Whole exome sequencing study of colorectal cancer in Chinese population reveals novel prevalently mutated genes and decreased mutation frequency of APC and Wnt signaling in lymph node positive cancer

Zhe Liu[1*], Chao Yang[2*], Min Yang[1§], Yong Zhou[2], Xiangchun Li[2], Wen Luo[2], Bhaskar Roy[2], Teng Xiong[2], Xiuqing Zhang[2], Huanming Yang[2,4], Jian Wang[2,4], Zhenhao Ye[3], Yang Chen[3], Xiaodong Fang[2#] & Jie Du[1#]

1 Beijing Anzhen Hospital, Capital Medical University, The Key Laboratory of Remodeling-Related Cardiovascular Diseases, Ministry of Education, Beijing Collaborative Innovation Center for Cardiovascular Disorders, Beijing Institute of Heart, Lung & Blood Vessel Disease, Beijing, China

2 BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China

3 Traditional Chinese Medicine Hospital of Guangdong Province

4 James D. Watson Institute of Genome Sciences, Hangzhou 310058, China

#Correspondence to: Jie Du (jdu@bcm.edu) or Xiaodong Fang (fangxd@genomics.cn).

*Zhe Liu and Chao Yang contributed equally to this work.

§ Current address: State Key Laboratory of Bioactive Substances and Function of Natural Medicine, Institute of Materia Medica, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100050, China

Abstract

Colorectal cancer is the fifth prevalent cancer in China. Nevertheless, a large-scale characterization of Chinese colorectal cancer mutation spectrums has not been carried out. In this study, we have performed whole exome-sequencing analysis of 98 patients' tumor samples with matched pairs of normal colon tissues using Illumina and Complete Genomics high-throughput sequencing platforms. Canonical CRC somatic gene mutations with high prevalence (>10%) have been verified, including *TP53*, *APC*, *KRAS*, *SMAD4*, *FBXW7* and *PIK3CA*. *PEG3* is identified as a novel frequently mutated gene (10.6%). *APC* and Wnt signaling exhibit significantly lower mutation frequencies than those in TCGA data. Analysis with clinical characteristics indicates that *APC* gene and Wnt signaling display lower mutation rate in lymph node positive cancer than negative ones, which are not observed in TCGA data. *APC* gene and Wnt signaling are considered as the key molecule and pathway for CRC initiation, and these findings greatly undermine their importance in tumor progression for Chinese patients. Taken together, the application of next-generation sequencing has led to the determination of novel somatic mutations and novel disease mechanisms in CRC progression, which may be useful for understanding disease mechanism and personalizing treatment for Chinese patients.

Keywords

Colorectal cancer, whole exome sequencing, disease etiology, APC, Wnt signaling

Introduction

Colorectal cancer (CRC) is the third most prevalent type of cancer worldwide. This fifth common tumor has nearly 4 millions new cases and caused 2 millions death in China each year [1, 2]. In the last decade, its incidence has increased constantly in China due to the diet and living habit change [1]. CRC patients usually attain promising clinical outcomes of early diagnosis, however, most of them failed to detect the tumor until a late stage. Around 60% of CRC reside on rectum in Chinese patients, while 40% occurs on rectum in Caucasian patients [3].

Currently, various genomic approaches have been devoted to investigate the molecular mechanisms of CRC initiation and progression. A pioneering study identified several somatic mutations in colorectal cancer, such as *TP53*, *KRAS*, *APC*, *PIK3CA* and *FBXW7* [4]. A later TCGA study has shed light on utilizing genomic data to elucidate mutation landscape of human CRC, and several novel somatic mutations were identified as well [5]. A follow-up study revealed several novel somatic mutations, such as *TCF7L2*, *TET2*, *TET3* and *ERBB3*, and illustrated possible treatment plan for colorectal cancer [6]. Whole exome sequencing (WES) study on American-African

patients discovered significant different somatic mutation genes, indicating alternative disease mechanisms in patients with different ethnic background [7]. Moreover, investigations on Iranian and Japanese patients uncover different somatic gene mutations and alternative mutation frequencies than Caucasian counterparts [8, 9]. Therefore, a whole exome sequencing of Chinese patients is essential for novel somatic gene mutation spectrums characterization, which may consequently change our understanding of disease etiology and precision medicine management.

The most relevant study of Chinese CRC using exome sequencing has used whole exome sequencing for the first CRC cancer prognostic study [10]. A few novel somatic mutation genes, such as *CDH10*, *FAT4* and *DOCK2*, are reported to be susceptible driver mutations. Notwithstanding, due to its limited sample size (22 samples) and low sequencing coverage (< 60X) strategy, its power for novel Chinese CRC gene discovery and mutation frequency characterization is limited for this high immune and stromal cell infiltration tumor [11]. Up to now, a genome-wide somatic gene mutation and frequency data are largely unknown for Chinese patients. Also, little was discussed towards analysis of the association between gene mutations and clinical characteristics for Chinese patients.

In this study, we used whole exome sequencing technology to study sporadic Chinese CRC, a single type of CRC taking for around 65-85% of the total CRC patients.

WES of 98 sporadic Chinese CRC patients' samples with matched controls is sequenced by Illumina and Complete Genomics platforms. First, we compared the somatic mutations and mutation frequency with TCGA data. Second, we carried out pathway analysis and compared them with TCGA data. Finally, we analyzed clinical characteristics by their association with somatic mutated genes.

Methods

Sample collection and preparation

During this study, 40 and 58 unrelated Chinese patients with colorectal cancer were recruited from Second Affiliated Hospital of Nanchang University and Wenzhou Medical University. They were referred for surgery from 2008 to 2010. These unrelated patients do not have a family history of colorectal cancer. All patients provided a written informed consent according to the research protocol approved by the Ethical Review Board at two hospitals. Fresh tissue samples were dissected and stored at liquid nitrogen. A description of the clinical characteristics is shown in Supplementary Table S1.

Illumina Exome Capture and Sequencing

Human genomic DNA was extracted from frozen tumor or matched colon tissue with QIAampDNA Mini kits (Qiagen). Covaris system is used to obtain the fragmented

genomic DNA between 150 to 200bp. Adaptors were ligated to both ends of the fragments and adaptor-ligated templates were then purified using AgencourtAMPure SPRI beads. Whole-exome enrichment was performed by the SureSelect Human All Exon 51M kit (Agilent). Captured DNA libraries are sequenced by HiSeq 2000 Genome Analyzer (Illumina), resulting of 90 base paired-end reads.

Complete Genomics Exome Capture and Sequencing

Whole-exome sequencing was performed by a sequencing-by-ligation method [12]. Briefly, fragmented genomic DNA was generated by Covaris system with length between 200-400bp. After ligation and PCR amplification, the whole exome is enriched by Exome 59M kit (BGI). Then, a second round PCR amplification is performed. The resulted DNA fragments are prepared for Complete Genomics Black Bird Sequencer, and 30–35 base paired-end reads are finally obtained.

Bioinformatics analysis

Illumina variant detection

The following pipeline was carried out with default parameters unless explicitly described. First, clean reads were obtained by removing adapter reads and low quality reads (≥10% of the bases are N, or ≥50% of the bases with Phred score ≤5). Burrows-Wheeler Aligner (BWA) was used to align the clean reads to the human

reference genome (UCSC Genome Browser hg19) with parameters "*-o 1 -e 50 -m 100000 -t 4 -i 15 -q 10 -I"* [13]. SAMtools was used to convert the SAM-formatted alignment results to BAM-formatted alignment files. Local read alignment re-calibration was performed by Genome Analysis Toolkit (GATK IndelRealigner) [14]. Finally, Picard toolkit was used to mark duplicates. [15]. **Somatic SNV detection**: Mutect v1.1.4 was used to compare tumor BAM files against their matched control BAM files for somatic single nucleotide variants (SNVs) identification [16]. They are filtered with the following requirements: *minimum coverage ≥ 10X, mutation allele fraction ≥10% and ≥5 reads*. **Somatic Indel detection**: GATK was used to detect somatic InDels with following parameters: *minCoverage=6, minNormalCoverage=4, minFraction=0.3*. False positive InDels were finally removed by in-house scripts.

Complete Genomics variant detection

The resulting mate-paired reads were aligned to the reference genome (hg19) and variants are called by the reported methods [17].

Statistical analysis

Qualitative variables were compared by Fisher's exact test. *T-test* was used for normal distributed data comparison, and Wilcoxon rank test was used for non-normal distributed data. All of the statistical analyses were performed in R or Bioconductor

environment.

Results

Sequencing statistics

To identify genetic variants involved in CRC, we performed whole-exome sequencing of the tumor and matched controls on genomic DNA from 40 (Second affiliated Hospital of Nanchang University) and 58 patients (Wenzhou Medical University) by Illumina HiSeq2000 (ILMN) and Complete Genomics (CG) Blackbird sequencers, respectively. The Illumina and CG platforms attained 117X and 168X average coverage for exon captured regions as shown in Figure 1a. All of the samples achieved 10X and 20X coverage rate more than 95% and 90% of the genome for Illumina and CG platforms. A description of sequencing and mutation statistics can be found in Supplementary Table S2.

It was reported that CG platform has high concordance rate for SNV detection with Illumina platform [18]. However, it is necessary to evaluate system bias from two sequencers by their ability to discover somatic mutations. This was verified by no segregation of mutation numbers in coding region from two platforms (Figure 1b). Also, no obvious differences were observed between the mutation counts ($p=0.29$, Wilcoxon rank test). Moreover, all types of SNVs (non-synonymous SNV, synonymous

SNV, stopgain SNV, splicing) and InDels (non-frameshift deletion, non-frameshift insertion, frameshift deletion, frameshift insertion) do not show statistical differences in terms of mutation counts between two platforms. Taken together, it is reasonable to merge these two datasets for further study. Mutation rates of tumors are around 3/Mb for each sample (Figure 1c), which is consistent with previous western and Chinese CRC whole exome sequencing results [5, 10, 19].

Overall, 13 patients have considerably higher mutation rate (Figure 1b), and we treat them as hypermutated samples. CG and Illumina platforms are able to identify 2.85M and 2.95M mutations in coding region for the non-hypermutated samples. A report of the detected SNVs and InDels are shown in Supplementary Table S3 and S4.

Somatic mutational spectrum

In general, dominant somatic SNVs in colorectal cancer are *CpG->T mutations as shown in Figure 2a. To investigate the origin of somatic mutations, we clustered samples into subgroups based on the six mutation types (Supplementary Appendix A). We stratified the patients into hypermutated and regularly mutated subgroups for Chinese data. Similar results are also derived from TCGA data, indicating the consistency of the clustering results. The differences between the mutation spectra of the two subgroups are quite obvious. Mutations in regularly mutated samples are mainly *CpG->T, while hypermutated samples are dominated by three mutation peaks

at TCT>TAT, TTT>TGT and TCG>TTG.

Somatic mutation

We performed somatic gene mutation analysis of the hypermutated samples due to its unique tumorigenisis process and clinical outcome. It is known in the previous research that most of the microsatellite instability (MSI) tumors are hypermutated. This is also validated in our study, and 8 out of 11 the hypermuated samples are MSI. The rest 3 hyper-mutated and microsatellite stability (MSS) samples displays mutations on *MSH4* and *POLE*. Mutations on *MSH4* and *POLE* may affect DNA binding and DNA replication, both of which will induce ultra-high mutation numbers. Moreover, canonical CRC genes, including *APC*, *PIK3CA*, *MSH6* and *FAT4* are observed in mutation status. It is interesting to discover *TP53* in lower mutation prevalence (2 out of 13) for hypermutated samples, which may indicate alternative disease mechanisms for hypermutated samples. A report of the prevalently mutated genes is shown in Supplementary Table S5.

We then discuss the somatic mutations in non-hypermutated samples. In general, Identification of driver mutations from passenger mutations is a challenging task due to following reasons: a) CRC is a type of cancer with high immune and stromal cells infiltration and somatic mutation signal will be diminished [20]. b) Patients may have sub-clone with different somatic mutated genes, and a mixture of the two or more

sub-clone will further decrease somatic mutation signal [21, 22]. c) Disease etiology for a fraction of patients may be induced by different somatic mutations and a long tail of genes was postulated to explain the CRC initiation and progression of entire population [11]. d) Consistency between different significantly mutated genes algorithms are not perfect, and genes shown up as statistical significant results from one algorithm may disappear in another one [23, 24].

In this study, we used the following rules to discuss the susceptible driver somatic mutations. 1) We used MutsigCV, one of the most widely used algorithms, for significantly mutated gene (SMG) detection [25]. 2) We reported prevalently mutated somatic genes (>5% of the entire population) due to its possibility as diagnostic and treatment targets on population level. 3) We discussed the mutated genes supported by additional literatures. 4) We used two gene expression data (GSE50760 and GSE18105) from GEO database as gene expression evidences. Genes with log fold change>1 and adjusted p-value<0.05 between tumors and matched controls are considered as differentially expressed genes [26, 27]. A report of these genes was presented in Supplementary Table S6 and Figure 3a.

Seven significant mutated genes were identified, including 6 classical CRC genes (*TP53*, *APC*, *KRAS*, *FBXW7*, *PIK3CA* and *SMAD4*[5, 6]) and a novel CRC gene (PEG3). *PEG3* accounts for 9 (10.6%) of the patients compared with 2.7% mutation

prevalence in TCGA data. Its aberrations on multiple levels, including gene expression, methylation and mutation have been reported in different types of tumors, including myeloma [28], ovarian cancers [29] and cholangiocarcinoma [30]. Moreover, expression of *PEG3* in colorectal patients shows statistically significant lower gene expression than matched controls from two independent studies using RNA-Seq and microarray platforms [26, 27]. Its down expression is also reported to be statistically significant associated with patients' survival in various types of tumors [31]. *PEG3* acts as a tumor suppressor gene by binding and promoting the degradation of β-catenin, which will consequently inhibit Wnt Signaling [32]. Its molecular role is to interact with *SIAH1* and induce *TP53* mediated apoptosis or bind with *TRAF2* to initiate *NFKB* and *MAPK* pathways [33, 34].

A few other genes with mutation frequency greater than 5% were reported in this study and their association with CRC is revealed previously, such as *HMCN1* [35], *SYNE1 [10]*, *NEB* [36], *OBSCN* [37], *MUC16*, *RYR2* [38] and *FAT4* [10]. A few genes with little study in CRC also show medium mutation prevalence, including *ABCA13*, *ABCC10*, *CEP192*, *DCHS2*, *DNAH5*, *DYNC1H1*, *PTPRT*, *MYO16*, *LRRK2*, *DYNC2H1*, *FBN1*, *FCGBP*, *FLG*, *FREM2*, *PAPPA2*, *RP1L1*, *TTN*, *ZFHX4* and *ZNF717*. These genes may interact with the canonical CRC genes in a network fashion as illustrated in Figure 3b.

Two genes display significantly different mutation frequencies between Chinese and TCGA data. The mutation rate of *APC* gene is 0.435 and 0.786 in Chinese and TCGA data (Figure 3d). This result is validated by previous Chinese exome-sequencing results (0.56 mutation frequency, p=0.0001 from Fisher' exact test) [10]. *APC* gene carries a hotspot mutation Q1429 in Chinese population, differs from the hotspot mutation (R1450) in TCGA data (Figure 3c). Two novel hotspot mutations, namely R273C/H, R282W on *TP53* gene and R265C/H on *SMAD4* gene were shown in this dataset (Supplementary Figure S1). Additionally, *KRAS* and *PIK3CA* genes are less frequently mutated but not statistically significant in Chinese patients were also observed in this dataset.

Pathway analysis of mutated genes

Significantly mutated pathways (SMP) in Chinese population agree with TCGA results, including canonical pathways, such as Wnt/beta-catenin signaling, Cell Cycle/Apoptosis, MAPK signaling, TGF-beta signaling and PI(3)K signaling [5] as shown in Figure 4a [5]. It should be noted that *APC* pathway is identified as significantly mutated pathway in TCGA data, but not in our data. This is consistent with gene mutation frequency alteration results. Different pathway mutation frequency is observed, such as lower mutations rate on Wnt/beta-catenin signaling (44.7% versus 67.2%, p-value= 1.514E-5) and MAPK signaling (41.2% versus 51.2%,

p-value=0.04485), and higher mutations rate on DNA damage control (64.7% versus 55.2%), Genome Integrity (62.4% versus 56.8%) and Cell Cycle/Apoptosis (65.9% versus 56.8%). Illustration of the pathway mutation frequency and associated genes are shown in Figure 4b and a report of the significantly mutated pathways is shown in Supplementary Table S7.

Discussion

We investigated the Chinese CRC tumorigenesis process by a whole exome sequencing in this study. We validated well-known somatic mutations in CRC, such as *TP53, APC, KRAS* and discovered a few high prevalent novel somatic mutations, including *PEG3* and *PTPRT.* Their mutation frequencies were also compared with that in TCGA data, and significant alternative frequency in *APC* or *PEG3* was observed. Pathway analysis of somatic mutations indicates a higher mutation frequency in cell cycle and lower mutation frequency in Wnt or MAPK signaling pathway.

The clinical characteristics association analysis was taken for genes with >5% mutation prevalence in regularly mutated samples and associations results of *TP53*, *APC*, *KRAS*, *PEG3* or *PIK3CA* genes are shown in Table 1. *RP1L1* and *SYNE1* genes were significantly enriched in male and female patients. *PEG3* and *RYR2* displayed a significant higher association with early colorectal cancer (onset age≤45). *NEC* and *SYNE1* were enriched in colon tumor, and *TP53* was enriched in rectum tumor. These

suggest a cause for higher rectum tumor incidence amongst Chinese patients than Caucasian patients. *APC* mutation frequency decreased with the increment of TNM stages, I+II (59.5%), III (27%) and IV (25%) (p-value = 0.006), and this pattern still persists when combining TNM III and TNM IV for this analysis. In contrast, TCGA data showed consistent mutation frequencies among TNM I+II (80%), III (76%) and IV (78.1%) patients (Supplementary Table S8). Our findings on alterations of the mutation frequency of *APC* in Chinese patients are consistent with the previous published results [39, 40]. Our identification on mutational pathways shows rates of Wnt/beta-catenin signaling in Chinese patients were decrease from 0.619 (TNM I+II) to 0.293 (TNM III+IV), while mutational analysis of pathways of DNA damage control, genome integrity and cell Cycle/Apoptosis did not show a significant change. Meanwhile, our analysis of the mutation rate of TCGA data showed a consistence between TNM I+II (0.829) and TNM III+IV (0.793) stages for Wnt/beta-catenin signaling.

It is widely accepted that sequential mutations on APC->KRAS->TP53 is a key CRC development pathway [41-43]. The reduction of APC mutations rate in TNM III+IV patients undermine its importance in Chinese patients tumor progression. This will greatly shape our understanding about tumor progression, and may have direct clinical impact on Chinese CRC precision medicine management, for example, APC has been recommended as prognostic gene for survival prediction in Caucasian patients and our finding of lower mutation rate of APC may diminish its utility in Chinese

patients [44]. The possible reason for lower APC mutation frequency in TNM III+IV tumor could be due to complementary mechanisms like PEG3, which has been reported to inhibit Wnt signaling by degrading β-catenin [32].

Acknowledgements

Thanks are given to Rongfa Yuan, Jianghua Shao and Chunmei Piao who have helped us to collect the patient samples, and Prasanth Bhatt who proofread this manuscript.

Conflict of Interest Statement

None declared

Funding

This project is supported by National High-tech R&D Program (863 Program) [2012AA02A201], Natural Science Foundation of China [81672818] and Guangzhou Science and Technology Program Key Projects [201604020005].

Authors' contributions

JD and XDF designed the project, MY initiated the project and collected the samples, CY, ZL, YZ, XCL and WL performed the bioinformatics analysis; ZL, YC,

YZ, XDF and JD wrote the paper. All of the authors agree the manuscript.

Abbreviations

CRC: colorectal cancer, TCGA: The Cancer Genome Atlas, SNV: Single Nucleotide variation, InDel: Insertion and deletion, COSMIC: the Catalogue Of Somatic Mutations In Cancer

References

1. Chen W, Zheng R, Baade PD, *et al.* Cancer statistics in China, 2015. *CA Cancer J Clin* 2016; **66**: 115-132.

2. Siegel RL, Miller KD, Fedewa SA, *et al.* Colorectal cancer statistics, 2017. *CA: A Cancer Journal for Clinicians* 2017; **67**: 177-193.

3. Xu AG, Yu ZJ, Jiang B, *et al.* Colorectal cancer in Guangdong Province of China: a demographic and anatomic survey. *World J Gastroenterol* 2010; **16**: 960-965.

4. Wood LD, Parsons DW, Jones S, *et al.* The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* 2007; **318**: 1108-1113.

5. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012; **487**: 330-337.

6. Seshagiri S, Stawiski EW, Durinck S, *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* 2012; **488**: 660-664.

7. Guda K, Veigl ML, Varadan V, *et al.* Novel recurrently mutated genes in African American colon cancers. *Proceedings of the National Academy of Sciences of the United States of America* 2015; **112**: 1149-1154.

8. Ashktorab H, Mokarram P, Azimi H, *et al.* Targeted exome sequencing reveals distinct pathogenic variants in Iranians with colorectal cancer. *Oncotarget* 2017; **8**: 7852-7866.

9. Dobbins SE, Houlston RS, Nagahashi M, *et al.* Genomic landscape of colorectal cancer in Japan: clinical implications of comprehensive genomic sequencing for precision medicine. *F1000Res* 2016; **8**: 136.

10. Yu J, Wu WK, Li X, *et al.* Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. *Gut* 2015; **64**: 636-645.

11. Giannakis M, Mu XJ, Shukla SA, *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma (vol 15, pg 857, 2016). *Cell Reports* 2016; **17**: 1206-1206.

12. Drmanac R, Sparks AB, Callow MJ, *et al.* Human genome sequencing using

unchained base reads on self-assembling DNA nanoarrays. *Science* 2010; **327**: 78-81.

13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754-1760.

14. Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078-2079.

15. 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**: 1297-1303.

16. Cibulskis K, Lawrence MS, Carter SL, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013; **31**: 213-219.

17. Carnevali P, Baccash J, Halpern AL, *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* 2012; **19**: 279-292.

18. Lam HYK, Clark MJ, Chen R, *et al.* Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology* 2012; **30**: 78-U118.

19. Seshagiri S, Stawiski EW, Durinck S, *et al.* Recurrent R-spondin fusions in

colon cancer. *Nature* 2012; **488**: 660-+.

20. Giannakis M, Mu XJ, Shukla SA, *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep* 2016; **17**: 1206.

21. 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**: 701-712.

22. Wu H, Zhang XY, Hu Z, *et al.* Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. *Oncogene* 2016.

23. Tokheim CJ, Papadopoulos N, Kinzler KW, *et al.* Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* 2016; **113**: 14330-14335.

24. Hofree M, Carter H, Kreisberg JF, *et al.* Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat Commun* 2016; **7**: 12096.

25. Lawrence MS, Stojanov P, Polak P, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013; **499**: 214-218.

26. Kim SK, Kim SY, Kim JH, *et al.* A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol Oncol* 2014; **8**: 1653-1666.

27. Matsuyama T, Ishikawa T, Mogushi K, *et al.* MUC12 mRNA expression is an

independent marker of prognosis in stage II and stage III colorectal cancer. *Int J Cancer* 2010; **127**: 2292-2299.

28. Bolli N, Avet-Loiseau H, Wedge DC, *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* 2014; **5**: 2997.

29. Feng W, Marquez RT, Lu Z, *et al.* Imprinted tumor suppressor genes ARHI and PEG3 are the most frequently down-regulated in human ovarian cancers by loss of heterozygosity and promoter methylation. *Cancer* 2008; **112**: 1489-1502.

30. Ong CK, Subimerb C, Pairojkul C, *et al.* Exome sequencing of liver fluke-associated cholangiocarcinoma. *Nat Genet* 2012; **44**: 690-693.

31. Li M, Sun Q, Wang X. Transcriptional landscape of human cancers. *Oncotarget* 2017; **8**: 34534-34551.

32. Jiang X, Yu Y, Yang HW, *et al.* The imprinted gene PEG3 inhibits Wnt signaling and regulates glioma growth. *J Biol Chem* 2010; **285**: 8472-8480.

33. Relaix F, Wei XJ, Li W, *et al.* Pw1/Peg3 is a potential cell death mediator and cooperates with Siah1a in p53-mediated apoptosis. *Proceedings of the National Academy of Sciences of the United States of America* 2000; **97**: 2105-2110.

34. Relaix F, Wei XJ, Wu X, *et al.* Peg3/Pw1 is an imprinted gene involved in the

TNF-NFkappaB signal transduction pathway. *Nat Genet* 1998; **18**: 287-291.

35. Lee SH, Je EM, Yoo NJ, *et al.* HMCN1, a cell polarity-related gene, is somatically mutated in gastric and colorectal cancers. *Pathol Oncol Res* 2015; **21**: 847-848.

36. Kortum KM, Langer C, Monge J, *et al.* Longitudinal analysis of 25 sequential sample-pairs using a custom multiple myeloma mutation sequencing panel (M(3)P). *Ann Hematol* 2015; **94**: 1205-1211.

37. Balakrishnan A, Bleeker FE, Lamba S, *et al.* Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. *Cancer Res* 2007; **67**: 3545-3550.

38. Bueno R, Stawiski EW, Goldstein LD, *et al.* Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nat Genet* 2016; **48**: 407-416.

39. Ko JM, Cheung MH, Kwan MW, *et al.* Genomic instability and alterations in Apc, Mcc and Dcc in Hong Kong patients with colorectal carcinoma. *Int J Cancer* 1999; **84**: 404-409.

40. May the APC gene somatic mutations in tumor tissues influence the clinical

features of Chinese sporadic colorectal cancers? *Acta Oncologica* 2007; **46**: 757-762.

41. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med* 2009; **361**: 2449-2460.

42. Jones S, Chen WD, Parmigiani G, *et al*. Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci U S A* 2008; **105**: 4283-4288.

43. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61**: 759-767.

44. Schell MJ, Yang M, Teer JK, *et al.* A multigene mutation classification of 468 colorectal cancers reveals a prognostic role for APC. *Nature Communications* 2016; **7**: 11743.

Table 1 Clinical association with somatic mutated genes

		TP53		APC		KRAS		PIK3CA		PEG3	
		Mutation	P-value	Mutation	P-value	Mutation	P-value	Mutation	P-value	Mutation	P-value
Age	<=45	6 (85.7%)	0.233	0 (0%)	0.016	2 (28.6%)	1.000	0 (0%)	1.000	3 (42.9%)	0.017
	>45	44 (57.1%)		37 (48.1%)		27 (35.1%)		8 (10.4%)		5 (6.5%)	
Sex	Male	22 (56.4%)	0.659	21 (53.8%)	0.124	14 (35.9%)	0.822	4 (10.3%)	1.000	3 (7.7%)	0.719
	Female	28 (62.2%)		16 (35.6%)		15 (33.3%)		4 (8.9%)		5 (11.1%)	
TNM stage	I+II	28 (65.1%)	0.554	26 (60.5%)	0.006	21 (48.8%)	0.015	5 (11.9%)	0.813	3 (7.1%)	0.645
	III	20 (54.1%)		10 (27%)		7 (18.9%)		3 (8.1%)		5 (13.5%)	
	IV	2 (50%)		1 (25%)		1 (25%)		0 (0%)		0 (0%)	
Lymph node	Negative	28 (65.1%)	0.377	26 (60.5%)	0.004	21 (48.8%)	0.011	5 (11.6%)	0.714	3 (7%)	0.473
	Positive	22 (55%)		11 (27.5%)		8 (20%)		3 (7.5%)		5 (12.5%)	
Tumor sites	Right colon	6 (35.3%)	0.050	5 (29.4%)	0.330	8 (47.1%)	0.234	2 (11.8%)	0.642	2 (11.8%)	1.000
	Left colon	10 (62.5%)		8 (50%)		3 (18.8%)		2 (12.5%)		1 (6.3%)	
	Rectum	30 (69.8%)		22 (51.2%)		15 (34.9%)		3 (7%)		4 (9.3%)	
Differentiation	Low	2 (28.6%)	0.119	3 (42.9%)	0.329	3 (42.9%)	0.867	1 (14.3%)	0.680	1 (14.3%)	0.628
	Medium	44 (63.8%)		34 (49.3%)		25 (36.2%)		7 (10.1%)		6 (8.7%)	
	High	3 (100%)		0 (0%)		1 (33.3%)		0 (0%)		0 (0%)	

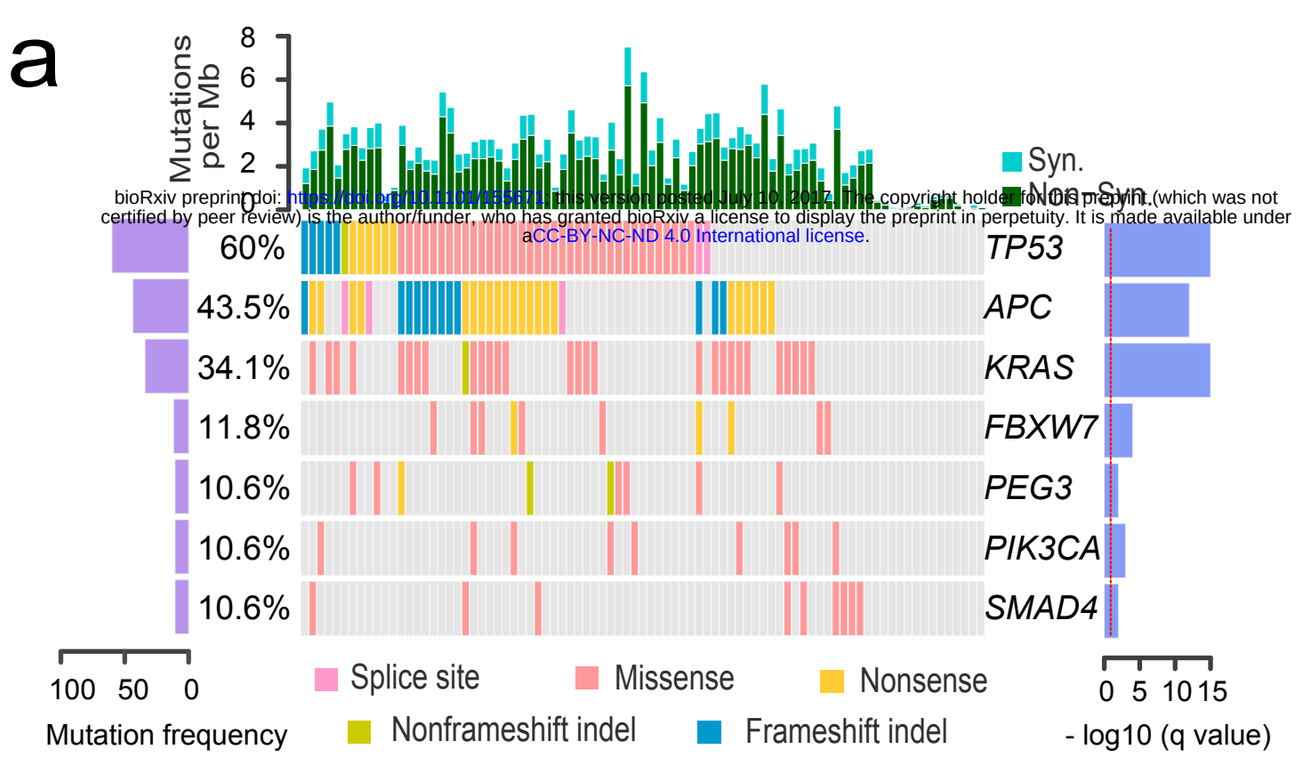
Figure 1. Sequencing statistics for Illumina and CG platforms. (a) This graph displays the average sequencing depth and exome coverage percentage with >10X and >20X for Illumina and Complete Genomics platforms. (b) This chart illustrates the mutations in coding regions for Illumina and Complete Genomics platforms. The dash line is used to separate samples into hyper-mutated and regularly mutated ones. (c) A display of the various categories of mutations across samples is shown for SNVs (non-synonymous SNV, synonymous SNV, stopgain SNV and splicing) and InDels (non-frameshift deletion, non-frameshift insertion, frameshift deletion and frameshift insertion).

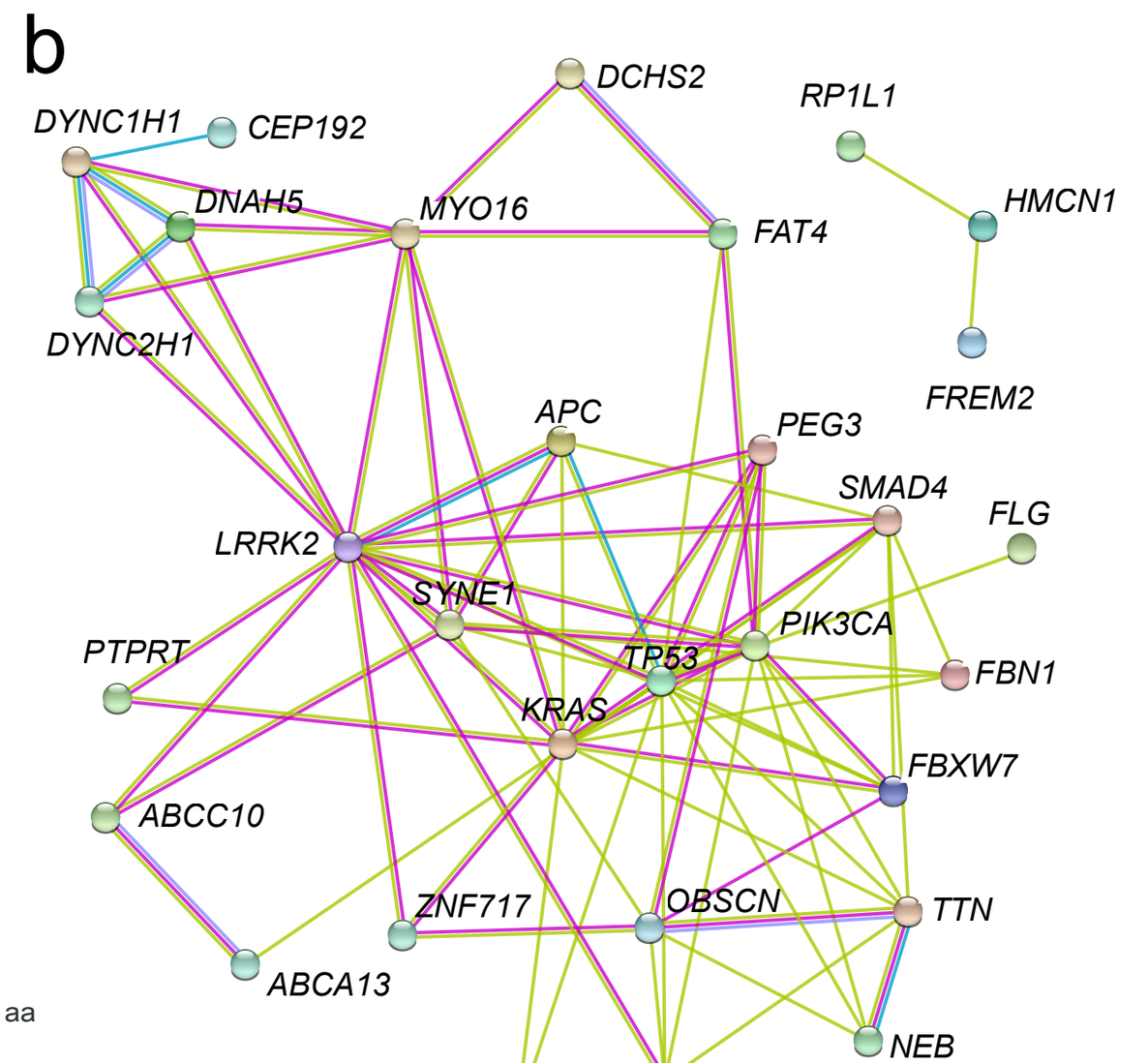
Figure 2. Somatic mutation spectrums. (a). Mutation spectrums for hypermutated and regularly mutated Chinese samples. (b). Mutation spectrums for hypermutated and regularly mutated TCGA samples.

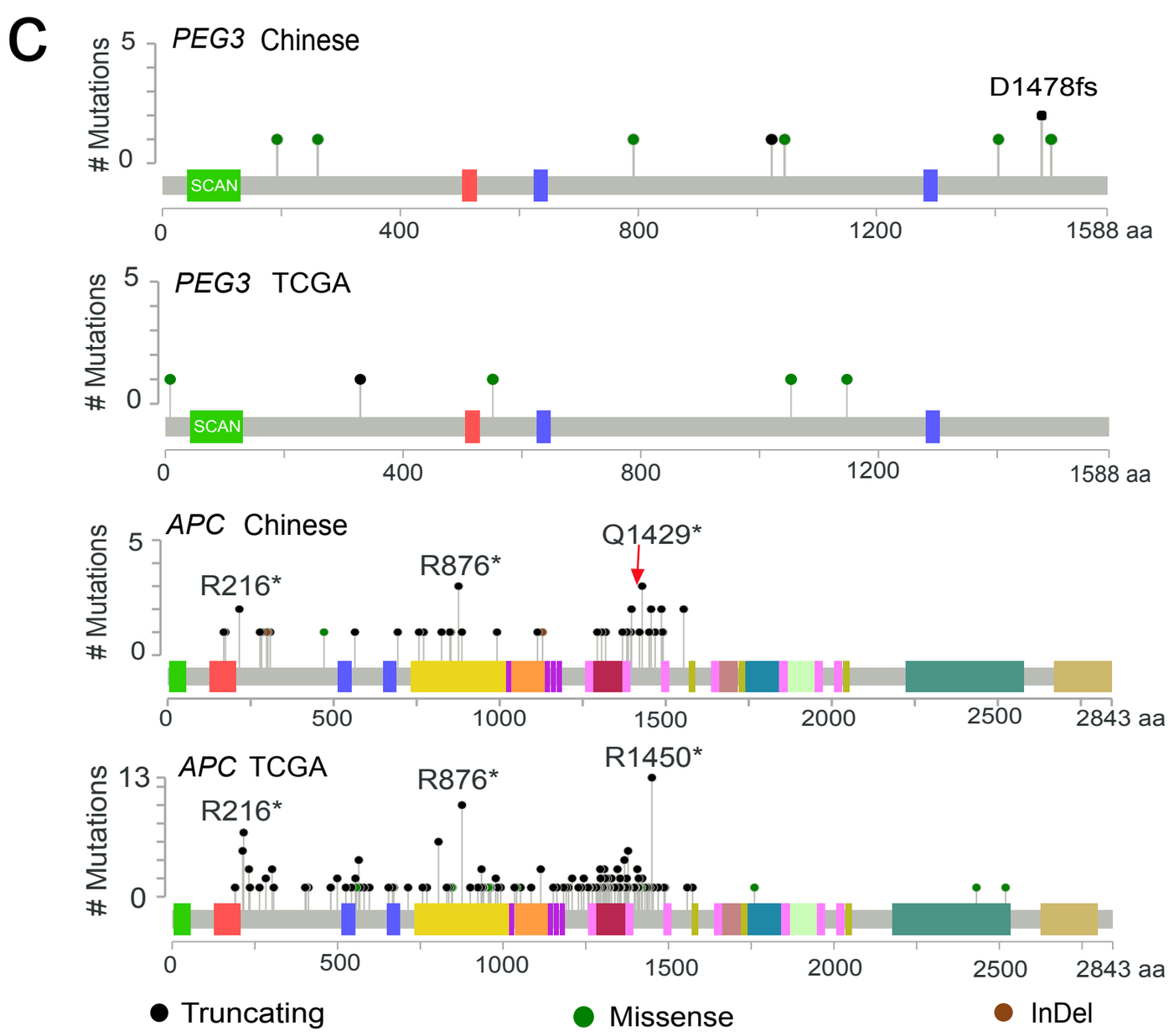
Figure 3. Illustration of prevalently somatic mutated genes. (a). Illustration of significantly mutated genes in non-hypermutated samples. The left axis shows mutation frequencies in 85 non-hypermutated samples. The right axis indicates the –log10 transformed q-value score from MutSigCV. (b) Protein interaction network from String database of mutated genes with >5% frequency. Green, red and blue edges represent the evidence from text-mining, experiment, and curated database, respectively. (c) Illustration of somatic mutations on PEG3 and APC genes in Chinese

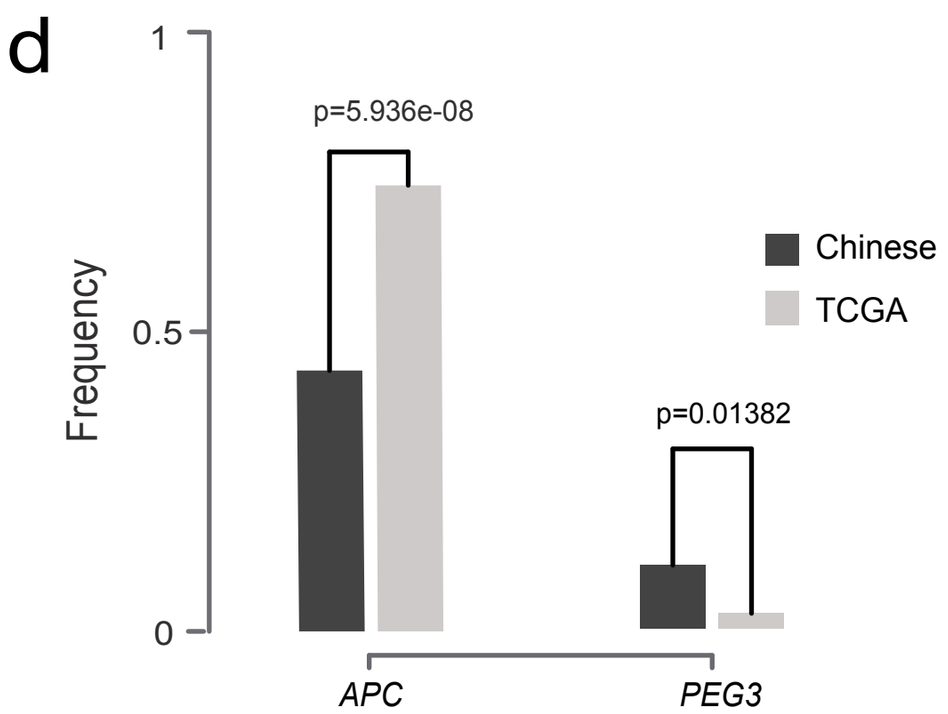
and TCGA data. (d) PEG3 and APC gene frequencies for Chinese and TCGA data (Chi-Squared test).

Fig. 4. Illustration of mutated genes in pathways. (a). Comparison of significantly mutated pathways between Chinese and TCGA data (q-value<0.01). (b). Illustration of the mutated genes on pathway level.

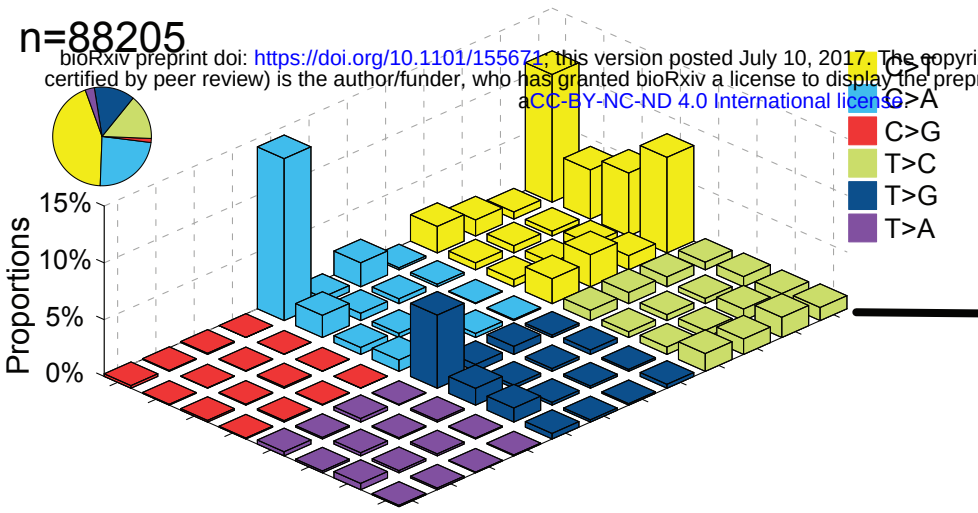




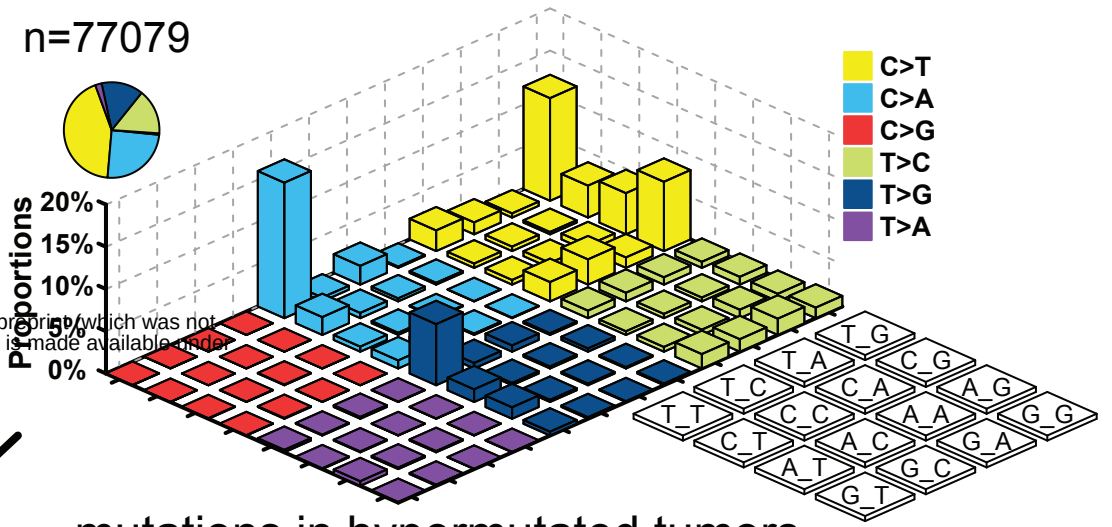




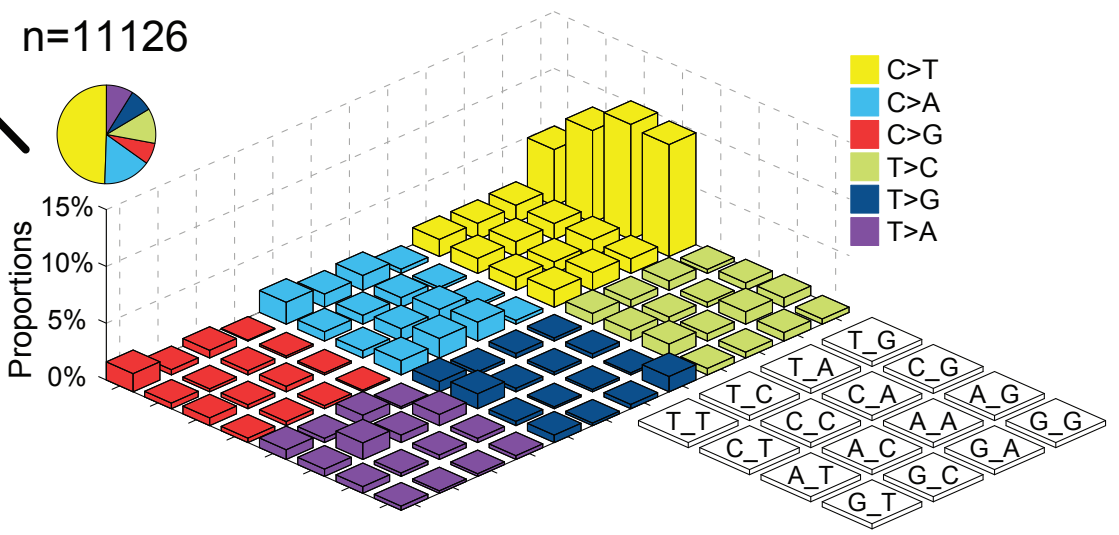
a Chinese



mutations in all tumors

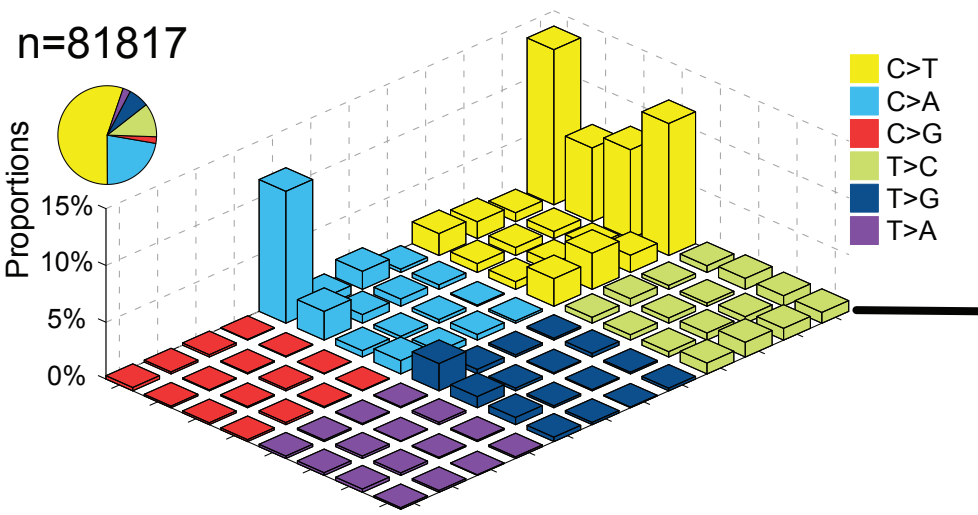


mutations in hypermutated tumors

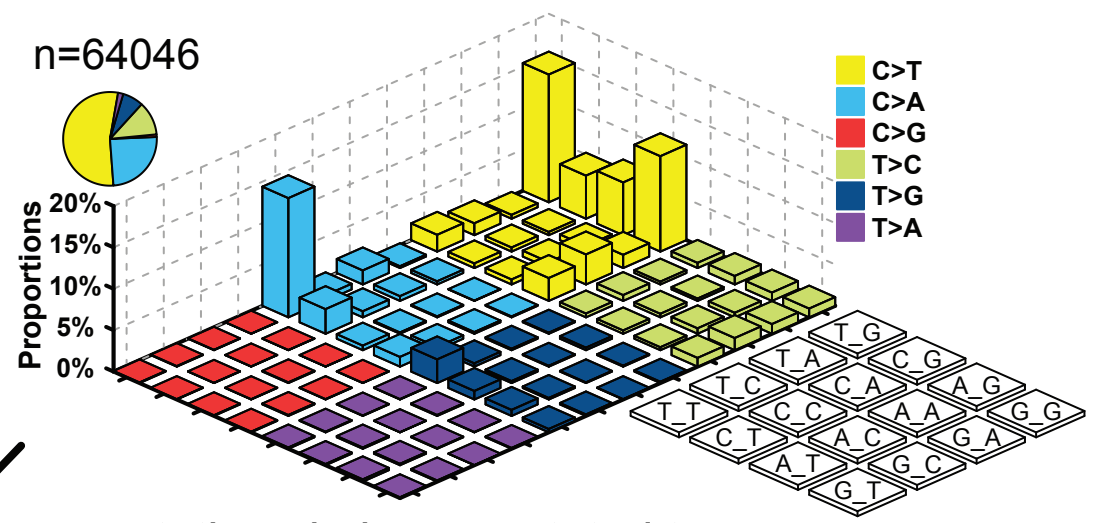


mutations in regularly mutated tumors

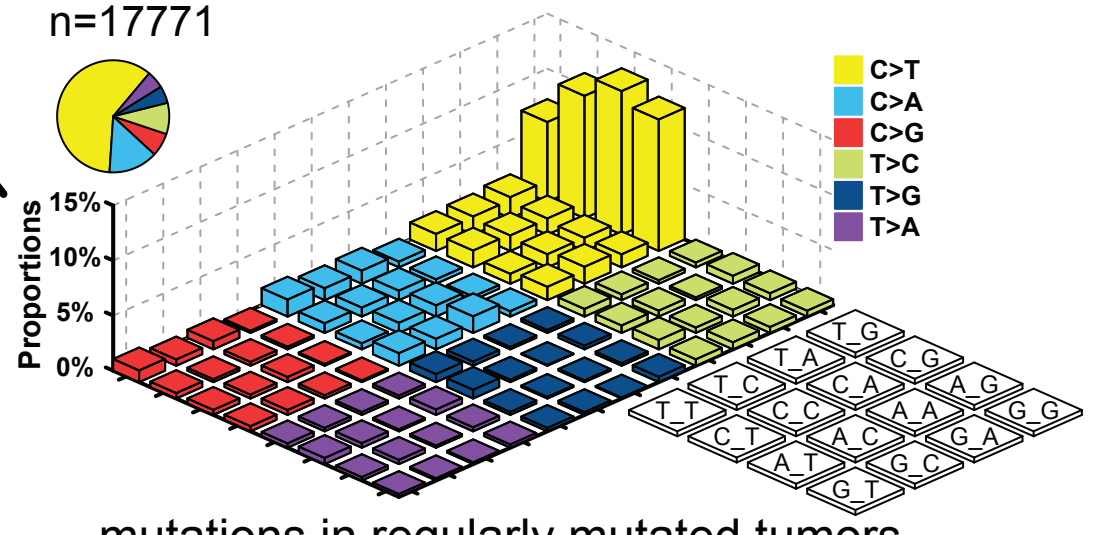
b TCGA



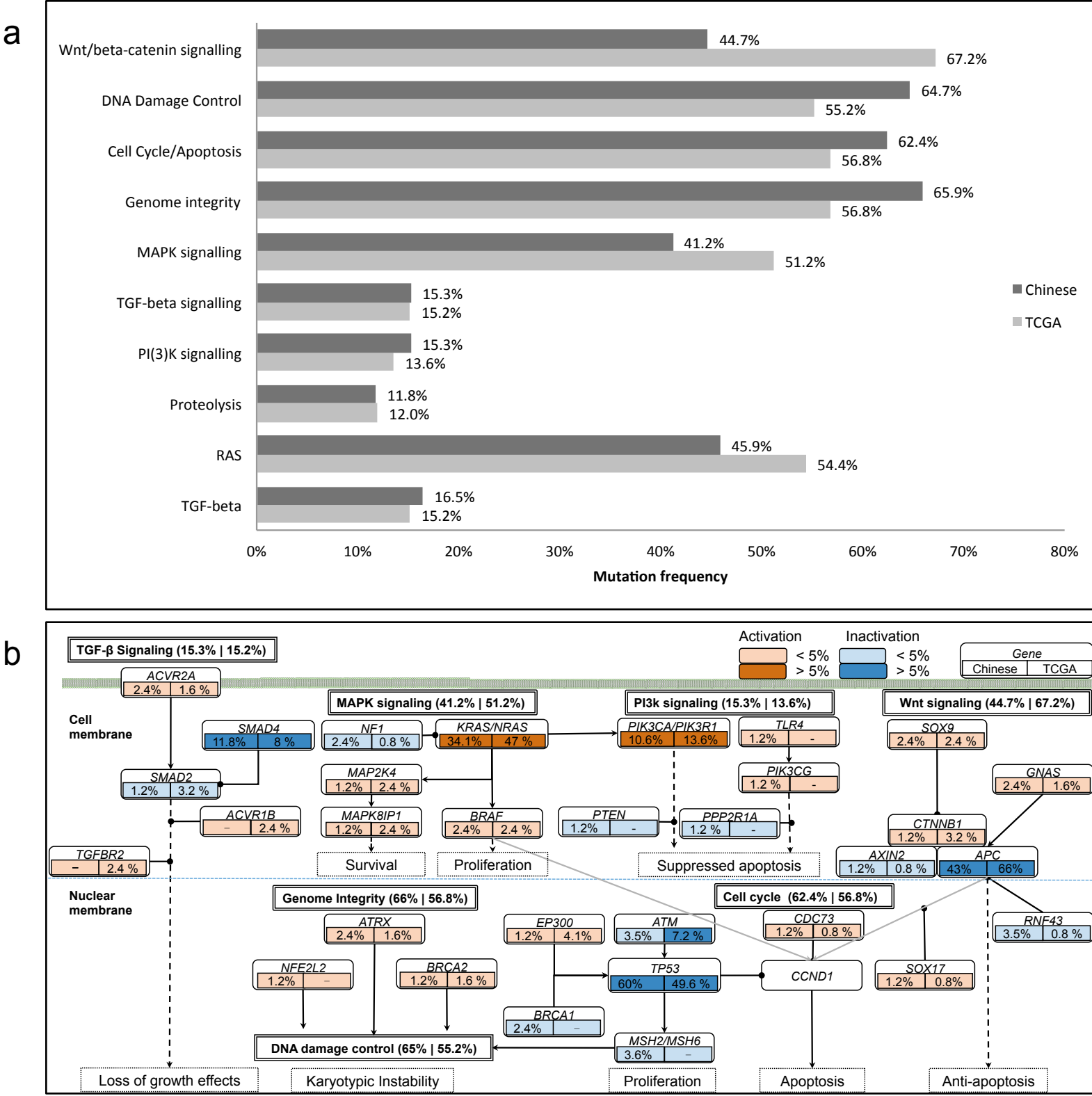
mutations in all tumors

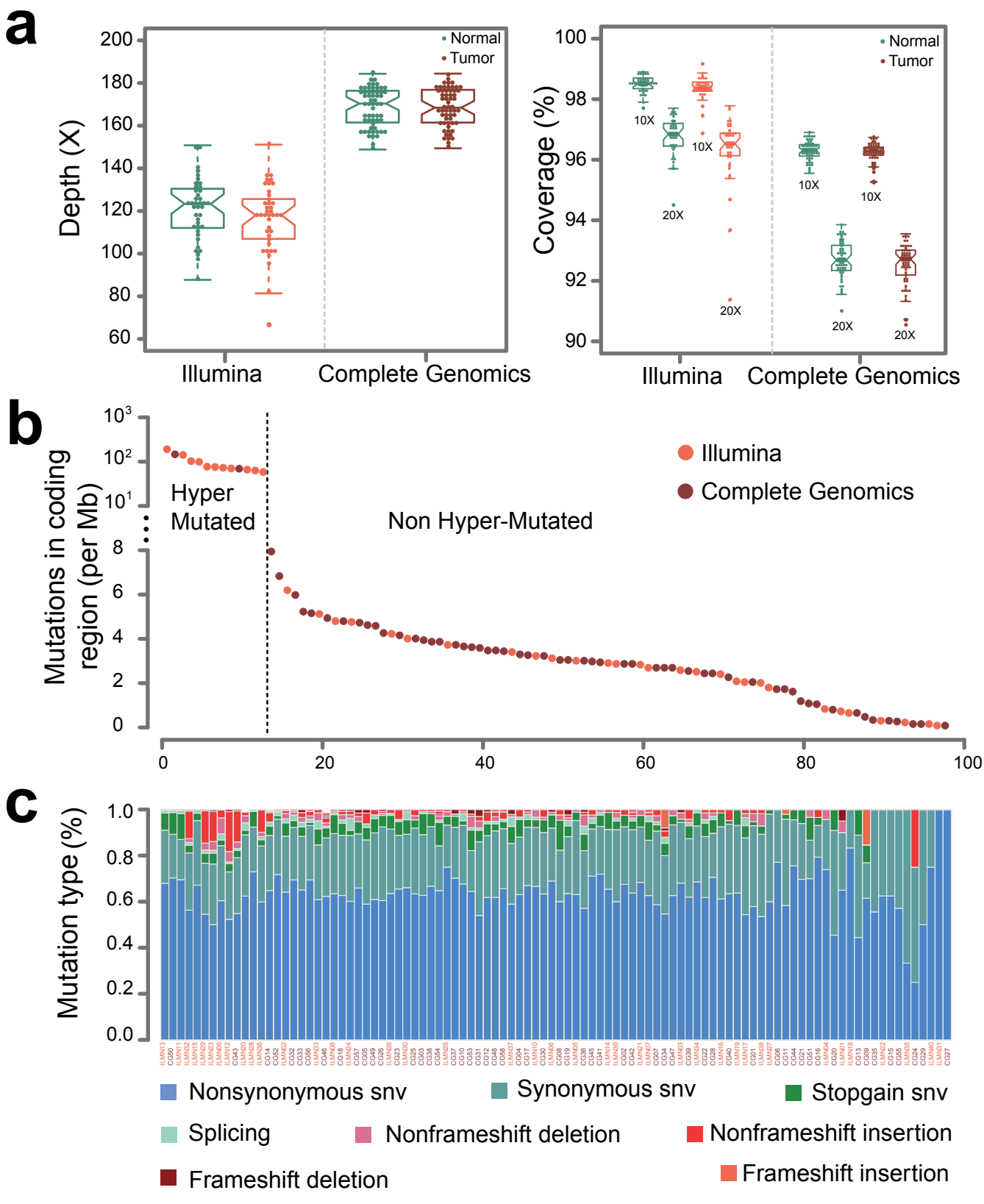


mutations in hypermutated tumors



mutations in regularly mutated tumors


a
Wnt/beta-catenin signalling
44.7%
67.2%
DNA Damage Control
64.7%
55.2%
Cell Cycle/Apoptosis
62.4%
56.8%
Genome integrity
65.9%
56.8%
MAPK signalling
41.2%
51.2%
TGF-beta signalling
15.3%
15.2%
PI(3)K signalling
15.3%
13.6%
Proteolysis
11.8%
12.0%
RAS
45.9%
54.4%
TGF-beta
16.5%
15.2%
Chinese
TCGA
0%
10%
20%
30%
40%
50%
60%
70%
80%
Mutation frequency
b
TGF-β Signaling (15.3% | 15.2%)
Activation
< 5%
> 5%
Inactivation
< 5%
> 5%
Gene
Chinese
TCGA
ACVR2A
2.4%
1.6 %
Cell membrane
MAPK signaling (41.2% | 51.2%)
PI3k signaling (15.3% | 13.6%)
Wnt signaling (44.7% | 67.2%)
SMAD4
11.8%
8 %
NF1
2.4%
0.8 %
KRAS/NRAS
34.1%
47 %
PIK3CA/PIK3R1
10.6%
13.6%
TLR4
1.2%
-
SOX9
2.4%
2.4 %
SMAD2
1.2%
3.2 %
MAP2K4
1.2%
2.4 %
PIK3CG
1.2 %
-
GNAS
2.4%
1.6%
ACVR1B
-
2.4 %
MAPK8IP1
1.2%
2.4 %
BRAF
2.4%
2.4 %
PTEN
1.2%
-
PPP2R1A
1.2 %
-
CTNNB1
1.2%
3.2 %
TGFBR2
-
2.4 %
Survival
Proliferation
Suppressed apoptosis
AXIN2
1.2%
0.8 %
APC
43%
66%
Nuclear membrane
Genome Integrity (66% | 56.8%)
Cell cycle (62.4% | 56.8%)
ATRX
2.4%
1.6%
EP300
1.2%
4.1%
ATM
3.5%
7.2 %
CDC73
1.2%
0.8 %
RNF43
3.5%
0.8 %
NFE2L2
1.2%
-
BRCA2
1.2%
1.6 %
TP53
60%
49.6 %
CCND1
SOX17
1.2%
0.8%
BRCA1
2.4%
-
DNA damage control (65% | 55.2%)
MSH2/MSH6
3.6%
-
Loss of growth effects
Karyotypic Instability
Proliferation
Apoptosis
Anti-apoptosis


a
Depth (X)
200
180
160
140
120
100
80
60
Normal
Tumor
Illumina
Complete Genomics
Coverage (%)
100
98
96
94
92
90
10X
20X
b
Mutations in coding region (per Mb)
10^3
10^2
10^1
8
6
4
2
0
Hyper Mutated
Non Hyper-Mutated
Illumina
Complete Genomics
0
20
40
60
80
100
c
Mutation type (%)
1.0
0.8
0.6
0.4
0.2
0.0
Nonsynonymous snv
Synonymous snv
Stopgain snv
Splicing
Nonframeshift deletion
Nonframeshift insertion
Frameshift deletion
Frameshift insertion