Science Bulletin 64 (2019) 236–244

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Characterization and validation of somatic mutation spectrum to reveal heterogeneity in gastric cancer by single cell sequencing

Lihua Peng [a,b,1], Rui Xing [c,1], Dongbing Liu [a,b,1], Li Bao [a,d,1], Wenxiang Cheng [e], Hongyi Wang [c], Yuan Yu [a], Xiaofeng Liu [f], Lu Jiang [a], Yan Wu [f], Zhongxue An [a], Qiaoyi Liang [g], Ryong Nam Kim [h], Young Kee Shin [h], Huanming Yang [a,i], Jian Wang [a,i], Jun Yu [g], Xiuqing Zhang [a,b], Xun Xu [a,b], Jiaan Yang [j], Kui Wu [a,b,k], Shida Zhu [a,b,k,*], Youyong Lu [c,*]

[a] BGI-Shenzhen, Shenzhen 518083, China
[b] China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen 518083, China
[c] Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing 100142, China
[d] Department of Drug Design and Pharmacology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
[e] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[f] Department of Histology and Embryology, Inner Mongolia Medical University, Huhhot 010110, China
[g] Department of Medicine and Therapeutics, State Key Laboratory of Digestive Disease, Institute of Digestive Disease and Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[h] Department of Pharmacy, College of Pharmacy, Seoul National University, Seoul 08826, South Korea
[i] James D. Watson Institute of Genome Sciences, Hangzhou 310058, China
[j] Micro Pharmatech, Ltd, Wuhan 430075, China
[k] Department of Biology, University of Copenhagen, Copenhagen N DK-2200, Denmark

ARTICLE INFO

Article history:
Received 21 September 2018
Received in revised form 12 October 2018
Accepted 2 November 2018
Available online 15 December 2018

Keywords:
Gastric cancer
Single-cell whole exome sequencing
SNV
Significant mutated gene
Heterogeneity

ABSTRACT

Gastric cancer (GC) is a highly heterogeneous disease with multiple cellular types and poor prognosis. However, the cellular evolution and molecular basis of GC at the individual intra-tumor level has not been well demonstrated. We performed single-cell whole exome sequencing to detect somatic single-nucleotide variants (SNVs) and significantly mutated genes (SMGs) among 34 tumor cells and 9 normal cells from a patient with GC. The Complete Prediction for Protein Conformation (CPPC) approach directly predicting the folding conformation of the protein 3D structure with Protein Folding Shape Code, combined with functional experiments were used to confirm the characterization of mutated SMGs in GC cells. We identified 201 somatic SNVs, including 117 non-synonymous mutations in GC cells. Further analysis identified 24 significant mutated genes (SMGs) in single cells, for which a single amino acid change might affect protein conformation. Among them, two genes (*CDC27* and *FLG*) that were mutated only in single cells but not in the corresponding tumor tissue, were recurrently present in another GC tissue cohort, and may play a potential role to promote carcinogenesis, as confirmed by functional characterization. Our findings showed a mutational landscape of GC at intra-tumor level for the first time and provided opportunities for understanding the heterogeneity and individualized target therapy for this disease.

1. Introduction

Gastric cancer (GC) is one of the major health burdens worldwide with various molecular characteristics and complex histological features. Previous genome studies on GC have revealed that some driver genes were differently mutated in diverse subtypes. For example, *ARID1A*, a chromatin remodeling gene, whose mutation spectrum differed among molecular subtypes of GC, was associated with prognosis [1]. In gastric adenocarcinoma, which accounts for more than 90% of GC, *FAT4* and *ARID1A* were found to exert a tumor suppressor activity [2]. Wang et al. studied the landscape and perturbations of the GC genome and epigenome, and demonstrated that the diffuse-type tumors had a lower number of somatic single-nucleotide variants (SNVs) and indels

* Corresponding authors.
E-mail addresses: zhushida@genomics.cn (S. Zhu), youyonglu@sina.com (Y. Lu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2018.12.014
2095-9273/

compared to the other subtypes, and that *RHOA* and *CDH1* were mutated more frequently in diffuse-type tumors [3]. Moreover, *RHOA* and *CDH1* were also found to frequently mutate in genetically stable type of GC, which are enriched for the diffuse histological variation [4]. In the recent studies, only a few mutated genes were frequently shared among GC patients, suggesting that rare mutations might play important roles in individual patients with GC. In addition, genomic level heterogeneity is the most important characteristic for cellular evolution during carcinogenesis. To address the intra-individual heterogeneity and genetic complexity, high-throughput single-cell sequencing has been widely employed in cancer research [5–11], such as colon and bladder cancer.

Here, we performed whole exome sequencing on tumor-normal paired tissues and 43 single cells from a GC patient. We identified 99 somatic SNVs at the tissue level, 201 somatic SNVs at the single-cell level, and 24 significantly mutated genes (SMGs) including *CDC27* and *FLG*. We also investigated the possible role of these events and demonstrated that mutations found in single cells but absent in the corresponding tumor tissue of this individual may also potentially contribute to the progression of GC. These results provided new insights into understanding the molecular heterogeneity of GC.

2. Materials and methods

2.1. Sample collection

In our study, the patient was a 59-year-old Chinese male with a sporadic poorly differentiated gastric adenocarcinoma and signet ring cell carcinoma on the gastric antrum, classified as stage III (T3N3M0) according to the seventh TNM staging system for gastric cancer of the American Joint Committee on Cancer/International Union against Cancer (AJCC/UICC). With written informed consent, the fresh tumor and adjacent normal tissue were obtained from this patient at Beijing Cancer Hospital in accordance with the Declaration of Helsinki II and with the approval of local Ethical Committees.

2.2. Preparation, lysis, and multiple displacement amplification (MDA) of single cells

Single-cell suspensions were extracted from fresh samples and immediately harvested in physiological saline. Single cells were isolated with a mouth-controlled microcapillary pipetting system and washed thrice in an elution buffer under an inverted microscope (Nikon Instruments Co., Ltd.). Each cell was immediately transferred into precooled PCR tubes containing cell lysis solution, and placed on ice. MDA of single cells was performed with the REPLI-g Mini Kit (Qiagen, Inc.), along with a physiological saline blank as the negative control. Samples were incubated in a thermocycler for 10 min at 65 °C. The 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. MDA products were then stored at −20 °C.

2.3. Concentration measurement, amplification coverage estimation, and sequencing

The DNA concentration of MDA products was measured on the Qubit™ Quantitation Platform (Life Technologies). A ten housekeeping-gene PCR was then performed to estimate the amplification coverage. All qualified MDA products in which DNA yields reached more than 30 ng/μL and at least 8 housekeeping genes were successfully amplified, were selected for further exome sequencing.

The libraries were prepared according to standard Illumina library preparation procedures. Exome capture was performed using the Agilent SureSelect Platform according to the manufacturer's protocol. Finally, massively parallel sequencing was performed on the Illumina Hiseq 2000 platform.

2.4. Reads mapping and sample selection

After removal of adapters and low quality reads using the SOAPnuke software (http://soap.genomics.org.cn/), all sequencing reads were mapped to the NCBI Build 37 Human Reference Genome using Burrows-Wheeler Aligner (http://bio-bwa.sourceforge.net/). PCR duplicates were also removed using Picard (http://broadinstitute.github.io/picard/). Then, local realignment and base recalibration was performed using the Genome Analysis Toolkit (GATK) [12]. The target region files of whole exome captured sequencing were downloaded from the Agilent website (http://www.genomics.agilent.com). Single cells which covered more than 50% of the target region were used for further bioinformatics analysis.

2.5. Detection of single nucleotide polymorphisms (SNPs)

SNPs and indels were detected by the GATK software. Variant quality score recalibration (VQSR) based on the gaussian model and hard filtering was applied to filter the false discovery mutations in the tissues and single cells.

There are two types of SNP sites in tissues: SNPs in the tumor tissue (T-SNPs) and SNPs in the normal tissue (N-SNPs). We summarized the base callings of T-SNPs or N-SNPs to the preprocessed aligned bam files of all samples (including all tissues and all single cells) in a pileup format. For tissues, the mutant allele frequencies of T-SNPs or N-SNPs were calculated as the variant-supporting reads divided by the total read depth. For single cells, the mutant allele frequencies of T-SNPs or N-SNPs were calculated as the number of cells with variant-supporting reads divided by the total number of cells.

2.6. Evaluation of allele dropout (ADO)

To estimate allele dropout, we used the similar method described by Li et al. without considering of sequencing depth [6]. First, a background heterozygous dbSNP subset and a background homozygous dbSNP subset in the normal tissue were built. We calculated the heterozygous ratio based on the background heterozygous dbSNP subset and the homozygous ratio based on the background homozygous dbSNP subset for each single cell. We also calculated the false negative rate (FNR) and false positive rate (FPR) for each single cell. Finally, the mean ADO of normal cells was selected as the ADO of the entire whole exome single-cell sequencing data.

2.7. Somatic mutation detection in the tissue samples

For the normal-tumor tissue paired sample, MuTect [13] based on Bayesian classifier was used to detect single nucleotide variants (SNVs).

These SNV sites identified as "alt_allele_in_normal" by MuTect will be filtered, except those with read depth greater than 200x, the frequency of normal-supporting reads less than 0.02 and the frequency of variant-supporting reads greater than 0.05. In total, we detected 99 SNVs in the tumor tissue.

2.8. Somatic mutation detection in single cells

MuTect was also used to detect the SNVs at single-cell level for each selected cell using the normal tissue as control. First, the SNV

sites identified as "alt_allele_in_normal" by MuTect will also be filtered, except those with read depth greater than 200x, the frequency of variant-supporting reads less than 0.02 in normal control and the frequency of variant-supporting reads greater than 0.05 in the single cell. Furthermore, the SNVs were filtered by the following criteria: 1) SNVs with low mutation frequency were filtered. SNVs that were supported by less than two independent reads or with a frequency of variant-supporting reads less than 0.04 were filtered. 2) We further filtered the SNVs to reduce the FPR caused by PCR. SNVs existing in three or more tumor cells but not in any normal cells, or supported by more than two independent reads in tumor tissue but not in normal tissue, were predicted to be somatic SNVs at the single-cell level. In total, we detected 201 somatic SNVs at the cellular level. ANNOVAR [14] was then used to functionally annotate the SNV sites.

2.9. Mutation spectrum analysis

According to their nucleotide context-specific exonic mutation rates, the fraction of 6 mutation types and 96 strand-collapsed trinucleotide context mutation signatures were calculated [15].

2.10. Pathway enrichment analysis

Somatic non-synonymously mutated genes were searched for over-representation in molecular Canonical pathways (CP), BIOCARTA, KEGG, and REACTOME pathway by Gene Set Enrichment Analysis (GSEA, http://www.broadinstitute.org/gsea/index.jsp). Pathways with a FDR (False discovery rate) q-value less than 0.05 were selected.

2.11. Prediction of significant mutated genes (SMGs)

To detect SMGs, we first used two methods: the method described by Youn et al. [16] and the MutSigCV method (version 1.4) [17]. The model described by Youn et al. evaluates both the functional impact and the mutation prevalence, and we selected genes with a Q-score greater than 1 as candidate SMGs. In the MutSigCV method, mutated genes with q value less than 0.1 were selected. *FLG* is a TCGA pan-cancer driver gene and *PTCH* was mutated in 5/34 tumor single cells. Although these two genes were not predicted by the two methods, we considered them as SMGs because of their importance. In total, we detected 24 SMGs.

2.12. Chromatin remodeling related genes, GCG genes, TCGA driver genes

All chromatin remodeling related genes were found in three databases: HIstome (http://www.actrec.gov.in/histome/), EpiFactors (http://epifactors.autosome.ru/) and CREMOFAC (http://www.jncasr.ac.in/cremofac/menuframe.html). CGC (Cancer Gene Census) genes and the TCGA pan-cancer driver genes were found in http://cancer.sanger.ac.uk/census [18] and Michael et al. [19], respectively.

2.13. Phylogenetic tree analysis

Somatic SNVs at the single cell level were used to construct a phylogenetic tree based on an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [20] using MEGA software (version 6.06) [21].

2.14. Protein conformation prediction

The Complete Prediction for Protein Conformation (CPPC) approach directly predicts the folding conformation of the protein 3D structure with Protein Folding Shape Code (PFSC) [22]. A set of 27 PFSCs, obtained through mathematical derivation, could cover the enclosed geometric space for any folding shape of 5 amino acids. From a sequence, the possible variations of conformation for each of five successive residues can be determined according to the structural data present in protein data bank (PDB). It also provides the complete description for the protein conformation. The prediction is based on a database that collects all folding shapes for any combination of 5 amino acids.

2.15. RNA isolation, reverse transcription (RT), and polymerase chain reaction (PCR)

Total RNA was isolated from cells using TRIzol Reagent (Invitrogen) and was reverse transcribed according to the manufacturer's instructions. Conventional PCR products were visualized on 1.5% agarose gel.

2.16. Immunofluorescence (IF)

Cells were fixed with 4% paraformaldehyde at 4 °C for 10 min. After washing and pre-blocking, the cells were incubated overnight at 4 °C with antibodies, followed by incubation with a FITC-conjugated secondary antibody for 1 h. DAPI was used for nuclear staining. Images were then analyzed by laser scanning confocal microscopy (Leica Sp5 Laser Scanning Confocal Microscope, GE).

2.17. Western immunoblotting

Sodium dodecyl sulfate–polyacrylamide gel electrophoresis and western blotting were performed using standard protocols. Antibody binding was detected using the Smartchemi image analysis system (Sagecreation, Beijing, China).

2.18. MTT assay

Cells were plated in 96-well culture plates, and 3-(4,5-dimethylthiazol-2-yl)-2, 5-diphenyltertrazolium bromide (MTT) was added at 24, 48, 72, and 96 h. After 4 h incubation at 37 °C in 5% CO_2, dimethylsulfoxide (DMSO) was added to solubilize the formazan product. Absorbance at 570 nm was determined using a microplate reader (Bio-Rad, Hercules, CA, USA).

2.19. Colony formation assay

Single cells were seeded into 60-mm culture dishes on day 0 and were allowed to attach at 37 °C for 24 h at the appropriate densities. Cells were then incubated in complete culture medium for 12–15 days. Colonies were stained with 1.0% crystal violet and 0.5% glacial acetic acid in ethanol.

3. Results

3.1. Whole-exome single cell sequencing in a GC patient

We obtained tumor-normal paired tissues from a 59-year-old Chinese male patient with GC, classified as T3N3M0 according to the seventh TNM staging system for this disease of the American Joint Committee on Cancer/International Union against Cancer (AJCC/UICC) (Fig. S1 online). We first performed whole-exome sequencing (WES) of this tumor-normal paired tissue with an aver-

Table 1
Exome sequencing coverage and depth for cancer and normal control tissue and cells.

Sample ID	Average depth (X) on target region	Depth range (X)	Average coverage of target region (%)	Coverage range (%)
34 tumor cells	63.12	29.30–113.16	74.08	50.00–96.50
9 normal cells	91.90	50.40–161.02	77.50	57.20–98.00
Tumor tissue	99.85	NA	99.00	NA
Normal tissue	115.39	NA	99.20	NA

Note: NA: not applicable.

age depth of 115.4 × (99.2% coverage) and 99.9 × (99.0% coverage) for normal and tumor tissue, respectively (Table 1).

To obtain the detailed cellular genetic information of this case, 100 tumor cells and 30 normal cells from the patient were sequenced by whole-exome single cell sequencing, as described by Hou et al. [8] (Fig. S2a, Table S1 online). After quality control, 43 single cells (≥50.0% coverage), including 34 tumor cells and 9 normal cells, were used for further analysis. The average sequencing depth was 63.1 × (average coverage is 74.1%) for 34 tumor cells, and 91.9 × (average coverage 77.5%) for 9 normal cells (Table 1). To assess the performance and reliability of our method, we estimated the false-negative rate (FNR) and false-positive rate (FPR) for the tumor and normal cells. The average FNR was 56.7% for tumor cells, and 53.6% for normal cells (Fig. S2b, Table S2 online). The average FPR was 0.59% and 0.34% for tumor and normal cells, respectively (Fig. S2c online). In total, the allele dropout ratio (ADO, see Method) of our single cell data was 53.6%, which is consistent with those of previous analyses [6,8,23]. Furthermore, a positive correlation was observed between the mutant allele frequencies of SNPs in tissues and those in single cells (Pearson correlation = 0.440 for the tumor tissue and Pearson correlation = 0.441 for the normal tissue, Fig. S2d, e online), suggesting that the SNPs in the tissue can also be replicated in single cells. These results indicated that a higher reliability of sequencing data could be achieved using our single-cell sequencing method.

3.2. Somatic mutation spectra among GC cells revealed heterogeneity

For this tumor-normal paired tissue, we detected 99 SNVs, including 58 missense, 5 splicing, 3 nonsense, and 11 synonymous SNVs (Table S3 online). C:G → T:A was the most common mutation type (Fig. S3 online), which was in accordance with the previously reported GC tissue sequencing data [1,2]. We found that *RHOA* and *CDH1*, which are also reported to be driver genes and were more significantly mutated in diffuse-type GC than in the other GC types [3,4], were mutated in this case, suggesting that the GC type of this patient may be more similar to the diffuse type.

Using normal tissue as the control, we also identified 20,795 SNVs in 43 single cells (Fig. S3a online). The most common mutation type of these SNVs was also C:G → T:A (Fig. S3b, c online). We observed that the numbers of SNVs in some cells were unusually high, especially in normal single cells, which was probably caused by a high false positive rate. To reduce the false positives and identify the potential somatic SNVs in tumor single cells, we selected SNVs that presented in at least three tumor cells or supported by more than two independent reads in the tumor tissue, but did not exist in any normal cell and were not supported by any reads in the normal tissue. Ultimately, 201 SNVs were selected as potential somatic SNVs and retained for further analysis, of which 151 were located in the coding region, including 112 missense, 34 synonymous, and 5 nonsense SNVs (Table S4 online). In single cells, 72.6% (146/201) somatic mutations were supported by reads from exome sequencing of the tumor tissue. C:G → T:A was also the most common mutation type for these somatic SNVs (Fig. S3b online). In addition, we performed pathway enrichment analysis of mutated genes that altered the protein sequence at the single-cell level and found several top enriched pathways, including the extracellular matrix (ECM) related pathway (Fig. S4, Table S5 online).

To explore the intercellular heterogeneity of GC, hierarchical clustering of 201 single-cell somatic SNVs was performed (Fig. 1a). However, the somatic SNVs were very sparse and no SNV presented in all tumor cells, possibly due to the intercellular heterogeneity of GC or the ADO. We then constructed a phylogenetic tree based on the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and found that tumor cells clustered separately from normal cells (Fig. 1b).

3.3. Identification of significant mutated genes (SMGs) in single cells

To further assess the likelihood of somatic mutant genes from single cells that might play an important role in GC development of this case, we identified 24 candidate somatic SMGs at the cellular level (Fig. 1c, Fig. S5, Tables S6 and S7, online). Of these, *SORD* (Sorbitol dehydrogenase), which has been reported in colorectal and prostate cancer [24,25], was the most frequently mutated gene of 24 SMGs and was detected in 9 of 34 tumor cells. Another gene, *REXO2,* which participates in DNA repair, replication, and recombination and may have a potential relationship with the cancer, was detected as mutated in 8 of 34 tumor cells. *REC8* (altered in 9/34 single cells) is related to the cell cycle, and mitotic and oocyte meiosis pathways. Furthermore, eight genes (*CDC27*, *FLG*, *NSD1*, *PDE4DIP*, *RAF1*, *ZNF483*, *ETV6*, and *SMO*) were reported in the Cancer Gene Census (CGC) or the TCGA pan-cancer driver gene list, whereas six other genes (*BAZ1A*, *NEK6*, *RNF2*, *RNF20*, *RBBP4*, and *NSD1*) were related to chromatin remodeling (Table S8 online).

To investigate whether somatic SMGs in single cells were also frequently mutated in tissues, we first compared all somatic mutated genes in tumor cells with those in the corresponding tumor tissue of this case and found that only 12 genes were shared (Table S9 online), suggesting the intercellular heterogeneity of GC. Moreover, we compared somatic mutated genes in single cells with those in an additional cohort of 54 paired GC tissues (whole genome sequencing, unpublished data), and found that 44 genes were shared (Fig. 2a). Of these, five genes (*CDC27*, *FLG*, *PDE4DIP*, *CLEC4M*, and *SPTA1*) had non-silent substitutions in at least three tumor tissues from the 54 GC patients but were not mutated in the tumor tissue of the current case (Fig. 2b, Table S10 online). Among them, *CDC27*, *FLG* and *PDE4DIP* were also somatic SMGs. These results indicated that genes with high mutation frequencies in our single-cell data but absent in the corresponding tissue, may also mutated frequently at the population level and tend to play an important role in GC single cells. We further validated this concept through fingerprinting of somatic SMGs and functional characterization of *CDC27* and *FLG* gene.

3.4. Protein conformation variations in SMGs

To investigate the noxiousness of mutations in 24 SMGs, we need to know the relevance between mutations and the variations in protein structure, such as the conformation changes induced by the replacement of residues. We found that conformation changes for 24 SMGs were predicted to change by the protein folding shape code (PFSC) [22], except for *REC8* (Table S11 online). Of which, we

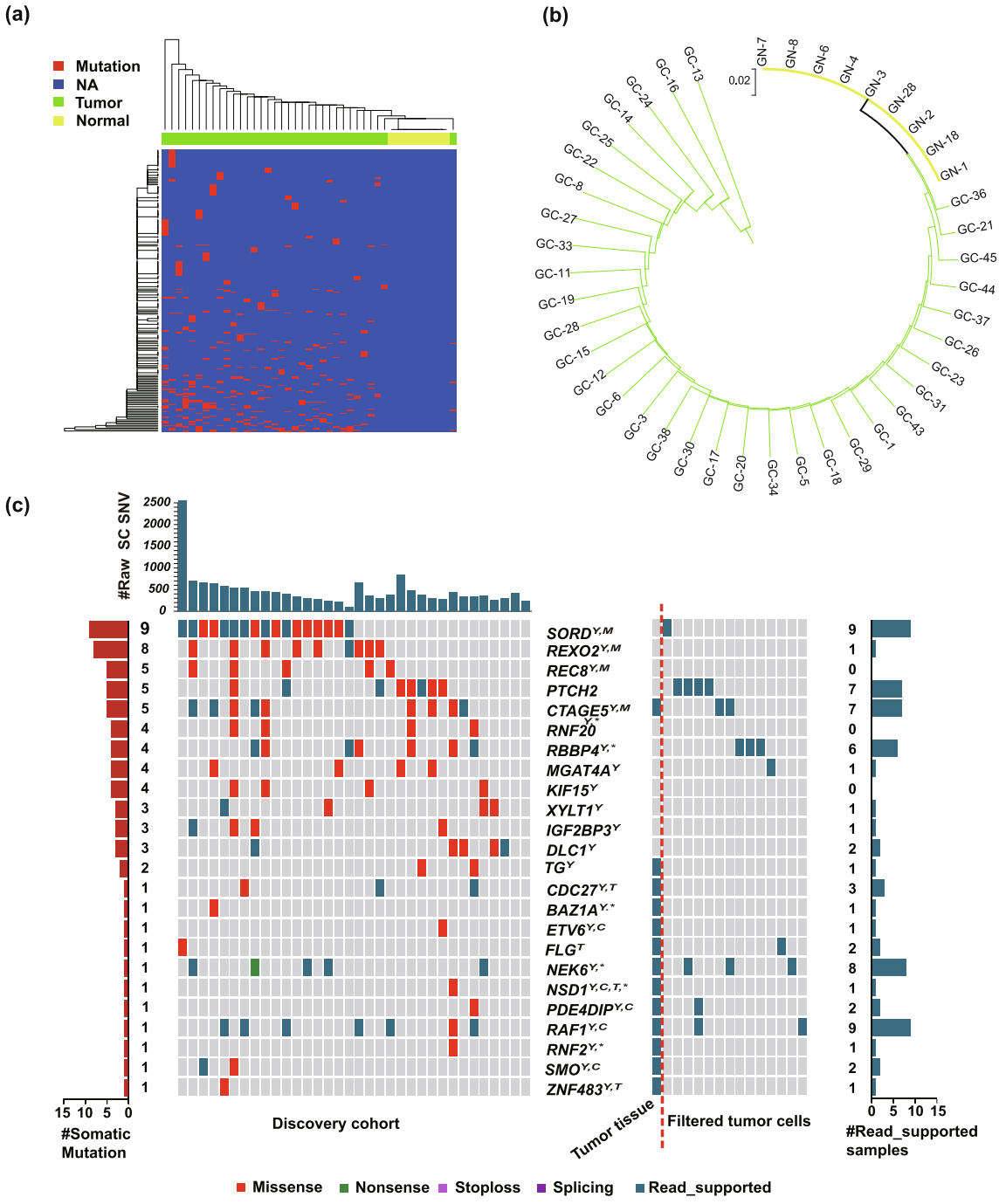


Fig. 1. The mutation spectrum of 43 single cells and their phylogenetic relationship. (a) Unsupervised two-way hierarchical clustering of 201 somatic SNVs in 43 single cells. Red and blue depicts the mutation or its absence, respectively. (b) Data of 201 somatic SNVs were used to build the phylogenetic three based on UPGMA. c Identification of SMGs in single cells. The mutant genes spectrum at the single-cell level. At the top, the column diagram depicts the number of raw SNVs in single cells; 24 SMGs were performed. The markers "Y", "M", "C", "T", and "*" indicate the genes predicted as SMGs by the Youn et al. method or MutSigCV, or presented in CGC or TCGA pan-cancer driver genes or chromatin remodeling-related genes, respectively. On the left, the horizontal bar depicts the mutation frequency of the selected genes and the horizontal bar on the right depicts the number of samples with variant-supporting reads of SMGs, including the remaining 66 GC single cells, which were filtered by QC.

performed the conformation changes induced by the point mutation of two genes, *CDC27* (E109D mutation) and *FLG* (T454A mutation) (Fig. 3a, b). Other than folding shape variations, we found that the folding number was reduced around position E109D after mutation in *CDC27*, suggested that the flexibility of this conformation change was restricted. A similar phenomenon was observed in *FLG* with the T454A mutation. These results suggested that these mutations could trigger characteristic changes in protein conformation, which would affect the biological functions of the protein, such as activity, inactivity, or noxiousness.

3.5. Characterization of mutated CDC27 and FLG in gastric cancer cells

In our single cell data, *CDC27* was mutated in 4 single cells (3 synonymous and 1 missense) and was supported by reads of mutant allele in another 8 tumor cells and the tumor tissue (Fig. 1c, Table 2, Table S12). To reveal the potential function of *CDC27*, we performed functional experiment with *CDC27 in vitro*. *CDC27* expression was assessed in gastric cancer cell lines by western blotting and higher *CDC27* expression was found in AGS and NCI-N87 cell lines compared with that in other cell lines

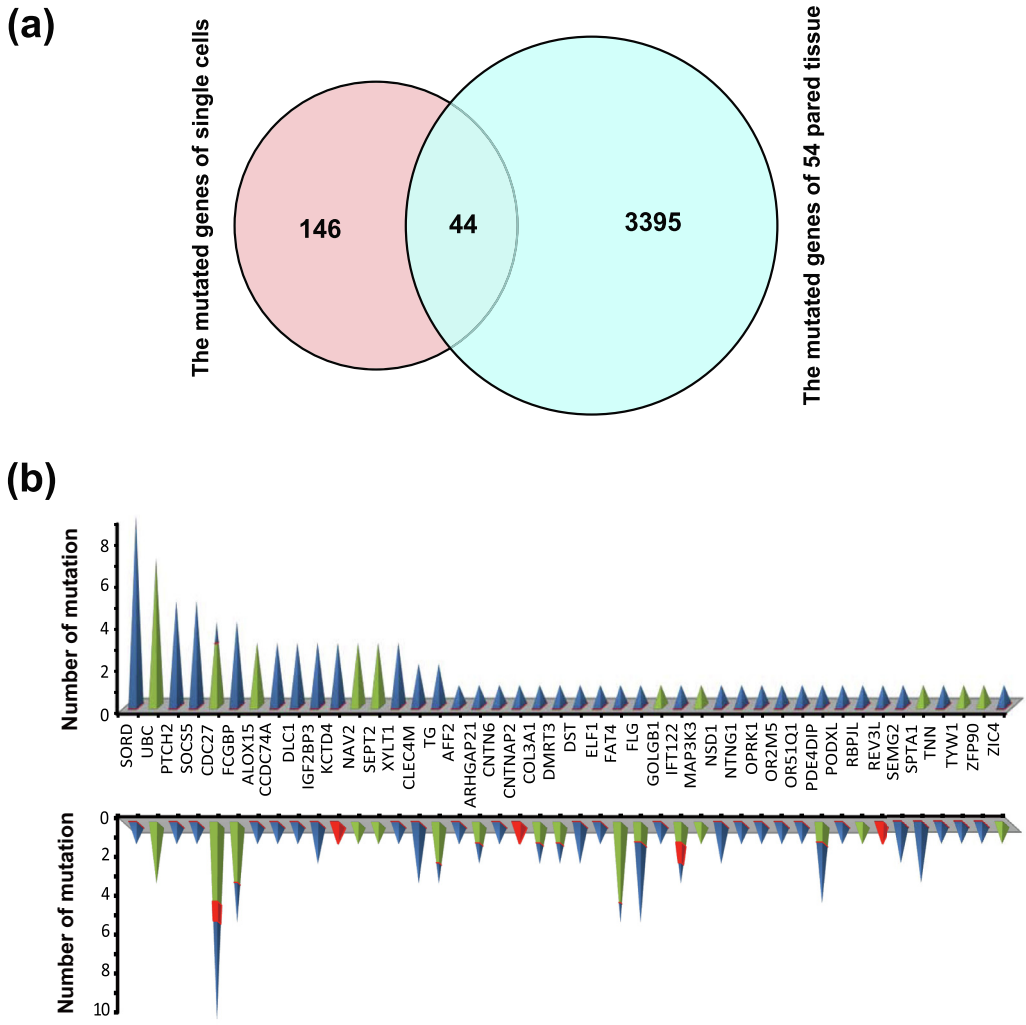


Fig. 2. Comparison between mutated genes in GC cells and those in 54 additional GC tissues. (a) The Venn diagram depicts the relationship between the number of mutated genes in GC cells and those in 54 additional GC tissues. There were 146 genes mutated in the coding region of GC cells, while 3395 mutated genes were detected in 54 additional tissues, with an overlap of 44 genes. (b) The pyramid depicts the mutation landscape of the overlapping genes between GC cells (upper panel) and the 54 additional tissues (lower panel). Blue, red, and green indicate missense, nonsense, and synonymous mutations, respectively.

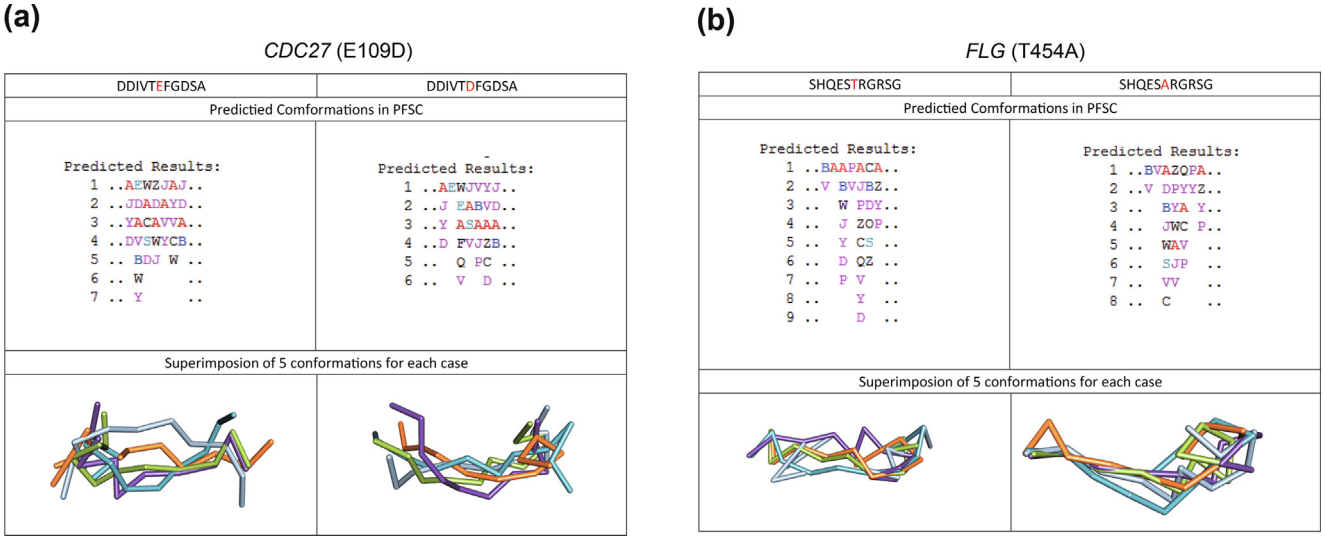


Fig. 3. Protein conformation of *CDC27* and *FLG* gene. (a) The conformation changes in *CDC27* by mutation of E109D are shown. (b) The conformation changes in *FLG* by mutation of T454A. The predicted conformations were described with PFSC. Each column lists the PFSC of the possible folding shape of 5 amino acids.

Table 2
Depths for missense sites in *CDC27* and *FLG* genes.

Gene	Chr	Position	Ref	Alt	SampleID	Depth	Ref_reads_count	Mutated_reads_count	SNV
CDC27	chr17	45,247,333	C	A	100611GC-T1	31	29	2	No
CDC27	chr17	45,247,333	C	A	GC-15	24	22	2	No
CDC27	chr17	45,247,333	C	A	GC-37	10	0	10	Yes
CDC27	chr17	45,247,333	C	A	GC-5	8	4	4	No
FLG	chr1	152,286,002	T	C	100611GC-T1	432	430	2	No
FLG	chr1	152,286,002	T	C	GC-24	18	12	6	Yes
FLG	chr1	152,286,002	T	C	GC-60	570	568	2	No

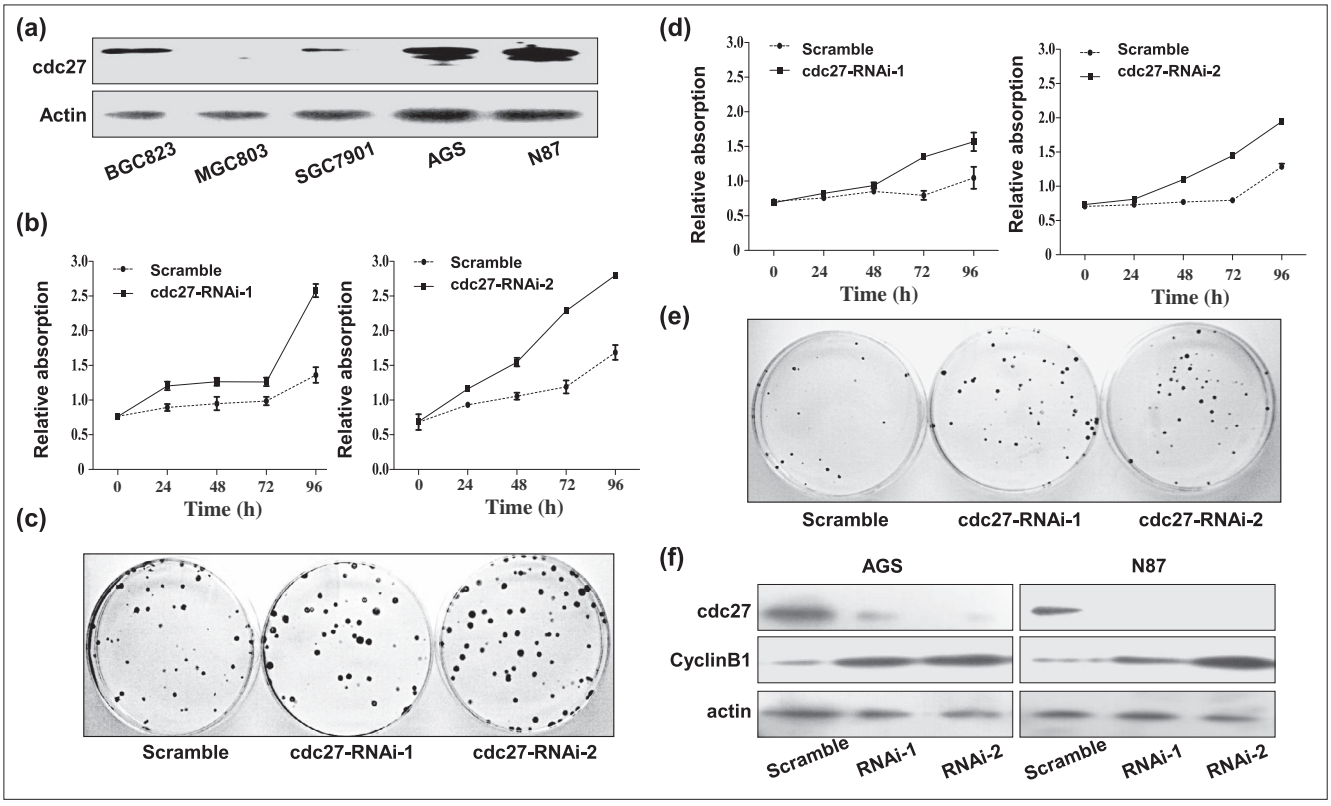


Fig. 4. Functional characterization of *CDC27*. (a) Expression of *CDC27* in GC cells detected by western blot. The effect of *CDC27* knockdown on cancer cell proliferation was detected by an MTT assay in AGS cells (b) a colony formation assay in AGS cells (c), an MTT assay in NCI-N87 cells (e) and a colony formation assay in NCI-N87 cells (d). (f) The effect *CDC27* knockdown on the expression of Cyclin B1 was determined by western blotting in AGS cells (left) and NCI-N87 cells (right).

(Fig. 4a). *CDC27* expression in AGS and NCI-N87 cells was then knocked down by shRNA. MTT and colony formation assays showed that deletion of *CDC27* promoted cell growth and colony-forming abilities compared to those in control cells (Fig. 4b-e). *CDC27* is a component of the anaphase promoting complex/cyclosome (APC/C) and a cell cycle-regulated E3 ubiquitin ligase, which can control progression through mitosis and the G1 phase of the cell cycle [26]; hence, the expression of Cyclin B1 was examined. As shown in Fig. 4f, Cyclin B1 expression was upregulated by *CDC27* deletion in both cell lines. Taken together, these results suggest that *CDC27* is a tumor suppressor that is inactivated upon mutation.

FLG (Filaggrin), which was reported as a TCGA pan-cancer driver gene, was mutated in one single cell and was supported by reads of mutant allele in another one tumor cell and the tumor tissue (Fig. 1c, Table 2). *FLG* expression was also detected in gastric cancer cell lines by western blot and higher *FLG* expression was found in MGC803 and SGC7901 cell lines compared with that in other cell lines (Fig. 5a). Immunofluorescence experiments revealed that *FLG* was localized in the cytoplasm of all the GC cell lines (Fig. 5b). *FLG* expression was then knocked down in SGC7901 cells by shRNA (Fig. 5c). MTT and colony formation assay showed that deletion of *FLG* promoted cell growth and colony-forming abilities compared with those in control cells (Fig. 5d, e). Our data showed that *FLG* knockdown resulted in up-regulated expression of IL8, which induced inflammation. These results suggest that knockdown of *FLG* might promote cell growth through induction of inflammation, which is an inducer of GC.

4. Discussion

Until date, the intra-individual variations of GC have remained unknown. This study presents the mutation landscape of GC at the single-cell level by whole-exome single-cell sequencing for the first time and demonstrates that 24 somatic significant mutated genes (SMGs) and highlighted *CDC27* and *FLG* might alter protein conformation and potentially contribute to promoting GC progression by inducing abnormal cell growth.

In recent studies on cancer at the single-cell level, researchers discovered that genes might mutate at a low frequency of the population level but at a high frequency of the single-cell level, indicat-

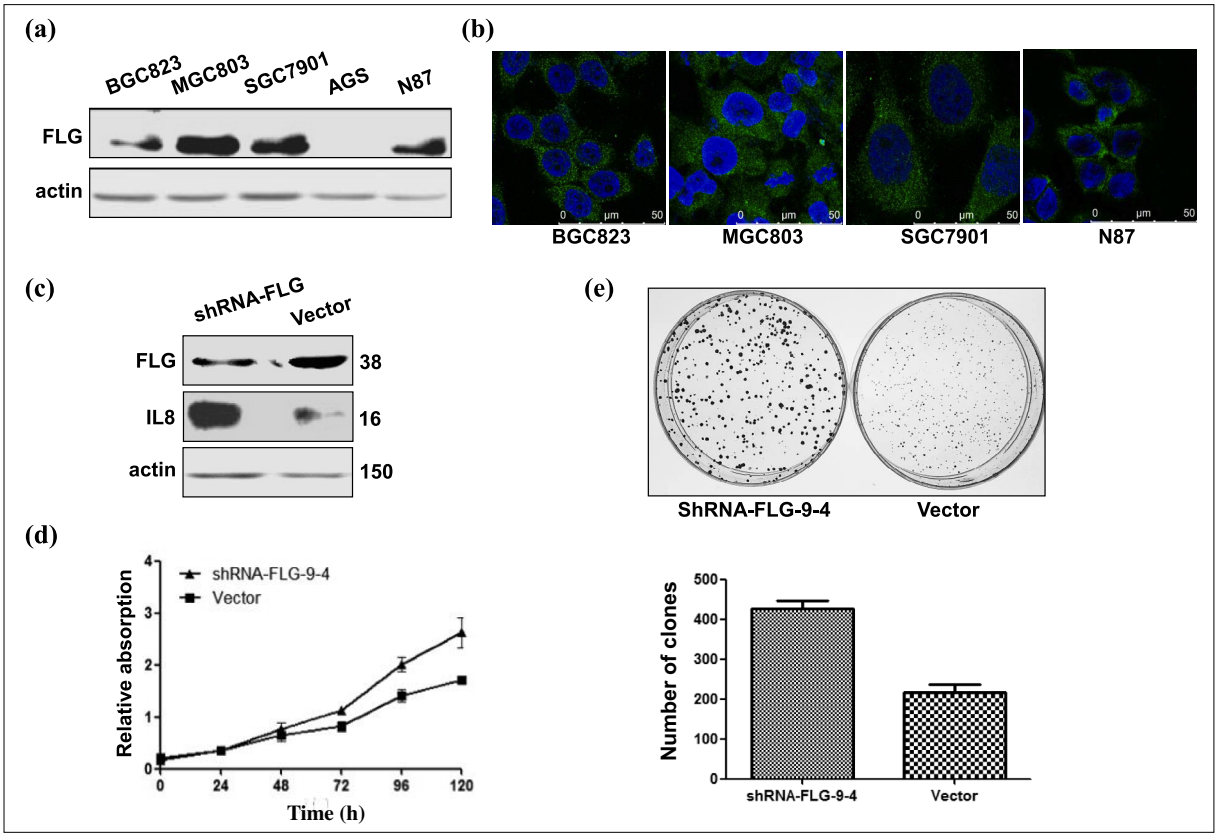


Fig. 5. Functional characterization of *FLG*. (a) Expression of FLG in GC cells detected by western blot. (b) Localization of FLG in GC cells detected by immunofluorescence. (c) The effect of *FLG* knockdown the expression of IL8 in SGC7901 cells. (d) The effect of *FLG* knockdown on cancer cell proliferation was detected by an MTT assay in SGC7901 cells and e a colony formation assay in SGC7901 cells.

ing the existence of intra-tumoral and inter-tumoral heterogeneity [5]. Analogously, we compared single cells with the corresponding tissue and 54 additional paired tissues, and found 4 groups of mutated genes with different mutation frequencies between the single-cell level and population level: high at both levels; high at the single-cell level but low among tissues, implying the driver event in an individual but not prevalent in the population; high at the population level but low among tumor cells; and low at both levels. The roles of these four groups of genes in GC were unclear and further investigations are necessary. In colon cancer, people have found a few driver events with high mutation frequencies both at single-cell and population level, such as *TP53* and *APC*, but also found *SLC12A5*, which frequently mutated among cancer cells but not population, was also potential driver [5]. In our study, we also found five genes with high mutation frequency at the single-cell level but absent in the corresponding tissue, which were also frequently mutated at the population level and might play an important role in single GC cells. As expected, *CDC27* and *FLG* promote GC progression through affecting cell growth by fingerprinting and functional characterization. Of these, *CDC27* was reported as a TCGA pan-cancer driver gene that is recurrently mutated in prostate cancer [27] and colon cancer single cells [5]. Low *CDC27* and high securin expression was reported in breast cancer patients and significantly correlated with clinical outcomes [28]. In present study, we focus to analysis of gene mutation correlated with intra-tumoral heterogeneity in GC. The protein structure determines the function of protein, hence, we used PFSC to predict the protein folding. Following gene mutation, it is not only the amino acid sequence changed, but also protein three-dimensional structure has been transformed. Future experiments in vivo, including animal models, would more definitely address the functional significance of these two genes in GC.

Single-cell genome sequencing is an effective way to provide new insights into heterogeneous human tissue samples and help to understand the progression of the cancer [29]. However, a critical question is how to design experiments that faithfully capture the true range of heterogeneity from samples of cellular populations, in which the balance between the number of cells sequenced and sequencing costs needs to be taken into account [29,30]. In early experiments and efforts [5–8], researchers have analyzed several cancer types with only one patient by no more than one hundred of single cell genome sequencing and successfully provided the intra-tumor heterogeneity, clonal evolution and driver mutations landscape of the specific cancer patients. For example, Hou et al. carried out whole exome single cell sequencing of 58 cells from a JAK2-negative myeloproliferative neoplasm patient and identified essential thrombocythemia (ET)-related candidate mutations such as SESN2 and NTRK1, which may be involved in neoplasm progression [8]. In this study, we also found the great intra-tumoral heterogeneity in a GC patient and identified functional relevant genes (such as CDC27 and FLG) in tumor cells which were missed in the SNV identification of tumor tissue based on the 43 single cells. However, with the recent experimental development, decreasing sequencing cost and huge intra-tumoral heterogeneity observed in GC, more patients and single cells should be included in order to capture not only abundant, but also rare cell types.

Gastric cancer has high heterogeneity with multiple cellular types and poor prognosis, especially diffuse-type. In our study, we found this case is more similar to diffuse-type and have *RHOA* and *CDH1* mutation. The somatic SNVs and high frequency mutations at single-cell level were sparse, possibly due to high intra-tumor heterogeneity [31]. A phylogenetic tree of cells demonstrated a large amount of "genetic noise" likely to be a stochastic progression of mutation in GC. In some single cell studies in cancer,

people have found a series of unanticipated discoveries, such as the high heterogeneity and stochastic changes in cancer-cell populations and the complicated clonal evolution mechanisms [10]. Future large-scale studies on GC single cells with more tumor stages are crucial for better understanding the intra-tumor heterogeneity and the evolution of this complex disease at the cellular level and providing novel insights for effective personalized therapies.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

We thank Fuqiang Li (BGI) and Leo J Lee for their insightful suggestions. This study was supported by the National Key Research and Development Program of China (2017YFC1308900), Beijing Municipal Commission of Health and Family Planning Project (PXM2018_026279_000005), National High-tech R&D Program of China (2012AA02A203, No.2012AA02A504) and Beijing talent fund.

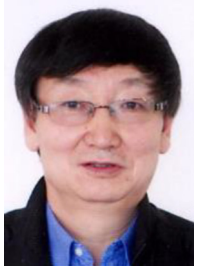
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2018.12.014.

References

[1] Wang K, Kan J, Yuen ST, et al. Exome sequencing identifies frequent mutation of arid1a in molecular subtypes of gastric cancer. Nat Genet 2011;43:1219–23.
[2] Zang ZJ, Cutcutache I, Poon SL, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet 2012;44:570–4.
[3] Wang K, Yuen ST, Xu J, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet 2014;46:573–82.
[4] Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–9.
[5] Yu C, Yu J, Yao X, et al. Discovery of biclonal origin and a novel oncogene slc12a5 in colon cancer by single-cell sequencing. Cell Res 2014;24:701–12.
[6] Li Y, Xu X, Song L, et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. GigaScience 2012;1:12.
[7] Xu X, Hou Y, Yin X, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 2012;148:886–95.
[8] Hou Y, Song L, Zhu P, et al. Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. Cell 2012;148:873–85.
[9] Demeulemeester J, Kumar P, Moller EK, et al. Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. Genome Biol 2016;17:250.
[10] Zhang X, Marjani SL, Hu Z, et al. Single-cell sequencing for precise cancer research: progress and prospects. Cancer Res 2016;76:1305–12.
[11] Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res 2015;25:1499–507.
[12] McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[13] Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–9.
[14] Wang K, Li M, Hakonarson H. Annovar: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
[15] Imielinski M, Berger AH, Hammerman PS, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 2012;150:1107–20.
[16] Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics 2011;27:175–81.
[17] Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
[18] Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
[19] Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501.
[20] Prager EM, Wilson AC. Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. J Mol Evol 1978;11:129–42.
[21] Kumar S, Tamura K, Nei M. Mega: Molecular evolutionary genetics analysis software for microcomputers. Computer applications in the biosciences. CABIOS 1994;10:189–91.
[22] Yang J. Comprehensive description of protein structures using protein folding shape code. Proteins 2008;71:1497–518.
[23] Spits C, Le Caignec C, De Rycke M, et al. Whole-genome multiple displacement amplification from single cells. Nat Protoc 2006;1:1965–70.
[24] Uzozie A, Nanni P, Staiano T, et al. Sorbitol dehydrogenase overexpression and other aspects of dysregulated protein expression in human precancerous colorectal neoplasms: a quantitative proteomics study. Mol Cell Proteomics: MCP 2014;13:1198–218.
[25] Wen D, Xu Z, Xia L, et al. Important role of sumoylation of spliceosome factors in prostate cancer cells. J Proteome Res 2014;13:3571–82.
[26] Lee SJ, Langhans SA. Anaphase-promoting complex/cyclosome protein cdc27 is a target for curcumin-induced cell cycle arrest and apoptosis. BMC cancer 2012;12:44.
[27] Lindberg J, Mills IG, Klevebring D, et al. The mitochondrial and autosomal mutation landscapes of prostate cancer. Eur Urol 2013;63:702–8.
[28] Talvinen K, Karra H, Pitkanen R, et al. Low cdc27 and high securin expression predict short survival for breast cancer patients. APMIS 2013;121:945–53.
[29] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet 2016;17:175–88.
[30] Grun D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell 2015;163:799–810.
[31] McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell 2017;168:613–28.

Lihua Peng is a research assistant in Cancer Institute at BGI-Research. Her research has been focused on multi-omics data analysis of cancer, including prostate, gastric cancer and hepatoma.



Youyong Lu is a professor and director of laboratory of molecular oncology in Peking University School of Oncology, Beijing Cancer Hospital/Institute. His research interests include cancer genomics and proteomics, genetic alterations of multiple-step carcinogenecisis in the gastric cancer; characterization of gene and protein expression profiling in the neopastic disease progression; cellular and molecular mechanism on the anti-carcinogenesis of garlic.

Shida Zhu is the Vice President of Precision Medicine at BGI-Research. Zhu's team is focused on developing and applying bioinformatics tools for NGS data mining to study diseases including birth defects, infections and cancer, aiming at identifying new biomarkers for early screening and precise diagnosis.