Intratumor Heterogeneity and Evolution of Colorectal Cancer

Santasree Banerjee[1,2,3†], Xianxiang Zhang[4†], Shan Kuang[1,3†], Jigang Wang[5†], Lei Li[1,3,6†], Guangyi Fan[1,2,3], Yonglun Luo[1,2,3,7], Shuai Sun[1,2,3], Peng Han[1,3], Qingyao Wu[4], Shujian Yang[4], Xiaobin Ji[5], Yong Li[1,3], Li Deng[1,3,8], Xiaofen Tian[2,3,9], Zhiwei Wang[1,2,3], Yue Zhang[1,3], Kui Wu[2,3], Shida Zhu[2,3], Shirley Victoria Hodgson[10], Samir Kumar Brahmachari[11,12], Lars Bolund[1,2,3,7,13], Huanming Yang[2,14], Xun Xu[1,2,3,15], Junnian Liu[1,2,3*], Yun Lu[4,16*], Xin Liu[1,2,3*]

[1] BGI-Qingdao, BGI-Shenzhen, Qingdao, 266555, China.
[2] BGI-Shenzhen, Shenzhen, 518083, China.
[3] China National GeneBank, BGI-Shenzhen, Shenzhen, 518120, China.
[4] Department of Gastroenterology, General Surgery Center, The Affiliated Hospital of Qingdao University, Qingdao, 266555, China.
[5] Department of Pathology, The Affiliated Hospital of Qingdao University, Qingdao, 266555, China.
[6] School of Future Technology, University of Chinese Academy of Sciences, Beijing, 101408, China.
[7] Department of Biomedicine, Aarhus University, Aarhus 8000, Denmark.
[8] State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China.
[9] MGI, BGI-Shenzhen, Shenzhen 518083, China.
[10] Department of Medical Genetics, St George's University of London, London, UK
[11] CSIR-Institute of Genomics and Integrative Biology, New Delhi, India.
[12] Academy of Scientific and Innovative Research, New Delhi, India.
[13] Lars Bolund Institute of Regenerative Medicine, Beijing Genomics Institute (BGI)-Qingdao, BGI-Shenzhen, Qingdao, China.
[14] James D. Watson Institute of Genome Sciences, Hangzhou 310058, Zhejiang, China
[15] Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Guangdong, China.
[16] Shandong Key Laboratory of Digital Medicine and Computer Assisted Surgery, Qingdao University, Qingdao, China.

[†] These authors contributed equally to this study.

[*] Correspondence

Xin Liu, BGI-Qingdao, BGI-Shenzhen, Qingdao, 266555, China. Tel: +86-18025460332; Email: liuxin@genomics.cn.

Yun Lu, Department of Gastroenterology, General Surgery Center, The Affiliated Hospital of Qingdao University, Qingdao, 266555, China. Tel: +86-18661802231; Email: cloudylucn@126.com.

Junnian Liu, BGI-Qingdao, BGI-Shenzhen, Qingdao, 266555, China. Tel: +86-18503088190; Email: chris.liu@genomics.cn.

Abstract

Identification of novel targets, and the development of target-based precision medicine for personalized cancer therapy is the biggest challenge in cancer research. Gradual and progressive genetic alterations drive the tumor from its benign stage to malignancies. Study of intra-tumor heterogeneity in terms of genetic alterations enable us to understand the clonal evolution which plays a key role in progression and metastases of tumors, and can lead to the development of personalized cancer therapeutic strategies. However, unlike other types of cancer, colorectal cancer (CRC) has been studied less, despite being a major threat world-wide, especially in the Chinese population. Here, in order to understand the intratumor heterogeneity and clonal evolution, we performed high depth whole exome sequencing followed by bioinformatic and statistical analysis for multi-region colorectal tumor samples. In this prospective study, we recruited different stages of 68 CRC patients with primary tumor at right-sided colon (18), left-sided colon (20) and rectum (30). In total, we sequenced 206 tumor regions including 176 primary tumors, 19 lymph node metastasis (LN) and 11 extranodal tumor deposits (ENTD) samples. Our result showed extreme intratumor heterogeneity with a Darwinian pattern of clonal evolution. We also identified that clonal evolution pattern of left-sided CRC was more complex and divergent than right-sided CRC, suggesting the evolutionary diversity in the initiation and progression of left-sided CRC. Genetic and evolutionary evidence found that both LN and ENTD were of polyclonal origin, and ENTD was an entity distinct from LN and evolves later. In addition, extensive heterogeneity was found in driver mutations in

KRAS and *PIK3CA* genes, suggesting major limitations of single biopsies in clinical diagnosis and for developing precision medicine for the patients with CRC. In conclusion, our study showed a Darwinian pattern of evolution in CRC with spatio-temporal differences in evolution between left-sided and right-sided CRC patients.

Key words: Colorectal cancer, intra-tumor heterogeneity, genetic alteration, lymph node metastasis, extranodal tumor deposits.

Introduction

Genomic instability in cancers is a continuous process involving genetic alterations at gene or chromosomal level. Extreme genetic heterogeneity drives a tumor from its benign state to malignancies. Tumor multi-region sequencing revealed intra-tumor heterogeneity and clonal evolution which play a key role in progression and metastases of the tumor, as well as identifying and developing novel targets for target-based precision medicine in personalized cancer therapy[1]. The development of effective target-based precision medicine and personalized cancer therapy is based on intra-tumor heterogeneity and the pattern of clonal as well as subclonal evolution in CRC tumors[2]. Therefore, patients with CRC may respond variably to the same treatment, due to intra-tumor heterogeneity, despite there being no significant differences identified in the tumor histopathology[3]. In addition, mono-sampling biopsies for clinical diagnosis are inadequate, due to intra-tumor heterogeneity[4]. Hence, study of intratumor heterogeneity is highly significant from both clinical and biological perspective, to understand the genomic changes driving the malignant process, which is fundamental to developing an effective personalized cancer therapy.

CRC is the third most common malignancy and the second leading cause of cancer death worldwide, with 18.1 million new cancer cases, and 9.6 million deaths in 2018[5]. According to the World Health Organization (WHO) GLOBOCAN database, there were 1,849,518 estimated new CRC cases and 880,792 CRC-related deaths in 2018[6]. In China, CRC is the second most common neoplasia, occupying the fifth position in mortality, accounting for an incidence of 521,490 new cases and 248,400 deaths in 2018[7].

Amongst CRC patients, the stage of the disease is one of the most important prognostic factors which is correlated with the disease survival rate[8]. Tumor Node Metastasis (TNM)/American Joint Committee on Cancer (AJCC) Cancer Staging system is the gold standard for determining the correct cancer stage, helping us for making appropriate treatment plans. Among CRC patients, the presence of cancer cells in lymph nodes is define as stage III disease[9]. To date, the molecular signature and evolutionary relationship between LN and ENTD has not been clear. Hence, the characterization of the molecular signature and evolution the primary tumor, LN and ENTD is very significant for TNM staging and therapeutic interventions for the patients with CRC.

The location of the primary tumor, either in the right- or left-side of colon, is also an important prognostic factor[10,11]. Clinical symptoms are also different between patients with right-sided and left-sided colon cancers. A possible explanation for this clinical heterogeneity might be due to the differences in their embryonic origin, genomic expression profiles and tumor microenvironment[12,13]. The differences in genomic expressions and subsequent alterations has not been studied well to explain the clinical heterogeneity between patients with right and left sided colon cancer.

Recently, tumor multiregion sequencing studies of primary tumor have demonstrated intratumor heterogeneity[14-21]. This multiregional sequencing approach, sequencing DNA samples from geographically separated regions of a single tumor, explores intratumor heterogeneity and cancer evolution[22-27]. Large-scale multiregional sequencing studies have systematically revealed intratumor heterogeneity as well as cancer evolution in patients with non-small-cell lung cancer and renal cancer[22-24]. However, large-scale multiregional sequencing studies of CRC have not been well reported. In addition, multiregional sequencing studies in CRC were performed at relatively shallow sequencing depths, making it difficult to assess intratumor heterogeneity, due to inability to detect somatic mutations with low frequencies[14-21].

In order to overcome the drawbacks of previous studies, we have comprehensively studied the intratumor heterogeneity and evolution of CRC, using high depth (median depth of 395×) whole exome sequencing of 206 multi-region tumor samples and 68 matched germline samples from 68 CRC tumors, determined the differences of intratumor heterogeneity, and the evolution

of CRC in patients with primary tumors in both right-sided and left-sided colon and characterized the molecular signature of the primary tumor, LN and ENTD, to define their evolutionary relationship.

Materials and methods

Patient recruitment

The study was approved by the Ethics committee of the Affiliated Hospital of Qingdao University. All the samples were collected after obtaining written informed consent from the patients. Patients were recruited based on the following criteria. (i). age over 18 years, (ii) patients diagnosed with CRC by enteroscopy, imaging, biopsy and followed by surgery, and histopathology performed with the resected tumor tissues. Based on the preoperative imaging studies, patients were included if they were presented with sufficient tissue available for the study.

Sample collection

A pathologist performed a macroscopic examination of all surgically resected specimens to guide the multi-region sampling in this study. Firstly, the pathologist performed routine pathological sampling for clinical diagnosis, and then multi-region sampling was performed by using the remaining samples. The pathologist dissected the tumor specimen with a new sterile scalpel blade on a clean surface to avoid contamination with other tissues. At least 2 regions of each tumor, which were at least 5 mm apart, were collected. Areas with significant necrosis, fibrosis, or hemorrhage were avoided to maximize the viability of tumor cells. Normal colorectal mucosa tissues were also sampled from areas remote from the primary tumor (at least 2 cm distant from the tumor edge). If possible, peri-intestinal nodules including lymph nodes present in the resected specimen were sampled. The pathologist examined the peri-intestinal lymph nodes in a "fresh" state. If there was malignancy appearance (the cut section appeared tan-gray and hard), a small portion was taken for frozen sections to confirm the diagnosis of malignancy. After confirming the malignancy, a portion of the lymph nodes was sampled for diagnostic requirements at first. If there was a surplus on the basis of satisfying the demands of

diagnosis, the remaining part was taken for this study. Each selected tissue block was split into two for snap freezing and formalin fixing respectively, such that the fresh frozen sample has its mirror image in the formalin-fixed sample which is subsequently paraffin embedded (mirrored FFPE sample). Fresh samples were placed in a 2 ml cryotube, labeled with the region number corresponding to the mirrored formalin-fixed paraffin embedded sample number, and snap frozen with immediate immersion into liquid nitrogen before transferred to -80°C freezer for storage. Peripheral blood was collected during surgery and processed into EDTA anticoagulation tube. The tumor tissue samples from 68 patients were sequenced and analyzed after filtering according to the filtering pipeline, schematically presented in the CONSORT diagram (CONSORT flowchart, **Supplementary Figure S1**). The workflow summarizing experiments and data analysis in our study was shown in **Supplementary Figure S2**.

Sample processing

Approximately 50 mm^3 of tumor tissue from each region was used for genomic DNA extraction using the QIAamp DNA Mini Kit (Qiagen, China) according to the manufacturer's instructions. 2 ml of peripheral blood was used for germline DNA extraction using the QIAamp DNA Blood midi kit according to the manufacturer's instructions. DNA was quantified by the Qubit Fluorometric Quantitation (Thermo Fisher Scientific, USA) and the quality of DNA was assessed by agarose gel electrophoresis.

Pathology diagnoses and review

Pathological diagnoses were established according to the WHO classification and independently reviewed by two professional gastrointestinal pathologists. Detailed clinical characteristics were summarized in **Supplementary Table S1**. Hematoxylin- eosin sections of mirrored FFPE samples for each region in each case (387 sections from 70 patients) were evaluated by the same pathologist. Only primary tumor regions with more than 30% tumor component were considered for sequencing. Then tumor regions with pathological heterogeneity were selected for final sequencing. In addition, pathologist distinguished LN and ENTD by reviewing hematoxylin-eosin sections of their mirrored FFPE samples in this study

were also sent for sequencing.

Whole exome library construction and sequencing

Tumor tissues and matched germline tissues were subjected to whole exome sequencing. Exome capture was performed on 1 μg of genomic DNA. Covaris (LE220) was used to randomly fragmented DNA into 150-250 bp. After fragmentation, these fragments were purified and connected through a PE Index Adaptor designed by BGI, and then captured by using the standard protocol of the MGIeasy Exome Capture V4 probe set (~ 59 Mb; MGI Tech Co., Ltd, China). All constructed libraries were loaded onto BGISEQ-500 (MGI Tech Co., Ltd, China) and the sequences were generated as 100-bp paired-end reads.

Sequencing reads containing sequencing adapters, more than 10% of unknown bases and low-quality bases (> 50% bases with quality <5) were removed by SOAPnuke (v1.5.6)[28]. The processed sequencing reads were then aligned to UCSC human reference genome (hg19) using BWA-MEM (v0.7.12)[29]. Picard (v1.137) (https://broadinstitute.github.io/picard/) was used to generate chromosomal coordinate-sorted bam files to remove PCR duplicates. After removing the duplicates, the median sequencing depth of the generated data for the tumor area were reached 391 (range 179-537), and the matched germline tissues were reached 414.5 (range 243-596). We then used the Genomic Analysis Toolkit (GATK v3.8.0)[30] to perform base quality score recalibration and local realignment of the aligned reads to improve alignment accuracy.

Quality control to prevent contamination, inter-patient sample swaps and removal of regions with extremely low mutation occurrence

ContEst[31], a GATK module, was used to estimate the cross-individual contamination level. Samples with contamination level more than 1% were deleted (3 samples failed the QC due to contamination as shown in **Supplementary Figure S1**. In order to avoid sample swaps between patients, we use BAM-matcher[32] to determine whether two samples from the same patient by comparing their genotypes.

The number of mutations in each tumor region was called independently. The median number of mutations across all regions for each tumor was calculated. A region in one tumor

was removed if less than 20% of the median mutation count of that tumor was identified in that region.

Somatic mutation detection and filtering

After processed the sequencing data, SAMtools (v1.2)[33] mpileup was used to locate non-reference locations in tumor and germline samples. Bases with phred scores less than 20 or reads with mapping quality (MAPQ) values less than 20 were deleted. Base-alignment quality (BAQ) computation was disabled with adjust mapping quality coefficient set of 50. Both VarScan 2 (v2.4.3)[34] in somatic mode and MuTect (v.1.1.7)[35] were used to call somatic mutations. The somatic variants called by VarScan 2 were filtered and the minimum coverage of the germline sample was set to 10, the minimum variant frequency was changed to 0.01, and tumor purity was set to 0.5. We further filtered the resulting single nucleotide variant (SNV) calls for false positives using Varscan 2 associated fpfilter.pl script. We used bam-readcount (v. 0.8.0) (https://github.com/genome/bam-readcount) to prepare input files for fpfilter and min-var-freq was set to 0.02. All insertions/deletions (INDELs) called in reads that VarScan 2 processSomatic classifies as "high confidence" were recorded for further downstream filtering. MuTect was used to detect SNVs using annotation files contained in GATK. The annotation files contained in GATK bundle v.2.8 were used in MuTect and variants were filtered according to the filter parameter 'PASS'.

Additional filtering was performed to reduce false positive mutation calls. If the variant allele frequency (VAF) is greater than 2%, and both VarScan 2 (with a somatic p-value <= 0.01) and MuTect called the mutation, then a SNV was considered as truly positive. Alternatively, if a SNV was called only in VarScan 2 with a somatic p-value <=0.01, a frequency of 5% is required. In addition, the sequencing depth supporting the variant call in each region requires >= 30, and the sequence reads requires >= 5. In contrast, the VAF value of the variant in the germline should be <= 1%. We filtered the INDEL using the same parameters as above, except that reads >= 10 are required to support mutation calls, somatic p-values <= 0.001 and sequencing depth >= 50.

ANNOVAR[36] was used to annotate mutations with COSMIC (v88)[37], SIFT[38], PolyPhen-

2[39], MutationTaster[40] databases. All mutations used in the analysis can be found in **Supplementary Table S2**. Mutations were classified as clonal or subclonal using PyClone (v0.13.1)[41]. Pyclone CCF (cancer cell fraction) value were calculated as described in the subclonal deconstruction section. Mutations with CCF>0.9 across all regions of a tumor were clonal mutations, otherwise they were considered as subclonal mutations.

Driver mutation identification

All variants were compared with all genes identified and enlisted in the COSMIC Cancer Gene Census (v88)[37]. Then, three types of mutations were classified as a driver mutation according to the following criteria. Firstly, if the gene was annotated as TSG (tumor suppressor gene) by COSMIC, and the non-silent variant was considered deleterious: either *loss of function* (stop-gain/stop-loss, frameshift deletion/insertion or non-frameshift insertion/deletion) or predicted deleterious in two of these three computational approaches applied – SIFT[38], PolyPhen-2[39] and MutationTaster[40], then the specific variant would be classified as a driver mutation. Secondly, if the variant was annotated as oncogene by COSMIC, then we tried to identify exact matches to non-silent variants in COSMIC. If an exact match was found ≥ 3 times, the variant was categorized as a driver mutation. Thirdly, if the gene was annotated as TSG (tumor suppressor gene) by COSMIC, and the variant is located at the canonical splice site, then the specific variant would be classified as a driver mutation. Finally, we compared all these three types of driver mutations to the CpG island location file on UCSC Genome Bioinformatics website (http://genome.ucsc.edu). We then deleted all mutations that occurred on the CpG island and finally got all driver mutations.

Copy number analysis

Sequenza (v3.0.0)[42] was used to detect the somatic copy number alterations (SCNAs) and evaluate the purity and ploidy of tumor cells as follows. Firstly, we used SAMtools (v1.2)[33] mpileup to convert the Bam file of whole exome sequencing data to Pileup format. Secondly, paired tumors and normal Pileup files were processed by sequenza-utils to extract the sequencing depth, determine the homozygous and heterozygous positions of variants in normal

samples, and calculate the variant alleles and allelic frequencies from tumor samples. The sequenza-utils output was further processed by using Sequenza R package to provide segmented copy number data, cellularity and estimated ploidy for each sample. All segmented copy number data has given in **Supplementary Table S3** in the Supplementary Appendix. Heatmap of genome-wide SCNAs is visualized by R package Copynumber (v1.24.0)[43].

The driver gene copy number variations (driver SCNAs) of all genes enlisted in the COSMIC cancer gene census were analyzed as follows. Firstly, if the gene is annotated as oncogene by COSMIC, gene level amplification was called if gene copy number >2x ploidy of that sample. Secondly, if the gene was annotated as TSG by COSMIC, gene level deletion was called if gene copy number = 0. To determine the intratumor heterogeneity (ITH) status of driver SCNAs, we called driver SCNAs across all regions from each tumor. If at least one region showed an amplified SCNA, we called a gene as clonal amplification if all other regions of this gene showed copy number > ploidy + 1. If at least one region showed a deleted SCNA, we called a gene as clonal deletion if all other regions of this gene showed copy number < ploidy -1. All other driver SCNAs were defined as subclonal amplification or deletion. In 8 polyclonal origin tumors (CRC32, CRC36, CRC42, CRC48, CRC49, CRC51, CRC52 and CRC60) without founder clusters (cluster with CCF > 0.9 across all regions of a tumor), all their driver SCNAs were subclonal. To correlate driver SCNAs with specific mutation clusters of PyClone, we first identified all clusters where >= 50% CCF was present in each tumor region. We then identified all the clusters present in the same regions as a given driver SCNAs. If more than one clusters were found, then chose the one which is closer to the trunk. If no clusters matched for a given SCNA, we re-called the clonality of the SCNA by using the following threshold. We called a gene as clonally amplified if all the regions of this gene showed copy number >2x ploidy while we called a gene as clonally deleted if all the regions of this gene showed copy number = 0. Then we repeated the association test above. If an SCNA still could not be associated with a mutant cluster, it was annotated as a subclone associated with no known cluster (NA cluster).

To determine the ITH status of global SCNA, all parts of the genome were considered independently and divided into the smallest contiguous segments that overlap in all the regions

within each tumor. The gains and losses of segment were determined as follows. Firstly, copy number data for each segment was divided by the sample mean ploidy and then converted to $\log_2$. Secondly, gain and loss were defined as $\log_2$ (2.5/2) and $\log_2$ (1.5/2), respectively. Thirdly, any segment of gain or loss that spanned across all the regions was defined as clonal and all other segments of SCNA were defined as subclonal. Within each tumor, we summarized the length of the genome that subjected to SCNA in any region (total SCNA), the length of the genome that subjected to clonal SCNA (clonal gain or clonal loss), and the length of the genome that subjected to subclonal SCNA (subclonal gain, subclonal loss or subclonal undetermined). The proportion of subclonal SCNAs were then defined as the percentage of genomes subjected to subclonal SCNA divided by the percentage of genomes subjected to total SCNAs.

Chromosomal arm level SCNAs were determined if at least one region has shown an increase or decrease of at least 97% in chromosomal arm. To determine the ITH status of chromosome arm gain and loss, we called clonal arm gain or loss if the same chromosomal arm showed at least 75% gain or loss in all the remaining regions. While we called subclonal arm gain or loss if at least one of the remaining regions showed less than 75% gain or loss. For 8 polyclonal origin tumors (CRC32, CRC36, CRC42, CRC48, CRC49, CRC51, CRC52 and CRC60) without founder clusters (cluster with CCF > 0.9 across all regions of a tumor), all their arm level SCNAs were subclonal. As previously described in the driver SCNAs part, we correlated arm level SCNAs with specific mutation clusters of PyClone in the same way.

Mirrored sub-clonal allelic imbalance analysis

Single nucleotide polymorphisms (SNPs) were called by using Platypus (v0.8.1)[44] and only SNPs with a minimum coverage of 20x were analyzed. The B allele frequency (BAF) of each SNP was calculated as the ratio of reads of reference base to variant. Heterozygous SNPs and BAFs were used as input and mirror subclone allelic imbalances (MSAI) were analyzed and visualized by RECUR[45].

Parallel evolution events for driver SCNAs were identified as follows. Firstly, driver SCNAs were identified as described in the "copy number analysis" section. Secondly, we annotated the regions of MSAI events in each tumor to the events of driver SCNAs. If two

events coincided with each other, then these driver SCNAs undergone parallel evolution.

Sub-clonal deconstruction

In order to estimate whether mutations were clonal or subclonal, and the phylogenetic trees of each tumor, the following formula were used[22,46]:

$$vaf = \frac{CN_{mut} \times CCF \times p}{CN_n \times (1 - \mathrm{p}) + CN_t \times p}$$

Where vaf is the mutated allele frequency of the mutated base; p is the estimated tumor purity; CNt is tumor locus specific copy number; CNn is normal locus specific copy number, assuming 2 for autosomal chromosomes; CCF is the fraction of tumor cells = carrying mutations. Considering that CNmut is the copy number of the chromosome harboring the mutation, the possible CNmut range is from 1 to CNt (integer). We then assigned one of the possible values to CCF: 0.01, 0.02, ..., 1, together with every possible CNmut to find the best fit CCF using maximum likelihood. In detail, for point mutations with alternative reads as "a" and sequencing coverage as "N", we used Bayesian probability theory and binomial distribution to estimate the probability of a given CCF:

$$\mathrm{P}\left(\mathrm{CCF} \middle| (\mathrm{a}|\mathrm{N})\right) \propto \mathrm{Binoma}\ (\mathrm{a}|\mathrm{N}, \mathrm{vaf}\ (\mathrm{CCF}))$$

Then, the distribution of CCF was obtained by calculating P (CCF) on 100 uniform grids with CCF values from 0.01 to 1 and dividing by their sum.

Then, we used PyClone (v.0.12.9)[41] Dirichlet process clustering to cluster all the mutations (SNVs and INDELs). For each mutation, we used the observed mutation count and set the reference count so that vaf equal to half of the CCF value calculated by maximum likelihood previously. We set the major allele copy number to 2, the minor allele copy number to 0 and the purity to 0.5 since they had been modified.

Since the vaf values of INDELs were potentially unreliable, we multiplied each estimated INDEL CCF with a region-specific correction factor, which was calculated by dividing the median mutation CCF of the ubiquitous mutations (mutations presented in all regions) in that region by the median INDEL CCF of the ubiquitous INDELs (INDELs presented in all regions) in that region. We ran PyClone with 10,000 iterations and a burn-in of 1000.

Phylogenetic tree construction

Phylogenetic trees were constructed using the published tool CITUP (v0.1.0)[47]. As input, CITUP requires mutation clusters and their mean cancer cell prevalence values which were collected from PyClone. All clusters with at least 5 mutations were used as input to CITUP. Clusters for phylogenetic tree construction were summarized in **Supplementary Table S4** in the Supplementary Appendix. The optimal phylogenetic trees for each patient from CITUP were illustrated using MapScape (v.1.8.0)[48].

Evolution subtype analysis

Evolutionary subtypes were clustered and visualized by REVOLVER (v.0.2.0)[49]. CCF values and cluster information of driver mutations, driver SCNAs and arm level SCNAs were processed as previously described, which were used as input to REVOLVER. REVOLVER requires a founder cluster for all the input tumors. Therefore, we artificially defined a founder cluster for 8 polyclonal origin tumors. For CRC32, CRC36, CRC42, CRC48, CRC49, CRC51, CRC52 and CRC60, the artificial founder cluster was cluster 2, cluster 2, cluster 1, cluster 2, cluster 3, cluster 1, cluster 2 and cluster 5, respectively.

Phylogenetic analysis

Phylogenetic distance between primary tumor, LN and ENTD were analyzed by using the binary matrix of mutations present or absent in each region of tumors with LN or ENTD. Private mutations of each region were discarded from phylogenetic tree building due to lack of information. Fake outgroups with no mutations were generated for each individual as a root. Phylogenies were constructed using the PHYLIP (v.3.697)[50] suite of tools. For each tumor, we used seqboot to generate 100 bootstrap replicates by resampling of the mutations with replacement.

Phylogenetic trees were then constructed for each bootstrap replicate by maximum parsimony using the Mix programme in Wagner method. The jumble = 10 option was used and the order of the input samples was randomized 10 times for each bootstrap replicate. Finally, the Consense program was used to build a consensus of all the phylogenetic trees by using the

majority rule (extended) option. Phylogenetic trees were redrawn by FigTree (v1.4.4)[51] with the length of trunks and branches, proportional to the number of mutations.

dN/dS analysis

Values of dN/dS were analysed for different types of mutations: missense (wmis), nonsense (wnon), essential splice site substitution (wspl), non-synonymous substitutions (wall) and truncating substitutions (wtru). Values of dN/dS were analyzed by dNdScv R package[52].

Mutation signature analysis

Mutation signatures were estimated by using the DeconstructSigs (v.1.8.0)[53] package in R. Mutational signature analysis was applied only in the presence of at least 15 mutations.

Statistical analysis

All analyses were performed in R statistical environment version >= 3.5.0. All statistical comparisons of two distributions used the Wilcoxon test (wilcox.test function in R).

Data availability

The sequencing data and supplementary tables has been deposited at the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa), under accession number CNP0000594.

Results

Comprehensive clinical description of these 68 patients are provided in **Supplementary Table S1**. Tumor multi-region high depth (median depth of 395×, range 179-596) whole exome sequencing (WES) was performed with 206 tumor regions including 176 primary tumor regions, 19 LN regions and 11 ENTD regions, (2-7 regions per tumor) as well as 68 matched germline samples from 68 tumors of 68 CRC patients. WES identified 6 hypermutated (mutation rates of each tumor region more than 10 mutations/1 Mb bases) CRC patients, of these four patients were identified with microsatellite instability (MSI). The remaining 62 CRC patients were microsatellite stable (MSS) and of these, 12 patients had right-sided colon tumors, 20 had left-sided colon tumors and 30 had rectal tumor. Hypermutated patients were analyzed separately.

Intratumor heterogeneity in colorectal tumors

WES of 62 tumors with 188 tumor regions identified 19454 somatic mutations including 17560 SNVs (14361 non-silent SNVs) and 1894 INDELs (**Supplementary Table S2**). The mutation rate of multi-region whole-exome sequencing was significantly more than single sample sequencing due to detection of subclonal mutations (median number of mutations/1MB bases, 4.61 vs. 3.23; $P=8.9\times10^{-9}$) (**Supplementary Figure S3**). In our study, the mutation rate of single sample sequencing was significantly higher than single CRC sample sequencing data from The Cancer Genome Atlas[54] (TCGA), probably due to the higher sequencing depth in our study (median number of mutations/1 MB bases, 3.23 vs. 2.07; $P=1.7\times10^{-22}$) (**Supplementary Figure S3**).

Then, identified somatic mutations were divided into clonal (mutations present in all cancer cells with cancer cell fraction (CCF) > 0.9 across all the regions of a tumor) and subclonal (mutations present in only a subset of cancer cells) mutations (**Figure 1A**). It is worth noting that 2 patients (CRC32 and CRC36) with left-sided colon tumors and 6 patients (CRC49, CRC42, CRC51, CRC48, CRC52 and CRC60) with rectal tumors had not identified with clonal mutations, suggesting that branched evolution was widespread in patients with left-sided colon tumors. However, patients with right-sided colon tumors had significantly more clonal mutations than the patients with rectal tumors (median number, 160 vs 119; P=0.035) (**Supplementary Figure S4**). There were no significant differences found in the number and percentage of mutations between early (stage I and II) and late (stage III and IV) stage of patients (**Supplementary Figure S5**).

Somatic copy number alterations (SCNAs) were measured as length of segments affected by either gains or losses (detailed copy number data has given in **Supplementary Table S3**). We summarized the total length of the genome that subjected to SCNAs and calculated the percentage of clonal and subclonal SCNAs (**Figure 1A**). Interestingly, in a patient (CRC43) with a rectal tumor, all SCNAs were subjected to subclonal SCNAs. There were no significant differences in the length and percentage of SCNAs among the patients with right-sided, left-sided and rectal tumors as well as between early and late stage of CRC tumors (**Supplementary Figure S6 and S7**).

In our study, we identified that the mutation frequency of 14 driver genes (*APC*, *TP53*, *KRAS*, *LZTR1*, *LRP1B*, *FBXW7*, *TCF7L2*, *FAT4*, *ARID1A*, *ATM*, *PIK3CA*, *AMER1*, *CSMD3* and *SMAD4*) were higher at patient-level than at sample-level, except *SMAD4* gene (**Figure 1B**). In addition, we also found that the mutation frequency was higher at patient-level compared to the TCGA study[54] except *CSMD3* gene (**Figure 1B**). Notably, the mutation frequency of the *LZTR1* gene was also much higher than TCGA study[54] (**Figure 1B**). We also identified that the frequency of SCNAs was higher than TCGA[54] study data, probably due to the identification of subclonal SCNAs in our study (**Figure 1C**).

Clonal architecture in CRC

All the mutations (SNVs and INDELs) were clustered according to their CCF values to understand the clonal architecture and evolutionary history of 62 CRC tumors. Each colored circle in the phylogenetic tree represented one cluster of the tumor (**Figure 2**). Phylogenetic trees for 62 tumors and 188 regions together with schematic diagram of 100 tumor cells representing distribution of clusters in each tumor region (**Supplementary Figure S8**). Driver mutations, driver SCNAs and their clusters were annotated beside the phylogenetic trees (**Supplementary Figure S8**). Detailed information of cluster numbers for each tumor is listed in **Supplementary Table S4**, with a median of 6 clusters per tumor (range, 1 to 13).

Patients with left-sided colon tumors possessed significantly more cluster numbers than patients with both right-sided colon tumors (median number, 7.5 vs. 6; P=0.028) and rectal tumors (median number, 7.5 vs. 5.5; P=0.025) (**Supplementary Figure S9**), which potentially reflected the more evolutional diversity in patients with left-sided colon tumors. There were no significant differences in cluster numbers between early and late stage of CRC tumors (**Supplementary Figure S9**). Only 22 out of 188 tumor regions (12%) were presented with subclones in all the branches of the phylogenetic tree (**Supplementary Figure S8**). This highlighted the limitations of single biopsy strategy of clinical diagnosis since mono-sampling was not enough to capture all the genetic information within a tumor.

Driver event alterations in CRC evolution

Identifying cancer driver events and their clonality might provide important evidences for developing the target-based effective therapeutic strategies. Therefore, driver mutations, driver SCNAs, arm level SCNAs and their clonality were analyzed for CRC tumors (**Figure 3**).

We identified 1373 driver events (405 driver mutations, 707 driver SCNAs and 261 arm level SCNAs) among 62 CRC tumors. Of these events, 55% of driver events (611 out of 1373) were found to be subclonal (41% of driver mutations, 41% of driver SCNAs and 60% of arm level SCNAs). However, significantly lower percentage of clonal driver events and more subclonal driver mutations were identified in rectal tumors than patients with both right-sided (median percentage, 56% vs. 72%; P=0.031) and left-sided colon tumors (median percentage, 56% vs. 74%; P=0.047) (**Supplementary Figure S10 and S11**), which potentially reflected the increased diversity in driver events existing amongst different tumor regions of patients with rectal tumors. Late stage CRC tumors possessed significantly more subclonal driver events than early stage CRC tumors (median number, 8 vs. 4; P=0.043) (**Supplementary Figure S12 and S13**), which suggested that late stage CRC tumors were more advanced in evolution than early stage CRC tumors.

In addition, no driver events were consistently clonal among 62 CRC tumors, suggesting high ITH status and evolutionary diversity among CRC tumors, which might be the reason of low efficiency of target-based precision medicine in CRC treatment (**Figure 3**). All the driver SCNAs and most of the driver mutations were early events while very few arm level SCNAs were early events, suggesting that genomic instability process occurred firstly at the driver SCNA level, then at the driver mutations level, and finally at the driver arm level SCNA level.

Driver mutations in *APC*, *TP53* and *KRAS* were mostly identified in CRC tumors with right-sided colon tumors, left-sided colon tumors and rectal tumors. Mutations in *APC*, *TP53* and *KRAS* genes were predominantly clonal and early among different locations of CRC tumors, suggesting their significance and key role roles in tumor initiation. However, except for *APC*, *TP53* and *KRAS*, other identified driver mutations were completely different between patients with right-sided and left-sided colon tumors (**Figure 3**). The genes of driver SCNAs identified were the same in patients with left-sided and rectal tumors while only 3 out of 24 genes of driver SCNAs (*CYSLTR2*, *FLT3* and *FOXO1*) were same in patients with right-sided and left-

sided colon tumors (**Figure 3**). The huge differences in both driver mutations and driver SCNAs between the patients with right-sided and left-sided colon tumors suggested that left-sided colon tumors were evolutionary closer to rectal tumors than to right-sided colon tumors. Chromosomal arm level gain of 13q and loss of 18q were mostly occurred and predominantly clonal and early in colon and rectal cancer patients except that gain of 13q were late events in rectal cancer.

Convergent features and parallel evolution in CRC

Evidence of convergent mutations in tumor driver genes may shed light on evolutionary selection, which may provide therapeutic targets for treatment. *APC*, *TP53* and *KRAS* were the most frequently mutated driver genes identified in our study, 80.6 % (50 / 62), 80.6 % (50 / 62) and 51.6 % (32 / 62) respectively (**Supplementary Figure S14 and S15**). Among these three genes, *APC* was the most often mutated gene in CRC patients. Among the 50 CRC tumors with *APC* mutations, 19 (38%) had 2 mutations, consistent with the two-hit hypothesis of *APC* genes in CRC tumorigenesis (**Supplementary Figure 15**).

Evolutionary selection was also exemplified by parallel evolution of driver mutations, in which different driver mutations in distinct regions of the same tumor converge on the same gene. In CRC36 (left-sided colon tumor), two mutations (one splicing and one nonsynonymous mutations) in *TP53* were detected in tumor region 1 and 4 while another nonsynonymous mutation of *TP53* was detected in tumor region 3, indicating parallel evolution of *TP53*.

Positive selection

In order to further evaluate evolutionary selection at the mutational level, we used the ratio of dN/dS, which could reflect the degree of enrichment of protein-altering mutations occurred.

As shown in **Supplementary Table S5**, evidence for positive selection (dN/dS>1 based on the 95% confidence intervals) was rejected when all the non-synonymous mutations were considered. However, when genes were narrowed to all driver genes identified in the COSMIC Cancer Gene Census (v88), positive selection was observed for all clonal non-synonymous mutations in rectal tumors, but not in colon tumors. These findings suggested that positive

selection happened in cancer driver genes in early stage of CRC evolution.

Mutation signature

We analyzed mutational processes for CRC evolution by using published mutational signatures[55]. We found that the age-related signature 1 was the predominant mutational process for CRC tumors, with a median percentage of age-related mutations of 70% (**Supplementary Figure S16**). Interestingly, in a patient (CRC66) was identified with all the mutations with age-related signature 1 while in another patient (CRC63) was identified with defective DNA mismatch repair-related signature 6 and 15.

The median percentage of age-related signature 1 in CRC tumors for clonal mutations was 73%, while it dropped to 53% for subclonal mutations (**Supplementary Figure S16**). This finding suggested that except for age, other mutational processes played more important roles in subclonal than clonal tumors, which accounted for ITH of CRC. Except for age, other main mutation processes were defective DNA mismatch repair-related signature 6, defective DNA double-strand break-repair-related signature 3 and defective DNA mismatch repair-related signature 15, suggesting that the main mutational process for ITH of CRC were age and defective of DNA repair system.

Chromosome instability

Previously we analyzed the length and clonality of SCNAs relatively to ploidy (Figure 1A), we then measured the absolute SCNAs in CRC tumors. SCNAs and ITH of SCNAs were ubiquitous, which described the continuing process of chromosome instability in CRC tumors **(Supplementary Figure S17 and S18)**. Left-sided colon tumor were found to have more loss type of SCNAs (total copy number = 0 or 1, or copy neutral loss of heterozygous) than rectal cancer (P=0.007) and have more SCNAs with total copy number equal to 1 than right-sided colon tumors (P=0.044) (**Supplementary Figure S17**). Late stage tumors were identified with significantly fewer SCNAs than early stage CRC tumors (P=0.016) (**Supplementary Figure S18**). Specifically, the late stage CRC tumors had significantly fewer SCNAs with total copy number equal to 4 than early stage CRC tumors (P=0.043) (**Supplementary Figure S18**).

Interestingly, we found that the late stage CRC tumors had significantly more subclonal driver SCNAs than early stage CRC tumors, which suggested that the loss of random SCNAs as well as enrichment of functional SCNAs in late stage CRC tumors (**Supplementary Figure S12**).

The SCNA frequency pattern in patients with left-sided colon tumors and rectal tumors were similar to each other, while right-sided colon tumors were very different (**Supplementary Figure S19**). Patients with right-sided colon tumors had more 9p gain, 3q gain, 19p loss and less 20q gain, 19p loss, 8p loss than all CRC tumors (**Supplementary Figure S19**). Late stage CRC tumors had more 13q gain, 9p gain, 21p gain, 11q loss, 21q loss and 12p loss than early stage CRC tumors (**Supplementary Figure S20**). Interestingly, both poor prognosis location (right-sided colon tumors) and stage (late stage of CRC tumors) of CRC tumors had more 9p gain.

Genome doubling

If the percentage of autosomal tumor genomes with a major copy number of two or more in a tumor were 50% or more than 50%, then this tumor was classified as genome doubling tumor[56]. Genome doubling events were identified in 76% of tumors (found in 47 out of 62 tumors, 9 right-sided colon tumors, 15 left-sided colon tumors, and 23 rectal tumors) and appeared to be clonal in 66% of tumors (found in 31 out of 47 tumors, 5 right-sided colon tumors, 11 left-sided colon tumors and 15 rectal tumors), which suggested that whole genome doubling was an early event in CRC evolution. In our study, we identified that the rate of whole genome doubling was much higher than 36% found in a previous study of CRC[56], and the high rate was likely to come from multi-region sequencing because 16 out of 47 CRC were identified with subclonal whole genome doubling.

We observed a strong positive correlation between genome doubling with the ITH of both mutations and SCNAs. (**Supplementary Figure S21 and S22**). These findings suggested that the genome doubling was important for the progression of chromosomal instability. Since tumors without genome doubling had less subclonal SCNAs and a significantly higher percentage of clonal SCNAs than subclonal genome doubling tumors (median percentage, 62% vs. 29%, P=0.037). Tumors with clonal genome doubling had significantly more clonal SCNAs

than both subclonal genome doubling tumors (median length × 1 MB, 545 vs. 420, P=0.017) and tumors without genome doubling (median length × 1 MB, 545 vs. 391, P=0.022) (**Supplementary Figure S22**).

Mirrored subclonal allelic imbalance

Recent studies identified parallel evolution of SCNA in NSCLC and renal cancer through mirrored subclonal allelic imbalance (MSAI)[22,24], which was defined as the maternal allele was gained or lost in a subclone in one region, yet the paternal allele was gained or lost in a different subclone in another region. We identified 53 MSAI events in 23 of 62 CRC tumors (37%, found in 5 right-sided colon tumors, 6 left-sided colon tumors and 12 rectal tumors) (**Supplementary Figure S23)**. MSAI parallel gain or loss events found in this study were summarized (**Figure 4A**). Chromosomal regions of 7p and 13q were identified with parallel gain events in 3 tumors and chromosomal regions of 21q were identified with parallel loss events in 4 tumors. We also analyzed parallel evolution of driver SCNAs, 5 tumors (4 tumors of parallel amplification and 1 tumor of parallel deletion) were found to have driver SCNAs which overlapped with MSAI events (**Figure 4B and C**). Interestingly, 2 of 5 patients (CRC12 and CRC59) were identified with parallel amplification of *FLT3* gene in chromosome 13 (**Figure 4C**).

Conserved evolutionary features in CRC

In order to understand the constraints and features of CRC evolution, we analyzed conserved patterns of driver events by REVOLVER[49] (**Figure 5**). Evolutionary trajectories were clustered by the CCF and cluster information of all the driver events in 62 CRC tumors and four clusters (cluster red, blue, green and purple) were found (**Figure 5**). In order to understand whether conserved patterns of CRC evolution correlated to distinct clinical phenotypes, clinical and genomic metrics were shown under 4 clusters. Genomic information for patients included ITH index, tumor mutation burden (TMB), SCNA index, genome doubling type and clinical parameters included tumor location, histology subtypes, stage, number of regions, tumor size and patient age (**Figure 5**).

We found that the red and blue clusters had relatively fewer driver events than green and

purple clusters. There were no specific genomic or clinical features for the tumors in red cluster. The purple cluster had the most driver events compared with the other three clusters. The blue, green and purple clusters had similar clinical and genomic features, which were enriched in left-sided CRC tumors and tumors with genome doubling type.

Phylogenetic distance between LN and ENTD

We analyzed 16 non-hypermutated stage III patients to understand the phylogenetic distance and evolutionary relationship amongst primary tumor, LN and ENTD. CRC21, CRC28, CRC43 and CRC48 were identified with both LN and ENTD samples which were sequenced (**Figure 6**). In CRC21, we identified that the clonal evolution of LN and ENTD was similar, while ENTD appeared evolutionarily later than LN (**Supplementary Figure S8**). In CRC28, two ENTD were clustered together while LN were far away from them, which indicated that the LN and ENTD were polyclonal in origin (**Figure 6 and S8**). In CRC 43 and CRC48, we identified that the ENTD were not clustered together with LN and evolved separately (**Figure 6 and S8**). In tumors with more than one LN sequenced (CRC01, CRC11, CRC29 and CRC33), some LN were clustered together while some LN were not (**Figure 6**). In tumors with two ENTD sequenced (CRC60), the two ENTD were far away from each other in the phylogenetic tree (**Figure 6**). These findings suggested that the LN and ENTD were polyclonal in origin.

LN were identified with significantly less TMB than primary tumors (P=0.035) (**Supplementary Figure S24**). LN were also presented with significantly more loss type of SCNAs (P=0.043) than primary tumors (**Supplementary Figure S25**) which suggested that the LN were distinguished from primary tumors and ENTD in absolute SCNA level. Regarding SCNA frequency, the biggest difference of gain events existed between LN and ENTD while the biggest difference of loss events existed between primary tumors and ENTD (**Supplementary Figure S26**). ENTD were identified with less 20p gain, 20q gain and 21p gain than LN while ENTD had more 20p loss, 19q loss and 19p loss than primary tumors (**Supplementary Figure S26**), which suggested that the ENTD were different from LN and primary tumor in SCNA frequency level. In conclusion, LN and ENTD were also different in both mutation and SCNA level.

Evolution landscape of hypermutated CRC tumors.

All 6 (CRC04, CRC05, CRC09, CRC13, CRC15 and CRC17) hypermutated CRC patients were identified with right-sided colon tumors, of these two (CRC09 and CRC13) were with MSS and remaining four (CRC04, CRC05, CRC15, CRC17) were with MSI tumors (**Supplementary Figure S27A**). All of the 6 hypermutated tumors had mutations in mismatch-repair genes, *POLE* or *POLD* gene family (**Supplementary Figure S27A**). CRC09 had one missense mutation and one nonsense mutation of *POLE*. CRC13 had one missense mutation of *POLE* (**Supplementary Figure S27A**). These findings were consistent with that the predominant mutational process in CRC09 and CRC13 was *POLE*-related signature 10 (**Supplementary Figure S27B**). Defective DNA mismatch repair-related signature 6, 15, or 26 contributed to the mutational process of 4 MSI tumors (CRC04, CRC05, CRC15, CRC17) (**Supplementary Figure S27B)**.

We next analyzed evolution landscape of hypermutated tumors in SCNA level. None of the tumor regions in 6 hypermutated tumors had genome doubling. Absolute SCNAs of hypermutated CRC tumors were less compared with non-hypermutated CRC tumors (**Supplementary Figure S27C and S17**), which suggested that hypermutated tumors were mainly mutation driven tumors. Interestingly, CRC04 had MSAI events in X-chromosome (**Supplementary Figure S28**).

Discussion

In this present study, we recruited 68 patients with CRC and collected 206 multi-region tumor samples including 176 primary tumor tissue samples, 19 tissue samples with LN and 11 tissue samples with ENTD. We performed high-depth whole exome sequencing and analyzed 206 multi-region tumor samples. Our result showed very clear evidence of ITH amongst these 206 multi-region tumor samples in respect of both mutations and somatic copy number alterations. Our result showed the specific temporal and spatial features of evolution of CRC, following a Darwinian pattern of evolution. In addition, left-sided CRC was structurally and functionally more complex and divergent in terms of evolutionary perspective. We also identified that both ENTDs and LNs were polyclonal in origin and ENTDs were a distinctive entity from LNs,

which appeared later in tumor evolution. In conclusion, our study identified a Darwinian pattern of evolution of CRC and provides a potential new perspective to distinguish both left-sided with right-sided CRC.

Evolution pattern: Darwinian pattern of evolution and neutral evolution

In this present study, we found predominantly Darwinian pattern of evolution (59 out of 62 tumors) and a small portion of linear evolution (3 out of 62 tumors). Previous studies proposed neutral evolution model for colorectal cancers[14,57,58], whilst our conclusion was different from them, based on two reasons. Firstly, clonal events of both mutations (SNVs and INDELs) and SCNAs were widespread, with a median percentage of 47% and 43% respectively. Secondly, our study demonstrated a clear selection for functional mutations rather than non-functional mutations in colorectal cancer. In addition, 59% of driver mutations were clonal while only 41% of non-driver mutations were clonal, which indicated the enrichment of clonal driver mutations in course of evolution. Furthermore, the dN/dS value was 1.03 (95% confidence interval, 0.983 to 1.07) for all non-synonymous mutations while it reached to 1.37 (95% confidence interval, 1.16 to 1.61) when we consider only cancer driver genes based on COSMIC database. These findings indicated that positive selection existed only for cancer related gene mutations. Thirdly, convergent and parallel events were also present for driver genes in both mutational and SCNA level, especially for genes *APC*, *TP53* and *KRAS*. Some studies showed that the evolution pattern for colorectal cancer was Darwinian evolution followed by neutral evolution[21,59]. In our study, according to the existence of mutations in different tumor regions, 28% of subclonal mutations were shared by tumor regions (branch or trunk mutations), which suggested the importance of branches in phylogenetic trees.

Right-sided colon, left-sided colon and rectal cancer: In the light of genomic evolution

Previous studies[11,60] had shown remarkable differences between right-sided (right-sided colon tumors) and left-sided colon cancer (left-sided colon tumors) and rectal cancer, based on histology, MSI status, genetic subtype and prognosis. Almost no research has been done to date for understanding the differences between right-sided and left-sided colon cancer from an evolutionary perspective. Our study demonstrated that ITH and evolution in different location

of CRC were different in the following aspects: mutation, SCNA, structure of polygenetic tree and driver events. Firstly, rectal cancers had shown fewer clonal mutations than right-sided colon cancers, indicating higher ITH in rectal cancer at mutational level. Secondly, the pattern and clonality of SCNA frequency in right-sided colon cancer were different from left-sided colon cancer, which addressed the evolutionary difference between them at SCNA level. Thirdly, the structure of phylogenetic trees in left-sided colon cancer were more complicated and branched than that of the right-sided colon cancer. Specifically, left-sided colon cancer had the most complicated structure of the phylogenetic tree, reflected by the more cluster numbers. In addition, only left-sided colon cancer had polyclonal origin. Fourthly, left-sided colon cancer was enriched in clusters (blue and purple clusters) which had more driver events. These findings indicated that left-sided colon cancer had more functional diversity in the course of evolution. Specifically, rectal cancer had less percentage of clonal driver events than colon cancer, indicating more functional diversity occurred in the process of evolution in rectal cancer. In conclusion, our data showed that left-sided colon cancer were more divergent and complicated in terms of evolution than right-sided colon cancer, not only structurally but also functionally, which indicated that the evolutionary diversity might play an important role in the initiation and progression of left-sided colon cancer. Moreover, the frequency and clonality of SCNA was a potential and significant biomarker to distinguish right-sided colon cancer from the left-sided colon cancer.

Primary tumor, LN and ENTD: In evolutionary perspective

To date, no systematic research studies have been done to understand the similarities and differences between ENTD and LN. In this study, we found that ENTD were later events in the evolution of the tumor than LN, which could be distinguished at mutational, SCNA and evolutionary levels. At the mutational level, the TMB of LN was significantly less than primary tumors. Also, LN and ENTD could not be clustered together in the polygenetic tree according to the occurrence of mutations. Unlike in previous studies[15,61,62], different LNs or ENTDs in the same tumor did not cluster together in all cases, indicating the polyclonal origin of both LN and ENTD. We also identified here that LN were different from ENTD at SCNA level. In addition, LN had significantly more SCNA loss than the primary tumor. However, the biggest difference

between LN and ENTD was gain of events in SCNA frequency. ENTD thus appeared to be as later event than LN according to the clonal evolution history in CRC21. In conclusion, our present study provided comprehensive evidences to prove that ENTD and LN were two distinctive entities. These findings suggest the 7th and 8th editions of TMN staging, should separate ENTD as a distinctive entity from LN.

Stage and ITH in CRC

There was no significant difference between early and late stage of CRC at both mutational and evolution (cluster number) level. It was well known that chromosomal instability was associated with a worse prognosis. Therefore, we analyzed the difference between early and late stage of CRC at SCNA level. There was no significant difference in length and clonality of SCNA relatively to ploidy between early and late stage of CRC in our study. However, early stage CRC were found to have significantly more absolute SCNAs than late stage CRC in our study. These two distinctive conclusions suggested the definition of chromosomal instability was important. Interestingly, late stage CRC had significantly more subclonal driver SCNAs than early stage CRC, which suggested ITH of functional SCNAs rather than total absolute SCNAs might be more closely related with CRC progress.

Germline mutations and ITH in CRC

Five patients in our study were found to have pathogenic or likely pathogenic germline mutations in DNA mismatch repair genes. Among them, two (CRC05 and CRC15) patients were hypermutated MSI patients identified with pathogenic and likely pathogenic mutations in *MLH1*. In addition, CRC33 were found to carry a likely pathogenic mutation in the *MLH1* gene, CRC37 were presented with a pathogenic mutation in *PMS2* gene and CRC44 were identified with a pathogenic mutation in *MLH3*. CRC33 and CRC37 were left-sided colon cancer patient and CRC44 was a rectal cancer patient. Interestingly, we found these 3 (CRC33, CRC37 and CRC44) non-hypermutated MSS patients had relatively divergent and complex clonal evolution with cluster number of 7, 6 and 8 respectively. These findings reminded us that germline mutations in MSI genes might accelerate the evolution process in CRC.

Personalized therapy: Target-based Precision medicine for CRC

Presently, the principle of treatment for CRC has shifted from 'one gene, one drug approach' paradigm to a 'multi-gene, multi-drug' model when making decisions for precision medicine[63]. It is of great clinical importance to define the mutation status, clonality of genes and biomarkers in CRC. Among 62 CRC tumors, 32 were identified with driver mutations in *KRAS* while 7 were subclonal. Similarly, 7 CRC tumors harbored clonal driver mutations in *PIK3CA,* while 3 were carrying subclonal driver mutations. Driver mutations in *NRAS* (1 tumor), *BRAF* (4 tumors) and *ERBB2* (1 tumor) were all clonal in origin. In addition, among 4 driver mutations in *BRAF*, 2 were V600E. Furthermore, we identified 5 CRC tumors with *EGFR* amplification whilst only 1 was clonal in origin. No amplification of *ERBB2* was found in our study. However, the occurrence of subclonal driver mutations in biomarker genes emphasized the limitations of the single biopsy strategy for the clinical diagnosis of CRC. In CRC59, we found all 4 tumor regions had wildtype *KRAS*, *BRAF* and *NRAS* genes, whilst only 1 region had the driver mutation in *PIK3CA*.

Acknowledgement

We are thankful to the proband and all the family members for participating in our study and we are thankful to the China National GeneBank.

Funding sources

The study was supported by the grants from Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011), National Natural Science Foundation of China (No. 81802473), key research and development plan of Shandong province (No. 2018GSF118206) and "Clinical medicine + X" project from Medical College of Qingdao University.

Conflict of interest

The authors confirm that there are no conflicts of interest.

Author contributions

Study concept and design: Santasree Banerjee, Shan Kuang, Junnian Liu, Yun Lu and Xin Liu; Patient recruitment and clinical sample collection: Xianxiang Zhang, Qingyao Wu, Shujian

Yang; Histology and histopathology: Jigang Wang and Xiaobin Ji; Experiments (DNA extraction and whole exome sequencing): Peng Han, Yong Li, Xiaofen Tian and Zhiwei Wang; Analysis and interpretation of data: Santasree Banerjee, Shan Kuang, Lei Li, Shui Shun, Li Deng and Yue Zhang; Drafting of the manuscript: Santasree Banerjee, Shan Kuang, Lei Li, Xianxiang Zhang and Jigang Wang; Critical revision of the manuscript for important intellectual content: Huanming Yang, Shirley Victoria Hodgson, Samir Kumar Brahmachari, Lars Bolund, Yonglun Luo, Kui Wu, Shida Zhu, Guangyi Fan and Xun Xu; Supervision of the study: Santasree Banerjee, Shan Kuang, Junnian Liu, Yun Lu and Xin Liu.

References:

1. Alizadeh, A.A. *et al.* Toward understanding and exploiting tumor heterogeneity. *Nature medicine* **21**, 846 (2015).
2. Turner, N.C. & Reis-Filho, J.S. Genetic heterogeneity and cancer drug resistance. *The lancet oncology* **13**, e178-e185 (2012).
3. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501 (2015).
4. Burrell, R.A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338 (2013).
5. World Health Organization. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed January 21, 2019. (2019).
6. International Agency for Research on Cancer. Cancer Today. https://gco.iarc.fr/today/. Accessed January 21, 2019. (2019).
7. Chen, W. *et al.* Cancer incidence and mortality in China, 2013. *Cancer letters* **401**, 63-71 (2017).
8. O'Connell, J.B., Maggard, M.A. & Ko, C.Y. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *Journal of the National Cancer Institute* **96**, 1420-1425 (2004).
9. Adjuvant therapy for patients with colon and rectal cancer. *JAMA* **264**, 1444-1450 (1990).
10. Loupakis, F. *et al.* Primary tumor location as a prognostic factor in metastatic colorectal cancer. *JNCI: Journal of the National Cancer Institute* **107**(2015).
11. Petrelli, F. *et al.* Prognostic survival associated with left-sided vs right-sided colon cancer: a systematic review and meta-analysis. *JAMA oncology* **3**, 211-219 (2017).
12. Missiaglia, E. *et al.* Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. *Annals of oncology* **25**, 1995-2001 (2014).
13. Iacopetta, B. Are there two sides to colorectal cancer? *International journal of cancer* **101**, 403-408 (2002).
14. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nature genetics* **47**, 209 (2015).
15. Wei, Q. *et al.* Multiregion whole-exome sequencing of matched primary and metastatic tumors revealed genomic heterogeneity and suggested polyclonal seeding in colorectal cancer metastasis. *Annals of Oncology* **28**, 2135-2141 (2017).
16. Mamlouk, S. *et al.* DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nature communications* **8**, 14093 (2017).
17. Roerink, S.F. *et al.* Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457 (2018).
18. Wu, H. *et al.* Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. *Oncogene* **36**, 2857 (2017).
19. Ulintz, P.J., Greenson, J.K., Wu, R., Fearon, E.R. & Hardiman, K.M. Lymph node metastases in colon cancer are polyclonal. *Clinical Cancer Research* (2017).
20. Kim, T.-M. *et al.* Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clinical Cancer Research* (2015).
21. Uchi, R. *et al.* Integrated multiregional analysis proposing a new model of colorectal cancer evolution. *PLoS genetics* **12**, e1005778 (2016).

22. Jamal-Hanjani, M. *et al.* Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine* **376**, 2109-2121 (2017).
23. Turajlic, S. *et al.* Tracking cancer evolution reveals constrained routes to metastases: TRACERx Renal. *Cell* **173**, 581-594. e12 (2018).
24. Turajlic, S. *et al.* Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell* **173**, 595-610. e11 (2018).
25. Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences* **112**, E6496-E6505 (2015).
26. Nikbakht, H. *et al.* Spatial and temporal homogeneity of driver mutations in diffuse intrinsic pontine glioma. *Nature communications* **7**, 11185 (2016).
27. Kumar, A. *et al.* Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nature medicine* **22**, 369 (2016).
28. Chen, Y. *et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1-6 (2018).
29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
30. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-303 (2010).
31. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601-2602 (2011).
32. Wang, P.P., Parker, W.T., Branford, S. & Schreiber, A.W. BAM-matcher: a tool for rapid NGS sample matching. *Bioinformatics* **32**, 2699-701 (2016).
33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
34. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568-76 (2012).
35. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213-9 (2013).
36. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
37. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805-11 (2015).
38. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073 (2009).
39. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* **76**, 7.20. 1-7.20. 41 (2013).
40. Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* **7**, 575 (2010).
41. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396-8 (2014).
42. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64-70 (2015).
43. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number

segmentation. *BMC Genomics* **13**, 591 (2012).
44. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912-918 (2014).
45. Jakubek, Y.A., San Lucas, F.A. & Scheet, P. Directional allelic imbalance profiling and visualization from multi-sample data with RECUR. *Bioinformatics* **35**, 2300-2302 (2019).
46. Zhang, H. *et al.* Sex difference of mutation clonality in diffuse glioma evolution. *Neuro-oncology* **21**, 201-213 (2019).
47. Malikic, S., McPherson, A.W., Donmez, N. & Sahinalp, C.S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* **31**, 1349-56 (2015).
48. Smith, M.A. *et al.* E-scape: interactive visualization of single-cell phylogenetics and cancer evolution. *Nat. Methods* **14**, 549-550 (2017).
49. Caravagna, G. *et al.* Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods* **15**, 707 (2018).
50. Falenstein, J. PHYLIP—Phylogeny inference packages (version 3.2). *Cladistics* **5**, 164-166 (1989).
51. Rambaut, A. FigTree, a graphical viewer of phylogenetic trees. (2007).
52. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
53. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
54. Network, C.G.A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330 (2012).
55. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
56. Bielski, C.M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature genetics* **50**, 1189 (2018).
57. Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature genetics* **48**, 238 (2016).
58. Loeb, L.A. *et al.* Extensive subclonal mutational diversity in human colorectal cancer and its significance. *Proceedings of the National Academy of Sciences* (2019).
59. Saito, T. *et al.* A temporal shift of the evolutionary principle shaping intratumor heterogeneity in colorectal cancer. *Nature communications* **9**, 2884 (2018).
60. Lee, M.S., Menter, D.G. & Kopetz, S. Right versus left colon cancer biology: integrating the consensus molecular subtypes. *Journal of the National Comprehensive Cancer Network* **15**, 411-419 (2017).
61. Hu, Z. *et al.* Quantitative evidence for early metastatic seeding in colorectal cancer. *Nature genetics*, 1 (2019).
62. Árnadóttir, S.S. *et al.* Characterization of genetic intratumor heterogeneity in colorectal cancer and matching patient-derived spheroid cultures. *Molecular oncology* **12**, 132-147 (2018).
63. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* **17**, 79 (2017).

Figures

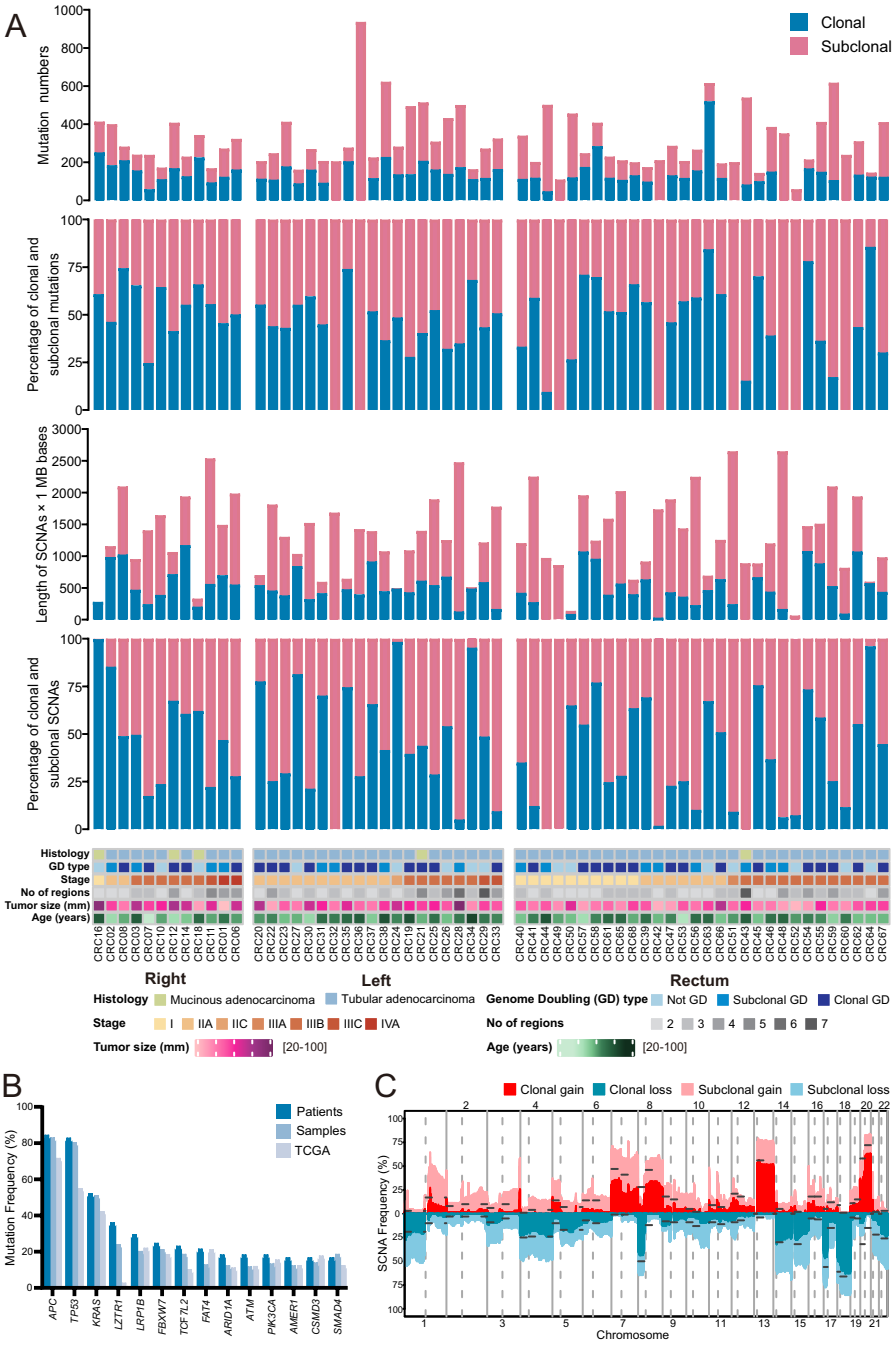

Figure 1 Overview of genomic heterogeneity in CRC tumors.

(A) Heterogeneity of mutations and somatic copy-number alterations (SCNAs). Tumors were sorted by location and stage.

(1) Number of all SNV and INDEL mutations (including coding and noncoding mutations) in CRC tumors. (2) The percentages of clonal mutations in CRC tumors. (3) Quantification of SCNAs in CRC tumors. (4) The percentages of clonal SCNAs in CRC tumors. (5) Demographic and clinical characteristics of the 62 CRC patients in this study (divided by histology; whole genome doubling status; stage; number of regions; tumor size; age and tumor location).

(B) Mutation frequency of driver genes (driver mutations occurred in not less than 10 patients) and comparison with TCGA data.

(C) Frequency of SCNAs in CRC tumors. The dotted lines are frequency of SCNAs in TCGA CRC samples.

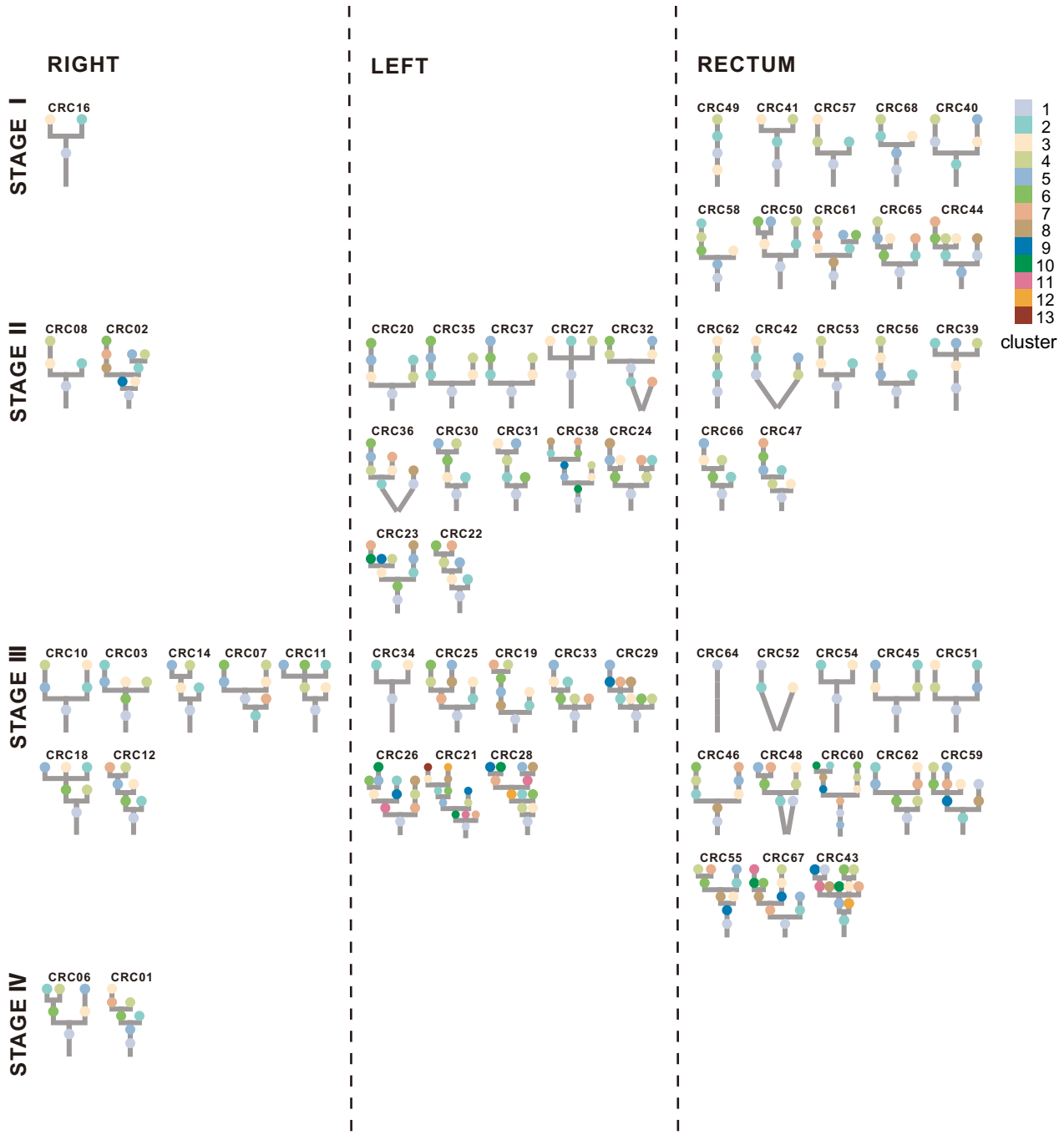

Figure 2 Phylogenetic trees.

Phylogenetic trees for each CRC tumor are shown. The trees are ordered by overall stage (I, II, III, IV) and position (right-sided colon, left-sided colon and rectum). The cluster number corresponding to the color is displayed in the upper right corner with largest cluster labeled "1". The lines connecting clusters does not contain any information.

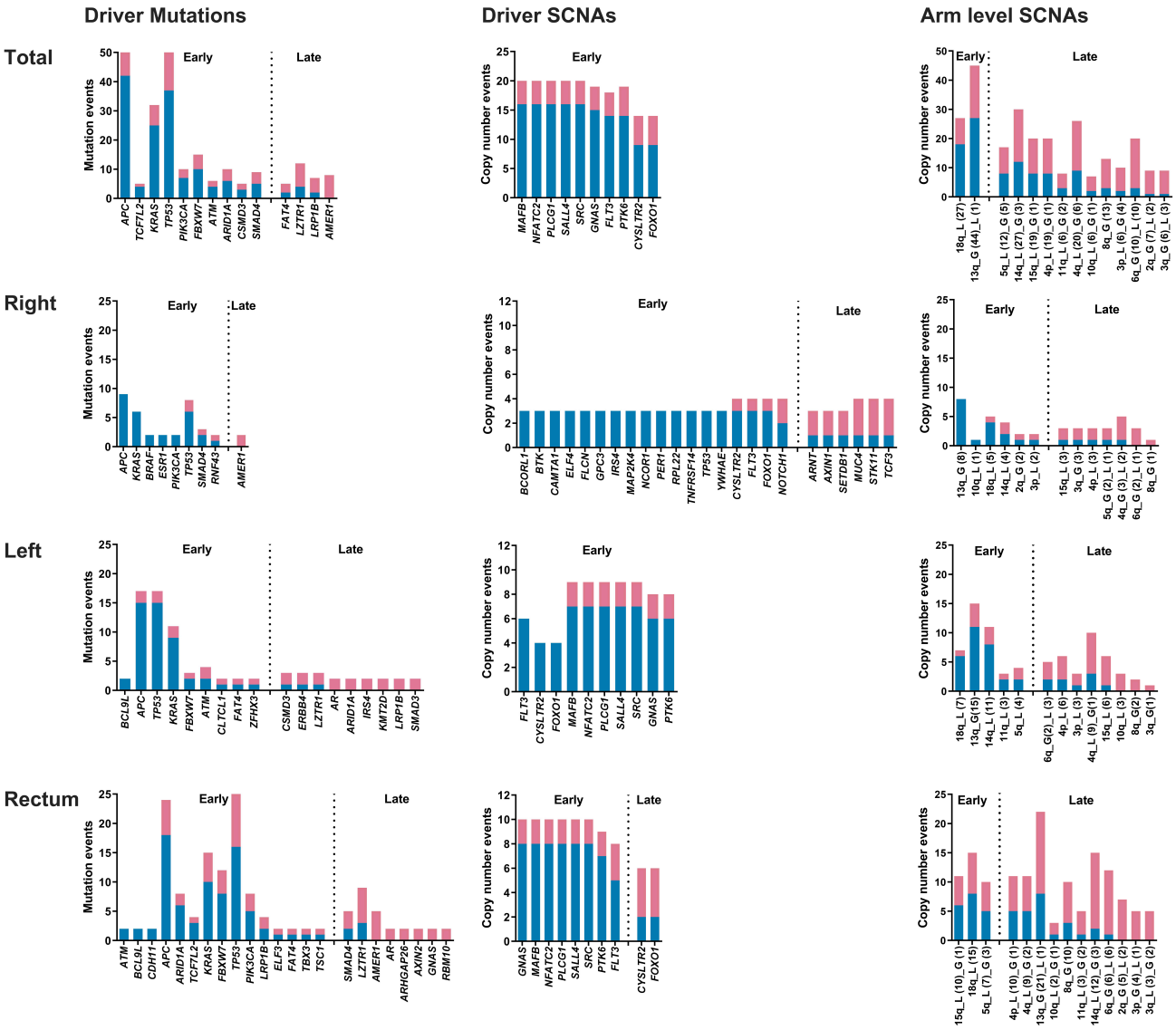

Figure 3 Summary of driver events in CRC evolution.
Mutations and somatic copy-number alterations (SCNAs) is shown as occurrence in patients indicating whether the events are clonal (blue) or subclonal (red). Only genes that were mutated in at least five patients in total or two patients in right-sided colon/left-sided colon/ rectum are shown. For SCNAs, driver SCNAs in at least 20% of the patients are shown while all the arm level SCNAs are shown. Driver events with more subclonal occurrence than clonal occurrence in tumors were late events, otherwise they were early events. In the arm level SCNAs part, “G” represented gain, “L” represented loss, and the numbers in parentheses represented the time of occurrence in tumors.

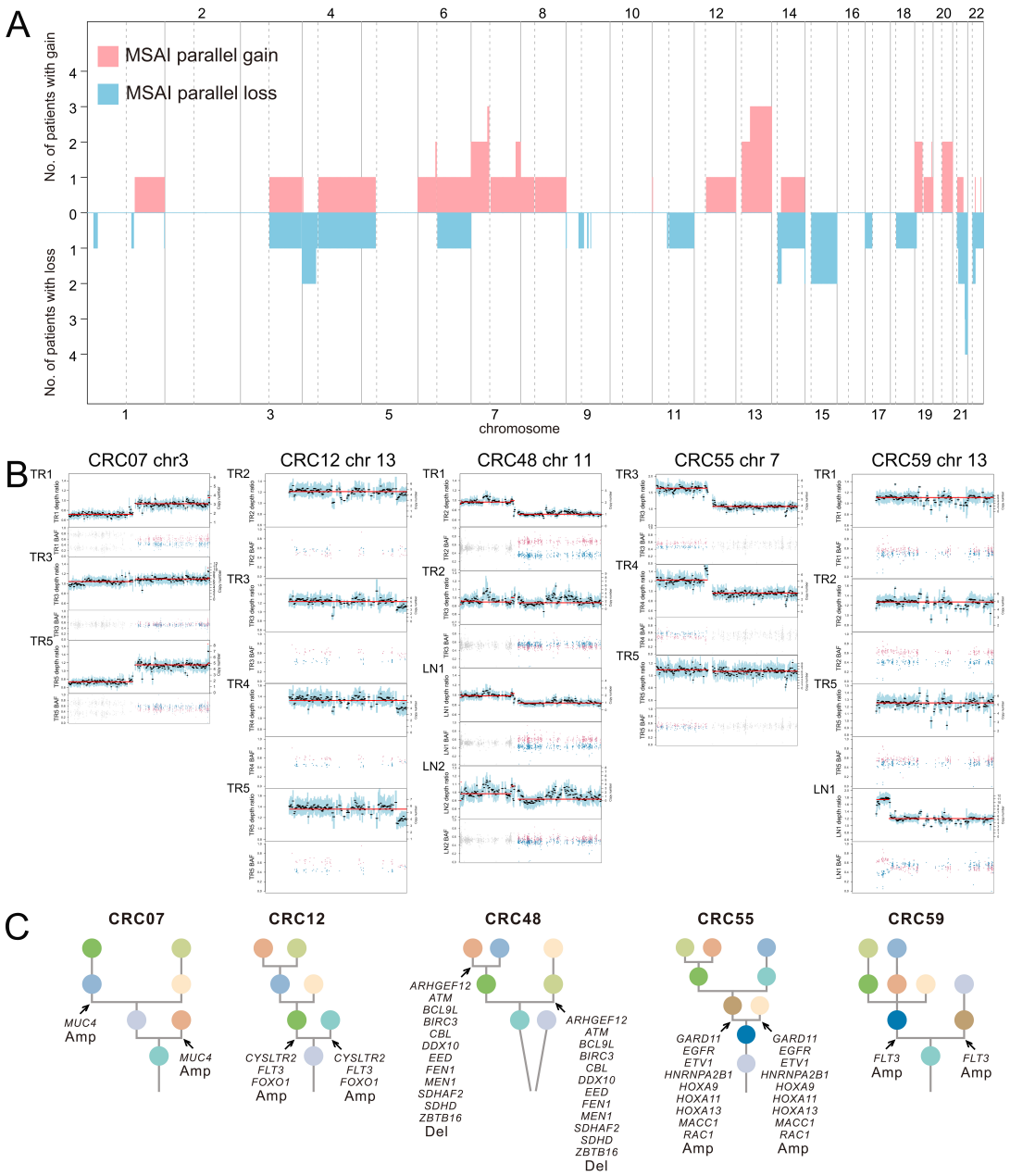

Figure 4 Parallel evolution.

(A) Genomic position and size of all MSAI parallel gain or loss events found in this study. This includes genome-wide copy number gains and losses which was subjected to mirrored subclonal allelic imbalance (MSAI) events and their occurrence in CRC tumors.

(B) Parallel evolution of driver SCNAs observed in 5 CRC tumors, indicted by the depth ratio and B-allele frequency values of the same chromosome on which the driver SCNAs (C) are located.

(C) Phylogenetic trees that indicate parallel evolution of driver amplifications (Amp) or deletions (Del) (Driver SCNAs) detected through the observation of mirrored subclonal allelic imbalance (arrows).

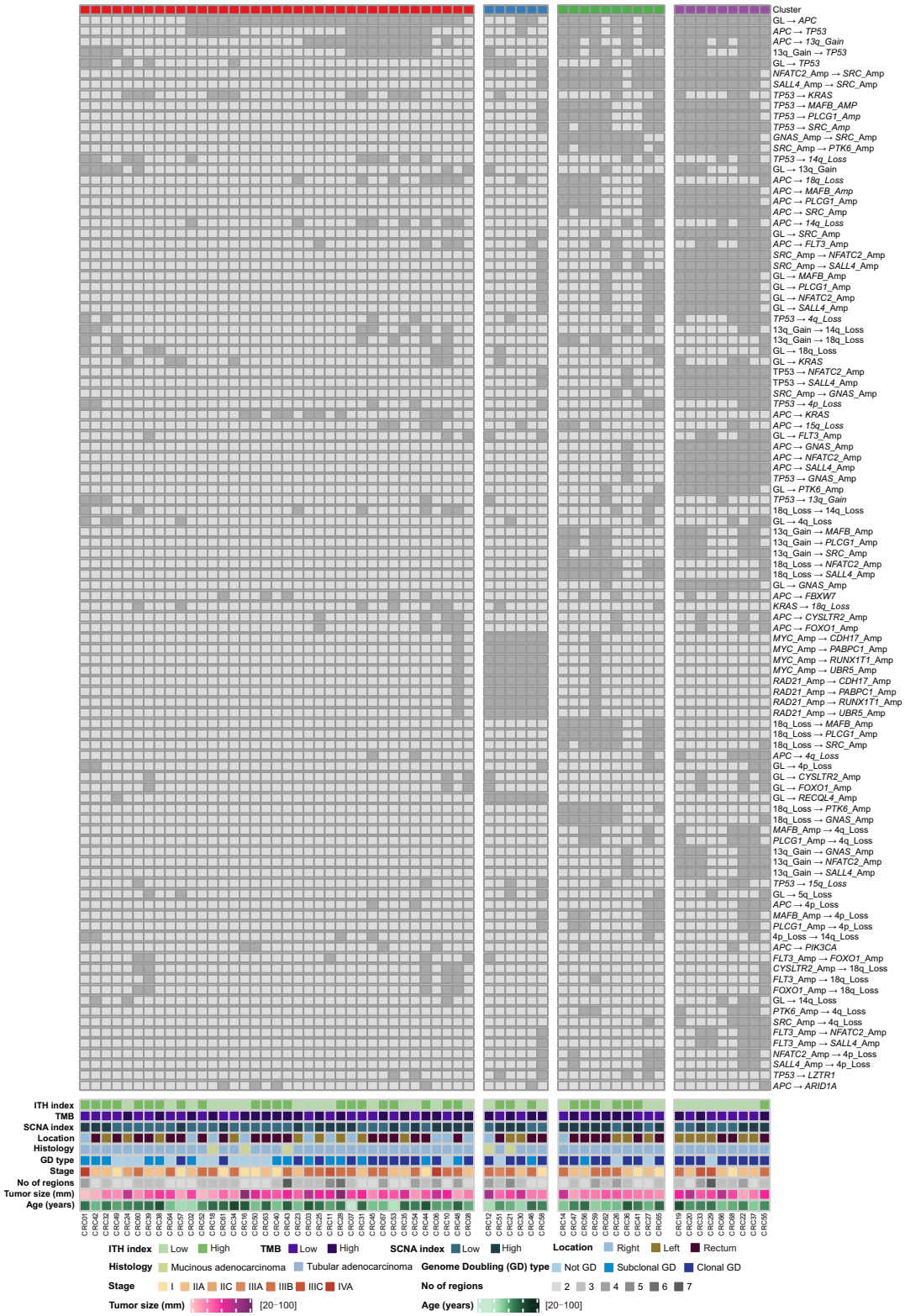

Figure 5 Evolutionary subtypes.

Evolutionary trajectories are clustered based on cancer cell frequency (CCF) value and cluster information of driver mutations, driver somatic copy-number alterations (SCNAs) and arm-level SCNAs. Heat maps showed the most recurrent evolution for the most recurrent driver mutations, driver SCNAs and arm-level SCNAs. Alterations are ordered by their frequencies in CRC tumors. CRC tumors are annotated by the following parameters: ITH index (high: half of the largest ITH index value; low: the other half), TMB (high > median, low = median), SCNA index (high > median, low = median), tumor location, histology, whole genome doubling status, stage, number of regions, tumor size and age.

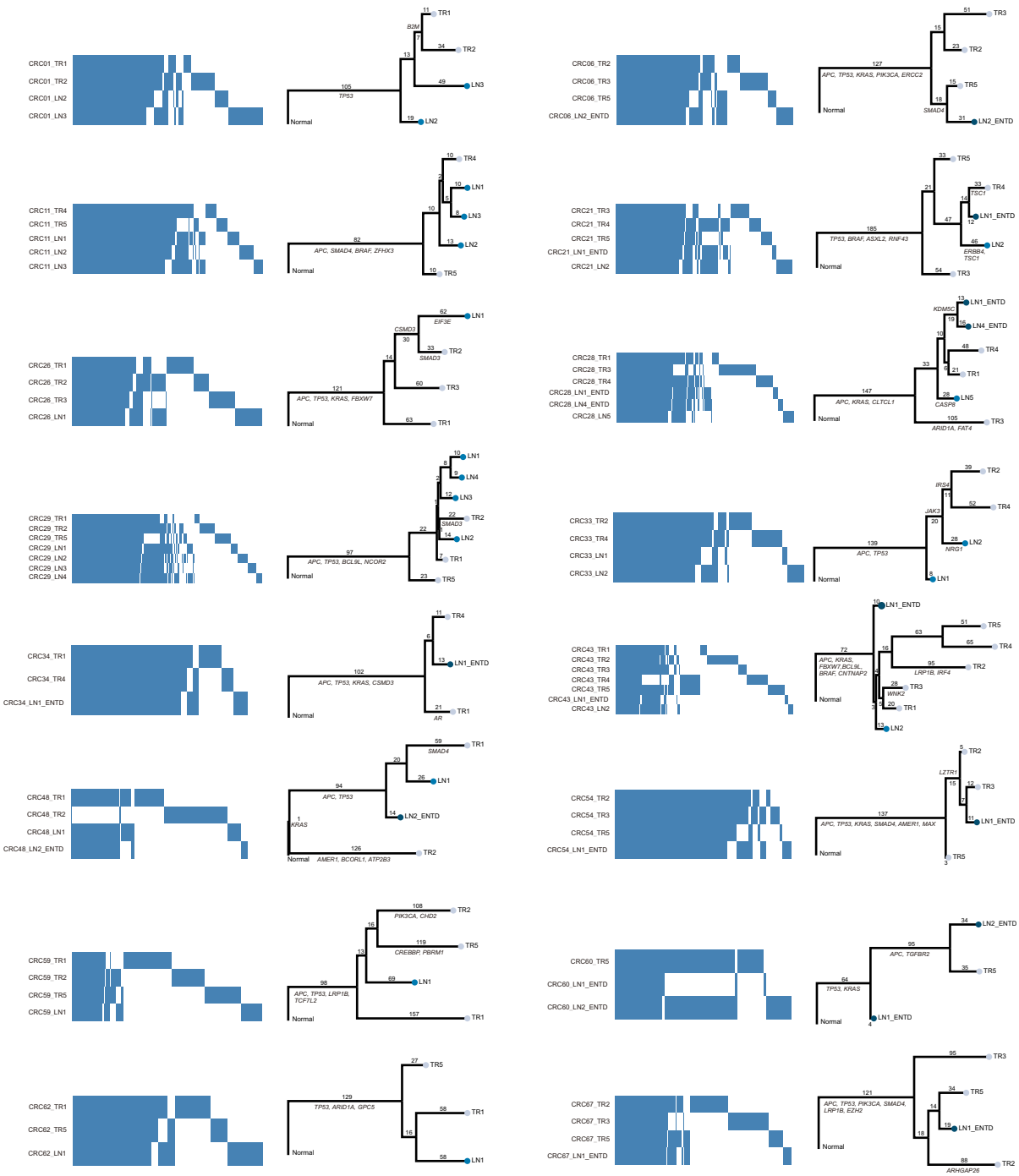

Figure 6 Phylogenetic distance between ENTD, lymph node metastasis and primary tumor. Heatmap showed the presence (blue) and absence (white) of all the mutations (SNVs and INDELs) among different tumor regions of the patients with ENTD or lymph node metastasis. Phylogeny reconstruction using maximum parsimony based on mutational presence or absence of all the mutations are shown beside heatmap. Driver genes are labeled in the phylogenetic trees.