SUPPLEMENTARY METHODS

Patient selection

As shown in the study flowchart (Fig. S1), medical records of 14,216 patients who underwent surgical resections for colorectal cancer (CRC) in West China Hospital (WCH) from January 2009 to June 2018 were systemically reviewed in a clinical database. To find patients with available metastatic CRCs samples, we cross referenced the clinical data with pathological database and identified 432 and 117 patients with liver or lung metastatic lesions resected respectively. Among these cases, six patients with available formalin-fixed paraffin-embedded (FFPE) samples from primary tumors, liver and lung metastases were included. Regional lymphatic metastases were also available for two patients (patient C01 and C02), and distance thoracic lymphatic metastases were available for one patient (patient C01). Multiple tumor regions of these patients were collected for sequencing and subsequent genomic evolutionary analysis. The computed tomography (CT) scan or magnetic resonance imaging (MRI) of metastatic event in each patient were shown (Data. S1). Meanwhile, data in the 543 patients with liver/lung metastases resected were matched with records of frozen samples in BioBank at our hospital. Individuals with no frozen metastatic samples were excluded. A total of 146 patients with available frozen samples from primary tumor and matched LIM (n = 136)/LUM (n = 10) were collected as a validation cohort (hereafter referred to as MET cohort) (Fig. S1). On the other hand, 210 consecutive CRC patients who underwent surgery in 2010 with available primary tumor samples only were included as another validation cohort (hereafter referred to as WCH cohort) and were further classified into metastatic CRCs (mWCH) and non-metastatic CRCs (nWCH) according to the presence of distant metastasis before or within three years after the initial operation. This study was approved by the Ethics Committee of West China Hospital, Sichuan University (2018(280); 2019(338)). Written informed consent was obtained from all patients or their guardians, as appropriate.

Patient and Public Involvement

This research was done without patients' involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

Cell lines

Humans CRC cell lines HCT15, HCT116, and SW480 were purchased from the ATCC. HCT15 cells were cultured in RPMI 1640 (Hyclone) medium containing 10% FBS (PAN Seratech) and 1% Penicillin/Streptomycin (Gibco). HCT116 and SW480 cells were cultured in DMEM (Hyclone) medium containing 10% FBS (PAN Seratech) and 1% Penicillin/Streptomycin (Gibco). All cell lines were grown at 37℃ in 5% CO2 and 95% humidified air. All cell lines have been authenticated by shore tandem repeat (STR) profiling. The mycoplasma contamination of all the cell lines was tested as negative.

Sample evaluation

To fully quantify the intratumor heterogeneity and determine the genomic evolution of distant metastasis in the six patients with multiple metastases, discrete tumor samples were obtained from the surgical-resected specimen of primary and metastatic tumors by gross evaluation and pathological review. A total of 205 FFPE tissue blocks were collected from the included six patients. Tumor purity and tissue quality were estimated based on histological sections of each sample by a pathologist. Samples with low tumor purity or confluent necrosis were excluded (Fig. S2A). The only exception was patient C06, for whom the tumor purities of all primary tumor samples are relatively low (range from 5% to 27.1%) due to the effect of neoadjuvant chemoradiotherapy. In this case, we included a biopsy sample from colonoscopy at the time of diagnosis. For all included tumor samples, macrodissection was performed to remove normal tissue as much as possible before genomic DNA extraction. In total, 74 patient-matched samples (normal tissue [NOM], primary tumor [PRM], regional lymph node metastasis [RLN], liver metastasis [LIM], lung metastasis [LUM], thoracic lymph node metastasis [TLN]) were obtained for subsequent analysis and sample contributions from each case were summarized in Fig. S2B. Digital slides of included samples were obtained using Pannoramic MIDI (3DHISTECH Ltd) and were reported in Data. S1.

Genomic DNA extraction

After macrodissection, 3-5 slides of 10 μm sections from the FFPE blocks were resuspended in 200 μl deparaffinization solution (Qiagen, 1064343) and incubated at 56°C for 3 min. Genomic DNA was extracted following the GeneRead DNA FFPE Kit (Qiagen, 180134) manufacturer's instruction and quantified through Qubit 2.0 fluorometer (Invitrogen, USA). Samples from the MET/WCH cohorts, including primary tumors, lung and/or liver metastasis tissues, were collected from frozen tissues, and nearly 20 milligram tissue was cut in liquid nitrogen and immediately disrupted and homogenized for genomic DNA isolation according to the Allprep DNA/RNA Mini Kit (Qiagen, 80204) manufacture's instruction. The extracted DNA was quantified through NanoDrop 2000 and agarose gel electrophoresis.

Whole-exome sequencing and raw data analysis

Sequencing libraries were generated using Agilent SureSelect Human All Exon kit (Agilent Technologies, CA, USA) following the manufacturer's recommendations and index codes were added to each sample. The captured pools were combined and sequenced on the Illumina Novaseq 6000 platform with paired-end 150-bp reads. Overall, whole-exome sequencing was performed on all samples, with a mean depth of 87.1× (range from 19.9× to 176.8×) (Table S2). Increased-depth resequencing mainly leads to duplicate results with no obvious increase in mean depth due to the DNA quality of FFPE samples.

Somatic mutation analysis

BAM files were obtained through mapping clean sequencing data to the reference human genome (GRCh37/hg19) by using Burrows-Wheeler Aligner (BWA) with default parameters [1]. Subsequently, GATK4.0 was applied for duplicate marking, local realignment, and base quality recalibration following best-practice protocol. Somatic mutations were obtained with Mutect2 based on the tumor-derived BAM file and the matched normal BAM file (matched call), as well as joint-calling with BAM files for all samples from each single patient. To obtain a high-confidence mutation set, we performed rigorous filtering of the mutations according to criteria listed as follows, the sites should be covered by ≥3 reads; the allelic frequency of mutations should ≥10%; at least 1 read supporting mutation was observed in reads of both directions. The suspected artifacts variants were examined on the Integrated Genomics Viewer manually based on incoherent allelic frequency between regions and variants with poorly aligned reads have been removed. Sequencing artifacts in each individual were predicted using a Bayesian inference model formalized as a Mixed Integer Linear Program (MILP) in Treeomics (see Phylogenetic analysis for details of parameters). These genuine mutations were annotated with ANNOVAR [2]. Finally, an average somatic mutation burden of 134 variants (85-237) was identified (Table S3).

Targeting sequencing

Somatic mutations identified in all samples with WES approach have been strictly filtered, and finally we got a highly conservative and confident list of potential mutations that defined the clonal mutation profile for each tumor sample. Totally 1,899 variants were used for designing a targeted sequencing capture array to increase the sequencing depth. Customer capture probes (3 overlapped probes with length of ~100 nt for each site) were designed to comprise the ~200 nt sequences surrounding these mutations. Target sequencing were performed on all WES samples, except 3 samples from the independent primary tumor lesion of patient C01 (i.e., PRM14-16) and 1 sample of primary tumor in patient C02 due to poor DNA quality. After sequencing on Illumina platform with 150 base paired-end reads and subsequent data analyses (same as described for WES above), we only focused on the 1,899 candidate mutations, and found that 100% of mutations have been captured with the median coverage of 384× (53× to 1121.5×). Reads of these mutations for each sample were used for subsequent evolutionary analyses.

DNA copy number aberration analysis

Somatic copy number alterations (SCNAs) were identified by CNVkit as described previously [3]. SCNAs from all samples of each patient were combined and illustrated (Fig. S5).

Phylogenetic analysis

In the scenario of metastatic seeding inference, we first used Treeomics to reconstruct phylogenetic trees [4]. Variant read data and coverage data of all samples in each case was prepared as input files. Treeomics analysis was performed with the following parameters: sequencing error rate e: 0.005, prior absent probability c0: 0.5, max absent variant allele frequency (VAF): 0.05, loss of heterozygosity (LOH) frequency: 0.0, false discovery rate: 0.05, false-positive rate: 0.005, and absent classification minimum coverage: 100. Tumor purity was estimated based on shared (none-private) variants present in multiple samples in each individual and take their median VAF in a given sample. All the 74 samples passed the purity control for the subsequent analyses. The genetic distance and Jaccard similarity coefficient between all pairs of samples in each patient were calculated based on significantly mutated or well-covered mutations. The detailed description on input data in each case including median coverage and number of variants can be found in Table S2.

To reaffirm the conclusion, we also conducted LICHeE with the default parameters except: maxVAFAbsent: 0.05, minVAFPresent: 0.05, maxClusterDist: 0.3, e: 0.2 to improve the performance of LICHeE in our data set [5]. To construct somatic copy number alteration (SCNA)-based phylogenetic trees, SCNAs obtained from CNVkit were subject to GISTIC2.0 for estimating the gain/loss of somatic copy numbers on gene level across samples [6]. Subsequently, we classified these events as 5 categories (-2: t < -1.3; -1: -1.3 < t < -0.1; 0: -0.1 < t < 0.1; 1: 0.1 < t < 0.9; 2: t > 0.9 where t represents the change of copy numbers). Using a Minimum Event Distance for Intra-tumor Copy-number Comparisons (MEDICC) algorithm [7], we constructed phylogenetic tree based on the alterations of copy number across all samples in each patient.

Driver gene mutation definition

Putative driver mutations were annotated on the basis of The Cancer Genome Atlas (TCGA) consensus [8], with a total of 299 genes. Genes in the COSMIC Cancer Gene Census [9] are further bolded in Fig. 5.

Subclone reconstruction

To explore the clonal composition, we first run MOBSTER with default settings [10] for each tumor sample to call neutral tail mutations based on the VAF data. Mutations clustered as neutral tails in all samples for each patient were considered as neutral mutations and were excluded from the subsequent analysis in order to avoid overfitting. Mutational cancer cell fractions (CCFs) were then estimated using PyClone with default settings [11] by integrating the reference/variant allele read counts of each remained somatic mutations in conjunction with copy-number profiles estimated from CNVkit. All samples from the same individual were analyzed jointly to track the clonal dynamic across samples. The CCFs were also adjusted by tumor purity as estimated in previous steps. Mutations with similar CCFs across all samples in each patient were clustered as a tumor clone. Private clusters with less than four mutations assigned were considered as artifacts of poorly called copy number and were excluded from the subsequent analysis. Output clusters have been filtered retaining those observed with inferred cluster clonality > 0.1 in at least 2 samples to simplify the clonal architecture and avoid overfitting.

Migration pattern and metastatic model inference

To infer the migration patterns among PRM, LIM, and LUM, MACHINA was run in parsimonious migration history with tree inference mode (i.e., pmh_ti) using the inferred clonal composition from PyClone and setting the colon/rectum as the primary anatomical site (Supplementary Fig. S7) [12]. Four researchers reviewed all possible solutions and chose the final solution based on the majority votes. For seeding patterns, polyphyletic seeding was considered if LIM and LUM originated from distinct clones of PRM, while the monophyletic seeding was concluded if LIM and LUM originated from the same clone of PRM. Based on the migration and seeding patterns, a total of three metastatic models were concluded eventually, including the branch-off model (LUM originated from PRM, polyphyletic seeding), the sequential model (LUM originated from LIM, monophyletic seeding), and the diaspora model (LUM originated from PRM, monophyletic seeding) (Fig. 7).

Mutation signature analysis

The mutation signature was analyzed using DeconstructSigs to map the mutation to the COSMIC signatures [13].

Build mutants

Given the RNA-binding functions of the tandem zinc finger domain (TZF) of the protein ZFP36L2, we evaluated the impact of mutant C206Y (C01) on the structure of the TZF domain. The initial three-dimensional geometric coordinates of the crystal structure of ZFP36L2 (PDB code: 1RGO)[14] were downloaded from the Protein Data Bank (RCSB)[15]. The C206Y mutation model was constructed using the Build Mutants protocol of Discovery Studio 3.5 (Accelrys Inc., USA).

Microsatellite instability estimation

Fluorescent polymerase chain reaction (PCR)-based assay was carried out to detect the microsatellite instability (MSI) in the tumor samples in MET and WCH cohort [16]. Briefly, five mononucleotide repeat microsatellite markers (NR-21, NR-24, BAT-25, BAT-26, and NR-27) with different fluorescence label (FAM, HEX, ROX) was independently amplified with Taq polymerase (Vazyme, P103) in 20 µl volumes containing 200 µM each dNTP, 1 x Taq Buffer, 0.4 µM forward primer, 0.4 µM reverse primer, 0.025 U Taq DNA Polymerase, 10 ng DNA per reaction. The PCR conditions consisted of an initial 5 min denaturation step at 94°C, followed by 35 cycles at 94°C for 30 s, 55 °C for 30 s and 72°C for 30 s, with a final extension at 72°C for 7 min. PCR products from each primer were mixed up with equal volume and run on an ABI3730XL DNA Analyzer at TSINGKE Biological Technology, the output data were analyzed with Genemapper® software 4.0 (Applied Biosystems) to determine MSI status of the colorectal tumor samples.

Sanger sequencing

Genomic regions containing the somatic mutations discovered via whole-exome sequencing were amplified by PCR and validated through Sanger sequencing. The PCR primers and corresponding sequencing primers were listed in Table S6. All the PCR reactions were performed using Phanta® Max Super-Fidelity DNA Polymerase (Vazyme, P505-d1) with different amplification conditions. Especially, biallelic mutations of *ZFP36L2* was discovered in patient C03, and we thus conducted TA cloning to validate the precise mutation pattern. In a word, the *ZFP36L2* mutation region was amplified as above and the PCR product was then cloned in pMD19-T vector (TAKARA, 6013) according to the manufacture's instruction after adding 'A' tail with Taq enzyme. Following transformation, colony PCR was conducted by using T5 Super PCR Mix (TSINGKE, TSE005). *ZFP36L2* mutations in expanded samples from MET/WCH cohorts were screened by Sanger sequencing. Two pairs of primers were designed to target the whole coding regions of *ZFP36L2* with two exons. The detail amplification primers and corresponding sequencing primers were presented in KEY RESOURCES TABLE. PCR was proceeded with the same condition described above. And three sequencing primers were used to sequence the whole region of exon2. The amplification products and positive insertion clones were sequenced at TSINGKE Biological Technology with forward primer and analyzed through Chromas software (Version 2.4.1).

Public data analysis

Somatic mutations of *ZFP36L2* for all cancer types were obtained from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) and Hartwig Medical Database (Hartwig, https://database.hartwigmedicalfoundation.nl/)[17]. Only non-silent variants (missense mutations, nonsense mutations, nonstop mutations, frameshift insertion/deletions, inframe insertion/deletions, and splice site mutations) were considered to illustrate the mutation frequencies across different cancer types. The MSI status for colon and rectal cancer (COAD and READ) in TCGA cohort were downloaded from the GDAC (https://gdac.broadinstitute.org) website as stratification in accordance with the process in the MET/WCH cohorts.

CRISPR vectors construction

The sgRNA/Cas9 expression vector lentiCRISPRV2 was obtained from Addgene (Cambridge, MA). The sequences for sgRNAs targeting *ZFP36L2* gene was designed from https://zlab.bio/guide-design-resources as shown below:

Human ZFP36L2 guide RNA targeting sequence1: 5'- CGCCGTTCTCGCTAAACGAG-3'

Human ZFP36L2 guide RNA targeting sequence 2: 5'- ACCCTTAAGGAGCCGTCGGG-3'

Human ZFP36L2 guide RNA targeting sequence 3: 5'- GCCGGCGGTCCGACCTCCTA-3'

Human ZFP36L2 guide RNA targeting sequence 4: 5'- TCGCGCCGGGATTCCTCCGA-3'

The construction of vector was in three steps. First, *ZFP36L2*-specific sgRNA were synthesized from TSINGKE, oligonucleotides of *ZFP36L2* were digested with T4 Polynucleotide Kinase. Second, lentiCRISPRV2 empty vector was incised with BsmBI enzyme and purified by SteadyPure Agarose Gel DNA Purification Kit. Third, annealed oligonucleotides and linearized vector connected with T4 ligase. All the constructed plasmids were confirmed by Sanger sequencing.

Immunohistochemical analysis

Immunohistochemical (IHC) staining assay was performed as described previously [18]. Briefly, FFPE tissue specimens were cut into 5 μm sections, place onto glass slides and then de-paraffinized in xylenes, followed by hydration in ethanol with gradiently decreased concentrations. Endogenous peroxidase was quenched by 3% hydrogen peroxidase treatment. After rinsing with deionized water, antigen retrieval was performed by boiling the sections in citrate buffer for 3 min within pressure cooker. Next, the sections were incubated with 5% fetal bovine serum for 30 min at 37°C, incubated with primary antibody: anti-ZFP36L2 (Abcam, ab70775, 1:50) at 4°C overnight in a moist chamber. After incubation, the tissue sections were rinsed with phosphate buffered saline (2 × 5min) and then incubated with an HRP-conjugated secondary antibody for 1.5 hours at room temperature, the sections were visualized with diaminobenzidine (DAB) kit according to the manufacturer's instructions, and counterstaining with hematoxylin.

Western Blotting

Antibodies used for Western Blot analysis were ZFP36L2 (ab70775) from Abcam and β-actin (3700) from Cell Signaling Technology. Cells were seeded and allowed to reach 70%-80% of confluence. Proteins were extracted using RIPA buffer, supplemented with protease inhibitors and phosphatase inhibitors (Beyotime), and quantified using the BCA assay (Beyotime). The relative molecular mass of the immunoreactive bands was determined using PageRuler Plus Prestained Protein Ladder (Thermo Fisher Scientific). The semiquantitative analysis was performed using β-actin as reference proteins for loading control.

Transwell assay

Cell migration experiments were performed using 24-well cell culture chambers (BD Biosciences) containing a PET membrane with 8 μm pores according to the manufacturer's instructions. CRC cells were seeded at a density of 5 ×10^5 (SW480 and HCT15) cells/well and 3 ×10^5 (HCT116) cells/well in the upper chamber with culture medium (200 μl) alone, while the bottom of the plate was filled with culture medium (800 μl) supplemented with 15% FBS or 10% FBS as a chemoattractant. Cell invasion experiments were performed. First, the membranes were covered with Matrigel (Corning) diluted in RPMI-1640 or DMEM. CRC cells were seeded at a density of 1 ×10^6 cells/well (HCT15 and SW480) and 6 ×10^5 cells/well (HCT116) in the upper chamber with culture medium (200 μl) alone, while the bottom of the plate was filled with culture medium (800μl) supplemented with 20% FBS or 10% FBS as a chemoattractant. After 24h (HCT15 and HCT116) or 48h (SW480), cells that invaded the underside of the membrane were in 4% methanol and stained by 1% crystal violet. Same number of cells were also plated onto plates separately to determine the total number of attached cells to normalize the relative cell migration and invention. For each experiment, the number of cells were counted using an inverted microscope in three random fields (magnification, 100×), and three independent filters were analyzed.

Quantification and statistical analysis

All statistical analysis was performed using R (version 3.5.1). Chi-squared test was used to compare mutation frequencies among validation cohorts. Paired t-test and Wilcoxon rank-sum test were used for the comparison of continuous variables. Pearson correlation test was used to analyze the correlation between the weight of mutation signature and the time elapsed from primary surgery to the diagnosis of metastasis. Overall survival was analyzed using the Kaplan-Meier method with p-value determined by a log-rank test. All hypothesis tests were 2-sided. A *p* value < 0.05 was considered statistically significant.

Contact for reagent and resource sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Heng Xu (xuheng81916@scu.edu.cn). Most data supporting the findings of this study are available within the article and the supplementary files. The raw sequencing data from this study have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number HRA000046 that can be accessed at http://bigd.big.ac.cn/gsa-human and BioProject number PRJCA001380 that can be accessed at https://bigd.big.ac.cn/bioproject/. All other data are available from the corresponding author upon reasonable request.

References

1 Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;**26**:589-95.

2 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;**38**:e164.

3 Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol 2016;**12**:e1004873.

4 Reiter JG, Makohon-Moore AP, Gerold JM, Bozic I, Chatterjee K, Iacobuzio-Donahue CA*, et al.* Reconstructing metastatic seeding patterns of human cancers. Nat Commun 2017;**8**:14114.

5 Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. Genome Biol 2015;**16**:91.

6 Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;**12**:R41.

7 Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol 2014;**10**:e1003535.

8 Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A*, et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018;**173**:371-85 e18.

9 Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature reviews Cancer 2018;**18**:696-705.

10 Caravagna G, Heide T, Williams MJ, Zapata L, Nichol D, Chkhaidze K*, et al.* Subclonal reconstruction of tumors by using machine learning and population genetics. Nat Genet 2020;**52**:898-907.

11 Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J*, et al.* PyClone: statistical inference of clonal population structure in cancer. Nat Methods 2014;**11**:396-8.

12 El-Kebir M, Satas G, Raphael BJ. Inferring parsimonious migration histories for metastatic cancers. Nat Genet 2018;**50**:718-26.

13 Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 2016;**17**:31.

14 Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. Nat Struct Mol Biol 2004;**11**:257-64.

15 Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L*, et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 2019;**47**:D464-d74.

16 Buhard O, Cattaneo F, Wong YF, Yim SF, Friedman E, Flejou JF*, et al.* Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. J Clin Oncol 2006;**24**:241-51.

17 Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C*, et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 2019.

18 Chen HN, Yuan K, Xie N, Wang K, Huang Z, Chen Y*, et al.* PDLIM1 Stabilizes the E-Cadherin/beta-Catenin Complex to Prevent Epithelial-Mesenchymal Transition and Metastatic Potential of Colorectal Cancer Cells. Cancer Res 2016;**76**:1122-34.