Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity

Tae-Min Kim, Seung-Hyun Jung, Chang Hyeok An, Sung Hak Lee, In-Pyo Baek, Min Sung Kim, Sung-Won Park, Je-Keun Rhee, Sug-Hyung Lee, and Yeun-Jun Chung

Author's Affiliations: Departments of Medical Informatics, Cancer Evolution Research Center, Integrated Research Center for Genome Polymorphism, Surgery, Hospital Pathology, Pathology and Microbiology, College of Medicine, The Catholic University of Korea, Seoul 137-701, South Korea

Running title: ITH of colorectal cancers with metastases

Keywords: Colorectal cancer, Heterogeneity, Cancer evolution, Metastasis, Multiregion sequencing

Grant Support: This study was supported by a grant from National Research Foundation of Korea (2012 R1A5A2047939).

Corresponding author: Yeun-Jun Chung, The Catholic University of Korea, 222 Banpo-daero Seocho-gu, Seoul 137-701, Korea. Phone: +82-2-2258-7343; Fax: +82-2-537-0572; Email: yejun@catholic.ac.kr or Sug-Hyung Lee, The Catholic University of Korea, 222 Banpo-daero Seocho-gu, Korea. Phone: +82-2-2258-7311; Fax: +82-2-537-6586; Email: suhulee@catholic.ac.kr

Disclosure of Potential Conflicts of Interest: No potential conflicts of interest were disclosed.

This draft contains 5043 words (Abstract to Discussion) with 6 figures.
Translation relevance

Metastasis, especially to liver, is a major cause of cancer-related death of CRC. Knowledge on genomic evolution from primary to metastatic CRC would be crucial not only to understanding CRC pathogenesis but also to biomarker development for diagnosis and prognosis. In this study, we identified that CRC showed branching evolutionary patterns rather than linear patterns during the progression. Also the genetic ITH was prevalent in CRC and might serve as a potential driving force to generate metastasis-initiating clones. The data suggest that a single regional biopsy may underestimate CRC mutations in clinic. We observed a relative paucity of recurrent metastasis-clonal event, suggesting that CRC distant metastases may not follow a uniform course of genomic evolution, which should be considered in the genetic diagnosis and the selection of therapeutic targets for advanced CRC.
Abstract

Purpose: The intratumoral heterogeneity (ITH) and the evolution of genomic architectures associated with the development of distant metastases are not well understood in colorectal cancers (CRCs).

Experimental design: We performed multiregion biopsies of primary and liver metastatic regions from five CRCs with whole-exome sequencing and copy number profiling.

Results: In addition to a substantial level of genetic ITH, multiregion genetic profiling identifies the subclonal mutational architecture leading to the region-based or spatial categorization of somatic mutations and the inference of intratumoral evolutionary history of cancers. The universal mutations (those observed in all the regional biopsies) are enriched in known cancer genes such as APC and TP53 with distinct mutational spectra compared to biopsy- or region-specific mutations suggesting that major operative mutational mechanisms and their selective pressures are not constant across the metastatic progression. The phylogenies inferred from genomic data show branching evolutionary patterns where some primary biopsies are often segregated with metastatic lesions. Our analyses also revealed that copy number changes such as the chromosomal gains of c-MYC and chromothripsis can be region-specific and the potential source of genetic ITH.

Conclusions: Our data shows that the genetic ITH is prevalent in CRC serving as a potential driving force to generate metastasis-initiating clones and also as a means to infer the intratumoral evolutionary history of cancers. The paucity of recurrent metastasis-clonal events suggests that CRC distant metastases may not follow a uniform course of genomic evolution, which should be considered in the genetic diagnosis and the selection of therapeutic targets for the advanced CRC.
Introduction

Colorectal cancers (CRCs) are the third most common human malignancy worldwide with a ~5% lifetime risk (1). Distant metastasis is a major cause of cancer-related morbidity and mortality of CRC, mainly due to liver metastasis that is observed in approximately 25% of primary diagnoses of CRC. In spite of the clinical successes in early detection and prevention of CRC that resulted in an overall reduction of CRC risks (2), metastatic CRC still remains one of the leading causes of cancer-related death, with few therapeutic options available. Therefore, a better understanding of the molecular and genetic mechanisms underlying the biological and phenotypic evolution of CRC during metastasis is crucial to reduce morbidity and mortality of this disease.

As previously proposed (3), cancer evolution is described as successive cycles of somatic mutation acquisition and the natural selection of the fittest subclones, eventually giving rise to metastatic subclones (4, 5). Assuming that the development and progression of distant metastasis are genetic events, there have been efforts to identify causal genetic determinants responsible for distant metastasis. Recently, massively paralleled sequencing of breast cancer revealed that the majority of mutations observed in the metastatic lesions were also present, but often in low allele frequencies in the primary tumors (6). This finding was also supported by a direct comparison of mutation profiles obtained from primary and metastatic breast cancers (7). These results support the presence of genetic evolution associated with the development of distant metastasis, also raising a possibility that the clonal ancestry of distant metastasis can be traced to primary tumors.

It is now well recognized that human cancers have a substantial level of intratumoral heterogeneity (ITH) and mutational ITH has been demonstrated across multiple human cancer types (8-15). Genetic ITH is the manifestation of a driving force in cancer genomes.
for the emergence of new subclones that have metastatic potential. Thus, ITH should be taken into account when identifying a segregation pattern of metastasis-related mutations in primary tumors (16). To account for the regional ITH of mutations, Yachida et al. performed multiregion biopsies in both primary and metastatic pancreatic tumors and identified that a primary tumor mass is a mixture of geographically distinct subclones, some of which might specifically give rise to distant metastases (17). However, the extent of genetic ITH in primary and metastatic CRC lesions and the subclonal architectures specifically associated with the development of distant metastasis are largely unknown.

To identify genetic alterations associated with the metastatic progression of CRC, we performed multiregion biopsies in both primary and metastatic CRC lesions from five CRC cases, and analyzed somatic mutations and copy number alterations (CNAs). In addition to the apparent genetic ITH in both primary and metastatic lesions, we found branched evolutionary patterns in the CRC genomes, in which preexisting subclones in primary lesions often appeared responsible for the development of distant metastases. The mutations showed distinct mutational spectra, enriched molecular functions and aberration types (e.g., chromothripsis) according to their regional distribution and intratumoral recurrences (i.e., spatial mutation categories) giving clues to the subclonal architectures and potential metastasis-associated genomic changes of the corresponding tumors.
Materials and Methods

Tumor specimen

Colectomy tissues from five CRC patients used for this study came from a university-affiliated hospital (Eujeongbu St. Mary Hospital, Korea). All of the patients were Koreans and we only collected sporadic cases without any positive family history of CRC. The hospital pathology department confirmed pathologic features of the CRC. After the surgery, two to six different tumor areas and one normal mucosal area were picked from each fresh colectomy specimen (Supplementary Fig. S1). Also, matched metastatic nodules in liver were picked from the same patients along with primary CRC. Three patients (CRC1, 3, 4) had liver metastases at initial surgeries. Two patients (CRC2, 5) had no liver metastasis at initial surgery, but later they were found to have liver metastasis. All of the picked tissues from tumor and normal areas were frozen and stained with hematoxylin & eosin. Two pathologists selected cases with rich tumor cell populations (at least 70%) which were subsequently used in the study. For genomic DNA extraction, we used the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer’s recommendation. To evaluate the status of microsatellite instability, we used five mononucleotide repeats (BAT25, BAT26, NR-21, NR-24, and MONO-27) (18). For all of the five CRC cases examined, no markers showed instability; thus, all cases were classified as microsatellite-stable (MSS). The other clinical information of five CRC cases is available in Supplementary Table 1.

Whole-exome sequencing and mutation analyses

Using genomic DNA from tumor and matched normal samples, we performed exome-capture sequencing. Using Agilent SureSelect Human All Exome 50Mb kit (Agilent
Technologies), whole exome-sequencing was performed with Illumina HiSeq2000 platform to generate 101 bp paired-end sequencing reads according to the manufacturer's instructions. Burrows-Wheeler aligner (19) was used to align the sequencing reads onto the human reference genome (hg19). We used Genome Analysis ToolKit for the local realignment and score recalibration of the sequencing reads (20). We employed Picard (http://picard.sourceforge.net) and Samtools (21) for the basic processing and the management of the sequencing data. MuTect (22) and SomaticIndelDetector (20) were used to call somatic point mutations and indels by comparing the sequencing reads of the tumor and matched normal genomes. ANNOVAR package was used to predict their functional consequences such as silent or non-silent variants for somatic variants located in coding sequences (23). The evolutionary analyses based on mutation calls can be substantially flawed when the mutations are independently called from each of the intratumoral biopsies and simply merged. This is mainly due to false negative calls of low-frequent mutations or those in sparsely covered regions leading to the overestimation of the intratumoral heterogeneity. To minimize such biases, we first collected index mutations with mutant allele frequencies > 5% and those called in the regions with sequencing coverage (> 20X). For each of the index mutations, we examined the rest of the intratumoral biopsies for the presence of any sequencing reads that support the corresponding mutation in other regions of the given case. The sequencing information including the depth of sequencing for 35 tumor and 5 normal genomes are available in Supplementary Table 1. Sequences have been deposited to the SRA database and can be accessed on Project ID of PRJNA271316.

**Spatial mutation categories with respect to the regional distribution**

We established schematics of mutation classification system as shown in Figure 1A. The
given example represents six regional biopsies from primary and metastatic lesions (P1-P3 and M1-M3, respectively) and the regional distribution of mutations which can be used to infer the phylogenetic relationship of regional biopsies. We classified the mutations into five spatial categories of 'universal', 'metastasis-clonal', 'primary-private', 'metastasis-private' mutations along with 'unclassified'. Universal mutations are those observed in all regional biopsies of a given tumor (red in Fig. 1A). The regionally clonal presentation (i.e., the regional commonality) of universal mutations suggests that these events have arisen earlier during tumor evolution and may contain potential cancer initiators. Metastasis-clonal mutations represent those that are regionally clonal only in the metastatic regions thereby excluding universal mutations. We used this mutation category in a broad sense to include the events with a regional clonality in metastatic lesions but partially present or absent in primary lesions (orange and yellow in Fig. 1A, respectively). The latter might have occurred after the physical separation of metastasis-originating clones; however, it is possible the not all mutations in primary masses can be identified in our experimental settings due to regional biases or low allele-frequencies. Thus, we collectively classified these events as metastasis-clonal mutations. These events correspond to 'metastasis progressors' in previous reports (17) and may include the genetic events responsible for the initiation and development of distant metastases. Hypothetically, the primary subclones from which the metastasis-originating clones arise can be identified (annotated as P2* with asterisk in Fig. 1A). The other mutational categories include 'metastasis-private' events that are present but not in all of the metastatic lesions (green in Fig. 1A). These events may have occurred after the distant metastasis is established and acquired during the expansion of metastatic clones. In addition, 'primary-private' events that are present in primary but absent in metastatic lesions (blue in Fig. 1A) representing the changes independently acquired during the expansion of the
primary tumor mass. Along with the four spatial mutational categories, we classified the mutations that cannot be classified into the mutational categories discussed above as an additional category of 'unclassified' (grey in Fig. 1A). We performed DAVID analyses (http://david.abcc.ncifcrf.gov) (24) to identify the molecular functions enriched in the spatial category-specific mutations (i.e., genes harboring non-silent mutations only in one spatial category but not in other categories). We selected the top three significant functional clusters and their functional annotations from the DAVID output for each spatial category.

**Copy number profiling**

We used Agilent SurePrint G3 Human CGH Microarray 180K for copy number profiling as previously described (12). The log₂ ratio (tumor biopsy/matched normal) was segmented using GLAD algorithm (25). We further employed ABSOLUTE algorithm to estimate the purity and ploidy for each copy number profile (26). The segmented log₂ ratios and probe-level intensities were adjusted using the estimated purity and ploidy as described previously (27). The microarray data have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE58512. To compare the array-based copy number calls with those from exome sequencing, we used VarScan2 (28) to obtain the read depth differences between the tumor and normal sequencing data. The concordance between two platforms was estimated by calculating the Pearson correlation coefficient \( r \) between the GC-corrected, log₂ transformed read depth differences and log₂ ratio from microarray data. A good concordance level was observed \( r = 0.75 - 0.98; \) median of 0.90) suggesting that DNA copy numbers estimated from two independent platforms were largely concordant.
Chromothripsis

The inference of chromothripsis was done as previously described (29). For each of the chromosomes with at least ten breakpoints, we first evaluated whether the sizes of individual segments are not substantially different from their neighboring segments given that the random occurrence of breakpoints in chromothripsis would generate multiple segments with their sizes roughly the same or at least at the same order of magnitude with their neighbors. If we suppose that a chromosome has \( n \) segments (sizes of \( s_1, s_2, \ldots, s_n \)), we can use \( R_i = |\log_2(s_i/s_{i-1})| \) as a measure for the size difference between the neighboring segments. Then, we used the score \( S = R/R_e \) (\( R \) as the median of the \( R_i \)'s and the \( R_e \) as the expected value of \( R \) under the hypothesis that breakpoints are randomly distributed in the chromosome) to filter out the chromosomes that are unlikely to be involved with chromothripsis (\( S < 2.0 \)). To further evaluate the oscillating pattern of copy number states, we counted the number of peak-and-valley (i.e., the segments showing copy number difference of \( \log_2 > 0.5 \) with their neighbors). The permutation was performed 10,000 times and \( P \) value was assigned as the proportion of the permutations whose peak and valley counts are greater than the observed counts of peak-and-valley. \( P < 0.05 \) was used to select the chromosomes with the oscillating or alternating copy number changes.

Phylogeny estimation from mutations and copy numbers

The phylogenetic relationship between the regional biopsies of a given sample was inferred by mutations and copy number profiles. For each CRC, maximum parsimony trees were inferred as previously described (11, 12) using branch-and-bound algorithm implemented in PHYLIP software package. The phylogenetic relationship of intratumoral biopsies was also inferred from the copy number changes using the TuMult packages (30).
Among the GLAD-based segments, we selected those with log₂ ratios of >0.2 or <-0.2 and used their boundaries as chromosomal breakpoints for the input of TuMult algorithm. As reference of Tumult, we downloaded the segmentation data of 470 colorectal cancer genomes as available in The Cancer Genome Atlas consortium (31). The copy number calls were also used to test for a star topology (i.e., whether a phylogeny agrees with a linear or a branched evolution) using MEDICCquant R package (32). The null hypothesis assuming a linear evolution was rejected ($P < 0.05$) so that the alternative hypothesis of a branched evolution was selected for four of the five cases examined (CRC1 to CRC4, except for CRC5).
Results
Multiregion sequencing and mutation analyses of CRC genomes

In this study, we adopted a scheme of multi-region exome sequencing combined with spatial mutational profiling (12, 17) to reveal the mutational ITH of primary and metastatic CRCs. Somatic mutational events were classified into five spatial mutation categories of 'universal', 'metastasis-clonal', 'metastasis-private', 'primary-private', and 'unclassified' (Fig. 1A) with respect to the regional distribution (i.e., the regional biases and the level of recurrences for given mutations). The details on the categorization system and the interpretation of different spatial mutation categories are available in Materials and Methods.

We analyzed 35 multiregion biopsies of the primary and metastatic lesions in five CRCs (n = 2–5 and n = 2–6 for primary and metastatic biopsies per case, respectively) (Supplementary Table S1). The geographical mapping of the biopsy spots in primary CRCs and liver metastatic lesions is shown in Supplementary Fig. S1. Overall, whole-exome sequencing of the 35 primary and metastatic CRC genomes identified 76–244 coding mutations and indels per biopsy (median of 104 somatic variants). Mutation numbers between the primary and metastatic biopsies of each patient were not significantly different except one case with the largest number of biopsies (CRC3; P = 4.8×10⁻⁵; t-test). The mutations observed from the five CRC cases are listed with the detailed information of spatial categories and functional consequences in Supplementary Tables S2. The regional mutation profiles for the five CRC cases with their spatial categories showed a substantial level of ITH (Fig. 1B). We observed that 19.8-53.9% of mutations in a given sample were universal, while 46.1–80.2% of mutations were subclonal. Among the subclonal mutations, metastasis-clonal, metastasis-private and primary-private categories comprised 0.7-15.6%, 2.4-40.7% and 13.8-
56.0% of mutations, respectively. In addition, 1.4–37.2% of mutations per biopsy were not observed in any other regions of the given sample (biopsy-specific mutations, see Supplementary Fig. S2).

We next investigated recurrent nonsilent mutations (i.e., those observed in ≥ 3 of the five CRC cases). Figure 2A shows the nine genes with such mutations including well-known the KRAS mutations. All of the APC, KRAS and TP53 mutations were identified as universal events except for a metastasis-clonal APC mutation in one case. Eight APC mutations were all truncating (either nonsense substitutions or frameshift indels) and were often biallelic (e.g., CRC2 harbored one nonsense mutation and another frameshift mutations). Two KRAS missense mutations occurred at amino acid residue of position 12 (p.G12S and p.G12D). Two of five TP53 mutations were truncating and the other three TP53 missense mutations occurred at (p.R213X and p.R273C) or near the known mutation hotspots (p.R267P) (33). See also Supplementary Tables S2 for the details on the functional consequences and the amino acid residues affected for other mutations. The other recurrent mutations besides APC, TP53 and KRAS were dispersed across different spatial categories. We also performed gene set enrichment analysis to infer the unifying molecular terms of the non-silent mutations that exclusively belong to each of the spatial categories (i.e., those observed in only one of five spatial categories) or the category-specific mutations. The functional annotations of 'cell adhesion' are commonly enriched in universal, metastasis-clonal and primary-private categories (Fig. 2B).

Distinct mutational spectrum of spatial mutation categories

We evaluated the properties of mutations according to the spatial categories (Fig. 3). Figure 3A shows the number of mutations across five spatial categories for five CRC cases. We
observed that frameshift and inframe indels were enriched to the metastasis-private category in two cases (CRC2 and CRC4, Fig. 3B). While C:G>T:A transitions at CpG dinucleotides were significantly enriched to universal category in three CRCs (Fig. 3C), the enrichment of C:G>A:T and T:A>A:T transversions were only observed in lesion-specific (metastasis-private and primary-private) categories in more than two cases. Analyses of mutant allele frequencies revealed that mutations from metastasis biopsies had significantly higher allele frequencies compared to those from primary biopsies ($P = 2.2 \times 10^{-16}$; $t$-test) (Supplementary Fig. S3). This finding suggests that metastatic lesions were relatively clonal due to the evolutionary bottleneck and restriction of diversity. Regarding spatial mutation categories, universal mutations had the highest allele frequencies, while primary- and metastasis-private mutations had the lowest (Supplementary Fig. S3).

**Phylogeny and copy number profiles of regional biopsies**

Phylogenetic trees inferred from genomic data may represent the evolutionary life history of individual tumors as discussed in previous literatures (11, 12). In the present study, the phylogenetic relationship between the regional biopsies of a given sample was inferred from mutation data using maximum parsimony and chromosomal breakpoints in DNA copy number profiles as described previously (11, 12). Figure 4 shows phylogenetic profiles (mutation- and copy number-based) of the five CRCs along with the genome-wide heatmap of CNAs. The inferred evolutionary patterns appeared as a branched evolution rather than a constant or linear evolution. For example, statistical analysis of the phylogeny patterns (32) indicated that the phylogeny of four (CRC1-4) of the five CRCs was branched; i.e. the null hypothesis of the test (a linear or constant evolution) was rejected for these four CRCs except for CRC5.
In either mutation- or copy number-based phylogenies, most regional biopsies from metastases were co-segregated apart from those of primary lesions indicating that they were more genetically similar to each other than the primary lesions (Fig. 4). Most CNAs were shared by both primary and metastatic lesions, representing early or universal genomic events. Interestingly, in two cases (CRC2 and CRC3) mutation-based phylogenies segregated some primary subclones to those of metastatic tumors rather than those of other primary tumors (e.g., CRC2-P1 and CRC3-P3/P4; those with asterisks in Fig. 4). This phenomenon was consistently observed in copy number-based phylogenies (Fig. 4). These observations suggest that the origins of distant metastases could also be traceable in the subclonal architecture of primary CRCs as previously reported in pancreatic cancers (17).

**Copy number alterations containing well-known cancer genes**

Well known cancer genes which were reported to be frequently altered in CRC genomes such as *APC, PTEN, IGF2* and *SMAD4* (31) were found to be universally altered across the whole biopsy sites in this study (Fig. 5A). We also observed the universal focal deletion in between *VTI1A* and *TCF7L2* in CRC2, which was reported to be responsible for the recurrent fusion between these genes in CRCs (34). Most of the universal focal changes involving potential driver genes showed exactly coinciding boundaries across the regional biopsies indicative of their clonal origins (Fig. 5A). By contrast, genes located in fragile regions such as *MACROD2, FHIT, WWOX* (all on CRC2), and *PARK2* (on CRC3) showed deletions whose boundaries were irregular across the regional biopsies of given tumors, although they were universal (Fig. 5B) indicative of ongoing genomic instability in these fragile loci. It is worth noting that the copy number gains involving *c-MYC* oncogene were observed in all five CRCs analyzed (one focal and four 8q arm-level change cases) (Fig. 5C). Universal or
metastasis-clonal copy number gains of 8q encompassing c-MYC were observed in CRC1, 2, 4, and 5, while a focal gain of c-MYC as a metastasis-clonal event was observed in CRC3. In CRC2, in addition to arm-level metastasis-clonal 8q gain, a focal gain encompassing c-MYC was also observed in a primary lesion (CRC2-P4) suggesting that the evolutionary path to c-MYC copy number gains may be diverse even within a single cancer mass.

**Evidences of region-specific chromothripsis**

Among the five CRCs examined, two candidate chromothripsis events were observed as one universal (on chromosome 2 of CRC3) and one metastasis-private event (on chromosome 16 of CRC5). In the chromothripsis of CRC3, breakpoints in deleted segments exactly overlapped across all the regional biopsies, suggesting that this event occurred early in tumor development (Fig. 5D). However, the intervening segments in between the copy number losses often showed neutral states or copy number gains in the primary tumors while they were consistently copy number gained in the metastatic lesions. This result suggests that additional copy number changes were acquired in the rearranged chromosomal segments during metastatic progression after the initial universal chromothripsis event. In CRC5, we observed a metastasis-private chromothripsis event that must have occurred after the divergence of metastatic clones from the primary cancer mass (Fig. 5E).

**Potential genetic determinants of CRC metastases**

Theoretically, metastasis-clonal events are potential candidates for metastasis determinants. We first investigated apparent loss-of-function events among the metastasis-clonal mutations, including frameshift indels (GDPD1, NPS and SLK) and nonsense mutations (APC, ASCC3, CDH12, ENTPD4, GYLTL1B, NEDD4L, PCDHB5, and TK2). Through capillary sequencing
of seven lesions (four primary and three metastatic) of the corresponding case (CRC1), the ITH of the NEDD4L nonsense mutation was confirmed (data not shown). In the following immunohistochemical analysis for the CRC1, primary lesions without the NEDD4L mutation showed positive NEDD4L immunostaining, whereas metastatic liver lesions with the NEDD4L mutation showed decreased NEDD4L immunostaining (Fig. 6A and B). Furthermore, in an independent set of 17 CRCs with liver metastasis, seven liver metastases (41.2%) even without NEDD4L mutation showed decreased NEDD4L expression compared to the corresponding primary CRCs (Fig. 6C and D), suggesting that liver metastasis-specific NEDD4L down-regulation might be caused by non-mutational as well as mutational events. Among the focal CNAs, we selected metastasis-clonal deletions of PCDH9 as a candidate genetic event for metastasis. The ITH of PCDH9 was confirmed by real-time qPCR in the corresponding case (CRC2). In this case, primary CRC sites without the PCDH9 deletion showed positive immunostaining, whereas two liver metastasis lesions showed much weaker immunostaining in the cancer cells, indicating that the metastasis-clonal pattern of the PCDH9 deletion was recapitulated at the protein level (Supplementary Fig. S4).
Discussion

In this study, multiregion biopsies from primary CRC and liver metastases revealed a substantial level of ITH in both primary and matched liver metastases. Our results, 46–80% subclonal mutation fractions (i.e., those not detectable across all the regional biopsies), are similar to the previous estimates of other cancer types (10-12), further supporting the notion that ITH is a general phenomenon in cancers. Although ITH at the genomic level has been observed across many cancer types, including a report of colon adenomas (35), to our knowledge, our study is the first report on the ITH of CRC genomes with a detailed view on spatial mutation categories and phylogenetic structures of primary and metastatic lesions obtained from the same individuals.

There have been concerns on previous metastasis-specific mutation identification that might generate bias. For example, studies based on ‘index’ mutations solely identified from metastatic tumors (17, 36) might not only ignore a substantial level of ITH, but also can be biased by the selection of metastatic biopsies used to generate the index mutations. In present study, we demonstrate that multiregion sequencing across multiple primary and metastatic lesions provides more information (e.g., the extent of ITH, the categorization of the mutations according to the different evolutionary stages and the evolutionary relationship between the primary and metastatic lesions) compared to a simple primary-vs.-metastatic lesions (37, 38).

The recurrent mutations on known cancer genes (e.g., APC, KRAS and TP53) (39) were almost exclusively observed in the universal category consistent with recent reports (37, 38). In case of APC mutations, three of five cases showed ‘second-hit’ mutations on APC, probably representing biallelic mutations or the occasional third hit on APC locus (40). To get a clearer understanding, biallelic APC mutations in one case (one universal frameshift indel and one metastasis-clonal nonsense mutation of CRC2) were validated using Sanger
sequencing (Supplementary Fig. 5). The nonsense APC mutation in this case was classified as metastasis-clonal because of the absence of the mutations in a primary region (CRC2-P1) that also exclusively harbors an entire loss of chromosome 5. Thus, it is expected that two APC mutations in this case were biallelic, but the subsequent chromosomal loss led to the loss of only one of the biallelic mutations (nonsense mutation) while retaining the other (frameshift indel). It is also notable that this lesion-specific chromosomal loss altered the spatial category of the mutation (universal to metastasis-clonal), which may be a potential source of unclassified mutations.

The mutations that cannot be explained according to the evolutionary model (Fig. 1A) were left in ‘unclassified’ category. We propose several assumptions for these evolutionarily unexplainable mutations (e.g., those subclonal both in primary and metastatic biopsies). First, the false negative calls of mutations due to a low sequencing depth or other technical issues may lead to a misclassification of mutations and can be the source of ‘unclassified’ mutations. Second, unclassified mutations may have arisen with subsequent chromosomal changes such as loss-of-heterozygosity as shown in the example of biallelic APC mutations in CRC2. Third, these mutations may have been acquired independently in primary and metastatic lesions representing a convergent evolution (homoplasy). Although homoplasy events have been previously reported in terms of gene-level convergent events (e.g., SETD2 frameshift indels and splicing mutations independently acquired by distinct biopsies in a cancer mass) (10), we believe that convergent mutations with identical genomic coordinates would be rare. But the mutations with known hotspots will be worthy of further investigation as the potential source of misclassification. Lastly, a recent report about the colorectal cancer evolution proposed that the subclonal mutational architecture as configured early in the evolution of cancers, may persist during the growth of cancer mass (41). The presence of ‘pervasive’ mutations (i.e.,
subclonal mutations present across the cancer mass), especially those with very low mutation frequencies, may be the source of error in the accurate inferring of the phylogeny.

The first issue may be overcome by using ‘high-confident’ somatic mutations (i.e., those called in genomic regions with sufficient sequencing coverage across all the regional biopsies). To test this, we selected somatic mutations called with sequencing depth >50X in all the regional biopsies and inferred the phylogeny (Supplementary Fig. 6). Although the phylogenies inferred from the mutations with or without this stringent filtering are largely concordant with each other, it should be further investigated whether the use of high-confident mutations may perform better in inferring the evolutionary relationship of intratumoral biopsies. In addition, low-confident mutations in particular biopsies can be selectively ignored to preserve the number of mutations as evolutionary markers.

We observed the preponderance of indels in late (metastasis-private and primary-private) genomic events suggesting that CRC genomes seem more vulnerable to different types of somatic variants during their evolution. It is also possible that these unique genomic footprints might reflect different fixation rates of the mutation types, e.g., highly deleterious events such as indels may be negatively selected in the early evolutionary phase. We also disclosed the mutation spectra associated with CRC development and progression stages (e.g., C:G>T:A transitions enriched in universal mutations), indicative of different mutational processes dominant in the early and late evolutionary phases during CRC progression.

The overall higher allele frequencies of mutations in the metastatic tumors compared to those in primary tumors was indicative of the presence of an evolutionary bottleneck. However, the allele frequencies of the five spatial mutation categories revealed that primary- or metastasis-private mutations had the lowest allele frequencies, indicating that these lesion-specific mutations could be largely late events. Cares should be taken for this interpretation
since this comparison as well as other analyses based on allele frequency can be biased by the purity and ploidy level of the selected tumor lesions. For example, low tumor purity with a high level of normal tissue contamination will lead to the underestimation of tumor allele frequencies of mutations. Moreover, we selected tumor-rich regions (>70%) for regional biopsy in this study, but additional source of low tumor purity (e.g., the selection of tumor invasive fronts) would still exist.

The phylogenetic trees inferred from somatic mutations and DNA copy number changes were largely concordant and provided the evolutionary history. First, the metastatic clones are co-segregated apart from those of primary lesions indicative of their genetic similarity. Second, mutation profiles in some primary lesions were genetically more similar to those of metastasis than the other primary lesions suggesting that they might be a potential metastatic source. In spite of the overall similarities, some discrepancies were observed between the trees from somatic mutations and DNA copy number changes. Potential causes for the discrepancies may include those of ‘unclassified’ mutations as previously addressed as well as technical issues in calling the CNAs. Among the unclassified mutations, those in CRC1 were all observed on chromosome 2 (i.e., four unclassified mutations that are present in all of the regional biopsies except for one metastatic biopsy of CRC1-M3). It is speculated that those mutations are universal events ab initio but subsequent chromosomal events such as loss-of-heterozygosity in CRC1-M3 led to the biopsy-specific loss of mutational calls. Moreover, the chromosome 2 of CRC1-M3 does not show apparent copy number changes suggesting that this chromosomal event is likely to be copy number neutral loss-of-heterozygosity. Supporting our speculation, all three universal mutation events in this chromosome showed significantly higher mutational allele frequencies ($P < 0.05$; paired $t$-test) for CRC1-M3 compared to those of other regional biopsies.
Although a prevailing view has supported early occurrences of chromothripsis, especially those that are cancer drivers (42), we found a CRC case that harbored a metastasis-private chromothripsis event supporting another report that chromothripsis does not necessarily represent an initiating event (43). Our results suggest that chromothripsis may occur at any developmental stage in CRC genomes, and the incidence of chromothripsis estimated from primary CRC cases might be underestimated. Additional complexity of chromothripsis in terms of the CNAs of affected chromosomal segments further suggests that this chromosomal event may be accompanied by subsequent chromosomal changes.

Systematic sequencing analyses of primary and metastatic cancer genomes have shown that a small number of mutations are required for an invasive CRC to metastasize (36). Moreover, even newly acquired mutations during metastatic progression are not overlapped among individual patient samples (17). Consistent with these findings, our study identified metastasis-clonal mutations only in 0.7–15.6% of total mutations in the given cases. These data suggest that essential mutations for cancer progression may be obtained in primary cancer genomes before metastasis. In line with this hypothesis, colon adenoma genomes were reported to be nearly as old as invasive cancers (44) suggesting the time required for malignant transformation may be relatively short in the entire life history of CRC. The driving force for CRC metastases may be beyond the scope of our study, e.g., mutations occurring in noncoding sequences (45) or epigenetic alterations (46).

Finally, we validated some of metastasis-clonal mutations, including a NEDD4 nonsense mutation and PCDH9 deletion (both as singletons) by resequencing or qPCR in our test sets along with immunostaining in an independent cohort. Metastasis-clonal nonsense mutation of NEDD4L is interesting, since the encoded protein NEDD4L interacts with Smad proteins (47) and can regulate TGF-β signaling that has been implicated in cancer invasion and metastasis.
(48). However, the potential role of NEDD4 as metastatic determinants along with that of PCDH9 requires further investigation. Although we did not find novel driver genes of CRC metastasis in the metastasis-clonal category, integration of these intermittent findings will be needed to elicit biologically relevant insights to improve diagnosis and treatment of CRC patients with metastasis.
References


Letouze E, Allory Y, Bollet MA, Radvantyi F, Guyon F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. Genome Biol 2010;11:R76.


Figure Legends

**Figure 1. Schematics of resolving somatic variants with respect to regional distribution and intratumoral recurrences.** A, Six regional biopsies from primary (P1-P3) and metastatic tumors (M1-M3) and their inferred evolutionary relationships are shown in this exemplary case. Asterisk (P2*) indicates the regional biopsy of the primary tumor as the potential ancestor of the metastatic clones. Mutations observed in all of the regional or metastatic biopsies are classified into 'universal' (red) or 'metastasis-clonal' (orange and yellow), respectively. Metastasis-clonal mutations include those with partial presence (orange) or absence in the primary lesions (yellow). Mutations observed in specific primary or metastatic biopsies are primary-private or metastasis-private (blue and green, respectively). Mutations that are not classified according to this scheme are unclassified (grey). B, The regional distribution of somatic variants for the five CRC genomes is shown. For each of the five CRC genomes (CRC1–CRC5), the somatic variants for at least one regional biopsy in a given case are shown as rows on the heatmap with the number of somatic variants shown at the top. The mutations of primary and metastatic lesions are shown as black and purple on the heatmap and barplots above. At the right of each heatmap, the classification of the somatic variants according to five spatial categories is shown with a color scheme below. Nonsilent mutations in cancer census genes are shown for missense (blue) and nonsense point mutations (red) as well as frameshifting (orange) and nonframeshifting indels (yellow).

**Figure 2. Recurrent mutations and pathway analyses.** A, Mutations with no less than three occurrences across the cases are shown with four cancer census genes (APC, TP53, KRAS, and MAML2) at the top (grey shaded). Subcolumns representing CRC1 to CRC5 are in each of the five columns corresponding to the spatial mutation categories. Nonsilent mutations are marked by a color scheme (right). Redundant mutations are also shown. For example, CRC3 harbored two universal mutations on APC, one nonsense mutation (red), and one frameshift indel (orange). B, The results of the DAVID pathway analyses are shown for the five spatial mutation categories. The top three functional clusters are shown for each spatial category with the x-axis representing the log-transformed P-values.

**Figure 3. The mutational features of five spatial mutation categories.** A, The number of mutations in five spatial mutation categories are shown for the five CRC genomes according to the color scheme shown in the inset. B, For each of the spatial mutation categories, the relative fraction of six functional categories (SC) is shown in percentage (%; y-axis). Asterisks indicate significant enrichment as measured by the Fisher's exact test ($P < 0.05$). For example, nonsense mutations were relatively enriched in the metastasis-clonal mutations of CRC2 given the total number of mutations and nonsense mutations for the other spatial categories in the sample. C, Similarly, significant enrichment is also indicated by asterisks for the sequence composition of mutations. C:G>T:A changes (C>T) are further distinguished into those specific to CpG dinucleotides or not.

**Figure 4. Phylogenetic trees of the regional biopsies.** For each case (CRC-1 to CRC-5 from top), two phylogenetic trees are shown (one on the left and the other in the middle) along with genome-wide plots of DNA copy number changes (right). Phylogenetic trees constructed using maximum parsimony from somatic mutations are shown on the left with the length of branches in the tree proportional to the number of mutations in the corresponding branches.
The phylogenetic trees inferred from chromosomal breakpoints in DNA copy number profiles using TuMult are shown in the middle. Asterisks indicate the primary biopsies that are more genetically similar to metastatic biopsies than those of remaining primary biopsies. The nodes representing normal (‘Norm’) and common precursors (‘Pre’) are shown with the biopsies from primary (P; grey) and metastatic lesions (M; red). The order of the regional biopsies in the heatmap (right) is according to the tree constructed from chromosomal breakpoints as shown with dotted lines. In the heatmap, chromosomal gains and losses are shown as red and blue, respectively.

**Figure 5. Copy number alterations in cancer genes and chromothripsis.** A, Focal and universal copy number alterations are shown for *APC, PTEN, VTI1A-TCF7L2, IGF2*, and *SMAD4*. Heatmaps of the selected loci are shown for probe-level log2 intensities. The case number and genomic coordinates are shown for each heatmap with arrows indicating the locus of genes. B, The probe-level intensities are shown for *FHIT, WWOX, MACROD2*, and *PAR2*. C, The copy number changes are shown for chromosome 8 for all the regional biopsies (left). Focal chromosomal gains involving *MYC* are marked and separately shown on the right. CRC-2 and CRC-3 cases harbored primary-private and metastasis-clonal focal chromosomal gains involving the *MYC* locus (arrow). D, Whole chromosome-scale chromothripsis on chromosome 2 is shown (CRC-3, left). Focal scale chromothripsis on chromosome 17 (CRC-1) (E) and biopsy-specific chromothripsis on chromosome 16 (CRC-5) are shown (F).

**Figure 6. Visualization of NEDD4L expression in colon cancer tissues and corresponding liver metastasis by immunohistochemistry.** A colon cancer case with *NEDD4L* nonsense mutation shows positive NEDD4L immunostaining in primary colon tissues (A), but very weak NEDD4L immunostaining in metastasized cancer cells from the liver of the same patient (B) (frozen tissues). Another colon cancer case in a validation set shows positive NEDD4L immunostaining in primary colon tissues (T in (C)), whereas metastasized colon cancer in the liver of the same patient shows very weak NEDD4L immunostaining (T in (D)). Adjacent normal colon mucosal (N in (C)) and liver (N in (D)) cells show positive NEDD4L immunostaining (FFPE tissues).
Figure 1

A

Primary tumor

Metastatic tumor

B

CRC1

CRC2

CRC3

CRC4

CRC5

Common in primary and metastatic lesions (universal)

Common in metastatic but private in primary lesions (metastasis-clonal)

Common in metastatic but private in primary lesions (metastasis-clonal)

Present (but not clonally) in metastatic lesions and not present in primary lesions (metastasis-private)

Present in primary lesions and not present in metastasis (primary-private)

Unclassified (not listed above)

---

Author Manuscript Published OnlineFirst on May 15, 2015; DOI: 10.1158/1078-0432.CCR-14-2413

Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.
Figure 2

A

![Genes and genetic alterations]

- **APC**
- **TP53**
- **KRAS**
- **MAML2**
- **CSMD3**
- **RYR1**
- **TANC1**
- **TMPRS13**
- **ZNF814**

The bars represent different types of genetic alterations:
- **Missense**
- **Nonsense**
- **Splicing**
- **Frameshifting indels**
- **Inframe indels**

B

Significance (-log10(P))

- **Glycosylation**
- **IPR013032: EGF-like region**
- **GO:007155: cell adhesion**
- **GO:007155: cell adhesion**
- **GO:0043588: skin development**
- **Nucleotide phosphate-binding**
- **Glycosylation**
- **Transmembrane protein**
- **GO:0005856: cytoskeleton**
- **IPR001611: Leucine-rich repeat**
- **Domain:PH**
- **IPR003961: Fibronectin, type III**
- **GO:0005578: proteinaceous**
- **Domain: Fibronectin type-III 2**
- **Cell adhesion**

Significance levels:
- **Universal**
- **Metastasis clonal**
- **Unclassified**
- **Metastasis private**
- **Primary private**
Figure 3

A

B

C

Mutational abundance

Relative fraction

Relative fraction

Universal
Metastasis clonal
Unclassified
Metastasis private
Primary private

Missense
Nonsense
Frameshift
Inframe
Splicing
Synonymous

T>G
T>C
T>A
C>A
C>G
C>T (nonCpG)
C>T (CpG)
Figure 4
Figure 5

A

CRC5
Chr5:100 - 140Mb

CRC4
Chr10:80 - 95Mb

CRC2
Chr10:114 - 115Mb

CRC5
Chr11:1.5 - 3Mb

CRC1
Chr18:46 - 50Mb

APC↑

PTEN↑

VT1A  TCF7L2

IGF2↑

SMAD4↑

B

CRC2
Chr3:59 - 62Mb

CRC2
Chr16:77 - 80Mb

CRC2
Chr20:13 - 17Mb

FHIT

WWOX

MACROD2

C

8p

CRC1

CRC2

CRC3

CRC4

CRC5

8q

D

Chromosome 2 (CRC3)

E

Chromosome 16 (CRC5)
Figure 6

A

Colon

B

Liver

C

N

Colon

T

D

N

Liver

T

Downloaded from clincancerres.aacrjournals.org on July 15, 2017. © 2015 American Association for Cancer Research.
Clinical Cancer Research

Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity


Clin Cancer Res Published OnlineFirst May 15, 2015.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-14-2413

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2015/05/20/1078-0432.CCR-14-2413.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.