Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor

Xun Xu, Yong Hou, Xuyang Yin et al.

Cell 148, 886-895, 2012

Triinu Kõressaar
Seminar in Bioinformatics

TARTU 2012

What is known about clear cell renal cell carcinoma (ccRCC)?

The most **common** kidney cancer

RCC accounts for ~209,000 new cancer cases and ~102,000 deaths worldwide per year –of which 80% are ccRCC

ccRCC has relatively low mutation rate

Has very **few mutations** that are **shared** between different patients (*including VHL and PBRM1, chr 3p*)

Why to search this carcinoma and why by single cell exome sequencing?

The intratumoral heterogeneity of ccRCC remains unknown

Quantification of the heterogeneity remains difficult

Not all mutations are in VHL and PBRM1

Single nucleotide resolution (*single nucleotide changes commonly underlie tumor development*)

Noncancerous cells as well as a mixture of cancer cells that may be at different mutational stages (*accumulation of mutations during cancer progression*) are analyzed usually in one sample

Allows to analyze tumor evolution in cancers

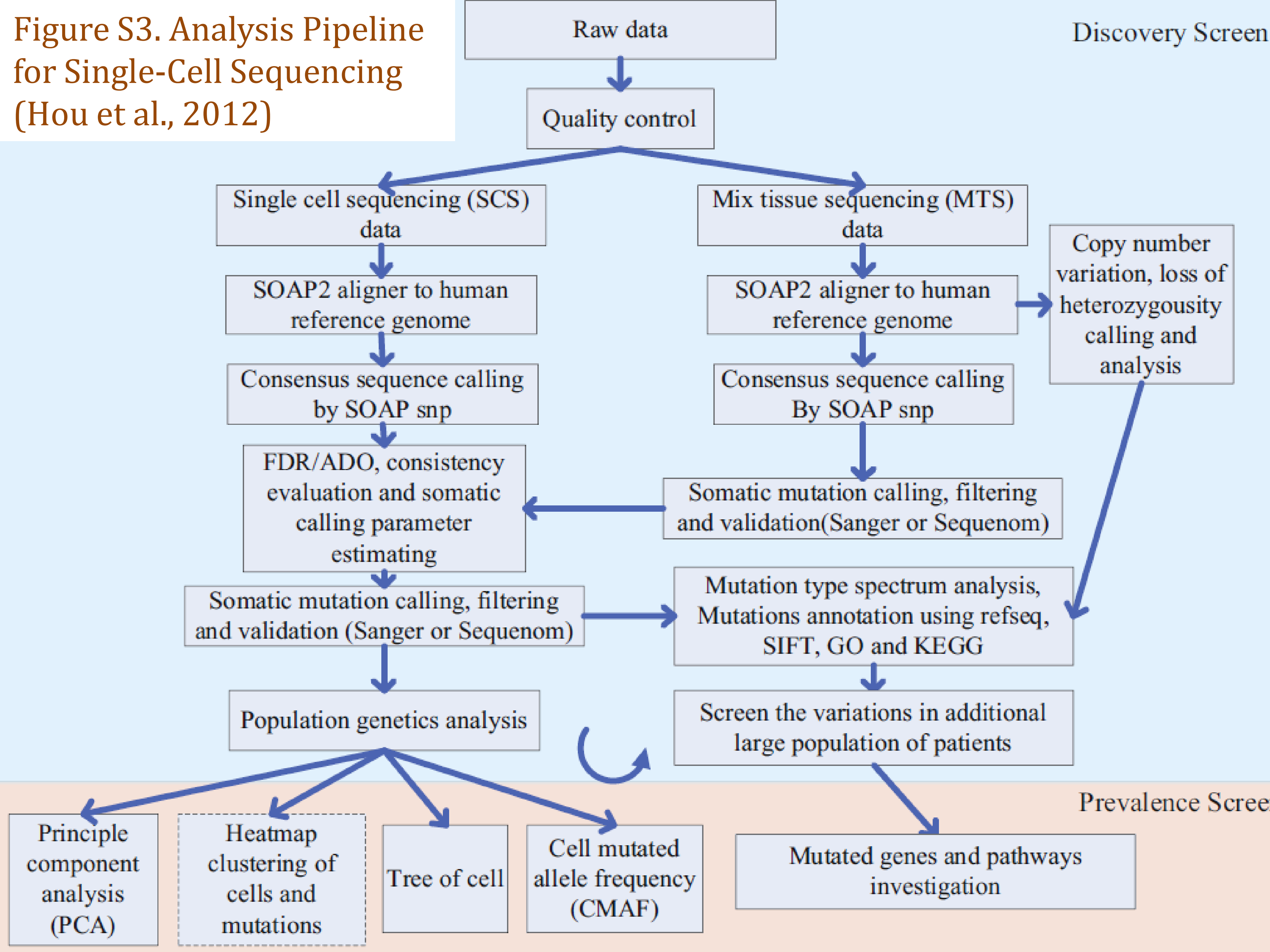


Figure S3. Analysis Pipeline for Single-Cell Sequencing (Hou et al., 2012)
Discovery Screen
Raw data
Quality control
Single cell sequencing (SCS) data
Mix tissue sequencing (MTS) data
SOAP2 aligner to human reference genome
SOAP2 aligner to human reference genome
Copy number variation, loss of heterozygousity calling and analysis
Consensus sequence calling by SOAP snp
Consensus sequence calling By SOAP snp
FDR/ADO, consistency evaluation and somatic calling parameter estimating
Somatic mutation calling, filtering and validation(Sanger or Sequenom)
Somatic mutation calling, filtering and validation (Sanger or Sequenom)
Mutation type spectrum analysis, Mutations annotation using refseq, SIFT, GO and KEGG
Population genetics analysis
Screen the variations in additional large population of patients
Prevalence Scree
Principle component analysis (PCA)
Heatmap clustering of cells and mutations
Tree of cell
Cell mutated allele frequency (CMAF)
Mutated genes and pathways investigation

Reads mapping

Multiple Displacement Amplification for WGA

Exome capture - Agilent SureSelect Platform

Reads - Illumina Hiseq 2000 platform

SOAPaligner 2.2

Human reference genome 36/Hg 18

Maximum of three mismatches, nongap mapping model, seed length 32

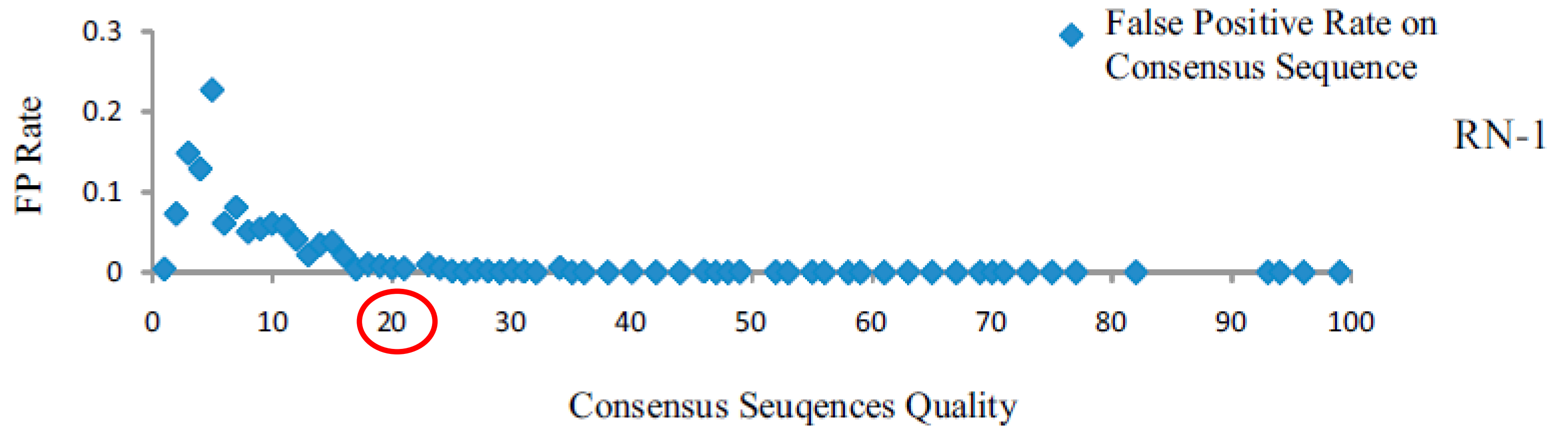
Insert size distribution of each library was checked by Eland (*Solexa Pipeline*)

Reads that could only be mapped to a unique exome capture target region were selected for consensus sequence identification

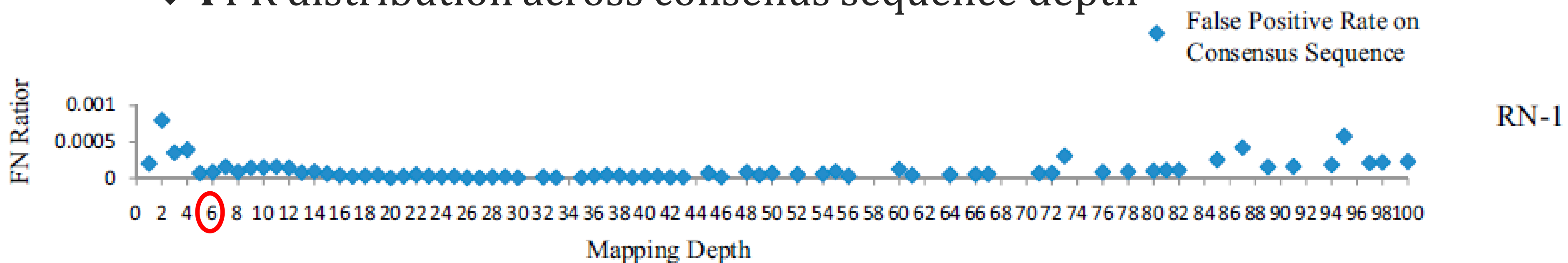
Consensus sequence calling

SOAPSNP 1.03

❖ **F**alse positive rate (FPR) distribution across consensus sequence quality



❖ **F**PR distribution across consenus sequence depth



Somatic mutation calling

Evaluated FP and FN rates

The average FP rate is $2.67*10^{-5}$
FN rate was 16.43%

The presence of three or more cells having a specific mutation in the cancer cells provided sufficient confidence to call a somatic mutation (*concluded from binomial distribution model*)

To avoid false positives, somatic mutation sites, which corresponding information in the normal mixed control was at sequencing depth <10, were removed

Exome sequencing

59-year-old Chinese male with ccRCC (*stage IV carcinoma - cancer has spread to another organ(s)*)

Exome sequencing of 25 single cells from the tumor and adjacent noncancer tissue

Table 1. Exome Sequencing Coverage of Cancer and Normal Control Tissue

	Human All Exons Coverage (%)			VHL Exons Coverage (%)			PBRM1 Exons Coverage (%)		
Sample ID	≥1×	≥10×	≥20×	≥1×	≥10×	≥20×	≥1×	≥10×	≥20×
RC-T	98.94	96.79	95.01	100	89.80	80.88	100	100	100
RN-T	97.05	80.72	65.70	100	66.67	64.25	100	99.65	89.76

"RC-T" and "RN-T" are cancer tissue and normal control tissue, respectively.

	Human All Exons Coverage (%)		
Sample ID	≥1×	≥10×	≥20×
RC-T	98.94	96.79	95.01
RN-T	97.05	80.72	65.70

260 (*229*) somatic mutation sites (*93.64% were covered by at least 10 reads*) in the coding region between the cancer and normal population (*average 78.9 mutations per single cancer cell*)

only 12 somatic mutations within the normal control population

Validation of somatic mutation calling accuracy (*PCR sequencing*):

- 35 somatic mutation sites randomly selected from three cells
- able to amplify 85 sites, 82 of these (*96.47%*) were confirmed by PCR-based capillary sequencing

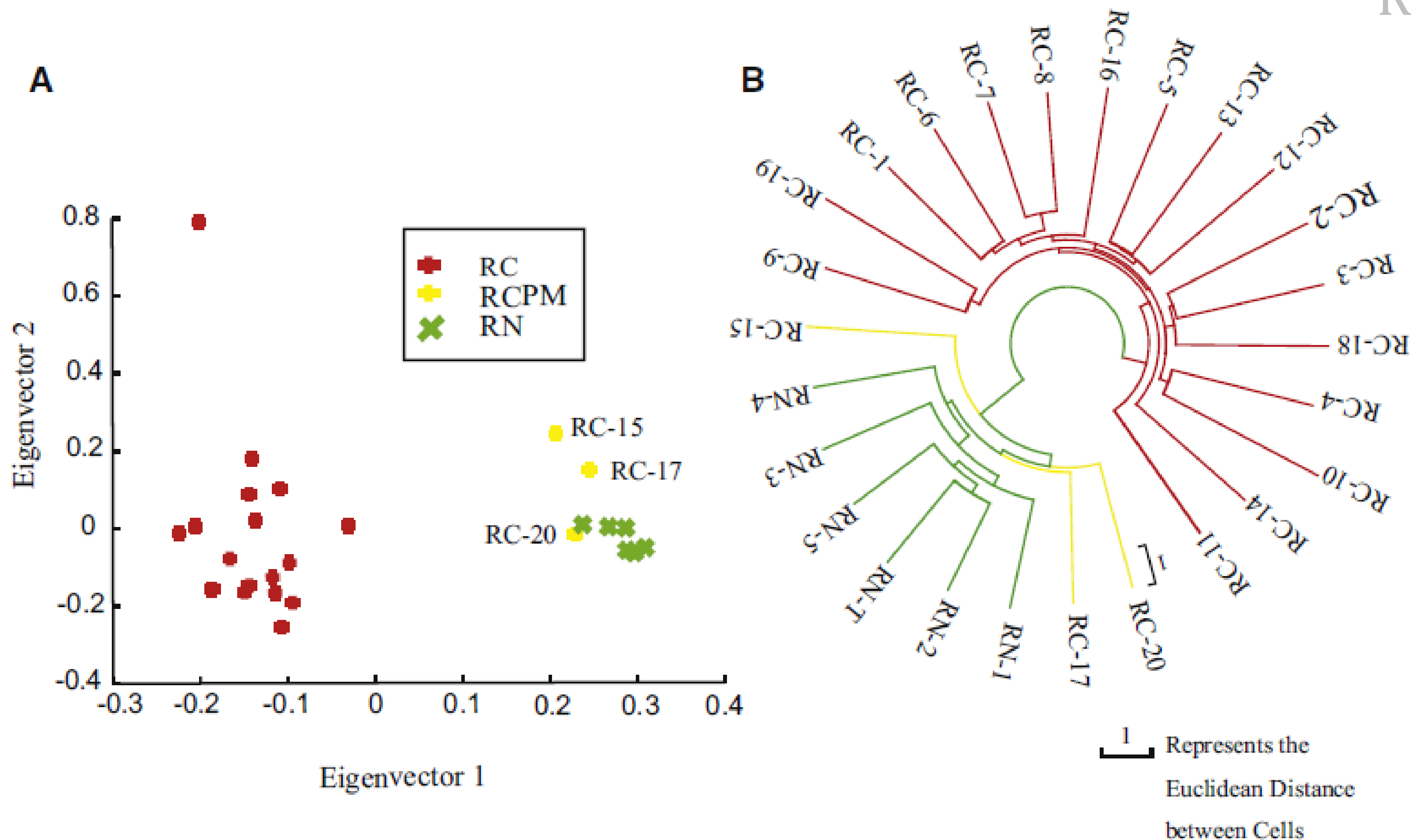


Figure 2. Somatic Mutation and Single-Cell Population Analysis in This ccRCC Patient

(A) Principle component analysis (PCA) of cancer cells (RC, red), normal control cells (RN, green), and normal cells picked as cancer cells (RCPM, yellow) based on principle component analysis (PCA).

(B) Neighbor joining phylogenetic tree constructed using sites of somatic mutation data by Euclidean distance; cancer cells (RC, red), normal control cells (RN, green), and normal cells picked as cancer cells (RCPM, yellow) are presented here. RN-T here represents the normal tissue DNA as control. After filtering the three normal cells picked as cancer cells, we identified 229 somatic mutations.

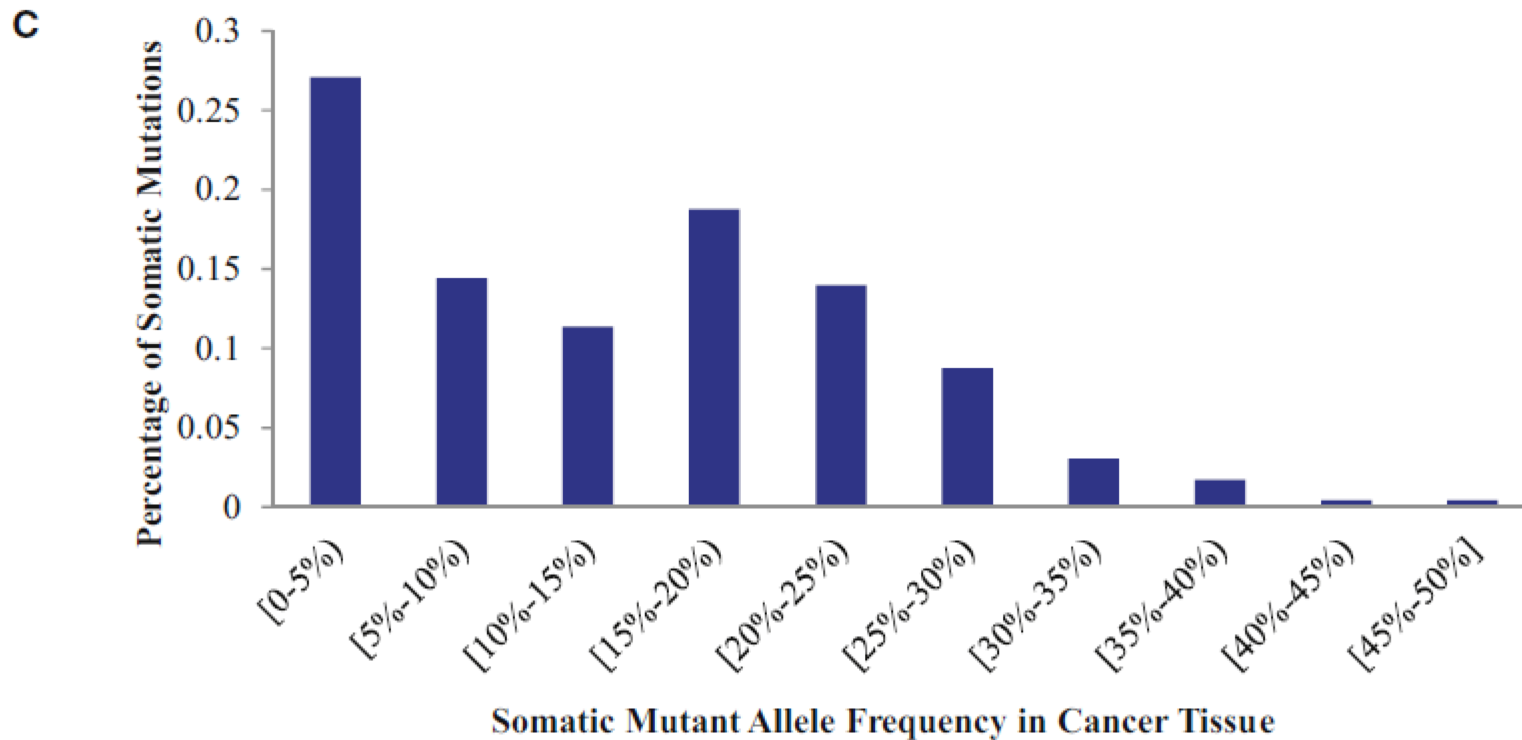


Figure 2. Somatic Mutation and Single-Cell Population Analysis in This ccRCC Patient

(C) Mutant allele frequency spectrum somatic mutations in 17 cancer cells. Based on Fisher's exact test, we separated the mutations into **COMMON** mutations *(>20% mutant allele frequency*) and **RARE** mutations (*all the rest*)

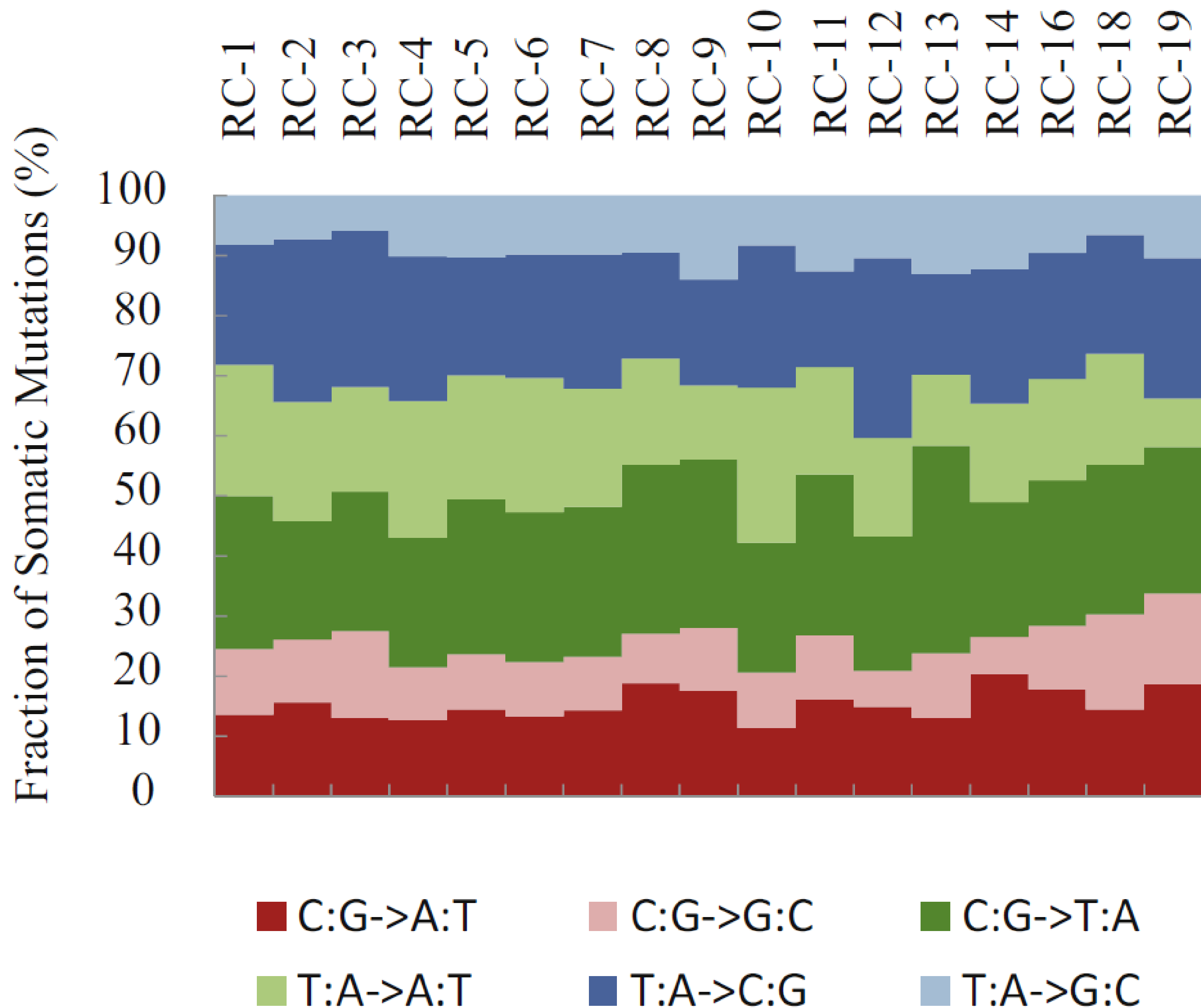


Figure 3. Somatic Mutation Pattern Spectrums

(A) Somatic mutation pattern spectrum of individual ccRCC cells

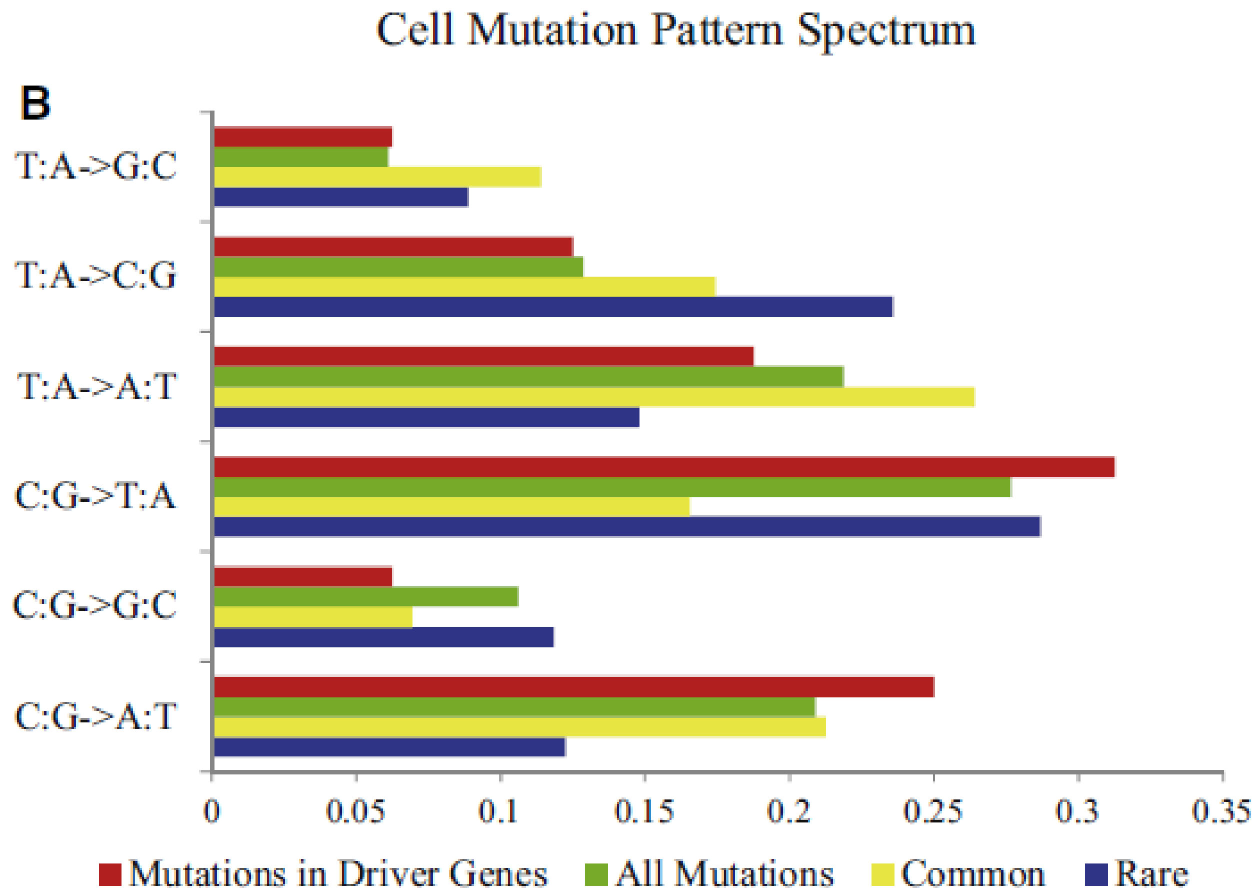
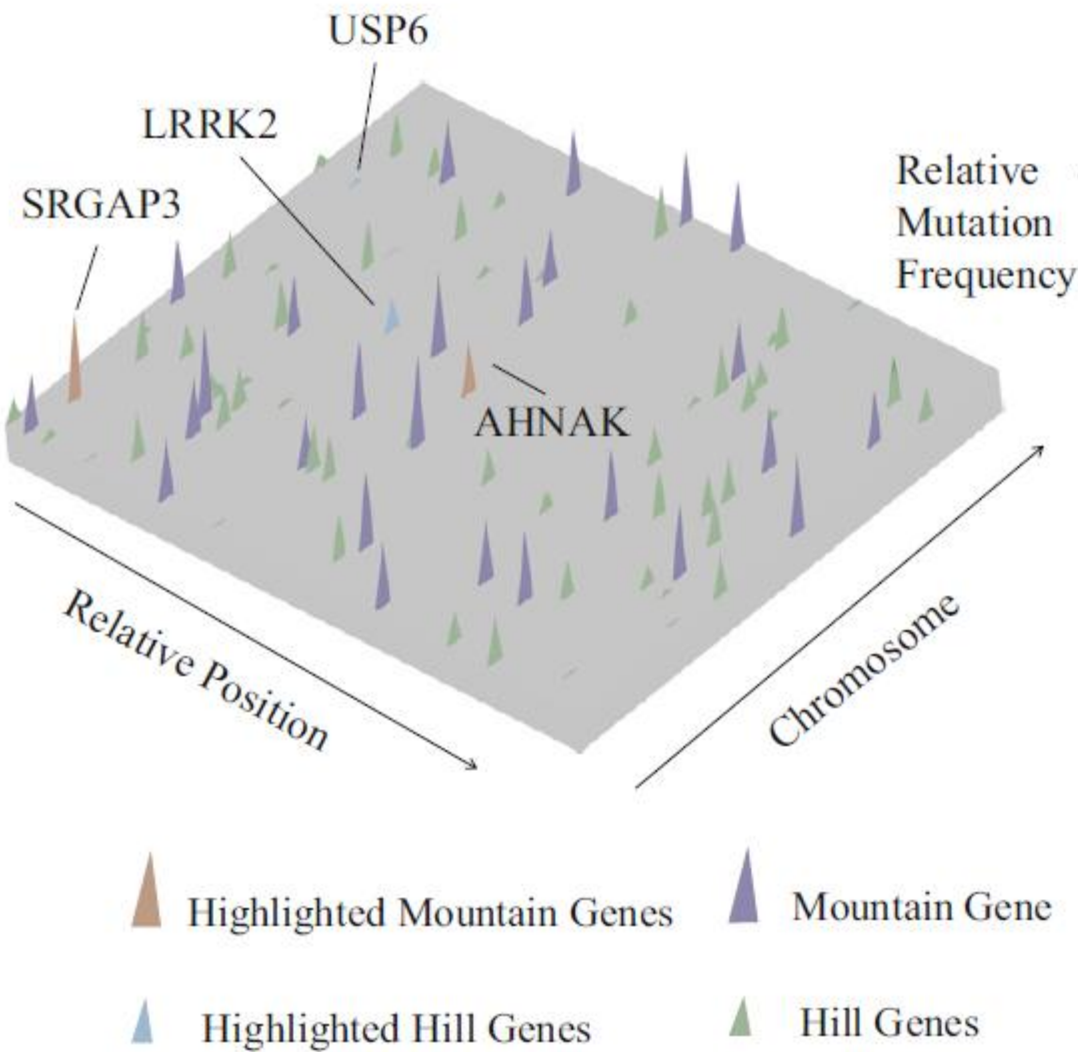


Figure 3. Somatic Mutation Pattern Spectrums

(B) Somatic mutation pattern spectrum of rare mutations (blue) and common mutations (yellow) compared with spectrum of driver mutations (red) and all nonsynonymous mutations (green) in the 98 patient cohort.



120 somatic mutations in the coding regions (NS/N=4.0)

Figure 4. Intratumoral Gene Mutation Landscape of an individual ccRCC Patient

Nonsynonymous somatic mutations are plotted in two-dimensional space, which represents chromosomal positions of mutant genes. Higher peaks (purple) —peak heights assigned a value of mutant reads ratio—indicate the **28** identified **mountain** genes. The shorter peaks (green), with peak heights assigned a value of mutant reads ratio, show the **66** identified **hill** genes. Genes recurrently mutated in the large patient cohort are marked in red (mountain) and blue (hill).

Table 2. Key Genes Identified in This Patient

Gene Name	Mutations	Patient Prevalence (%)[a]	P Value[b] (Passenger Probability)	Mutant Allele Frequency in Cancer Tissue	Mutant Cell Number	Mountain/Hill[a]
AHNAK	g.chr11:62042132G > A; p.P5445 > S	5%	9.29×10^{-9}	20%	12	M
LRRK2	g.chr12:38985956A > G; p.I1294 > V	4%	4.28×10^{-4}	8%	8	H
SRGAP3	g.chr3:9041948T > A; p.R535[a]	2%	2.92×10^{-1}	34%	16	M
USP6	g.chr17:4976948C > G; p.T72 > R	2%	3.26×10^{-1}	1.99%	3	H

[a]Patient prevalence means the mutant genes recurred in the 99 ccRCC patients (including this patient); M/H represents mountain or hill gene.
[b]Significance of the observed mutation rate over the expected mutation rate in Guo et al. (2012).

Summary

The first intratumoral genetic landscape at a single-cell level
Provides information that can lead to new ways to investigate individual tumors

Tumor did not contain any subpopulations
Different genes are characteristics of this tumor referring to genetic complexity
Common and rare mutations could be found in quantification analysis