Cell Research (2014) 24:701-712.
© 2014 IBCB, SIBS, CAS All rights reserved 1001-0602/14 $ 32.00
www.nature.com/cr

ORIGINAL ARTICLE

Discovery of biclonal origin and a novel oncogene *SLC12A5* in colon cancer by single-cell sequencing

Chang Yu[1, *], Jun Yu[2, *], Xiaotian Yao[1, *], William KK Wu[2, *], Youyong Lu[3, *], Senwei Tang[1, 2], Xiangchun Li[1], Li Bao[1], Xiaoxing Li[2], Yong Hou[1, 4, 5], Renhua Wu[1], Min Jian[1], Ruoyan Chen[1, 6], Fan Zhang[1, 7], Lixia Xu[2], Fan Fan[1], Jun He[1, 2], Qiaoyi Liang[2], Hongyi Wang[8], Xueda Hu[1], Minghui He[1], Xiang Zhang[2], Hancheng Zheng[1], Qibin Li[1], Hanjie Wu[1], Yan Chen[1], Xu Yang[1], Shida Zhu[1], Xun Xu[1], Huanming Yang[1], Jian Wang[1], Xiuqing Zhang[1], Joseph JY Sung[2], Yingrui Li[1], Jun Wang[1, 9, 10, 11]

[1]BGI-Shenzhen, Shenzhen, Guangdong 518083, China; [2]Institute of Digestive Disease and the Department of Medicine and Therapeutics, State Key Laboratory of Digestive Disease, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China; [3]Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University School of Oncology, Beijing Cancer Hospital/Institute, Beijing 100142, China; [4]School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu, China; [5]State Key Laboratory of Bioelectronics, Southeast University, Nanjing, Jiangsu 210096, China; [6]Department of Peadiatrics & Adolescent medicine, The University of Hong Kong, Hong Kong, China; [7]Department of Computational Medicine and Bioinformatics, Medical School, University of Michigan, Ann Arbor, USA; [8]Department of Surgery, Peking University School of Oncology, Beijing Cancer Hospital/Institute, Beijing 100142, China; [9]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark; [10]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia; [11]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China

Single-cell sequencing is a powerful tool for delineating clonal relationship and identifying key driver genes for personalized cancer management. Here we performed single-cell sequencing analysis of a case of colon cancer. Population genetics analyses identified two independent clones in tumor cell population. The major tumor clone harbored *APC* and *TP53* mutations as early oncogenic events, whereas the minor clone contained preponderant *CDC27* and *PABPC1* mutations. The absence of *APC* and *TP53* mutations in the minor clone supports that these two clones were derived from two cellular origins. Examination of somatic mutation allele frequency spectra of additional 21 whole-tissue exome-sequenced cases revealed the heterogeneity of clonal origins in colon cancer. Next, we identified a mutated gene *SLC12A5* that showed a high frequency of mutation at the single-cell level but exhibited low prevalence at the population level. Functional characterization of mutant *SLC12A5* revealed its potential oncogenic effect in colon cancer. Our study provides the first exome-wide evidence at single-cell level supporting that colon cancer could be of a biclonal origin, and suggests that low-prevalence mutations in a cohort may also play important protumorigenic roles at the individual level.

***Keywords*:** single-cell sequencing; colon cancer; *SLC12A5*; biclonal; oncogene
Cell Research (2014) **24**:701-712. doi:10.1038/cr.2014.43; published online 4 April 2014

Introduction

Colon tumorigenesis follows a distinct progression pattern involving a spectrum of histopathological stages, ranging from single crypt lesions to invasive adenocarcinomas [1]. The most prevalent genetic alterations identified in colorectal cancer are *APC*, *KRAS* and *TP53* mutations and loss of 18q [2, 3]. However, the delineation of genetic abnormalities in malignancy is often confounded

*These five authors contributed equally to this work.
Correspondence: Jun Wang[a], Yingrui Li[b], Joseph JY Sung[c]
[a]Tel: +86 10 80481662; Fax: +86 10 80498676
E-mail: wangj@genomics.org.cn
[b]Tel: +86 10 80481662; Fax: +86 10 80498676
E-mail: liyr@genomics.cn
[c]Tel: +852-37636103; Fax: +852-21445330
E-mail: jjysung@cuhk.edu.hk
Received 17 September 2013; revised 22 January 2014; accepted 26 February 2014; published online 4 April 2014

by genetic heterogeneities, which can occur at both intratumoral and population levels. The former may give rise to phenotypic heterogeneity that is pertinent to disease progression (e.g., metastasis) and chemotherapy response [4-6]. In this regard, tumor progression has long been conceived as genetic evolution that follows the Darwinian logic (i.e., the diversification of heritable types tested by natural selection) [7]. According to this model, tumor arises as a consequence of the sequential accumulation of genetic or epigenetic alterations. When subpopulations of cells within the tumor have acquired additional mutations to key driver genes, the fast-growing subclones will outgrow other cells and dominate the tumor, progressing towards an invasive phenotype. The complexity of intratumoral heterogeneity is further increased by the possible violation of the paradigm of monoclonal tumor origin, which stipulates that multiple clones of cells in a malignancy are all derived from a single progenitor or "stem" cell. By examination of X-linked markers in aberrant crypt foci, colonic polyps and colorectal cancer, polyclonality in colon tumorigenesis has been suggested [8-10]. In 2010, Thirlwell *et al.* [11] reported that, through sequencing of *APC*, both familial and sporadic colorectal adenomas could be derived from polyclonal origins. Nevertheless, genome- or exome-wide evidence of polyclonality in colorectal cancer has not yet been provided.

In addition to intratumoral variations, the occurrence of these genetic alterations varies greatly among individuals, signifying the genetic heterogeneity at the population level [12]. To this end, only a few mutated genes (e.g., *APC*, *KRAS* and *TP53*), known as gene mountains, are freqeuntly shared among patients with colorectal cancer, whereas many individual patients harbor their own private mutations. Despite advances in our knowledge of its etiology and pathogenesis, challenges remain to understand the inter-individual variations in the molecular pathogenesis of colorectal cancer and to determine whether private mutations could take part in tumor initiation, progression or regulation of responsiveness or resistance to anticancer agents.

The newly developed single-cell sequencing technology is a promising systematic and comprehensive approach for unraveling intratumoral genetic changes, reconstructing tumor clonal architectures and identifying common mutations located in tumor subgroups [4, 13, 14]. In the present study, we applied multiple displacement amplification (MDA)-based high-throughput single-cell sequencing to carry out single-cell analysis of a colon cancer specimen in order to reveal the intratumoral genetic complexity and delineate the clonality of the tumor of this patient. We also determined the functional importance of a novel activating mutation in *SLC12A5*, a gene mutated at low frequency at the population level but at high frequency among tumor cells within our single-cell sequenced sample. The results support our proposition that private mutations could be of functional significance.

Results

Quality assurance for single cell isolation and sequencing

Fresh tumor and adjacent normal tissues were taken from a 60-year-old Chinese male patient with colon adenocarcinoma classified as T3N0M0. To obtain detailed cellular genetic information on this tumor, we carried out single-cell sequencing in individual cells from the tumor and adjacent normal samples as previously described [14]. Exome capture was performed on the whole-genome amplification (WGA) products of each cell. The resulting libraries were then subject to Illumina sequencing. In total, 63 single tumor cells that passed the criteria were selected for subsequent analysis. An average sequencing depth in exome regions of the single cells derived from the tumor specimen is 16.6-fold, totaling a comprehensive dataset of approximately 1 061-fold coverage, which enables genotype calling on most sites in exome regions for each cell (Figure 1A). Other isolated tumor-derived cells were discarded because of amplification and/or hybridization failures to avoid errors in results analysis. To avoid variation calling biases in single-cell sequencing, DNA from tumor tissue and adjacent normal tissue were subject to conventional whole-tissue exome sequencing, with 135-fold in tumor and 88-fold in normal tissue of haploid exome data generated to both cover more than 99% of target regions. With specific stringent criteria based on our false-positive and false-negative estimations (Figure 1B) to eliminate bias from the MDA process, 192 somatic mutations were ascertained from single-cell exomes (Supplementary information, Table S1). After subtraction of somatic mutations that occurred at dbSNP v130 loci, 176 somatic mutations were regarded as novel. Ninety-eight percent mutant alleles ascertained from single cells were supported by reads from exome sequencing of whole-tissue sample. Comparison of somatic mutation frequency from the single-cell sequencing data with that of the whole-tissue sequencing showed a high correlation ($r^2 = 0.7$, $P < 2.2e{-}16$) (Figure 1C). These data collectively indicate that a high accuracy of mutation calling could be achieved using the single-cell dataset.

Identification of two subsets in tumor cell population by population genetics analyses

To explore the intercellular heterogeneity structure in

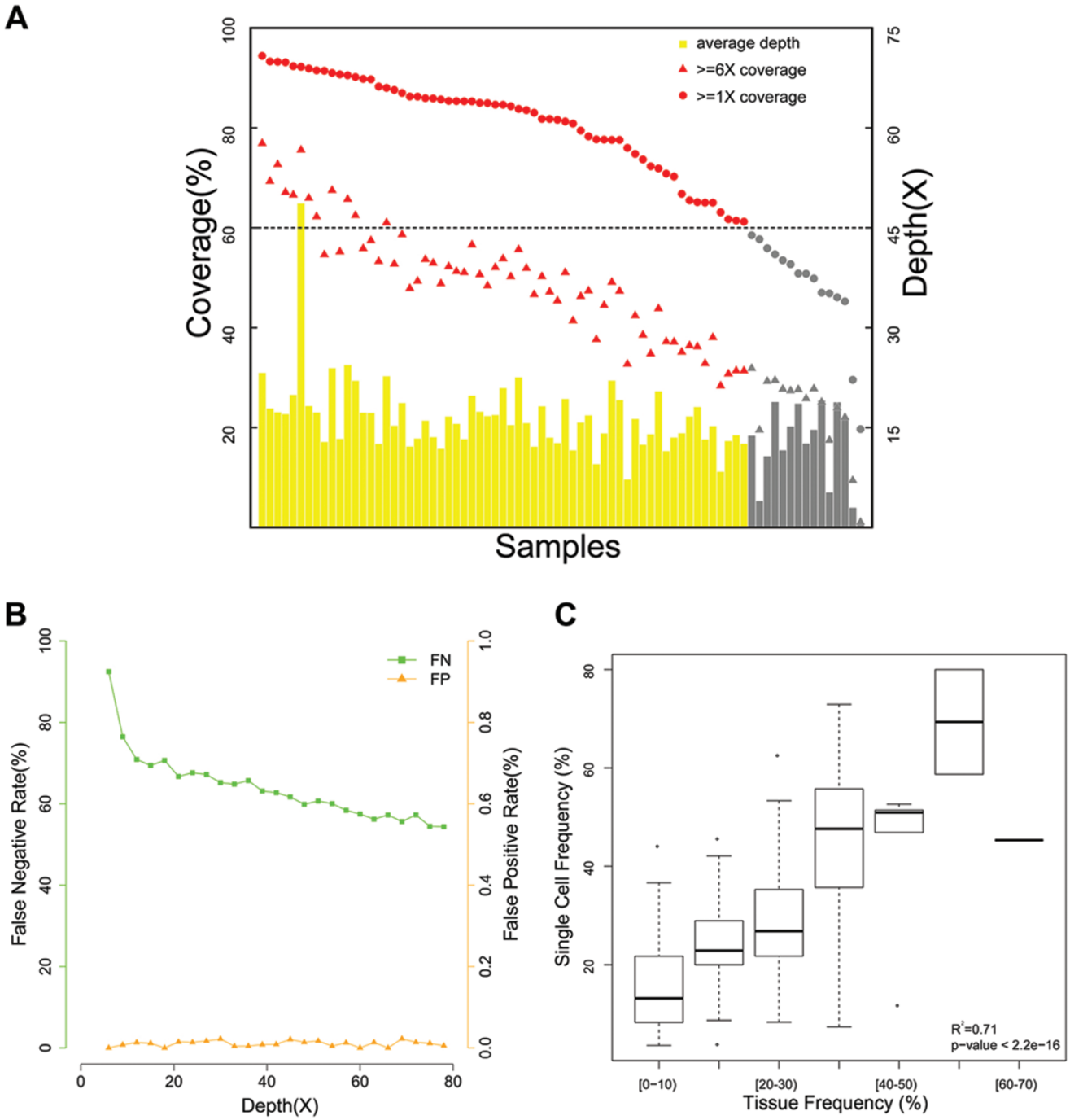

Figure 1 Quality assurance of single-cell sequencing data. **(A)** The distribution of sequencing depth and coverage of isolated colon cancer single cells. Sixty-three single cells that reached at least 60% coverage were chosen for somatic mutation calling and population genetics analysis. **(B)** False-positive and false-negative rates of single-cell sequencing data were estimated by homozygous and heterozygous allele dropout rates, respectively. **(C)** Box plots of somatic mutation frequency showing positive correlation between allele frequencies obtained from single-cell and whole-tissue sequencing data.

this colon cancer tissue, we applied population genetics analyses on the comprehensive dataset. A total of 106 non-synonymous mutations were subject to principal component analysis (PCA) to characterize the genetic heterogeneity of this tumor (Figure 2A). We found that two first components (PC1 and PC2), which represented 45% of the whole inertia (Supplementary information, Table S2), clearly divided isolated single cells into three

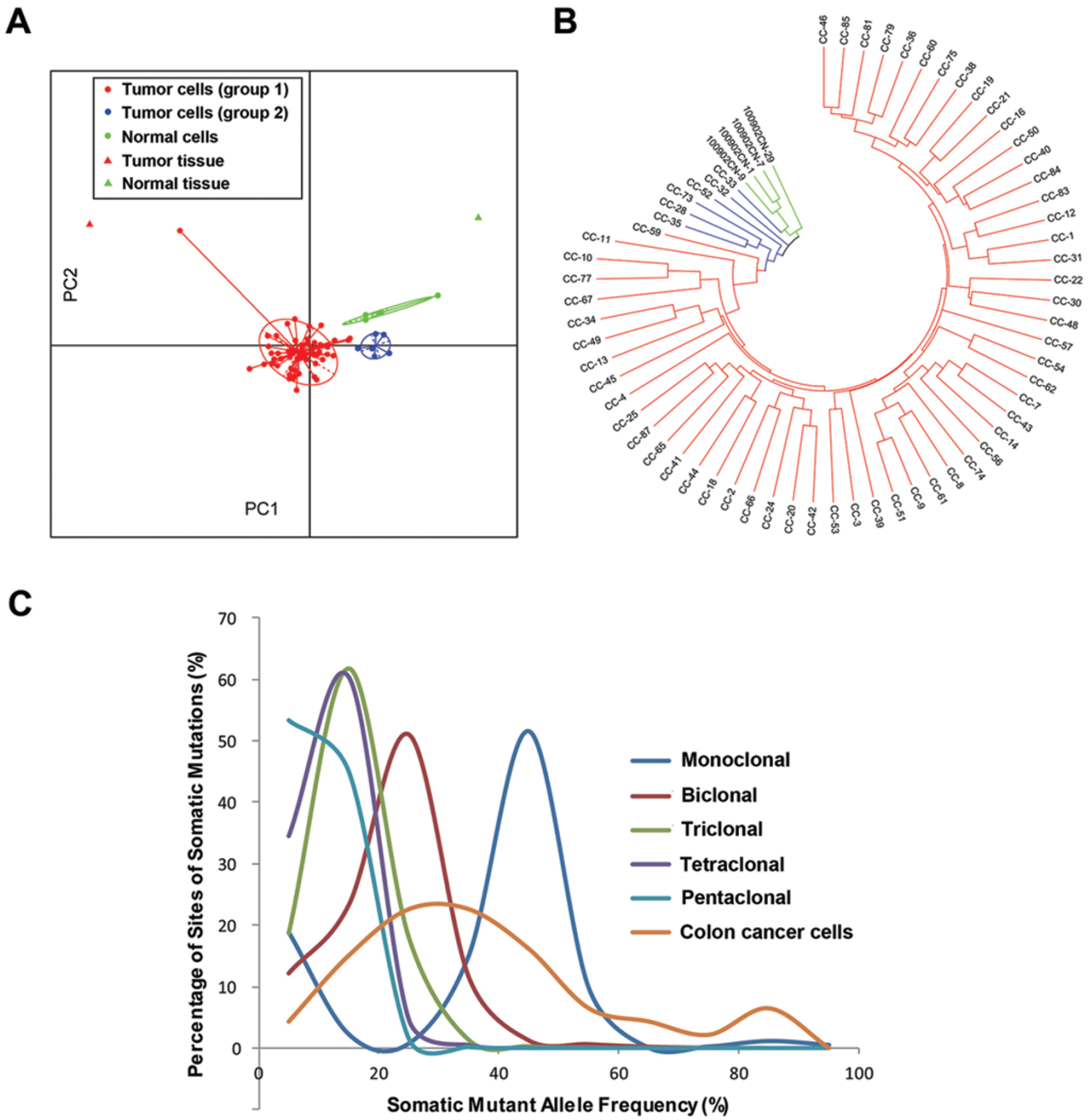

Figure 2 Reconstruction of clonal structure in a case of colon cancer by population genetics analyses. **(A)** Principle component analysis (PCA) of non-synonymous mutations in isolated cells and whole tissues was performed. The value of $\log_2$(Pr(mutated)/Pr(normal)) for each non-synonymous mutated site was used. **(B)** The same data were used to compute the phylogenetic relationship of each cell pairs to build the UPGMA tree. **(C)** The simulated SMAFS of tumors originating from 1 (monoclonal) to 5 (pentaclonal) progenitor cells were compared with the observed SMAFS of sequenced single tumor cells (orange), which was similar to the predicted SMAFS pattern of a tumor of biclonal origin.

non-overlapping subgroups, namely, normal cells and two tumor cell subsets. This observation indicated that the tumor cell population was composed of two groups of cells, each with distinct genetic features. Next, to infer the population substructure and estimate the appearance of somatic mutations in individual cells relative to one another, we constructed a phylogenetic tree based on UPGMA (Unweighted Pair Group Method with Arithmetic

Mean) [15], which also clustered the isolated cells into three subgroups (Figure 2B). To better understand the intratumoral architecture and infer the potential clonal relationship, we turned to derive the somatic mutant allele frequency spectrum (SMAFS) of isolated tumor cells (Figure 2C). We simulated the SMAFS of tumors originating from one to five progenitor cells as previously described [13] and compared them with the actual data, which revealed a peak in the frequency window of 20%-30% resembling the predicted pattern of a tumor of biclonal origin.

Distinct cellular origins of two subsets revealed by clustering of driver events

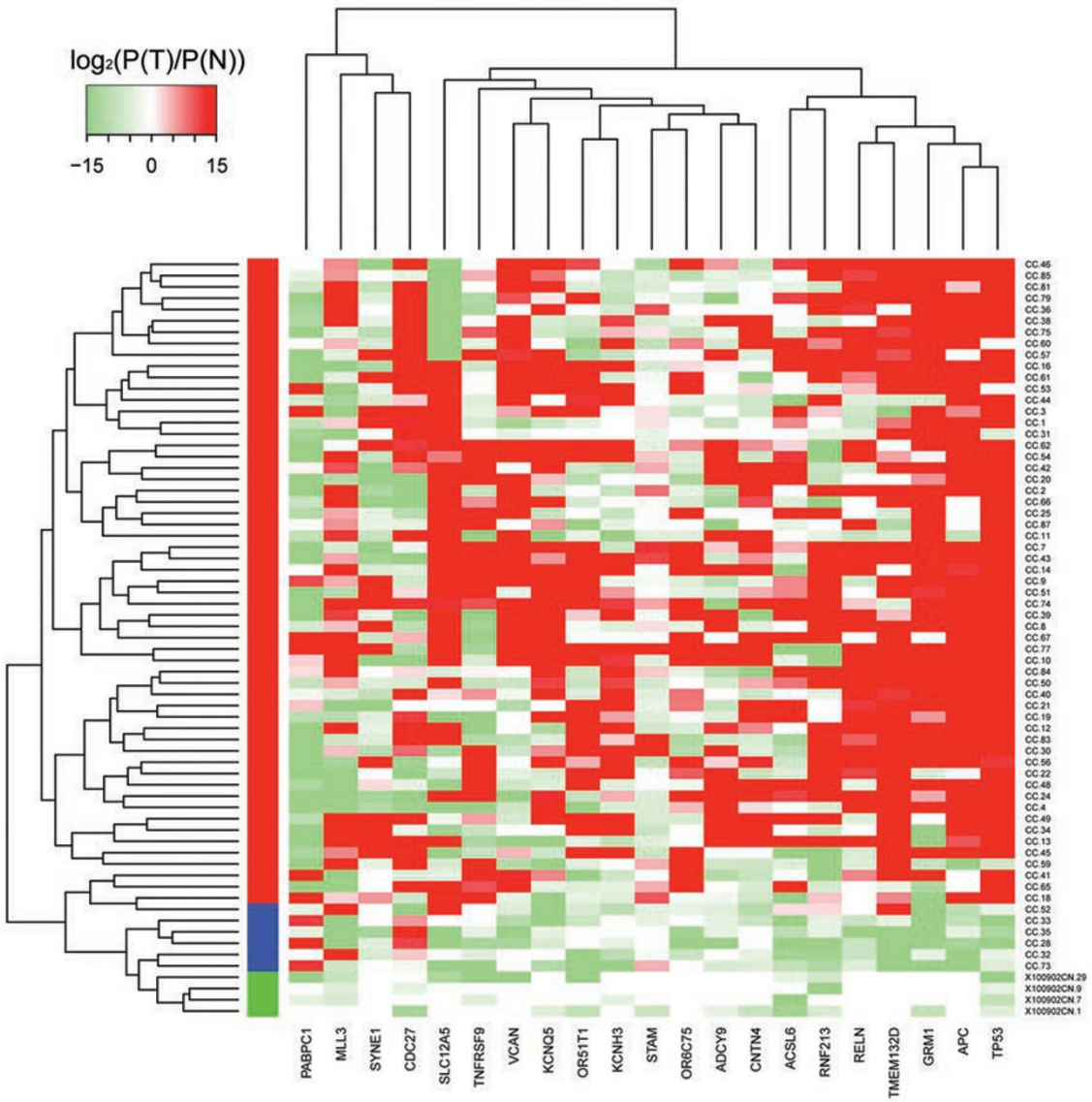

Figure 3 Potential driver events in two independent clones of tumor cells isolated from a case of colon cancer. Unsupervised two-way hierarchical clustering of potential cancer driver genes identified by both single-cell sequencing and a separate cohort study of 21 colon cancer cases distinguished two clones of tumor cells. The major clone (red) was characterized by prevailing *TP53* and *APC* mutations, whereas the minor clone (blue) contained mutations in two cancer-related genes, *CDC27* and *PABPC1*. Isolated normal cells (green) did not harbor any detectable mutation. The *TP53* and *APC* mutations were shared by cells in the major clone (red) but not by cells in the minor clone (blue), which strongly suggests that the tumor is of biclonal origin. The clustering analysis *per se* does not imply a known ancestral root. Color intensity in the clustering analysis corresponds to the $\log_2$ ratio of the probability of being mutated to that of being wild type. Red color denotes the high likelihood of being mutated whereas green color refers to the opposite.

To identify potential cancer driver genes that play a causal role in colon tumorigenesis in this individual, we compared our single-cell data with that of an additional cohort study (unpublished data) comprising whole-tissue exome-sequencing analyses of 21 Chinese colon cancer patients. Hierarchical clustering of overlapped genes divided the isolated cells into three subgroups (i.e., normal cells and two tumor cell subsets). The major tumor cell subset harbored *TP53* and *APC* mutations as early events. Although mutations of *TP53* and *APC* were not detected in all cells in the major clone, this could be caused by the failure to cover these two genes during sequencing. A high proportion of these cells also contained somatic mutations in other cancer-related genes, such as, *GRM1* and *RELN* whose genetic alterations have been reported in other human tumors [16, 17]. In contrast, the minor tumor cell subset predominantly contained *CDC27* and *PABPC1* mutations (Figure 3). Concordantly, both *CDC27* and *PABPC1* have been reported as tumor suppressors in other cancer types [18, 19]. The absence of *TP53* and *APC* mutations as well as other genetic alterations in the minor cell subset strongly supports that these

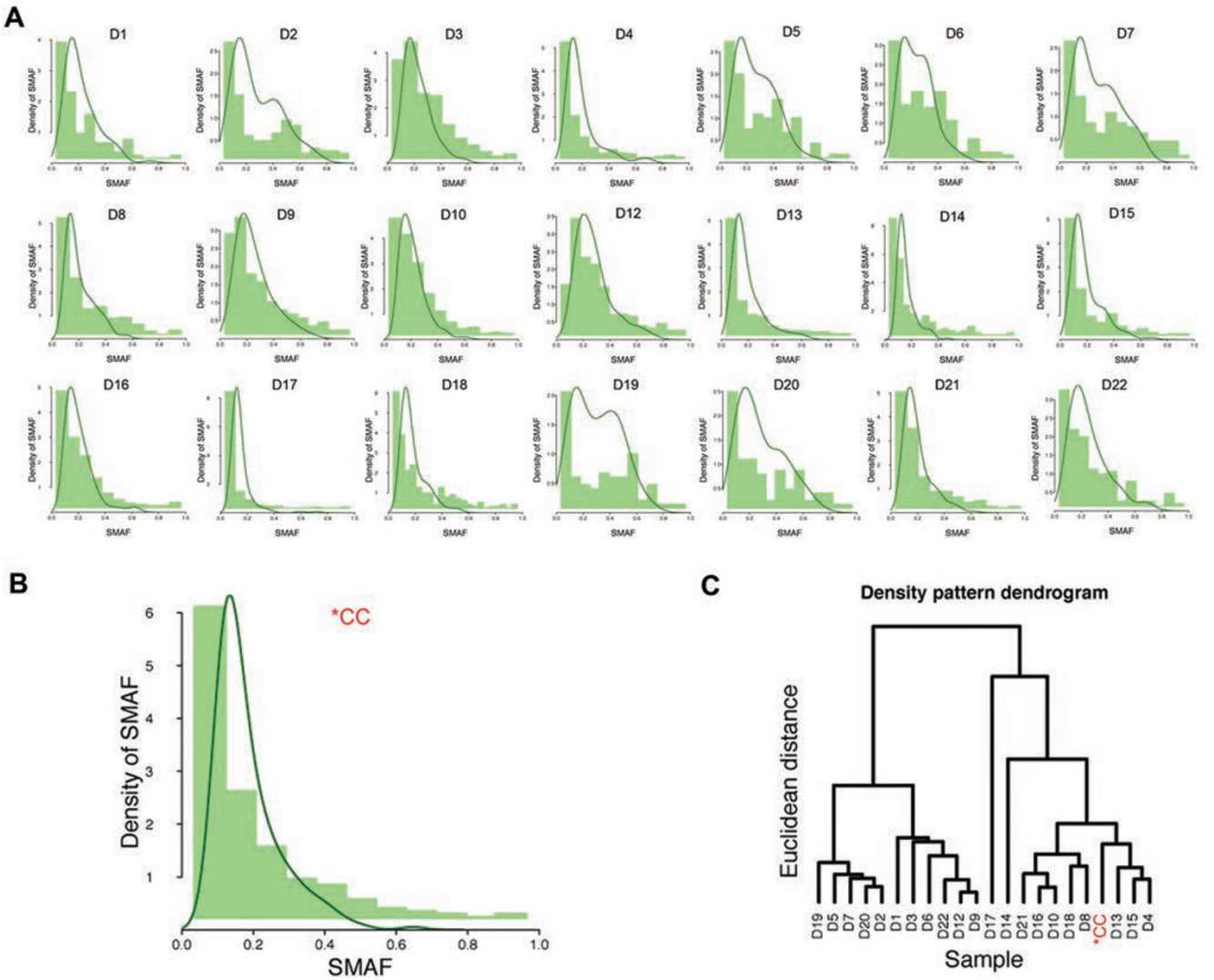

Figure 4 Heterogeneity of clonal origin in colon cancer. **(A)** The density distribution of SMAFS of 21 whole-tissue exome-sequenced colon cancer was analyzed by kernel density estimate (KDE). Tumors of different clonality were expected to give rise to different KDE patterns. **(B)** KDE plot of the same tissue specimen (*CC) that was subject to single-cell sequencing was extracted from whole-tissue exome-sequencing data. Similar to SMAFS derived from single-cell sequencing, this specimen exhibits a peak in the frequency window of 20%-30% in the KDE plot of whole-tissue exome-sequencing data. **(C)** Hierarchical clustering was performed to segregate colon cancer samples into subgroups based on KDE patterns. It is inferred that specimens D4, D13 and D15 clustered along with *CC are of possible biclonal origin.

two tumor cell subsets were indeed derived from two cellular origins at a point before the occurrence of *TP53* and *APC* mutations.

Heterogeneity of cellular origin in colon cancer

To further examine whether polyclonality is a common feature of colon cancer, we derived the SMAFS patterns from the 21 exome-sequenced colon cancer cases. Although the SMAFS patterns of whole-tissue exome-sequencing data could not be used to infer the exact clonal origins in these samples, we observed multiple recurrent SMAFS patterns among the 21 cases (Figure 4A). Consistent with the single-cell dataset, the SMAFS pattern derived from whole-tissue exome sequencing of the same specimen (denoted as *CC in Figure 4B) was characterized by a distinct peak in the frequency window of 20%-30%. Hierarchical clustering of these SMAFS patterns revealed that clonal origin was heterogeneous at the population level in colon cancer (Figure 4C).

Identification of cancer driver genes

Next, we sought to identify somatic mutations that

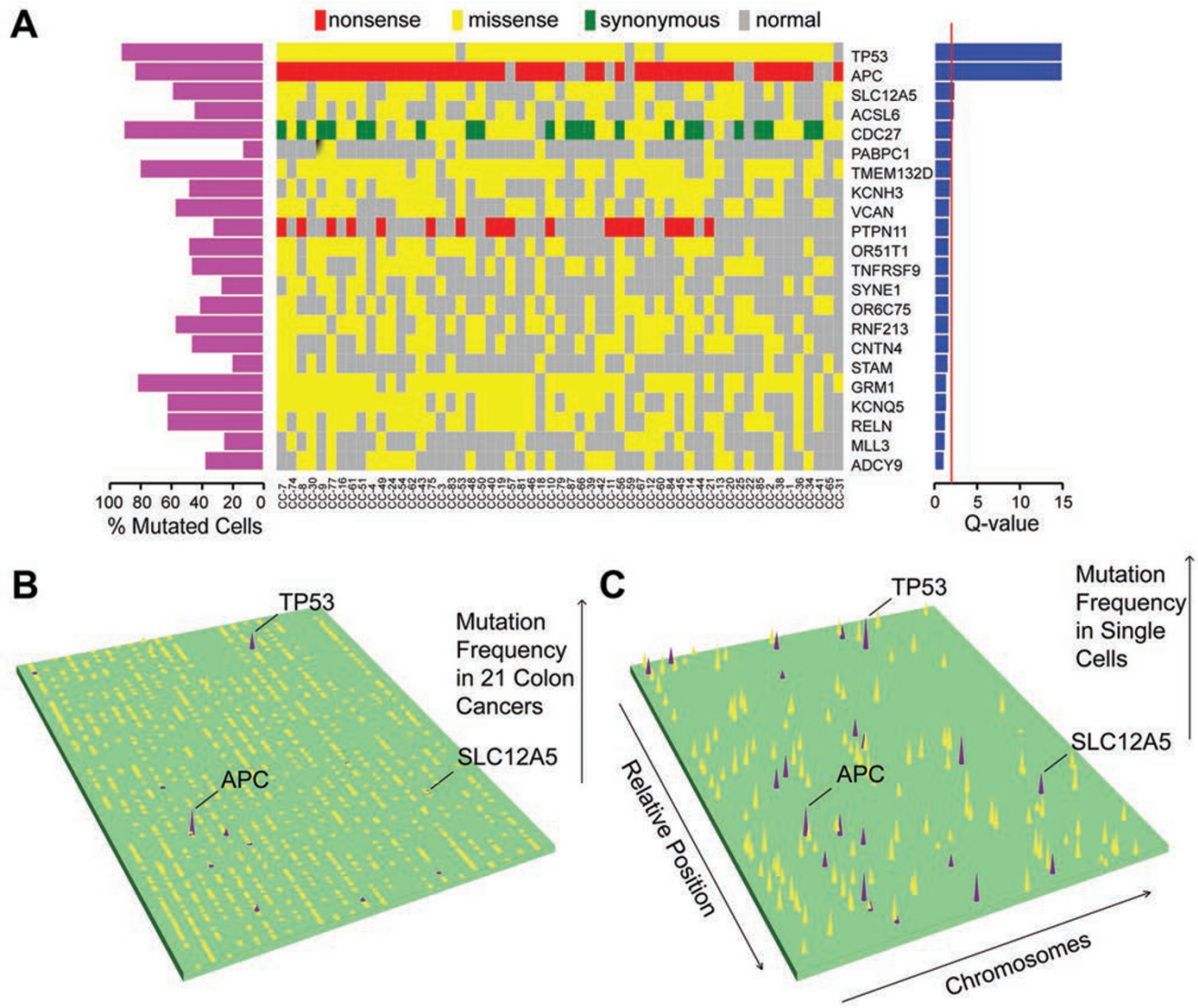

Figure 5 The mutation landscape and driver-gene prediction. **(A)** The matrix in middle shows mutations grouped by mutation types, which were identified in the major clone. The length of left bar represents the mutation frequency of single cells. The length of right bar represents the predicted driver gene Q-score. **(B)** The somatic mutation landscape in the additional colon cancer cohort study. The height of each gene reflects the mutation frequency in these cancers. **(C)** The somatic gene mutation landscape in single cells. The genes with estimated Pr(mutated)/Pr(normal) > 9 were considered as mutated genes.

had a higher likelihood of having functional impacts. To do so, we calculated the statistical significance of the observed mutation prevalence for each gene in the context of the background mutation rate and gene sequence length (Figure 5A). The mutational landscapes at both individual and population levels were depicted (Figure 5B and 5C). Such an approach shortlisted several genes, including two genes (i.e., *APC* and *TP53*) that have been

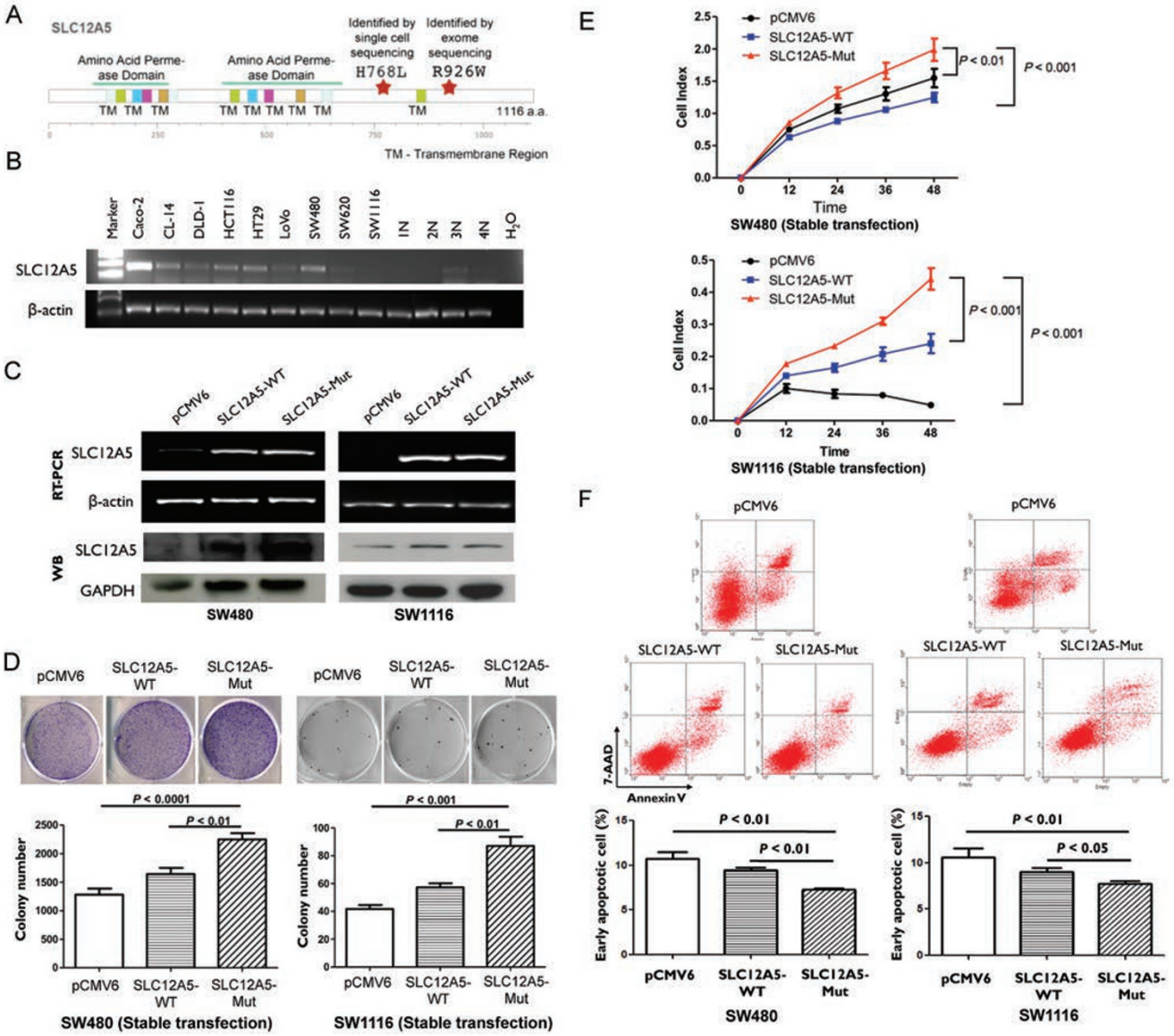

Figure 6 Oncogenic function of SLC12A5 H768L mutant. **(A)** A schematic view of the domain structure of SLC12A5 and the site of point mutation (H768L) detected by single-cell sequencing. Another mutation site R926W detected by whole-tissue exome sequencing in 1 out of 21 cases of colon cancer was also illustrated. **(B)** The overexpression of SLC12A5 mRNA in colon cancer cell lines compared with normal colonic tissues (1N-4N) was demonstrated by RT-PCR. **(C)** Ectopic expression of SLC12A5 in colon cancer cell lines SW480 and SW1116 stably transfected with pCMV6 (vector), wild-type pCMV6-SCL12A5 (WT) and pCMV6-SLC12A5 H768L (Mut) was evidenced by RT-PCR and western blot. **(D)** The effect of stable overexpression of wild-type or mutant SLC12A5 on cancer cell proliferation was determined by colony formation assay. Mutant SLC12A5 significantly enhanced clonogenicity in SW480 (left) and SW1116 (right) cells compared with the wild-type SLC12A5 and empty vector control. **(E)** The effect of stable expression of wild-type or mutant SLC12A5 on colon cancer cell proliferation was determined by the xCELLigence System. SLC12A5 H768L mutant significantly promoted cell proliferation in SW480 and SW1116 cells compared with the wild-type SLC12A5 and empty vector control. **(F)** Annexin V/7-AAD analysis by flow cytometry revealed that overexpression of mutant SLC12A5 significantly reduced basal levels of early apoptotic cells in SW480 and SW1116. Data are means ± SD of three separate experiments in triplicates.

well characterized as driver genes in colon carcinogenesis. Consistent with earlier colon cancer studies [20], *TP53* had the highest prevalence of protein-altering mutations identified in cell population of the major clone (53/57 = 93%), in which an arginine-to-cysteine mutation was detected at codon 273, a hotspot in colon cancer [20]. A novel truncation mutation G1288* in *APC* was also detected in 48 (84%) cells in the major clone. The mutation site was in close proximity to the "mutation cluster region" (codon 1300), in which truncation mutations have been shown to abrogate the β-catenin-binding activity of APC [21]. When we assessed the landscape at an individual-cell level within the single tumor tissue, we found that only a few genes (13/186 = ~7.0%) were shared with the cohort results. Most other mutated genes highly shared among cancer cells within this individual tumor were not present in the cohort, indicating that a driver event at an individual level may not be prevalent at the population level. Next, we validated this concept through further functional characterization.

Functional characterization of a low-prevalence mutated gene SLC12A5

To elucidate the potential oncogenic functions of rare mutations discovered by single-cell sequencing in colon cancer, we investigated a mutant gene *SLC12A5* of unknown oncogenic or tumor-suppressing function. This gene is frequently mutated among cancer cells within this individual tumor and thus predicted as a driver gene. Mutation of *SLC12A5* was also detected in 1 out of the additional 21 cases of colon cancer but occurred at a different site. The biological function of the *SLC12A5* H768L mutation as identified by single-cell sequencing was evaluated (Figure 6A). The expression of *SLC12A5* was first determined in normal colon tissues and a panel of colon cancer cell lines by RT-PCR. *SLC12A5* mRNA was overexpressed in colon cancer cell lines, suggesting a potential oncogenic role (Figure 6B). Stable ectopic expression of the mutant SLC12A5 increased colony-forming abilities and cell proliferation in two colon cancer cell lines, SW480 and SW1116, when compared with the wild-type (WT) control (Figure 6C-6E). Ectopic expression of the H768L *SLC12A5* mutant also reduced apoptosis in these two cell lines (Figure 6F). Taken together, these results suggest that the *SLC12A5* mutation identified in this study may potentially promote colon carcinogenesis through altering the balance between cell proliferation and cell death.

Discussion

Although the paradigm of monoclonal tumor origin has been upheld for decades, accumulating evidence supports that many human tumors could have a polyclonal origin [22]. Sakurazawa *et al.* [8] using a method based on X-chromosome inactivation found that aberrant crypt foci, the putative premalignant lesions of colon cancer, were a mixture of monoclonal and polyclonal lesions. Beutler *et al.* [9] also reported a case of colon cancer in which the primary tumor had a multicentric origin, whereas the liver metastases had a unicentric origin. Examination of electrophoretic variants of glucose-6-phosphate dehydrogenase isolated from colonic polyps of three patients with Gardner's syndrome, a subcategory of Familial adenomatous polyposis disease, also revealed that these lesions are multiclonal in origin [10]. Here, we presented a genome-wide evidence at single-cell level that supports a biclonal origin in this case of colon cancer. First of all, two non-overlapping subgroups were identified by population genetics analyses (i.e., PCA and phylogenetic tree analysis) in the tumor cell population. Moreover, SMAFS of isolated tumor cells was consistent with the predicted pattern of a tumor of biclonal origin. Although a slight deviation between the observed and predicted SMAFS patterns was noted, it could be caused by unequal dominance of the two tumor clones. The two clones of tumor cells also harbored distinct mutation patterns, which probably explained the unequal dominance of these two tumor cell subsets. Importantly, we observed that polyclonality in the origination of colon cancer is common based on the heterogeneity of SMAFS patterns derived from the exome-sequencing data of additional colon cancer cases.

Understanding the clonal origin of cancer is of paramount importance to disease modeling. In particular, in this case of colon cancer, two or more colonocytes in close proximity might have been exposed concurrently to the same source of insult (e.g., mutagens) that had lead to distinct genetic alterations in individual cells. It is likely that the colonocyte with *APC* and *TP53* inactivation had more oncogenic potential when compared with colonocytes with other genetic alterations (e.g., *CDC27* and *PABPC1* mutations) because *APC* and *TP53* are potent tumor-suppressors, which could explain why two tumor cell clones of unequal dominance could have arisen in a single tumor at the same time. Polyclonality also has therapeutic implications because different clones of tumor cells would have different chemosensitivity and response to targeted therapy. Further experiments, including multiple sample analysis using single-cell sequencing, would more definitely address whether polyclonality is a common phenomenon in colon cancer.

Another important feature of this work is the successful verification of the functional impact of a rare mutated gene. Historically, the vast majority of large-scale cancer genomics researches have focused on highly prevalent

mutants, most of which are shown as "mountains" in the genetic landscape [2, 3]. However, the underestimation of the importance of rare mutants may confound our understanding of the complexity of tumor evolution within individual cancer patients. The ability to grasp the non-prevalent mutants in individual tumors by single-cell sequencing may contribute to the study of biological basis of genetic heterogeneity in cancer. To explore the potential oncogenic activity of an infrequent mutant, we carried out a series of functional studies on a somatic mutation *(H768L)* in *SCL12A5*, which encodes a chloride-potassium symporter of the solute carrier family 12. This ion transporter is responsible for establishing the chloride ion gradient through maintaining low intracellular chloride concentration and plays important roles in synaptic inhibition and neuron protection against excitotoxicity [23]. Mutations in *SLC12A5* have been described in cases of ovarian and skin carcinomas (COSMIC database). We demonstrated here that activation of *SLC12A5* is a potential oncogenic driver event in colon cancer, through promoting cell proliferation and inhibiting apoptosis. Concordantly, a recent study demonstrated that *SLC12A5* could promote cervical cancer cell migration and invasion [24]. Interestingly, enforced expression of wild-type *SCL12A5* in colon cancer cells by itself could exert protumorigenic effects, which could be further enhanced by the H768L mutation. This observation suggests that *SLC12A5* might be an oncogene of which *H768L* mutation is a gain-of-function mutation. However, the mechanism by which this chloride-potassium symporter mediates its oncogenic effect in the wild-type form and in relation to the *H768L* mutation warrants further investigation. Our findings also support our proposition that non-prevalent mutations could play crucial roles in cancer development at the individual level and highlight the application of single-cell sequencing as a tool to identify key oncogenic events which would be essential to the implementation of personalized cancer therapy.

In conclusion, we demonstrated the biclonal origin in a case of colon cancer by single-cell sequencing. Rare driver events such as activating mutation of *SLC12A5* could also be identified at the single-cell level with this platform. Our results provide new insights into the specific molecular events of colon cancer development.

Materials and Methods

Sample collection and preparation of cell suspension

The fresh tumor and adjacent normal tissues were taken from a 60-year-old male patient with adenocarcinoma of the colon classified as T3N0M0 at Peking University Cancer Hospital. A signed written consent was obtained before recruitment for the study according to the regulations of the institutional ethics review boards. Single-cell suspension was prepared by harvesting fresh single cells from carcinoma and adjacent normal tissues in physiological saline.

Collection of single cells and preparation of cell lysates

A manual-controlled pipetting system was used to isolate single cells under an inverted microscope (Nikon Instruments Co., Ltd.). Each cell was transferred into the precooled PCR tube containing the cell lysis solution. The samples were incubated in a thermocycler for 10 min at 65 °C. A physiological saline blank was included as a negative control.

Multiple displacement amplification (MDA) and storage

Whole-genome amplification (WGA) was achieved using REPLI-g Mini Kit according to the manufacturer's manual (Qiagen GmbH). All samples were amplified by MDA. A reaction in a total volume of 50 μl was performed at 30 °C for 16 h and then terminated at 65 °C for 10 min. Amplified DNA products were then stored at –20 °C.

Concentration measurement, amplification coverage estimation and sequencing

The Qubit™ Quantitation Platform (Life Technologies) was used to measure the concentration of MDA products to check if the MDA was successful. Then all successfully amplified products of which DNA yields reached more than 30 ng/μl were examined by housekeeping-gene PCR to estimate amplification coverage. The MDA products with at least eight housekeeping genes successfully amplified were selected for further exome sequencing.

Somatic mutation detection and genotype likelihood estimation

After removal of adapters and low quality reads, all the sequencing reads were mapped to the hg18 reference genome, with no more than two mismatches by SOAP2. Then the PCR duplicates, low-quality ($Q < 20$) and non-uniquely mapped reads were also removed. Somatic mutations were predicted by SOAPsnv based on the reads-supported information for different alleles. A candidate somatic mutation was called if the following criteria were met: (1) The read depth of high-quality ($Q > 20$) sequencing was ≥ 6 in normal/cancer data set; (2) The reads-supported mutated allele was not a result of sequencing error (Binomial test, $f = 0.1$, $P < 0.01$); (3) The scores of sequencing quality for mutated alleles were not lower than those of normal alleles (Wilcoxon rank sum test, $P > 0.01$); (4) Mutant allele frequency change between cancer sample and adjacent normal ≤ 20% and a Fisher's exact test $P < 0.01$; (5) Mutated alleles were not significantly different among repeatedly aligned reads (Fisher's exact test, $P < 0.01$); (6) Mutant alleles were not significantly enriched within 10 bp of 5′ or 3′ ends (Fishers exact test, $P > 0.01$).

For each single nucleotide variant (SNV) site detected in each sequenced cell, a Bayes calibration which takes allele dropout (ADO), sequencing data and *a priori* probability into consideration was applied to estimate whether the site was mutated. The probability of each possible genotype Ti can be estimated by the following formula,

$$\Pr(T_i \mid D) = \frac{\Pr(T_i)\Pr(D \mid T_i)}{\sum_{x=1}^{n} \Pr(T_x)\Pr(D \mid Tx)}$$

In the formula, n represents the number of possible genotypes. Here, we considered three genotypes (AA, Aa, aa), where "A" represents a normal allele and "a" represents a mutant allele. Other alleles are considered as sequencing error.

$$Q=\log_2\frac{\Pr(Aa)+\Pr(aa)}{\Pr(AA)}$$

The Q-score calculated as above was used to indicate whether the site was mutated. Generally, an unknown site should have a Q-score of zero. In such case, the *a priori* probability was calculated as follows:

$$\Pr(T_i)=\begin{cases}\frac{1}{2} & \text{if } T_i=AA\\ \frac{1}{3} & \text{if } T_i=Aa\\ \frac{1}{6} & \text{if } T_i=aa\end{cases}$$

Taking ADO into consideration, the probability for each genotype will be:

$$\Pr(T_i\mid ADO)=\begin{cases}\Pr(AA)+0.5\times\Pr(Aa)\times\Pr(ADO) & \text{if } T_i=AA\\ \Pr(Aa)-\Pr(Aa)\times\Pr(ADO) & \text{if } T_i=Aa\\ \Pr(aa)+0.5\times\Pr(Aa)\times\Pr(ADO) & \text{if } T_i=aa\end{cases}$$

For a genotype T_i, the likelihood Pr(D | Ti) at a site could be calculated from the sequenced alleles. For a total sequence depth m, the likelihood was calculated as:

$$\Pr(D\mid T_i)=\prod_{k=1}^{m}\Pr(d_k\mid T_i)$$

In the formula, d_k is the k – the observed allele. Thus, the Q-score or likelihood for each genotype was calculated.

Principal component analysis (PCA) and clustering of single cells

The mutation probability of each non-synonymous mutation for each cell was calculated according to the genotype likelihood estimation. The $\log_2$ ratio of the probability of being mutated to that of being wild type was taken to perform the PCA using the ade4 package in R program. The same data for potential cancer driver genes in all cells were used for clustering. The hierarchical clustering method using the Euclidean distance was performed using gplots in the R package.

Clonal heterogeneity in colon cancer

To quantify the heterogeneity of clonal origin in each colon cancer sample, somatic SNVs in diploid regions were subject to kernel density estimate (KDE) analysis. Copy number variants (CNV) were detected by VarScan and processed by DNAcopy. SNVs found in CNV regions as well as those on chrX, chrY and chrM were excluded in this analysis, giving rise to copy-number-neutral heterogeneous somatic SNVs. The KDE plot of each colorectal cancer (CRC) was drawn. Hierarchical clustering over the KDE pattern that is characterizing each CRC sample was also performed to identify CRC subgroups.

Prediction of cancer driver gene

We used the method described by Youn A and Simon R [25] to compute the significance of observed mutations on each gene. The statistical model takes both mutation prevalence and functional impact into consideration. Functional impact was evaluated by mutation scores assigned in the following order: missense < mutation in splice sites < nonsense. Different types of missense mutations were also assigned different scores based on BLOSUM80 matrix. This method assumes that passenger mutations, including silent and non-silent mutations, were generated from the same background mutation process. By incorporating different background mutation rates of each sample into consideration, background distribution of mutation score for each gene was computed. The *P*-value for each gene was calculated from this background distribution and the test statistics from the observed mutation scores across samples. The *P*-value was adjusted using the Benjamini-Hochberg method to estimate the false discovery rate (FDR). Significantly mutated genes were selected if FDR ≤ 10%.

Construction of SLC12A5 expression vector

The *SLC12A5* expression vector was generated by PCR-cloning with pcDNA3.1 TOPO TA Expression Kit (Invitrogen). cDNA corresponding to the open-reading frame of *SLC12A5* transcript (NCBI reference number: NM_020708) was obtained by RT-PCR amplification of normal human colon RNA using the following primers for *SLC12A5* (forward: 5′-GCCACCATGCTAAACAACCTG-3′ and reverse: 5′-GGTTCTCAGGAGTAGATGGTGAT-3′). PCR aliquots were subcloned into the pcDNA3.1 TOPO vector. Clones were screened and sequenced using vector-specific primers.

Site-directed mutagenesis

A single point mutation of *SLC12A5* was introduced by inverted PCR using the GENERAT site-directed mutagenesis system according to the manufacturer's instruction (Invitrogen). Briefly, the methylated plasmid was amplified in a mutagenesis reaction with two complementary mutagenic oligonucleotide primers encoding a single base substitution A > T located at chr20:44113842 (mutant-forward: 5′-GGGCTGCAGCTCAACACTGTGCTTGTTGGCT-3′; mutant-reverse: 5′-AGCACAGTGTTGAGCTGCAGCCCCCCGAGG-3′). *In vitro* recombination reaction was carried out using the amplified plasmid, which was then transformed into DH5α-T1 competent *E. coli* cells for clone selection. The *SLC12A5* mutant constructs were validated by sequencing.

Gene transfection, colony formation and cell proliferation assays

The expression vector was transfected into colon cancer cell lines (SW480 and SW1116) using Lipofectamine LTX (Invitrogen). Cells transfected with wild-type or mutant pcDNA3.1-SLC12A5 or pcDNA3.1 empty vector were selected with G418 (0.2 mg/ml; Merck, Darmstade, Germany) for 2 weeks. Stably transfected cells were then seeded on six-well plates and the colonies (with > 50 cells per colony) were counted after 2 weeks. Colony formation was analyzed after staining with crystal violet.

Cell proliferation was measured by cell viability assay using the xCELLigence System (Roche Applied Science). Briefly, SW480 and SW116 cells stably transfected with wild-type or mutant pcDNA3.1-SLC12A5, or pcDNA3.1 empty vector were seeded in E-plate 16 in triplicates (1×10^4 cells/well) under 200 μl complete RPMI1640 medium. Dynamic cell proliferation of cells was monitored in 30-min intervals until the end of experiment. Cell index values for cells were analyzed with the Prism software version 5.01 (GraphPad Software, CA, USA).

Apoptosis assay

Apoptosis was assessed by flow cytometry after staining with Annexin V (APC-conjugated; BD Biosciences, Erembodegem, Belgium) and 7-amino-actinomycin (7-AAD; BD Biosciences). Cells stably transfected with wild-type or mutant pcDNA3.1-SLC12A5, or pcDNA3.1 empty vector were stained and incubated for 20 min. Samples were immediately analyzed by fluorescence-activated cell sorting on a FACScan (BD Biosciences). Cell populations were counted as viable (Annexin V-negative, 7-AAD-negative), early apoptotic (Annexin V-positive, 7-AAD-negative) and late apoptotic and necrotic cells (Annexin V-positive, 7-AAD-positive).

Acknowledgments

This work was supported by the National Basic Research Program of China (973 program; 2011CB809202, 2011CB809203), the Hi-Tech Research and Development Program of China (863 program; 2009AA022707), the Theme-based Research Scheme of the Hong Kong Research Grants Council (T12-403-11), the Shenzhen Municipal Government of China (ZYC201005250020A), the Key Laboratory project Supported by Shenzhen City (CXB200903110066A, CXB201108250096A), and Shenzhen Key Laboratory of Gene Bank for National Life Science. This work was also supported by the Innovative Research Team Project of GuangDong, the GuangDong Enterprise Key Laboratory of Human Disease Genomics, the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research and Cancer Genome Project, Scheme B CUHK 2009. We also thank the Danish Natural Science Research Council for the Ole Rømer grant and the Shenzhen Municipal Government and the Local Government of Yantian District of Shenzhen for providing funding.

References

1 Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61**:759-767.
2 Sjoblom T, Jones S, Wood LD, *et al*. The consensus coding sequences of human breast and colorectal cancers. *Science* 2006; **314**:268-274.
3 Wood LD, Parsons DW, Jones S, *et al*. The genomic landscapes of human breast and colorectal cancers. *Science* 2007; **318**:1108-1113.
4 Navin N, Kendall J, Troge J, *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* 2011; **472**:90-94.
5 Kreso A, O'Brien CA, van Galen P, *et al*. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* 2013; **339**:543-548.
6 Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 2013; **501**:338-345.
7 Nowell PC. The clonal evolution of tumor cell populations. *Science* 1976; **194**:23-28.
8 Sakurazawa N, Tanaka N, Onda M, Esumi H. Instability of X chromosome methylation in aberrant crypt foci of the human colon. *Cancer Res* 2000; **60**:3165-3169.
9 Beutler E, Collins Z, Irwin LE. Value of genetic variants of glucose-6-phosphate dehydrogenase in tracing the origin of malignant tumors. *N Engl J Med* 1967; **276**:389-391.
10 Hsu SH, Luk GD, Krush AJ, Hamilton SR, Hoover HH Jr. Multiclonal origin of polyps in Gardner syndrome. *Science* 1983; **221**:951-953.
11 Thirlwell C, Will OC, Domingo E, *et al*. Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. *Gastroenterology* 2010; **138**:1441-1454.
12 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**:330-337.
13 Hou Y, Song L, Zhu P, *et al*. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* 2012; **148**:873-885.
14 Xu X, Hou Y, Yin X, *et al*. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 2012; **148**:886-895.
15 Prager EM, Wilson AC. Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. *J Mol Evo* 1978; **11**:129-142.
16 Govindan R, Ding L, Griffith M, *et al*. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 2012; **150**:1121-1134.
17 Esseltine JL, Willard MD, Wulur IH, Lajiness ME, Barber TD, Ferguson SS. Somatic mutations in GRM1 in cancer alter metabotropic glutamate receptor 1 intracellular localization and signaling. *Mol Pharmacol* 2013; **83**:770-780.
18 Takashima N, Ishiguro H, Kuwabara Y, *et al*. Expression and prognostic roles of PABPC1 in esophageal cancer: correlation with tumor progression and postoperative survival. *Oncol Rep* 2006; **15**:667-671.
19 Pawar SA, Sarkar TR, Balamurugan K, *et al*. C/EBP{delta} targets cyclin D1 for proteasome-mediated degradation via induction of CDC27/APC3 expression. *Proc Natl Acad Sci* 2010; **107**:9210-9215.
20 Iacopetta B. TP53 mutation in colorectal cancer. *Hum Mutat* 2003; **21**:271-276.
21 Kohler EM, Derungs A, Daum G, Behrens J, Schneikert J. Functional definition of the mutation cluster region of adenomatous polyposis coli in colorectal tumours. *Hum Mol Genet* 2008; **17**:1978-1987.
22 Parsons BL. Many different tumor types have polyclonal tumor origin: evidence and implications. *Mutat Res* 2008; **659**:232-247.
23 Boulenguez P, Liabeuf S, Bos R, *et al*. Down-regulation of the potassium-chloride cotransporter KCC2 contributes to spasticity after spinal cord injury. *Nat Med* 2010; **16**:302-307.
24 Wei WC, Akerman CJ, Newey SE, *et al*. The potassium-chloride cotransporter 2 promotes cervical cancer cell migration and invasion by an ion transport-independent mechanism. *J Physiol* 2011; **589**:5349-5359.
25 Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* 2011; **27**:175-181.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)