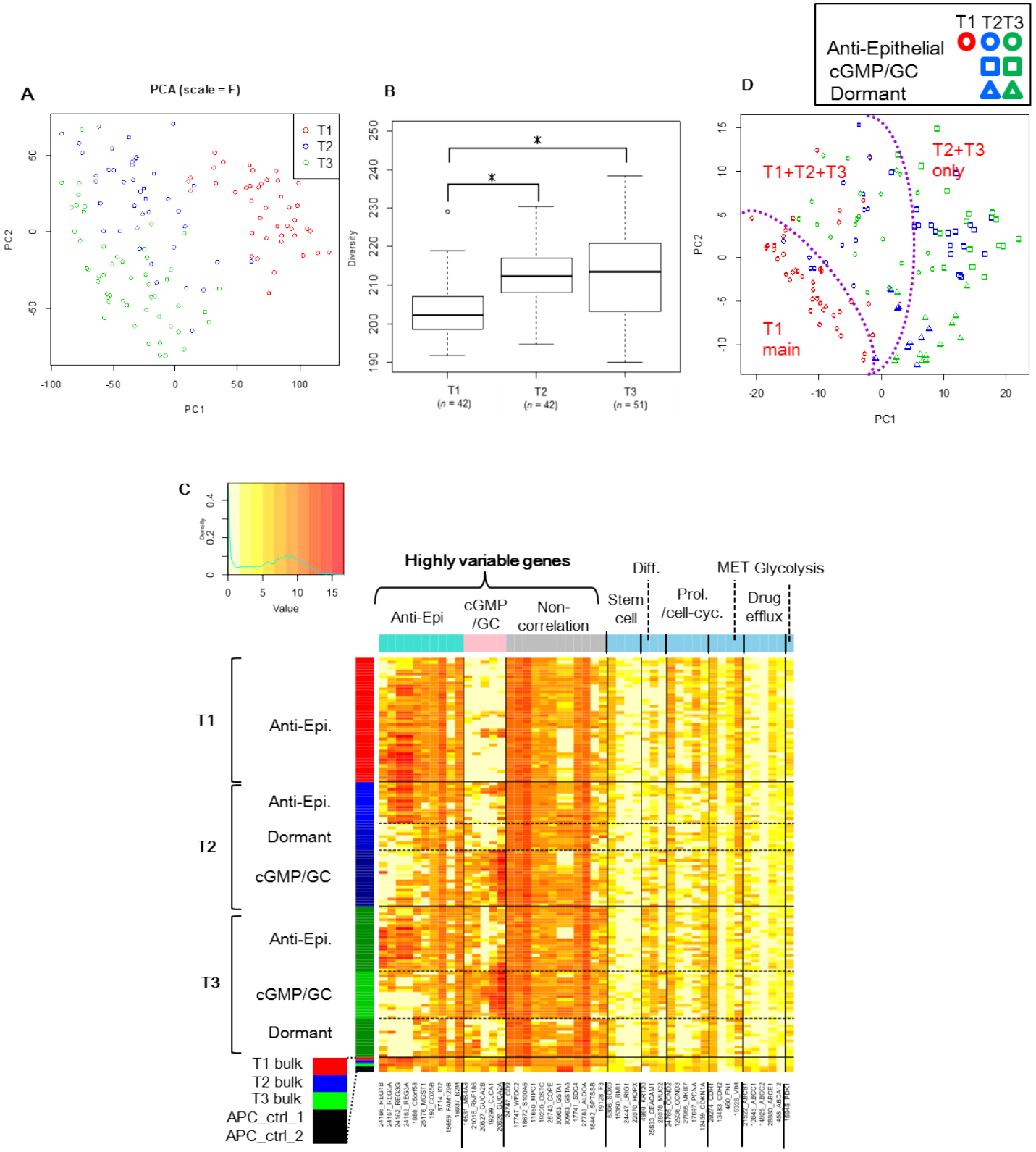


Fig. 2. Transcriptome analysis. (A) PCA plot of single cells based on expression levels (genes with TPM ≥ 10 in at least one cell). T1, at the time of 3D culturing; T2 and T3, after the first and second transplantations, respectively. (B) Euclidean distance from the centroid in the PCA space (using full dimensions). *$P < 0.01$

(two-sided Wilcoxon rank sum test). (C) Heatmap of gene expression levels (in TPM). The rows represent single cells or bulk-cell samples (in the bottom), and the columns represent highly variable genes and several types of marker genes. The cell and gene groups were determined as shown in S1 Appendix: **Figure S4**. The red, blue, and green codes in the rows correspond to T1, T2, and T3. "Diff." and "Prol./cell-cyc." represents differentiation and proliferation/cell cycle. "APC_ctrl" indicates control samples that were cultured in our 3D culture system and derived from normal cells without *APC* knockdown. (D) PCA plot of cells grouped based on expression levels of highly variable genes.

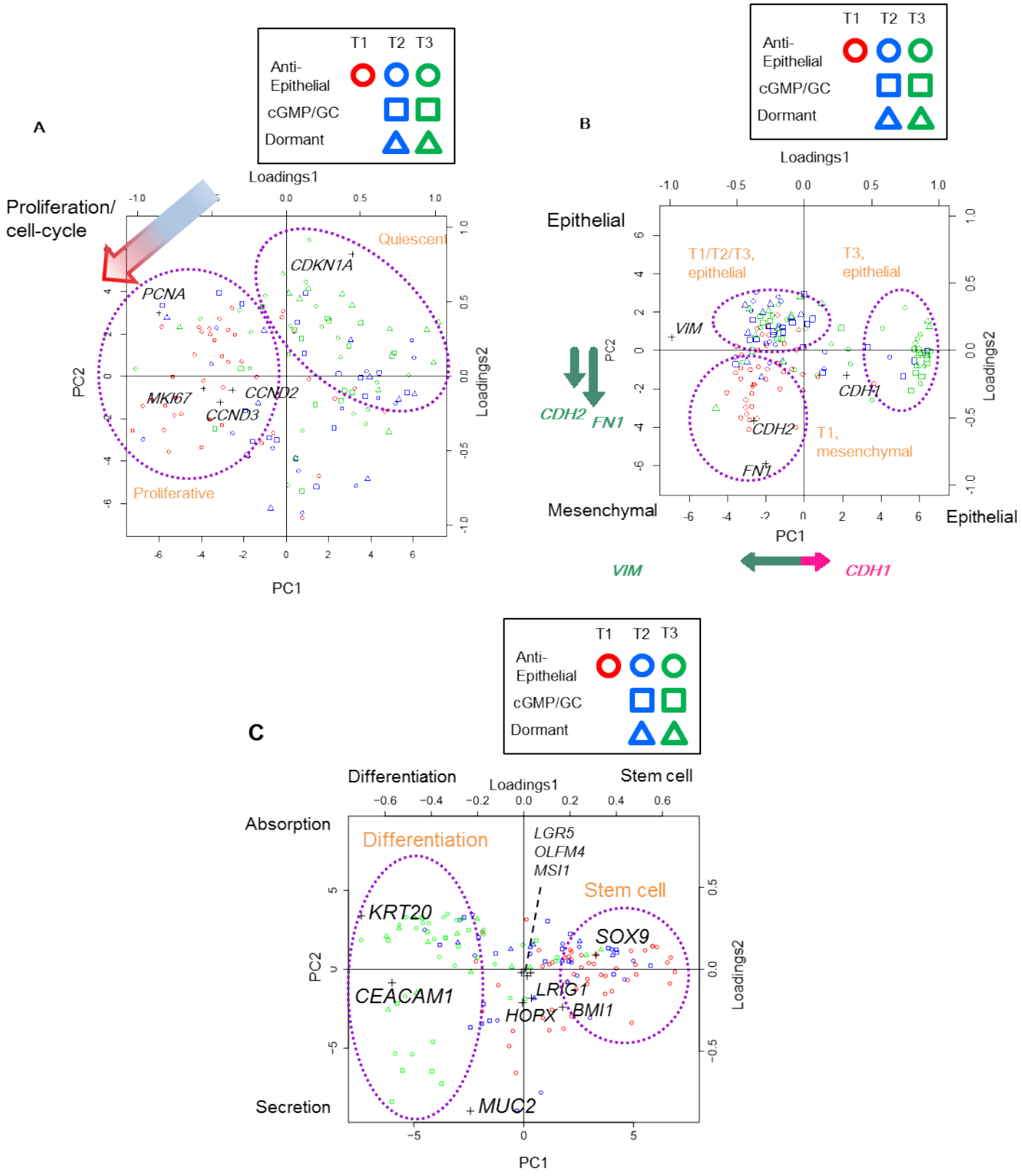


Fig. 3. PCA and overlaid loading plots based on expression levels of markers (A) about the proliferation/cell cycle. The arrow indicates the direction from negative to positive markers in the loading plot; cells positioned in that direction in the PCA plot had higher expression levels of positive marker genes. (B) About

763 the epithelial and mesenchymal. The arrows along the *x* and *y* axes represent projected loadings in the
764 loading analysis, where cells positioned in that direction in the PCA plot had higher marker gene expression
765 levels. (C) About stem cell and differentiation.
766

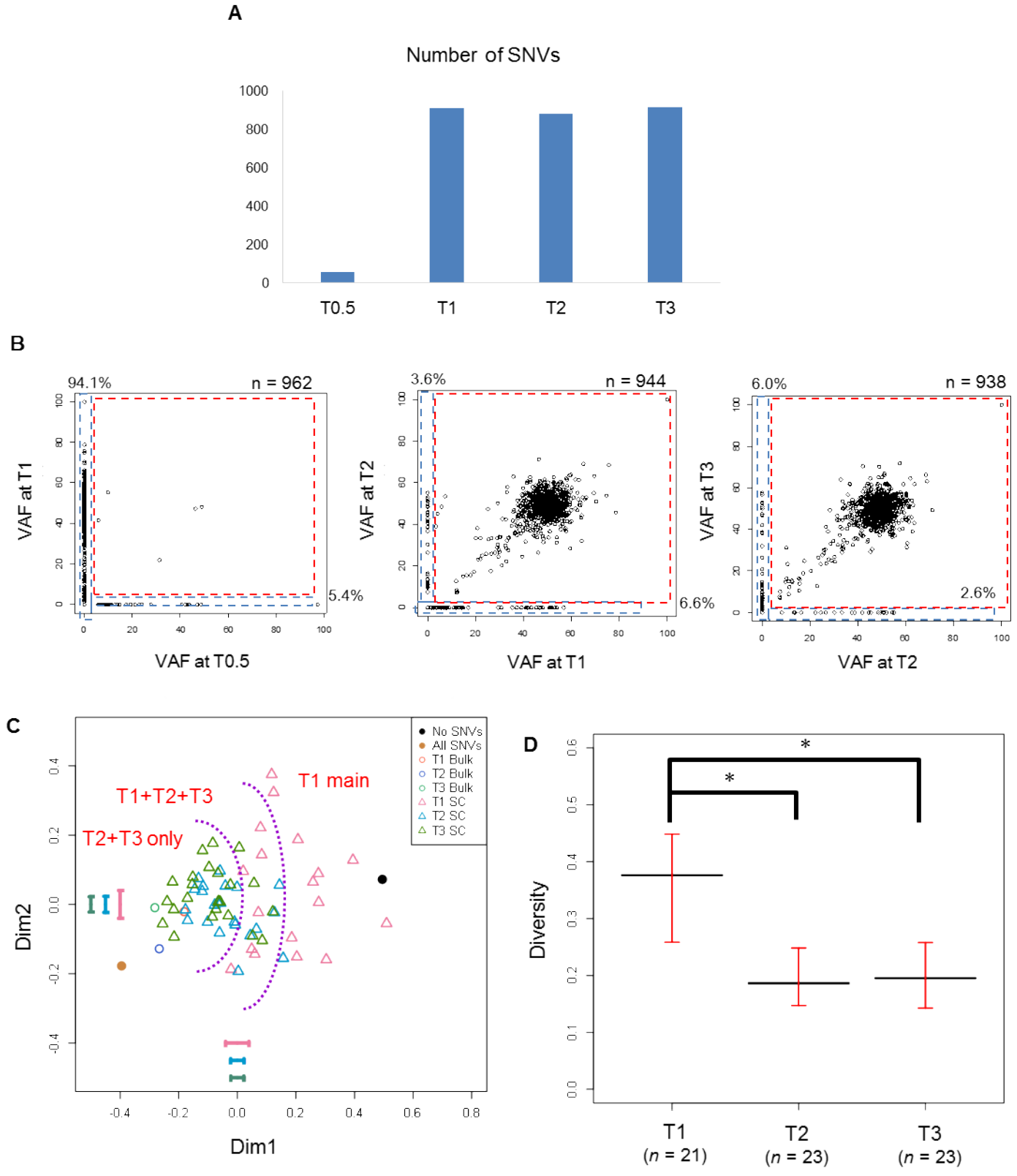
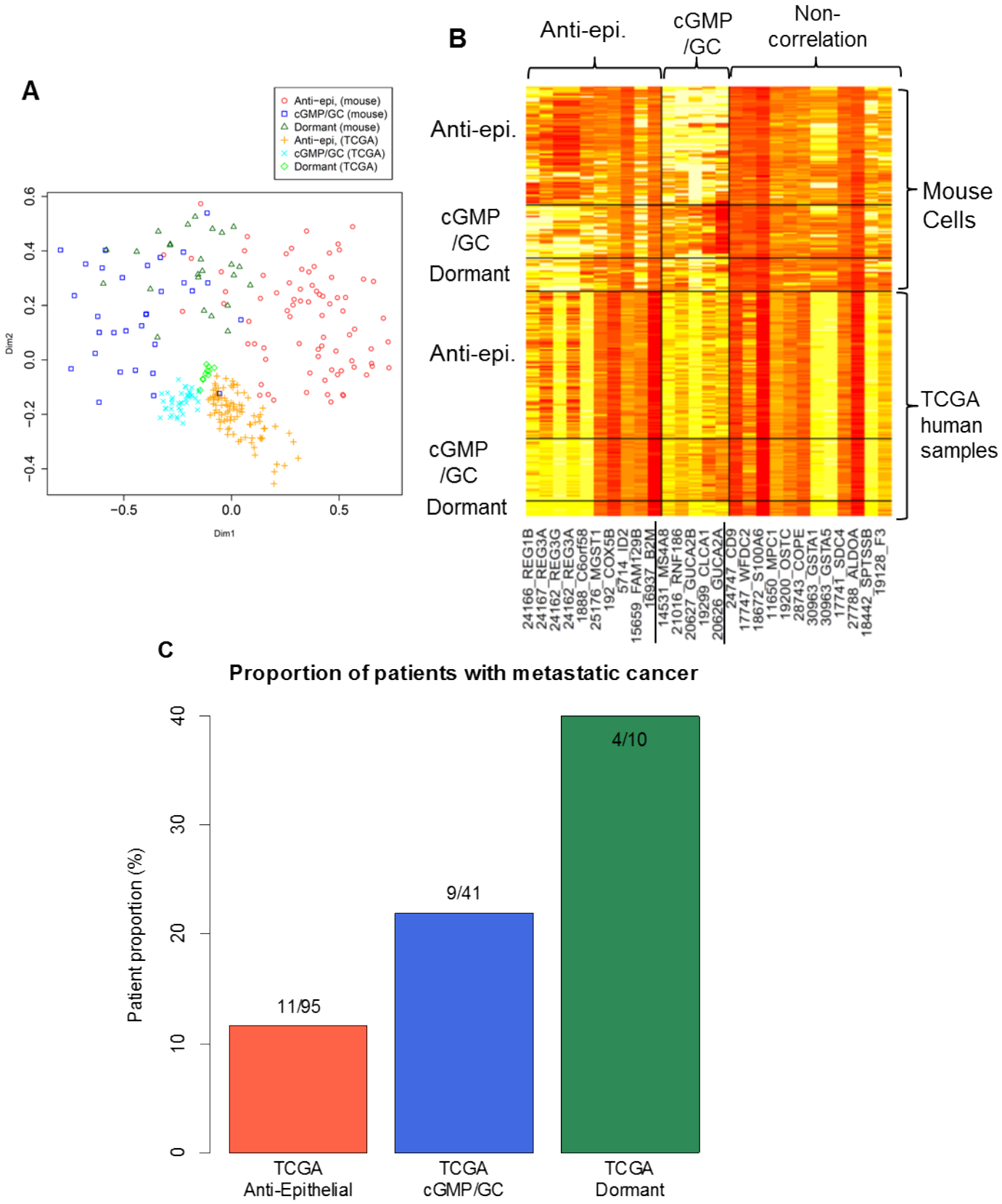


Fig. 4. Exome analysis. (A) Number of SNVs called in bulk-cell sequencing. (B) Comparison of VAFs of SNVs called in bulk-cell sequencing at successive time points. One point indicates one SNV. Numbers represent the number of points. (C) MDS plot based on single-cell exome sequencing. “No SNV” and “All

771 SNV" represent sequences with no SNVs and with SNVs at all sites, respectively, which were artificially
772 generated as a reference. Error bars represent the standard deviation for each dimension calculated with a
773 bootstrapping approach that took into account ADO rates. (D) Median Euclidean distance from the centroid
774 over cells in the MDS space. The black and red bars represent the observed value and 95% confidence
775 interval calculated with the bootstrapping approach. *$P < 0.05$ (bootstrapping test).
776



777

778 **Fig. 5.** Analysis of TCGA human samples with gene expression patterns similar to mouse cell groups. (A)

779 MDS plot of mouse single-cell samples and such TCGA samples on the basis of a similarity of gene
780 expression patterns. (B) Heatmap of the samples. Genes are highly variable genes shown in **Fig. 2C**. (C)
781 The fraction of patients with metastatic tumor in TCGA samples with expression patterns similar to mouse
782 cell groups

783

784

Supplementary Results

Estimation of false positive rate

We first estimated the number of SNV sites that differed between two individual mice of the pure C57BL/6J strain. For normal intestinal tract samples obtained from the two mice, we called SNVs in bulk-cell sequencing data using each of the two samples as the foreground data and the other as the background: the numbers were 1.0 and 4.5×10^{-7} per chromosomal position for the two sample pairs, respectively. When we called SNVs in half-split sequencing data used as the fore- and background data for the same sample, the number of SNVs per position was 0 and 0.4×10^{-7} for the two samples, respectively. Taken together, the false positive rate in bulk-cell sequencing was estimated as 1.0–4.9 ([1.0 + 0.0]–[4.5 + 0.4]) $\times 10^{-7}$. Because we called SNVs in single cells only at SNV sites called in bulk-cell sequencing data, the false-positive rate in single cells was not more than that in bulk-cell sequencing. Since 10–23% of chromosomal positions were called by our loose criteria for sequencing data from four single cells obtained from normal intestinal tract tissue, the false-positive rate per position in single-cell sequencing was estimated as 0.1–1.1 $\times 10^{-7}$.

Association with human cancers

We investigated the features of human colorectal cancer that correspond to those of our mouse cancer model using TCGA human colorectal cancer data and our mouse bulk sequencing data (39). We first examined individual molecular features. SNV density in the mouse model was closer to the hypermutation type of human colorectal cancer (S1 Appendix: **Figure S7**). The expression of *MLH1*, the dysregulation of which causes hypermutation, was repressed with the levels decreasing over time (from T1 to T3) (S1 Appendix: **Figure S7**). The average copy number across the mouse genome was closer to the hypermutation type, indicating low chromosomal instability (S1 Appendix: **Figure S7**). Taken together, these results suggest that the mouse model was closer to the hypermutation type (albeit not extremely hyper) of human cancer.

We then analyzed clinical features in a machine learning approach (Random Forest) using a

812 clinical feature as the objective variable and omics (SNV/indel/RNA) data as explanatory variables. Of the
813 three histological types, including colon and rectal mucinous adenocarcinoma, our mouse model was
814 closest to human colon adenocarcinoma and was closer to the MSI-high than MSI-low and microsatellite-
815 stable types (S1 Appendix: **Figure S8**). Thus, our mouse model represented the MSI-high hypermutation
816 (although, not extremely hyper) type of human colon adenocarcinoma.
817

S1 Appendix: Supplementary Figures

A

HumanGene	Chr	Start	End	Mut_type	Ref	Alt	Notion	Reference
KRAS	6	145,169,253	145,169,253	indel	-	A	intronic	TCGA SMG
TP53	11	69,402,151	69,402,151	snv	A	T	nonsynonymous SNV	TCGA SMG and COSMIC
CLTC	11	86,520,656	86,520,656	snv	A	T	nonsynonymous SNV	COSMIC
ALK	17	72,952,883	72,952,883	snv	A	C	nonsynonymous SNV	COSMIC
LRP1B	2	40,724,718	40,724,718	snv	C	T	nonsynonymous SNV	COSMIC
GRIN2A	16	9,579,188	9,579,188	snv	T	C	nonsynonymous SNV	COSMIC
MSH2	17	88,079,144	88,079,144	snv	C	T	nonsynonymous SNV	COSMIC
SALL4	2	168,580,005	168,580,005	snv	C	T	nonsynonymous SNV	COSMIC

B

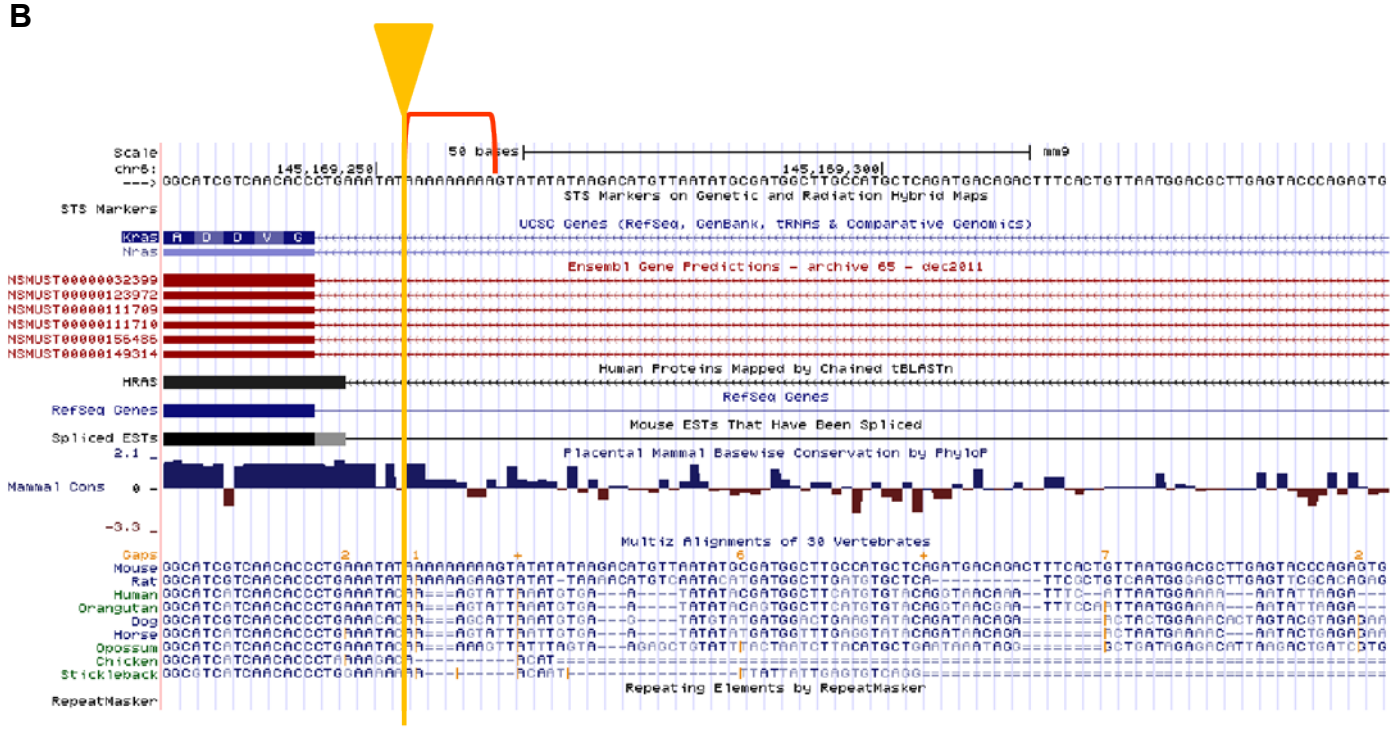
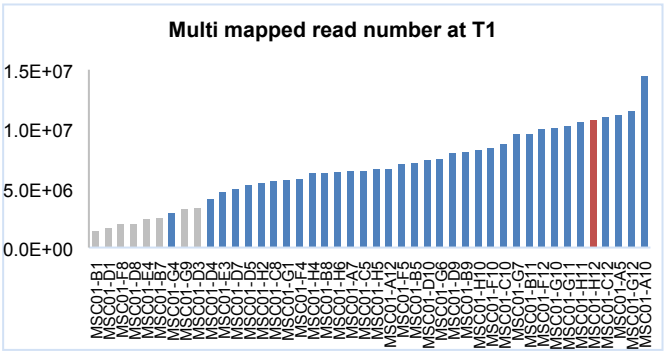


Figure S1 Details of mutations found in the significantly mutated genes of TCGA colorectal cancer and in cancer-related genes referred in COSMIC by bulk-cell DNA sequencing.

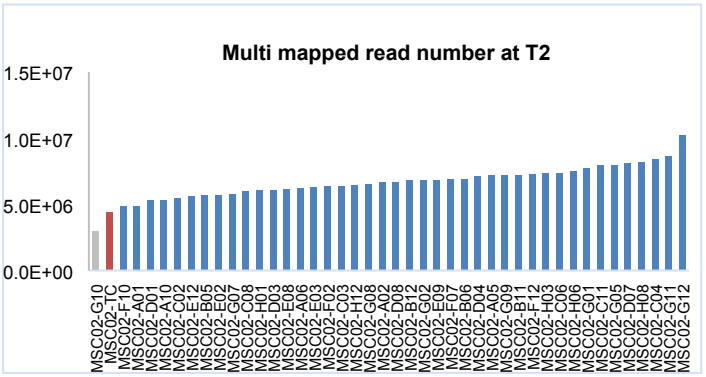
(**A**) Annotations of genes found in the significantly mutated genes of TCGA colorectal cancer and in COSMIC cancer-related genes by bulk-cell DNA sequencing. (**B**) The *KRAS* mutation in the mouse genome by the UCSC genome browser. The reversed U symbol in red indicates a mono-repeat of A. The arrow and line in gold indicate the position of mutation.

831

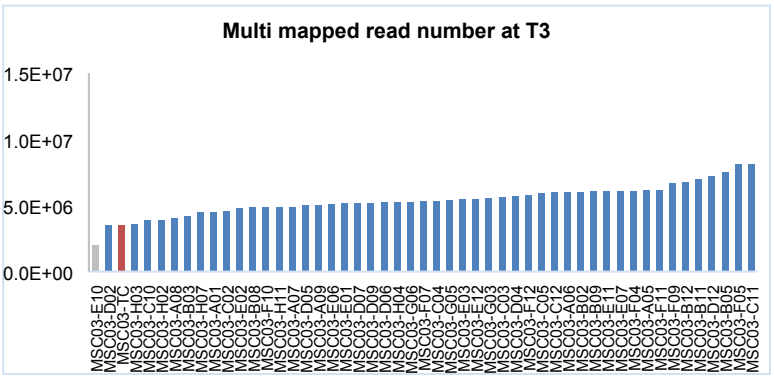
A



832



833



834

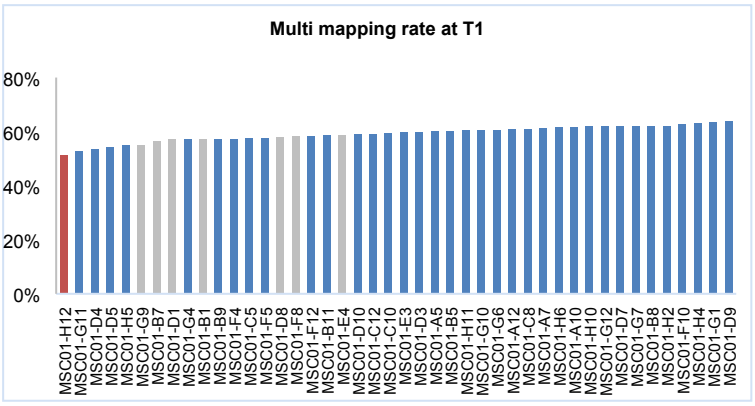
835

836

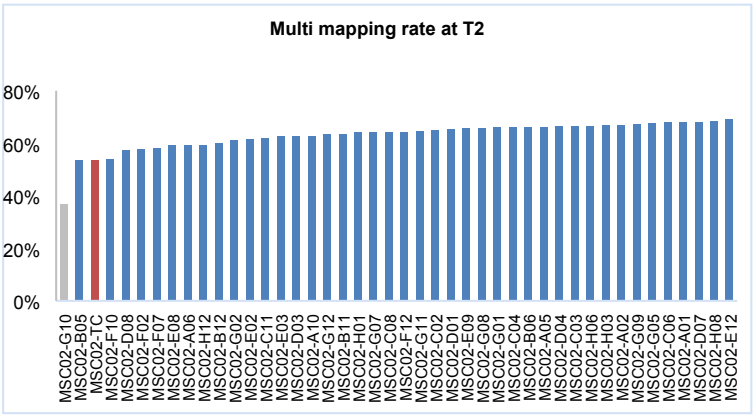
837

838

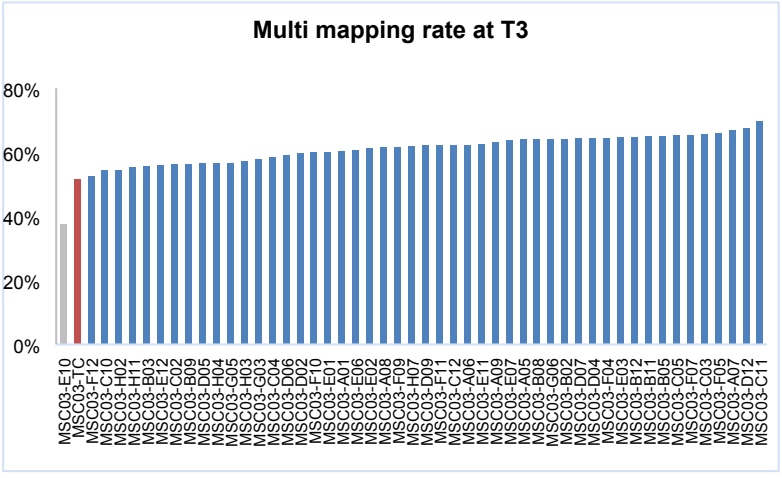
B



839



840



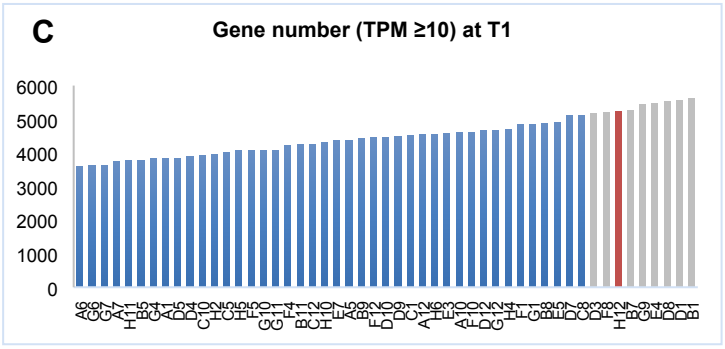
841

842

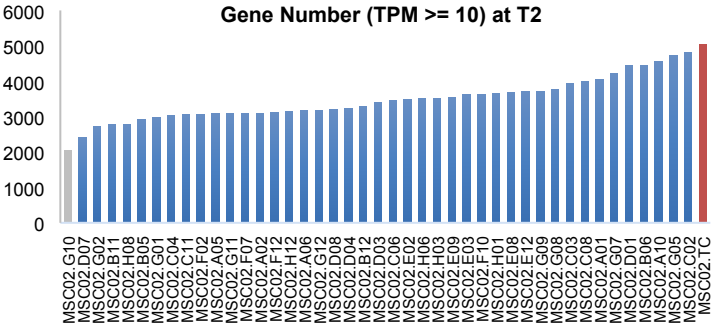
843

844

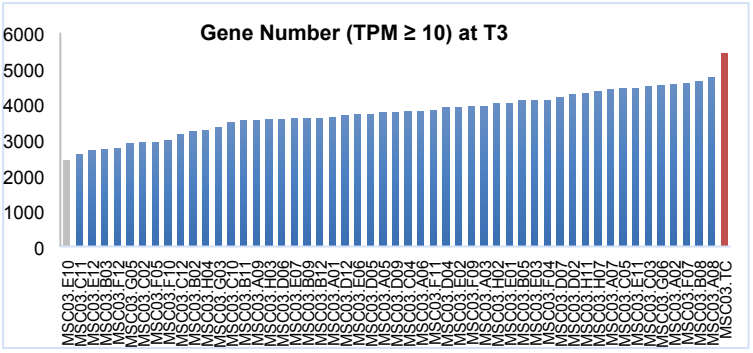
845



846



847



848

849

850

851

852

853

854

855

856

857

858

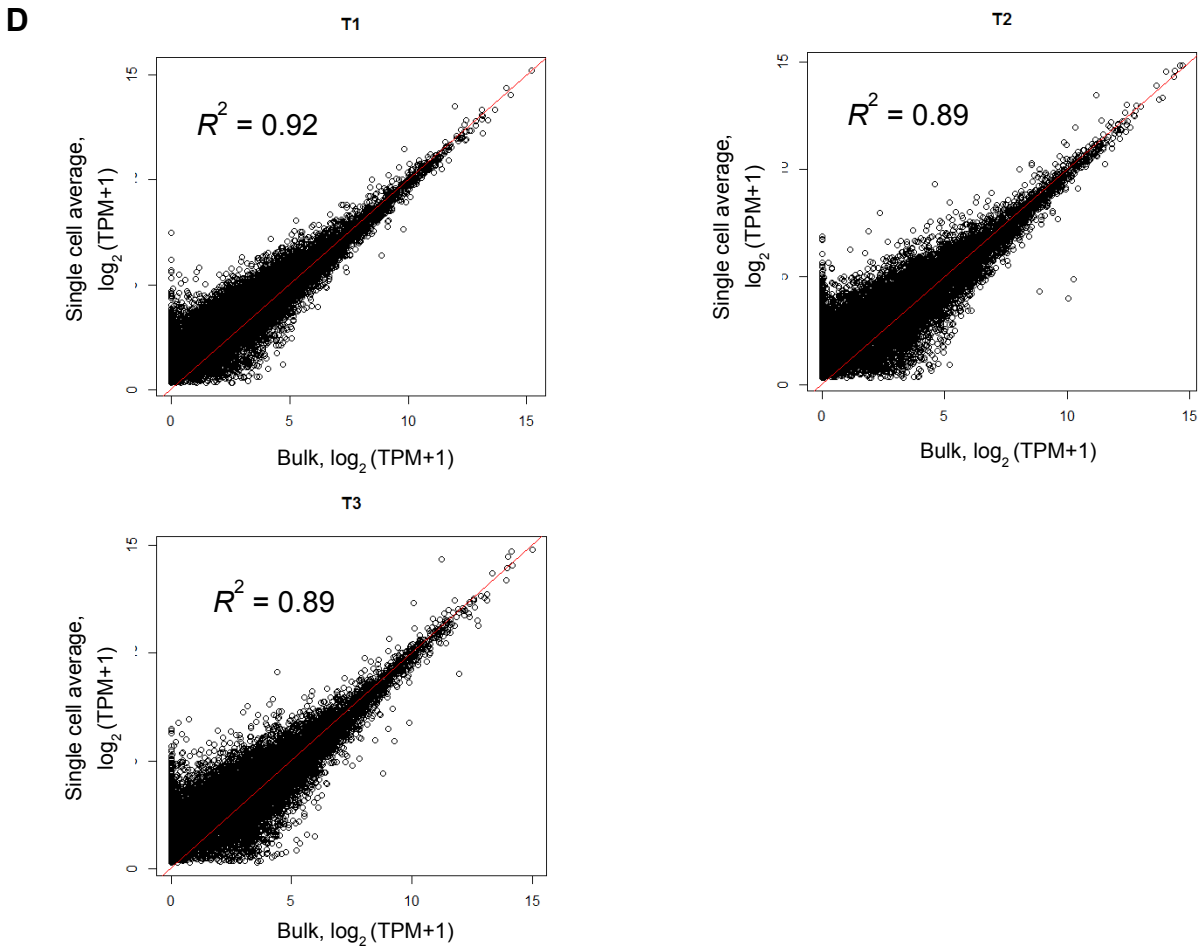
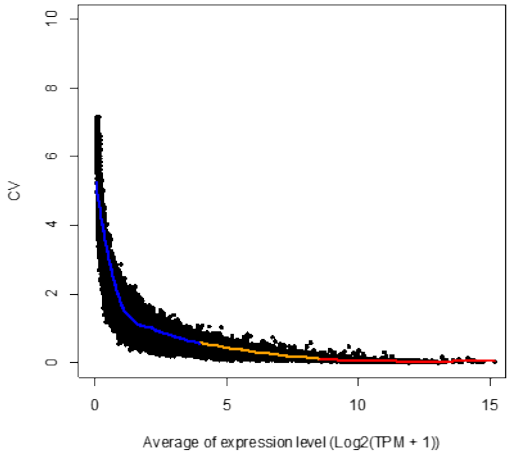
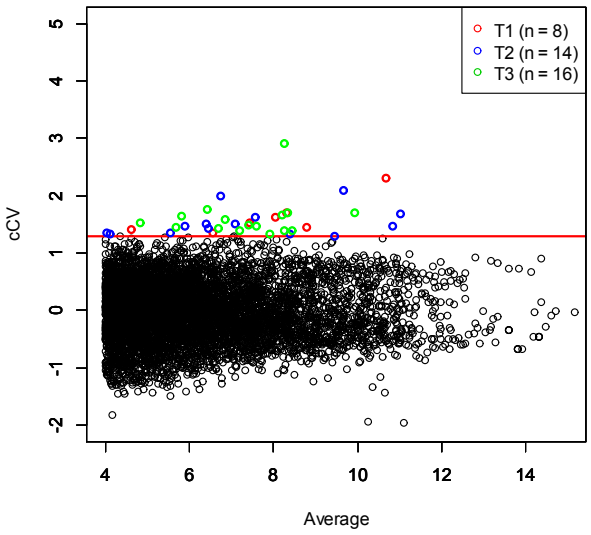


Figure S2 Quality check of single-cell transcriptome sequencing data. (A) Number of mapped reads, (B) mapping rate, and (C) number of expressed genes (TPM ≥ 10). We removed outliers (gray) based on the combination of the number of expressed genes (≤5200) and number of mapped genes (≤2.2×10^6), and mapping rate (≤20%). Blue and orange bars represent single-cell samples that were ultimately used and bulk samples, respectively. (D) Scatter plot of gene expression levels from a bulk sample versus expression levels averaged across the single cells that were ultimately used.



A

B



C

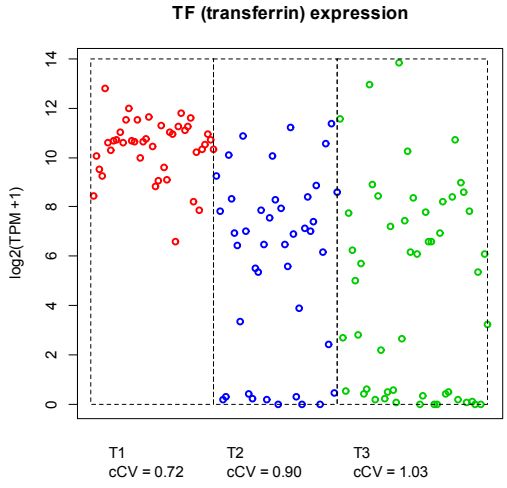


Figure S3 c*CV* and highly variable genes. (A) *CV* versus gene expression levels averaged across single

913 cells. Regression analysis was performed to obtain the locally weighted scatterplot smoothing (LOWESS)
914 curve within the range indicated by each color (blue, yellow, and red). (B) cCV and average expression
915 levels. Highly variable genes are shown above the red line. (C) cCV and distribution of gene expression
916 levels across single cells, illustrated with the transferrin gene. Each circle represents gene expression level
917 in a single cell.
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939

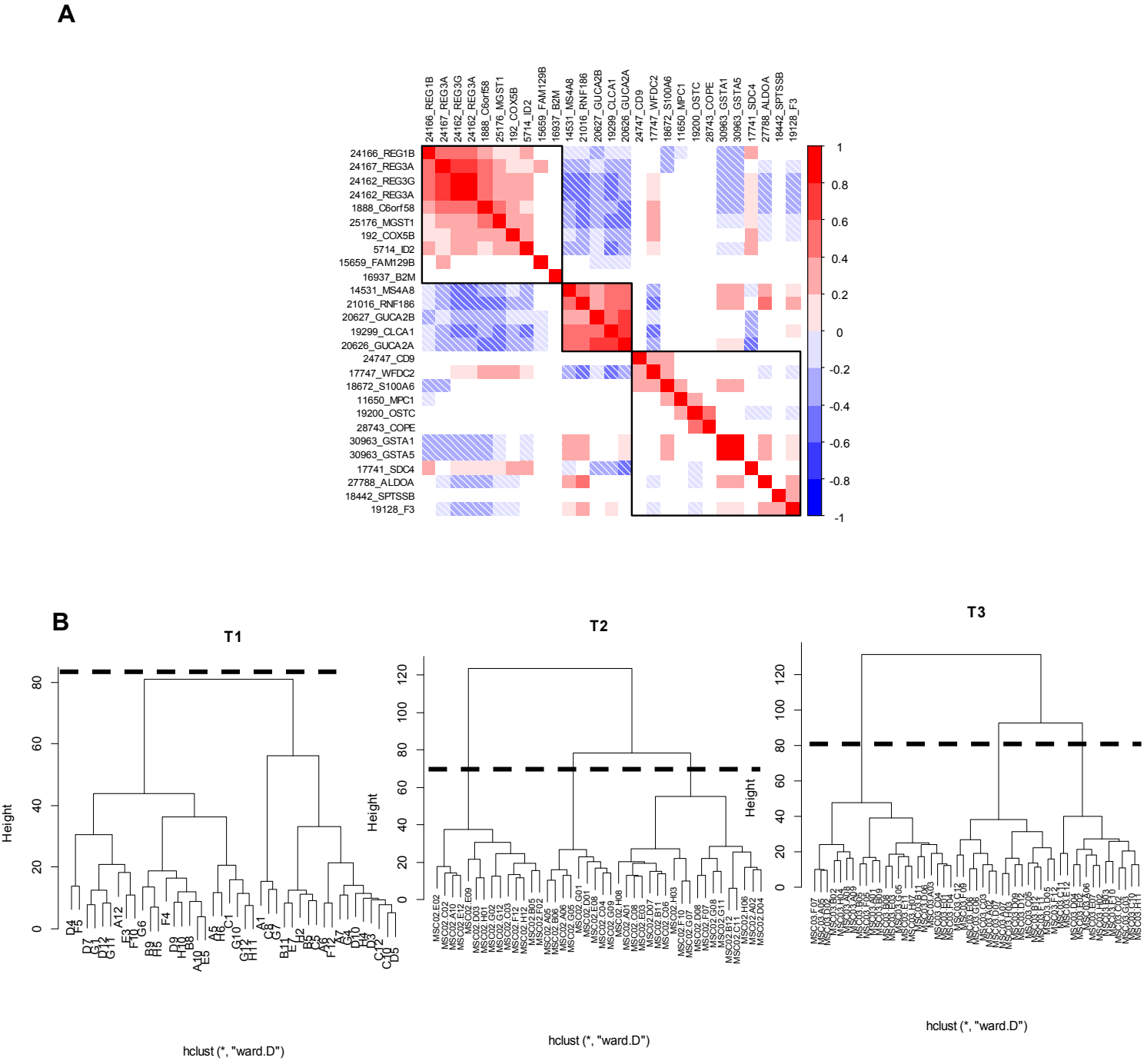
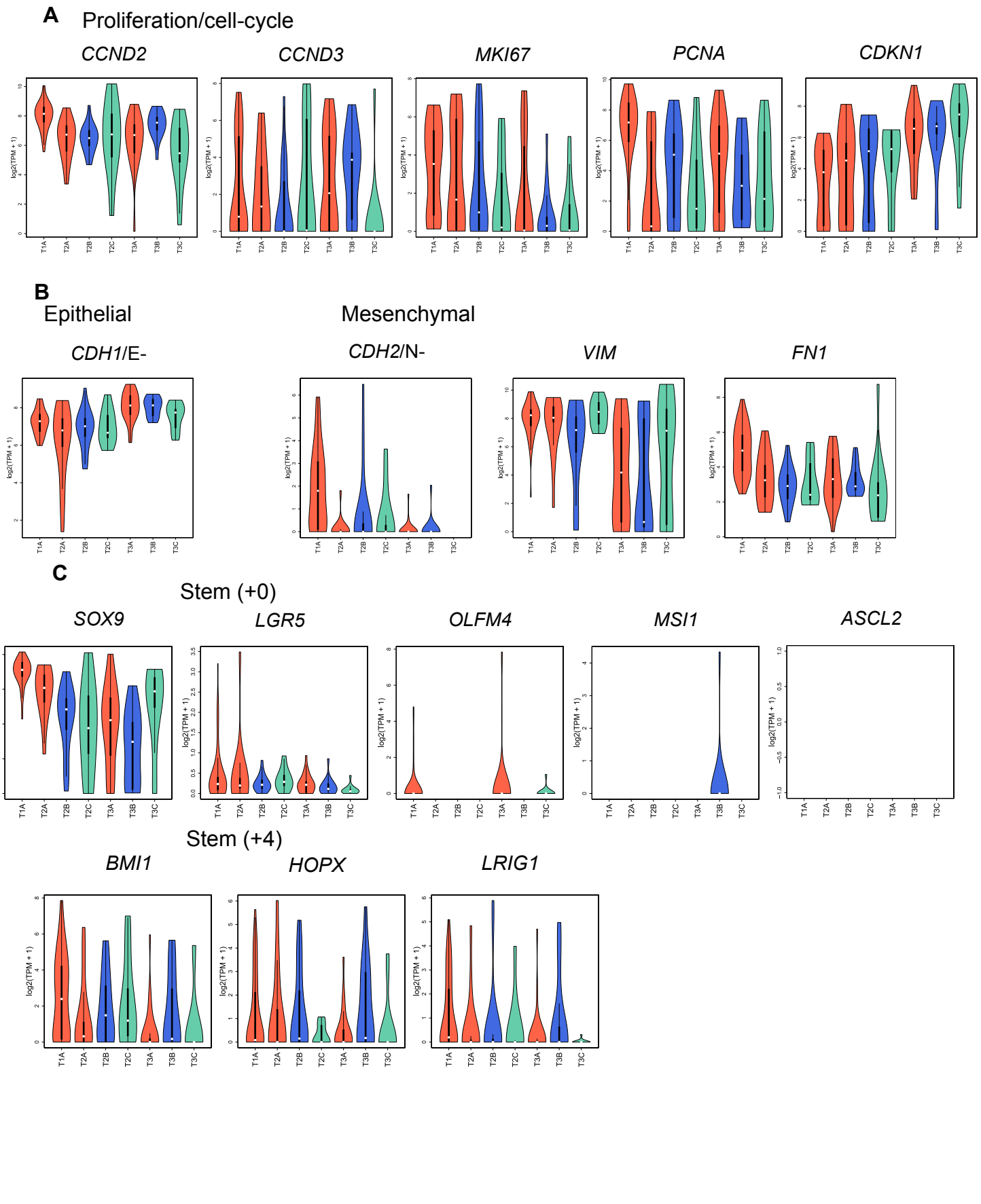
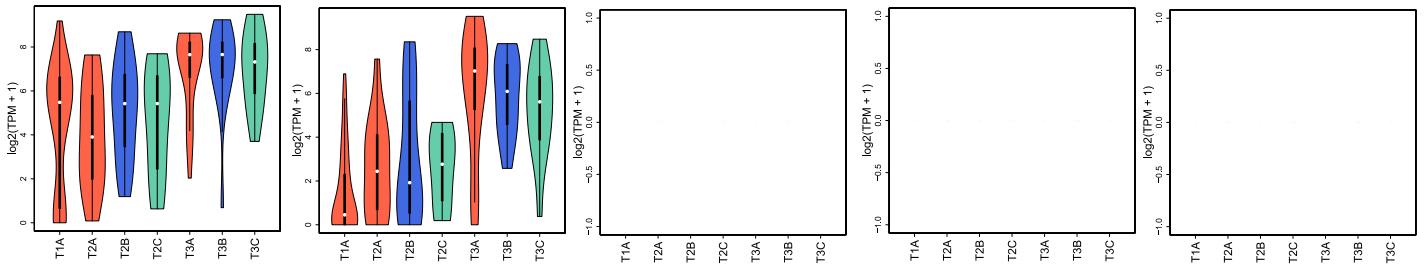


Figure S4 Determination of gene and cell groups in single-cell RNA sequencing. (A) Correlation plot of highly variable genes. (B) Dendrogram of single cells.



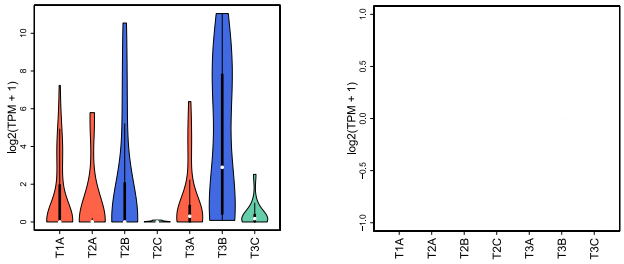
Differentiation (Absorption)

CEACAM1 *KRT20* *AQP8* *CA1* *SLC26A3*



Differentiation (Secretion)

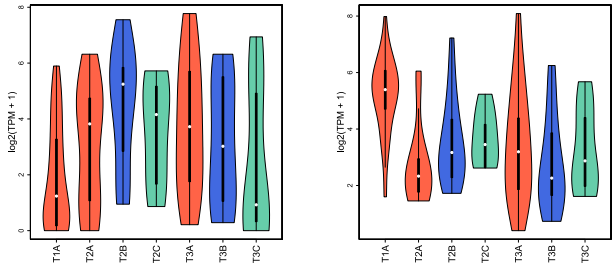
MUC2 *SPINK1*



D

Drug efflux

ABCB1 *ABCE1*



E

Glycolysis

PDK1

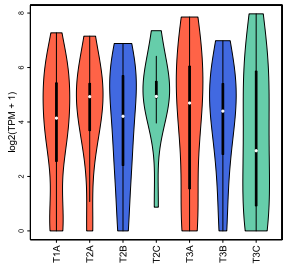
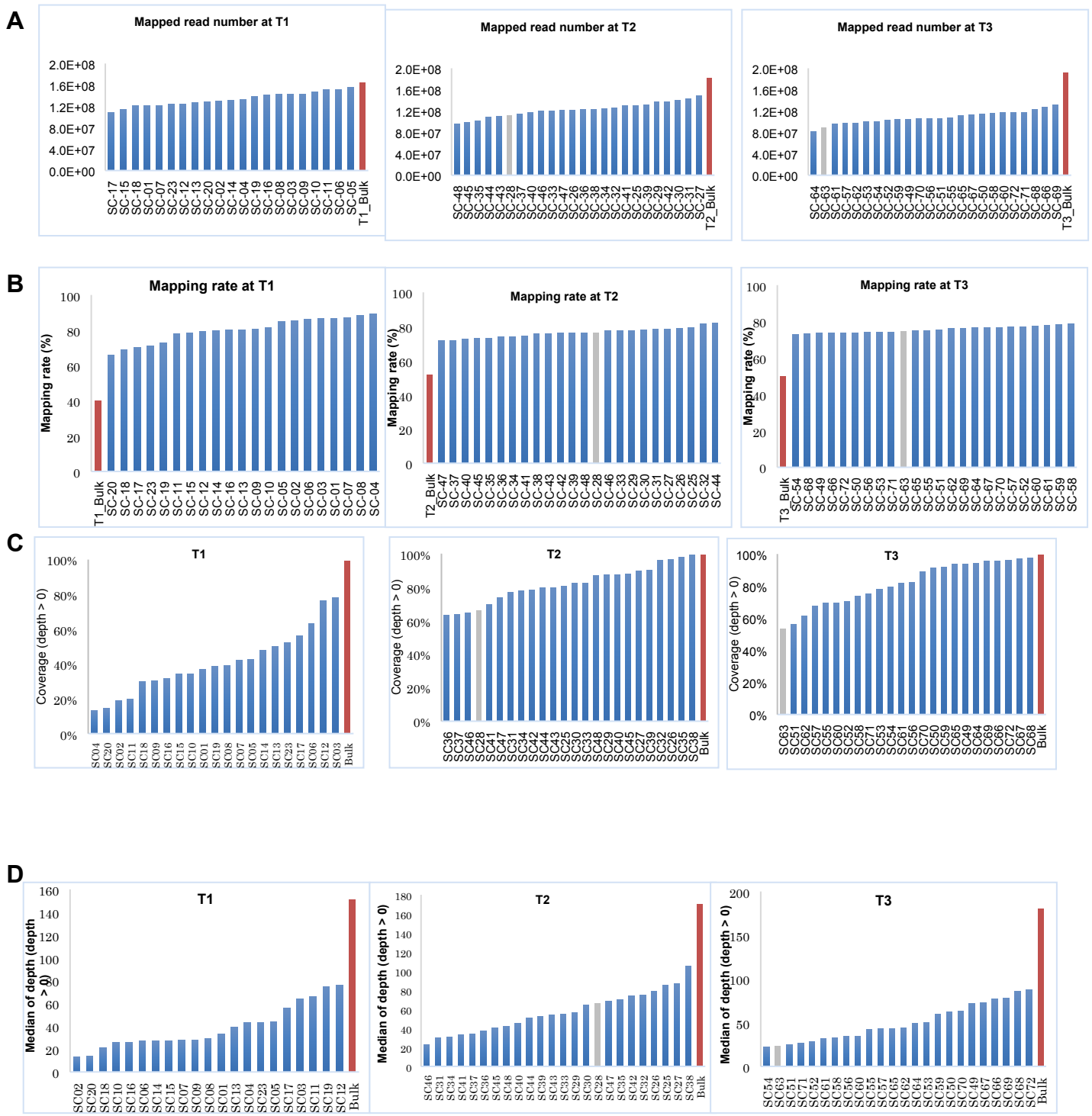
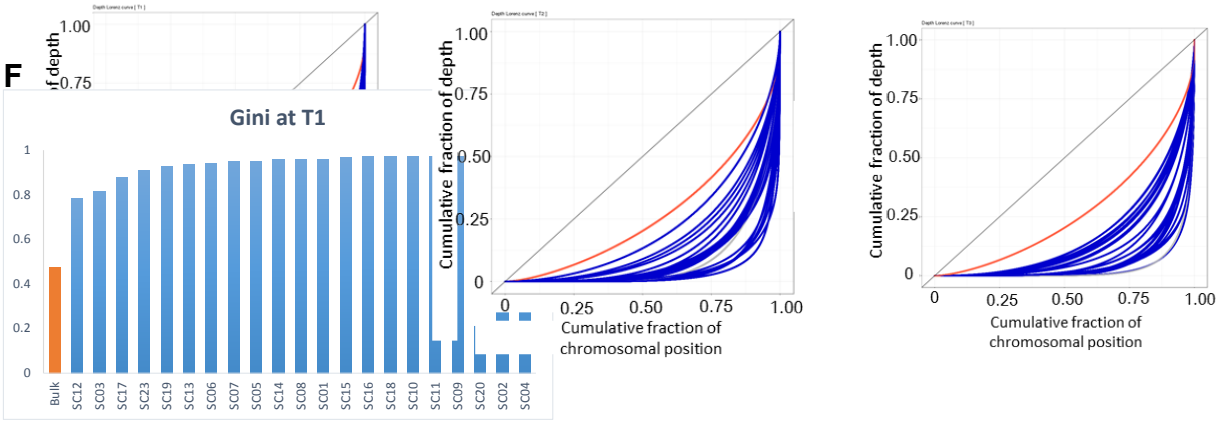


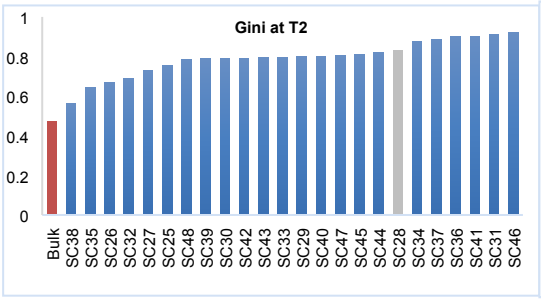
Figure S5 Violin plots of the expression levels of the marker genes. "A," "B," and "C" followed by T1/T2/T3 represent Anti-Epithelial, cGMP/GC, and Dormant cell groups, respectively (*n*: 42 for T1A, 14 for T2A, 19 for T2B, 9 for T2C, 22 for T3A, 16 for T3B, and 13 for T3C). Some genes such as *ASCL2* were not expressed in any category. (A) Proliferation/cell-cycle markers, (B) epithelial and mesenchymal markers, (C) stem cell and differentiation markers, (D) drug efflux markers, and (E) glycolysis markers.

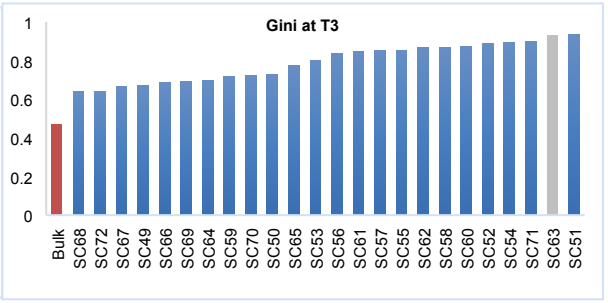


E

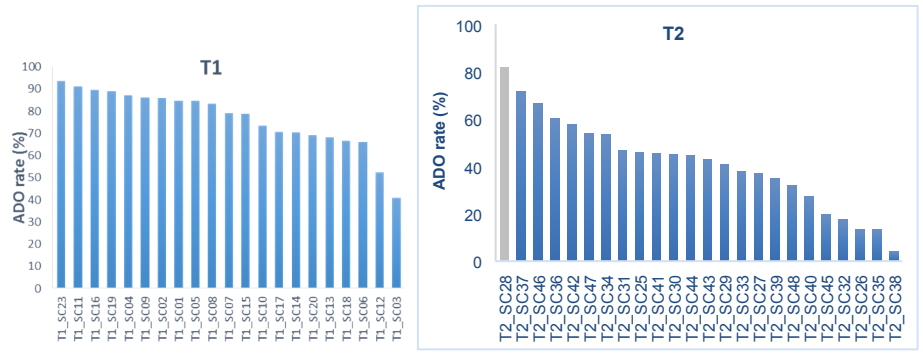


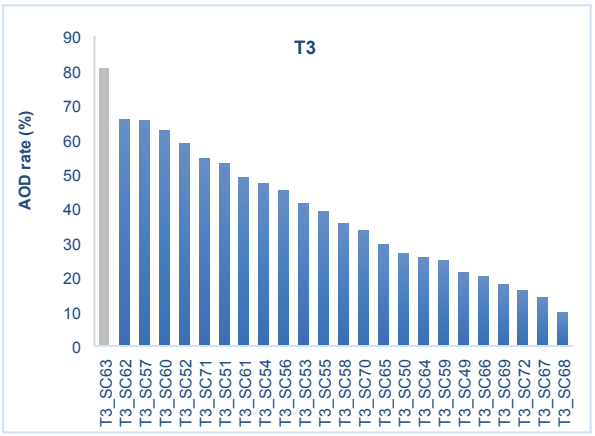
F





G





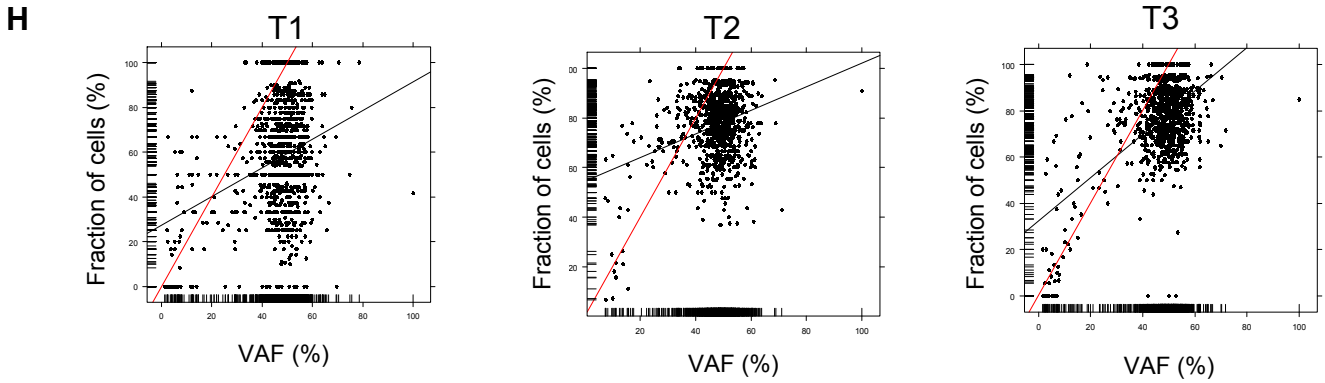


Figure S6 Quality check of single-cell exome sequencing data. Bars/curves in orange, blue, and gray represent bulk-cell, single-cell, and filtered-out data, respectively. Shown are the (A) number of mapped reads; (B) mapping rate; (C) coverage of genome with depth > 0; (D) median depth, in which regions with depth = 0 were excluded; (E) Lorenz curve of depth (including regions with depth = 0); (F) Gini coefficients of depth (including regions with depth = 0); (G) ADO rate; and (H) Scatter plot of SNVs between VAFs in bulk-cell sequencing and fractions of single cells with SNVs called in single-cell sequencing. Black and red lines represent the linear regression and theoretically expected lines, respectively.

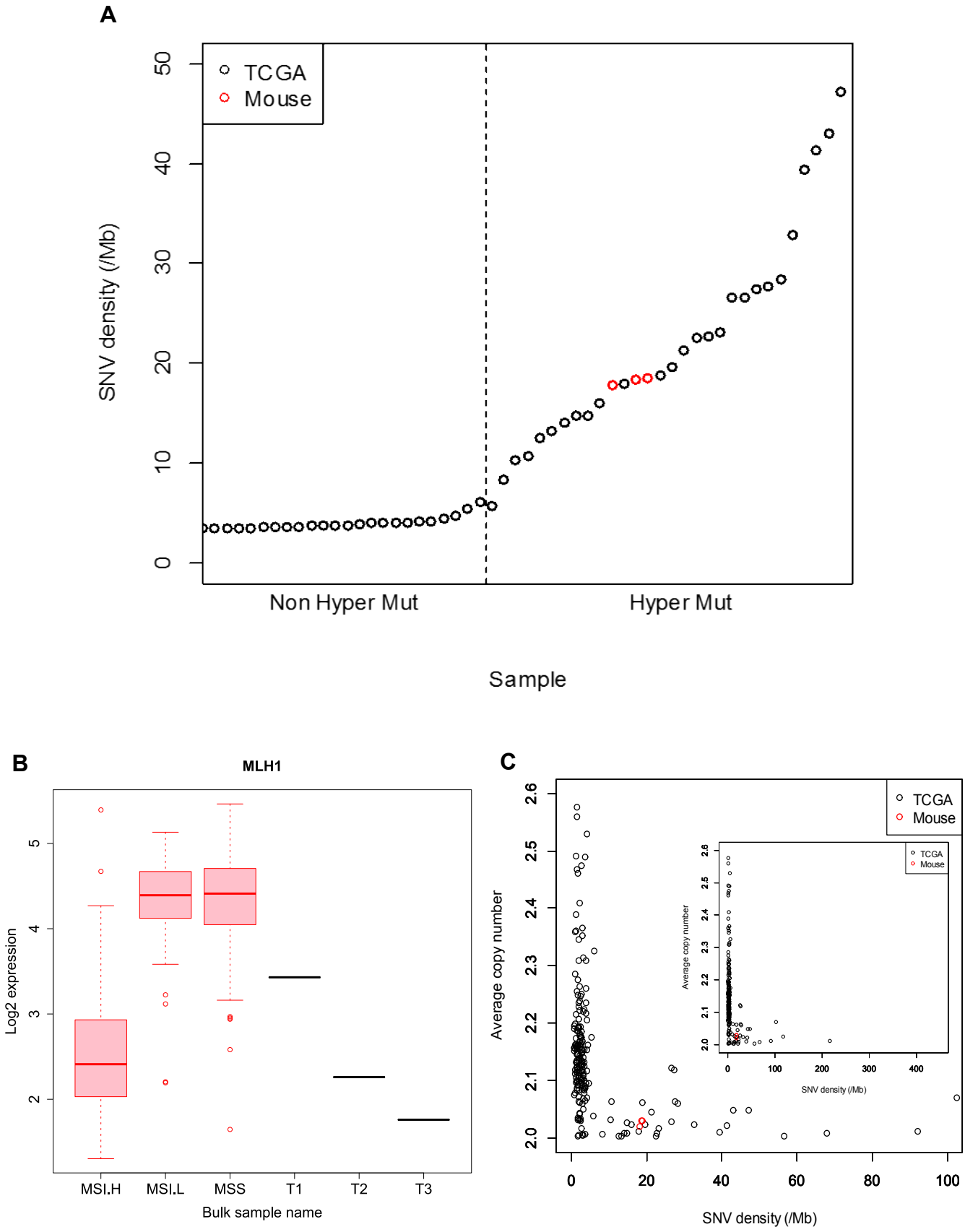
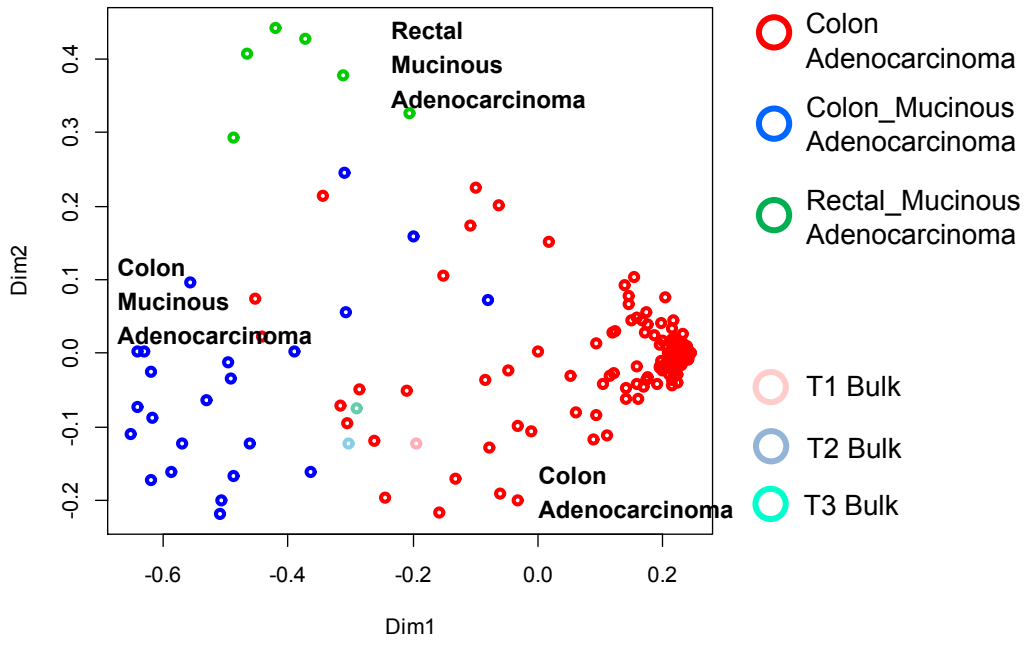


Figure S7 Human cancer counterpart to our mouse model according to molecular features. (A) SNV density in human colorectal cancer and in the mouse model. Black and red circles represent TCGA human colorectal cancer samples and mouse samples at T1, T2, and T3, respectively. Broken lines separate hyper and non-hyper mutation types. (B) *MLH1* expression in TCGA and mouse samples. MSI.H, microsatellite instability high (n = 35); MSI.L, microsatellite instability low (n = 42); MSS, microsatellite stable (n = 166). (C) Average copy number across the genome versus SNV density. Insets in panels A and C show zoomed-out views.

A



B

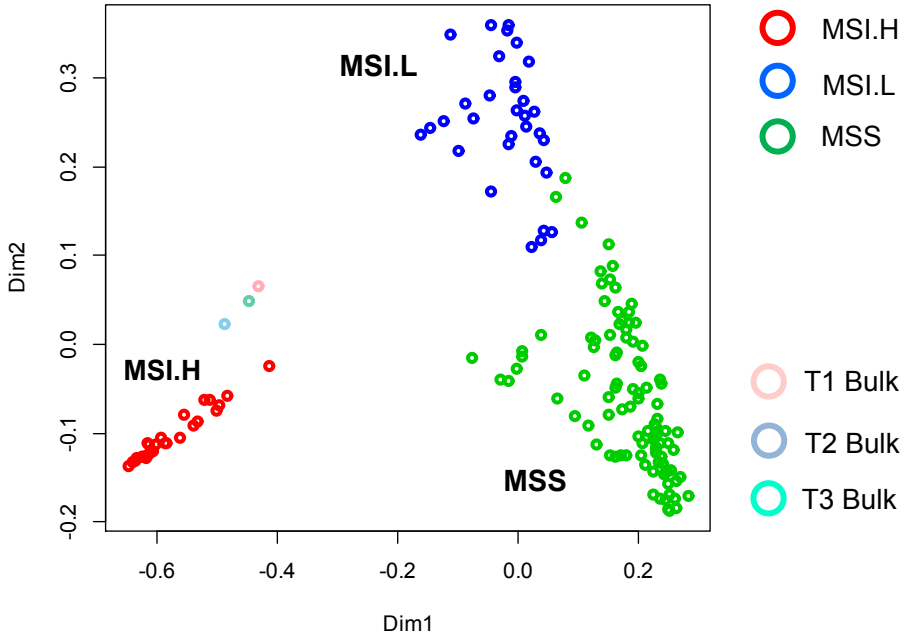


Figure S8 Human cancer counterpart to our mouse model according to clinical features. Multidimensional scaling plots generated by Random Forest based on the proximity matrix are shown. (A) For histological type. (B) For microsatellite instability. MSI.H, microsatellite instability high; MSI.L, microsatellite instability low; MSS, microsatellite stable.

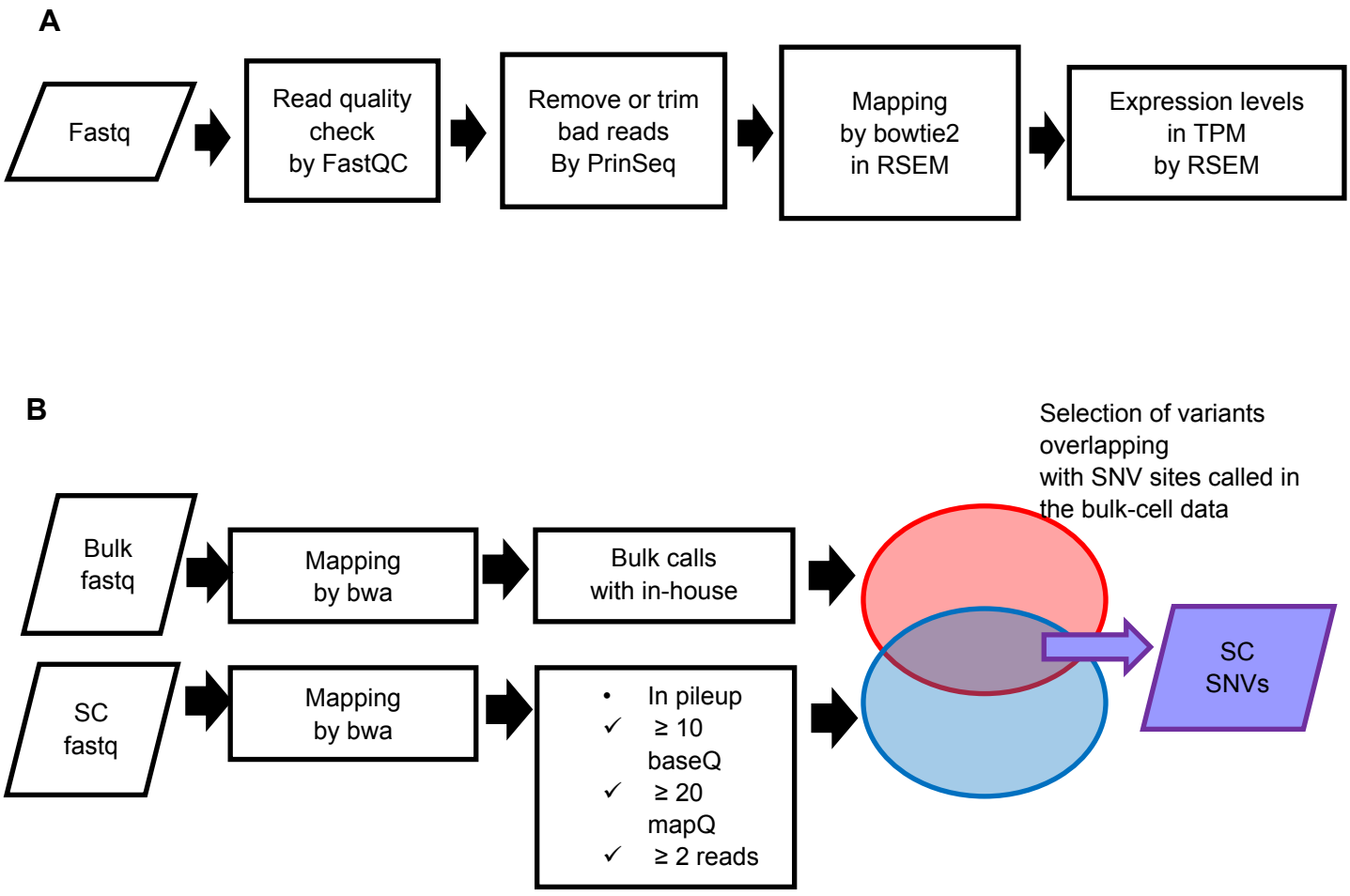


Figure S9 Procedure for calculating expression levels and for calling SNVs in single-cell sequencing. (A) Procedure for calculating expression levels (TPM). (B) Procedure for calling SNVs in single cells (SCs).