nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0423-x

Quantitative evidence for early metastatic seeding in colorectal cancer

Zheng Hu[1,2,3], Jie Ding[1,2,3,12,13], Zhicheng Ma[1,2,3,13], Ruping Sun[1,2,3], Jose A. Seoane[1,2,3], J. Scott Shaffer[3], Carlos J. Suarez[4], Anna S. Berghoff[5,6,7], Chiara Cremolini[8], Alfredo Falcone[8], Fotios Loupakis[9], Peter Birner[5,10], Matthias Preusser[5,6], Heinz-Josef Lenz[11] and Christina Curtis[1,2,3*]

Both the timing and molecular determinants of metastasis are unknown, hindering treatment and prevention efforts. Here we characterize the evolutionary dynamics of this lethal process by analyzing exome-sequencing data from 118 biopsies from 23 patients with colorectal cancer with metastases to the liver or brain. The data show that the genomic divergence between the primary tumor and metastasis is low and that canonical driver genes were acquired early. Analysis within a spatial tumor growth model and statistical inference framework indicates that early disseminated cells commonly (81%, 17 out of 21 evaluable patients) seed metastases while the carcinoma is clinically undetectable (typically, less than 0.01 cm³). We validated the association between early drivers and metastasis in an independent cohort of 2,751 colorectal cancers, demonstrating their utility as biomarkers of metastasis. This conceptual and analytical framework provides quantitative in vivo evidence that systemic spread can occur early in colorectal cancer and illuminates strategies for patient stratification and therapeutic targeting of the canonical drivers of tumorigenesis.

Metastasis is the primary cause of cancer-related death in patients with cancer, but the timing and molecular determinants of this process are largely uncharacterized[1–3]. In particular, when and how metastatic competence is specified are of clinical importance. The prevailing linear progression model posits that metastatic capacity is acquired late following the gradual accumulation of somatic alterations, such that only a subset of cells evolve the capacity to disseminate and seed metastases[4–7]. However, at odds with this model, gene-expression signatures from primary tumors are predictive of distant recurrence, indicating that metastatic cells constitute a dominant subpopulation in the primary tumor[8,9]. In addition, disseminated tumor cells have been identified in patients with early breast lesions[10] and in mouse models of early breast and pancreatic cancers[11–13]. However, the timing of metastatic dissemination has not been evaluated in human cancers due to the challenge in obtaining paired primary tumors and distant metastases and the limitations of applying phylogenetic approaches to bulk tissue samples.

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and leading cause of cancer death[14], as well as a suitable model for studying tumor progression given that the initiating driver alterations are well-characterized[4]. The site and resectability of CRC metastases dictate treatment options and prognosis[15,16]; the liver is the most common site of metastasis, presumably because of venous drainage, and one-third of patients with metastatic CRC (mCRC) have liver-exclusive metastasis[16]. By contrast, brain metastasis is a rare (less than 4% of mCRCs), but devastating diagnosis with limited therapeutic options and a median survival of three to six months[17]. In CRC, metastasis is assumed to be seeded by genetically advanced cancer cells that have evolved through a series of sequential clonal expansions[4,18]. However, CRC progression is not necessarily linear. Rather, we described a Big Bang model of tumor evolution, in which after transformation some CRCs grow as a single expansion populated by heterogeneous and effectively equally fit subclones, and from which most detectable intratumor heterogeneity arises early[19]. These data suggest that some CRCs may be 'born to be bad', wherein invasive and even metastatic potential is specified early[19,20]. Effectively neutral evolution has since been reported in other primary tumors[21–24], but the mode of evolution (effective neutrality versus subclonal selection) has not been evaluated in paired primary tumors and metastases.

Although the metastatic process is largely occult, spatio-temporal patterns of genomic variation are embedded in the evolutionary histories of paired primary tumors and metastases. Here we analyze exome-sequencing data from 118 biopsies from 23 patients with mCRC who had paired distant metastases to the liver or brain to delineate the timing and routes of metastasis and to define metastasis-competent clones (Fig. 1). The data show that primary tumor–metastasis genomic divergence (PMGD) is low and that genomic drivers were acquired early. Moreover, through simulation studies, we establish that low PMGD in bulk-sample sequencing data is indicative of early dissemination, contrary to current assumptions[2]. Phylogeny reconstruction and analysis of the mutational cancer cell fraction (CCF) revealed the early divergence of metastatic lineages and their monoclonal origin. To overcome the limitations of phylogenetic approaches—which cannot resolve the timing of

[1]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [3]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [5]Comprehensive Cancer Center CNS Tumor Unit, Medical University of Vienna, Vienna, Austria. [6]Division of Oncology, Department of Medicine I, Medical University of Vienna, Vienna, Austria. [7]Institute of Neurology, Medical University of Vienna, Vienna, Austria. [8]Department of Oncology, University Hospital of Pisa, Pisa, Italy. [9]Unit of Medical Oncology 1, Department of Clinical and Experimental Oncology, Istituto Oncologico Veneto, IRCCS, Padua, Italy. [10]Department of Pathology, Medical University of Vienna, Vienna, Austria. [11]Department of Medical Oncology, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [12]Present address: Veracyte Inc, South San Francisco, CA, USA. [13]These authors contributed equally: Jie Ding, Zhicheng Ma. *e-mail: cncurtis@stanford.edu

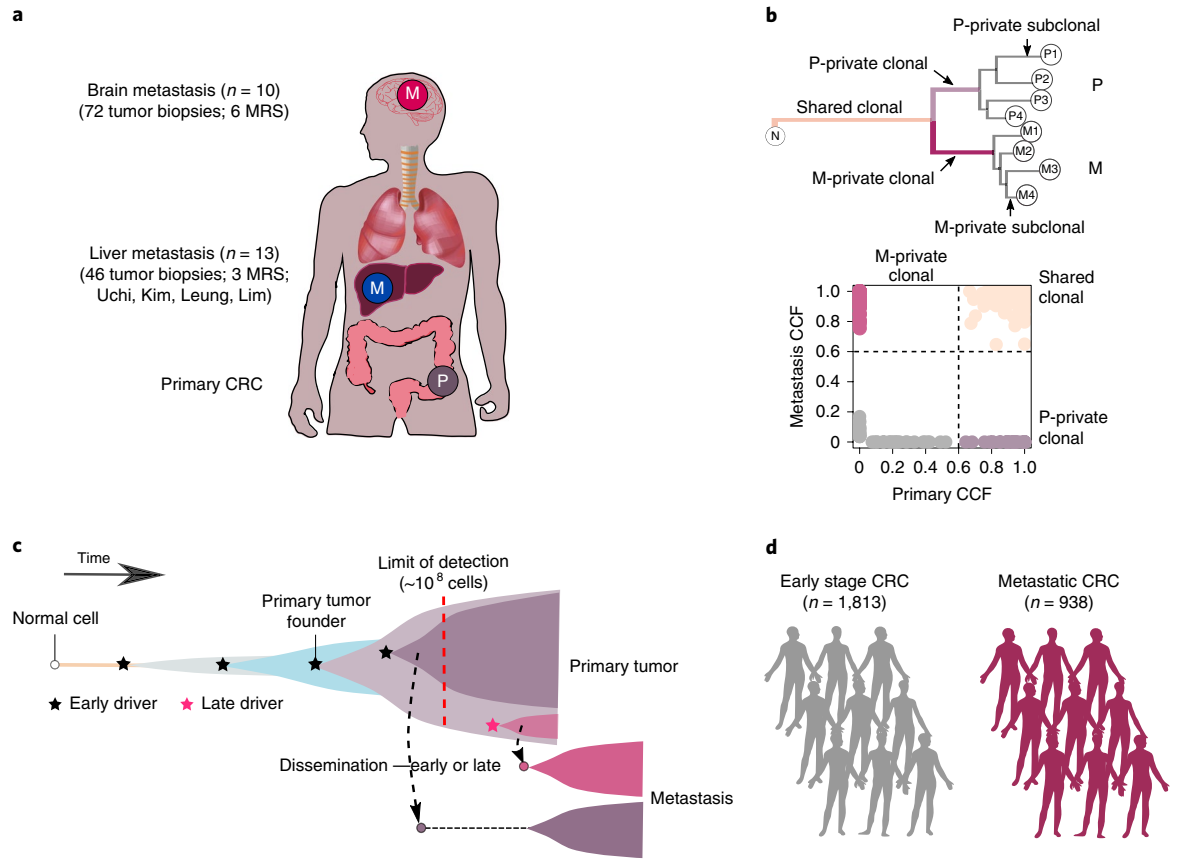

Fig. 1 | Study overview. a, The cohort of patients with mCRC includes 118 tumor biopsies from 23 patients. Paired CRCs with metastases to the brain and other sites (liver, lung or lymph nodes) from 10 patients and 72 tumor biopsies were whole-exome sequenced, including 6 cases with MRS of 3–5 regions, each from the primary CRC and metastases. Additionally, four publicly available cohorts with paired CRCs and liver metastases from 13 patients and 46 tumor biopsies were reanalyzed within the same bioinformatics framework, including 3 cases with MRS. **b**, Tumor phylogenies were reconstructed from somatic alterations (sSNVs and indels). The mutational CCF was computed for each primary CRC and metastasis pair. **c**, Schematic illustration of tumor evolution starting from a normal cell that acquires mutations leading to malignant transformation, growth of the primary tumor and metastatic dissemination, seeding and outgrowth. It is unknown whether dissemination occurs early from a dominant subclone when the size of the primary tumor is below the limits of clinical detection (10^8 cells or 1 cm^3) (early dissemination) or later from a minor subclone after the acquisition of additional driver alterations (late dissemination). To address this question, we developed a three-dimensional model of tumor growth and statistical inference framework to time metastasis from patient genomic data. **d**, We further leveraged a large collection of metastatic ($n = 938$) and non-metastatic ($n = 1{,}813$) CRCs with targeted sequencing data to evaluate the association between specific combinations of early driver genes (modules) identified in the mCRC cohort.

dissemination[2,25–28]—we developed a spatial computational model of tumor progression and Bayesian statistical inference framework to time dissemination in a patient-specific fashion. Furthermore, we validated the association between combinations of early driver genes and metastasis in an independent cohort of 2,751 CRCs, demonstrating their utility as biomarkers of aggressive disease. These results provide quantitative in vivo evidence for early metastatic seeding in mCRC with implications for systemic therapy and earlier detection.

Results

Overview of clinical cohorts. Patients with mCRC exhibit varied progression paths, of which liver-exclusive metastasis and brain metastasis represent extreme scenarios with distinct prognoses[15,16]. We therefore characterized the genomic landscape, routes and timing of metastasis in mCRC by analyzing exome-sequencing data from 118 biopsies from 23 patients with paired distant metastases to the liver or brain (referred to as the mCRC cohort, Fig. 1a, Supplementary Fig. 1, Supplementary Table 1 and Methods). To investigate these patterns, we sequenced 72 samples from a unique cohort of 10 patients with mCRC who had paired brain metastases and some of whom had additional metastases to the liver ($n = 1$), lung ($n = 1$) and lymph nodes ($n = 4$). Five patients had brain-exclusive distant metastasis (V402, V514, V855, V953 and V974), which is estimated to occur in only 2–10% of patients with brain metastasis[16]. For six patients, multi-region sequencing (MRS) of the paired primary tumor and metastasis (P/M pairs) was performed (3–5 regions each), enabling the detailed reconstruction of tumor phylogenies (Fig. 1b). Additionally, we included 46 tumor biopsies from 13 patients with mCRC who had paired liver metastases after excluding cases with low tumor cell purity (<0.4; Supplementary Fig. 2) from four published datasets[21,29–31], analyzed using the same unified bioinformatics framework (Methods). No other sites of metastasis were reported for these patients and MRS was available for 3 P/M pairs ($n = 2$–9 regions each). As we have previously shown, MRS enables more accurate estimation of the cancer cell fraction (CCF) of sSNVs and discrimination between clonal and subclonal mutations relative to single-sample sequencing[23] (Fig. 1b and Supplementary Fig. 3). Additionally, we leveraged an independent collection of 2,751 patients with CRC, including 938 patients with metastatic disease (stage IV) and 1,813 patients with early-stage disease (stages I–III) for whom targeted sequencing data from

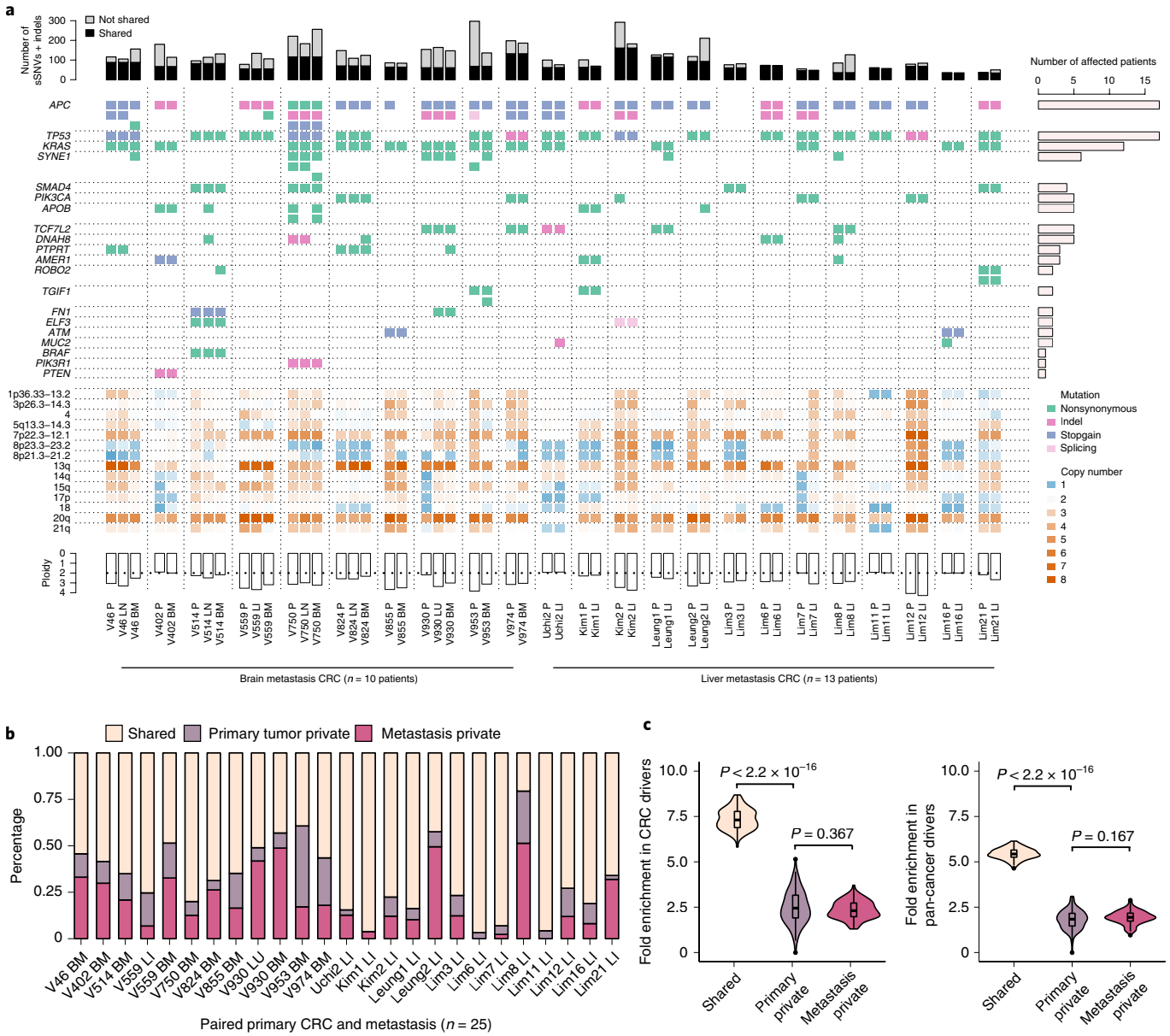

Fig. 2 | The mutational landscape and patterns of genetic divergence in paired primary CRCs and metastases. a, Concordance among somatic alterations (sSNVs, indels and CNAs) in known CRC driver genes between paired primary CRCs and metastases. Stacked bar plots illustrate the total number of sSNVs and indels in exonic regions with a lower cut-off of variant allele frequency = 0.1 in the corresponding site (primary tumor or metastasis). BM, brain metastasis; LI, liver metastasis; LN, lymph node metastasis; LU, lung metastasis. **b**, The percentage of clonal sSNVs/indels that are shared, primary tumor-private or metastasis-private out of all clonal sSNVs/indels with CCF > 60% in any of paired primary tumors and distant metastases. **c**, Violin plots illustrate the probability density of driver gene fold enrichment among shared, primary tumor-private and metastasis-private clonal non-silent sSNVs/indels based on known CRC or pan-cancer drivers. The inset box corresponds to the 25th to 75th percentile (interquartile range); the horizontal line indicates the median; and the vertical line includes data within 1.5× the interquartile range. A test statistic was computed based on $n = 100$ downsamplings among patients (Methods). *P* value, two-sided Wilcoxon rank-sum test. P, primary tumor; LN, lymph node metastasis; LI, liver metastasis; LU, lung metastasis; BM, brain metastasis.

the MSK-Impact[32] and GENIE[33] studies were available in order to evaluate the association between specific combinations of early driver genes (modules) defined in the mCRC cohort (Fig. 1d and Methods).

Genomic heterogeneity in CRCs and paired metastases. High concordance among putative driver genes was observed in the mCRC cohort (Fig. 2a), consistent with previous studies[21,29,34–37]. For instance, mutations in *KRAS*, *TP53*, *SMAD4*, *TCF7L2*, *FN1*, *ELF3* and *ATM* were completely concordant between P/M pairs (Fig. 2a and Supplementary Table 2). On average, 70% of high-frequency somatic single-nucleotide variants (sSNVs) and small insertions and deletions (indels) with CCF > 60% (Methods) in any primary tumor or metastasis were shared by both lesions (Fig. 2b). Among genes that were mutated in more than five patients, *SYNE1* (four out of six patients) and *APOB* (three out of five patients) tended to be private to the primary tumor or metastasis and thus likely arose after transformation. Although metastases usually had more

private high-frequency sSNVs than the primary tumor ($P=0.020$, Wilcoxon rank-sum test; Fig. 2b), they were not enriched for CRC drivers (defined based on IntOGen[38] and The Cancer Genome Atlas (TCGA)[39]) or a published list of pan-cancer drivers[40] (Fig. 2c, Supplementary Table 3 and Methods). Similar results were obtained when stratifying by brain or liver metastases (Supplementary Fig. 4). These data reflect limited driver-gene heterogeneity between P/M pairs and suggest that few additional private genomic drivers were required for metastasis when the primary CRC is already advanced. Somatic copy-number alterations (CNAs) were also generally concordant, with chromosomes 7p22.3–12.1, 13 and 20q11–13 exhibiting recurrent amplification and chromosomes 8p23.3–23.2, 8p21.3–21.2 and 18 exhibiting recurrent deletion in P/M pairs[41] (Fig. 2a and Supplementary Fig. 5). Several putative oncogenes, including *PIK3CA*, *GNAS*, *SRC*, *FXR1*, *MUC4*, *GPC6* and *MECOM* were recurrently (≥4 patients) amplified in metastases relative to paired primary tumors. Notably, *HTR2A* (encoding 5-hydroxytryptamine receptor 2A)—which encodes a receptor for the neurotransmitter serotonin (which also functions as a regulatory factor in the gastrointestinal tract[42])—was amplified more frequently in brain (4 out of 10) than liver (1 out of 13) metastases (Supplementary Fig. 5).

We defined the number of metastasis-private clonal sSNVs as L_m (merged CCF > 60% in the metastasis samples and <1% in the primary tumor samples) and the number of primary tumor-private clonal sSNVs as L_p (merged CCF > 60% in the primary and <1% in the metastasis), where a cut-off of 60% accurately distinguished clonal and subclonal sSNVs (Fig. 1b and Supplementary Figs. 6, 7a,b). Therefore, we used a merged CCF value of 60% as the cut-off to distinguish clonal and subclonal mutations throughout. Brain metastases exhibited higher L_m than liver metastases (median = 24.5 compared to 9.5, respectively, $P=0.01$, Wilcoxon rank-sum test), whereas no difference was found for L_p (median = 8.5 compared to 6.0, respectively, $P=0.70$, Wilcoxon rank-sum test; Supplementary Fig. 7c), potentially reflecting longer progression times (and more cell divisions). Neither L_m ($P=0.68$, Wilcoxon rank-sum test) nor L_p ($P=0.95$, Wilcoxon rank-sum test) differed significantly in chemotherapy-naive versus treated cases despite a slight shift in mutational spectra (A|T>C|G) after chemotherapy (Supplementary Fig. 8).

Gene ontology analysis showed enrichment for cellular adhesion terms among both brain and liver metastasis-private non-silent clonal mutations, but not primary tumor-private clonal or subclonal mutations (Supplementary Table 4). Nervous system development and neuronal differentiation terms were enriched among brain and liver metastasis-private clonal mutations and primary tumor-private mutations, consistent with hijacking of the enteric nervous system in gastrointestinal malignancies[43]. By contrast, primary tumor-private non-silent clonal mutations were enriched for metabolic processes, DNA repair and damage, suggestive of more general deregulation and resource constraints during tumor expansion.

Phylogenetic reconstruction of metastatic CRC. The MRS data revealed extensive intratumor heterogeneity both within tumors and between P/M pairs (Fig. 3a,b, Supplementary Fig. 9 and Supplementary Table 2) and ample mutations for phylogeny reconstruction. We used the F_{ST} statistic[44] to quantify intratumor heterogeneity within tumors (primary tumor or metastasis) in the mCRC cohort based on subclonal sSNVs[23] (Methods). Clonal mutations present in all samples do not contribute to intratumor heterogeneity and were excluded from F_{ST} calculations. Both the primary tumor (median $F_{ST}=0.180$, range = 0.150–0.430) and paired metastases (median $F_{ST}=0.178$, range = 0.123–0.271) exhibited high F_{ST} values, consistent with rapid genetic diversification (Supplementary Fig. 10a). Proliferative indices based on Ki-67 staining were also similar between paired CRCs and metastases ($P=0.765$, Wilcoxon signed-rank test, Supplementary Fig. 10b).

Tumor phylogenies were reconstructed using sSNVs and indels across multiple regions of each P/M pair using the maximum parsimony method[45]. Distant metastases corresponded to monophyletic clades in all but one (Kim1) case (eight out of nine cases with MRS; Fig. 3c, Supplementary Fig. 9 and Methods), consistent with the unique origin of the metastatic lineage. Inspection of the phylogeny for Kim1 indicated that the liver metastasis preceded the primary tumor, which is improbable and likely due to metastasis-specific loss of heterozygosity (LOH) spanning multiple mutations. In most patients, the metastatic lineage diverged before genetic diversification of the primary tumor (V402, V930, V953, V974 and Uchi2; early divergence), whereas divergence occurred during diversification of the primary tumor in patients V750, V824 and Kim2 (late divergence). All brain metastases and most liver metastases contained many private clonal sSNVs, but lacked shared subclonal sSNVs with the primary tumors, consistent with monoclonal seeding (Supplementary Figs. 11, 12), as demonstrated by simulation studies (Supplementary Fig. 13). Two liver metastases (Lim6 and Lim11) exhibited enrichment of shared subclonal mutations, but lacked metastasis-private clonal mutations, consistent with polyclonal seeding (Supplementary Figs. 12, 13). These data suggest that distant metastases are often seeded by a single clone (a single cell or a group of genetically similar cells). Notably, the phylogenetic tree for case V930 indicates that the brain metastasis derived from the lung metastasis, in line with the clinical history of the patient (Fig. 3). Brain metastases and regional lymph node metastases formed separate clades in the two cases in which they were profiled (V750 and V824), indicative of their independent clonal origin from the primary tumor (Fig. 3c and Supplementary Fig. 9) and consistent with polyguanine-repeat analysis[46].

The finding that paired CRCs and metastases formed separate phylogenetic clades in most patients suggests that metastatic dissemination may occur early during cancer development, such that the primary tumor has sufficient time to accumulate many unique clonal mutations after dissemination. However, phylogenetic divergence may occur much earlier than dissemination (Supplementary Fig. 14) and phylogenetics cannot resolve the timing of dissemination[2,25–28]. As such, we next investigated the determinants of PMGD and quantified the timing of metastasis.

The timing of dissemination and PMGD. To model the evolutionary dynamics of metastasis, we developed a three-dimensional agent-based computational model to simulate the spatial growth, progression and lineage relationships of realistically sized patient tumors under varied parameters[19,23] (Fig. 4a, Supplementary Fig. 15, Supplementary Table 5 and Methods). We modeled the growth of a primary CRC starting from a single founder cell and assumed that the metastasis was seeded by a random single cell from the periphery of the primary tumor, yielding primary and metastatic tumors composed of approximately 10^9 cells (around 10 cm^3). To account for distinct modes of tumor evolution, we simulate effective neutrality and stringent subclonal selection[19,23], resulting in four evolutionary scenarios for P/M pairs: neutral/neutral (N/N), neutral/selection (N/S), selection/neutral (S/N) and selection/selection (S/S) (Fig. 4a, Supplementary Figs. 15, 16 and Methods). Using this simulation framework, for which ground-truth values are known, we evaluated the relationship between the number of metastasis-private clonal sSNVs (L_m) and the primary CRC size at the time of dissemination (N_d) in hundreds of virtual paired P/M tumors, for which size is a surrogate measure for time, as cell division rates are unknown (Methods).

To define L_m, we first evaluated metastasis-private clonal sSNVs with relatively high-frequency sSNVs in the whole primary tumor (CCF > 1%). Therefore, any clonal sSNVs in the metastasis will be private to the metastasis if CCF < 1% in the primary tumor. We found that L_m is positively correlated with N_d under all four

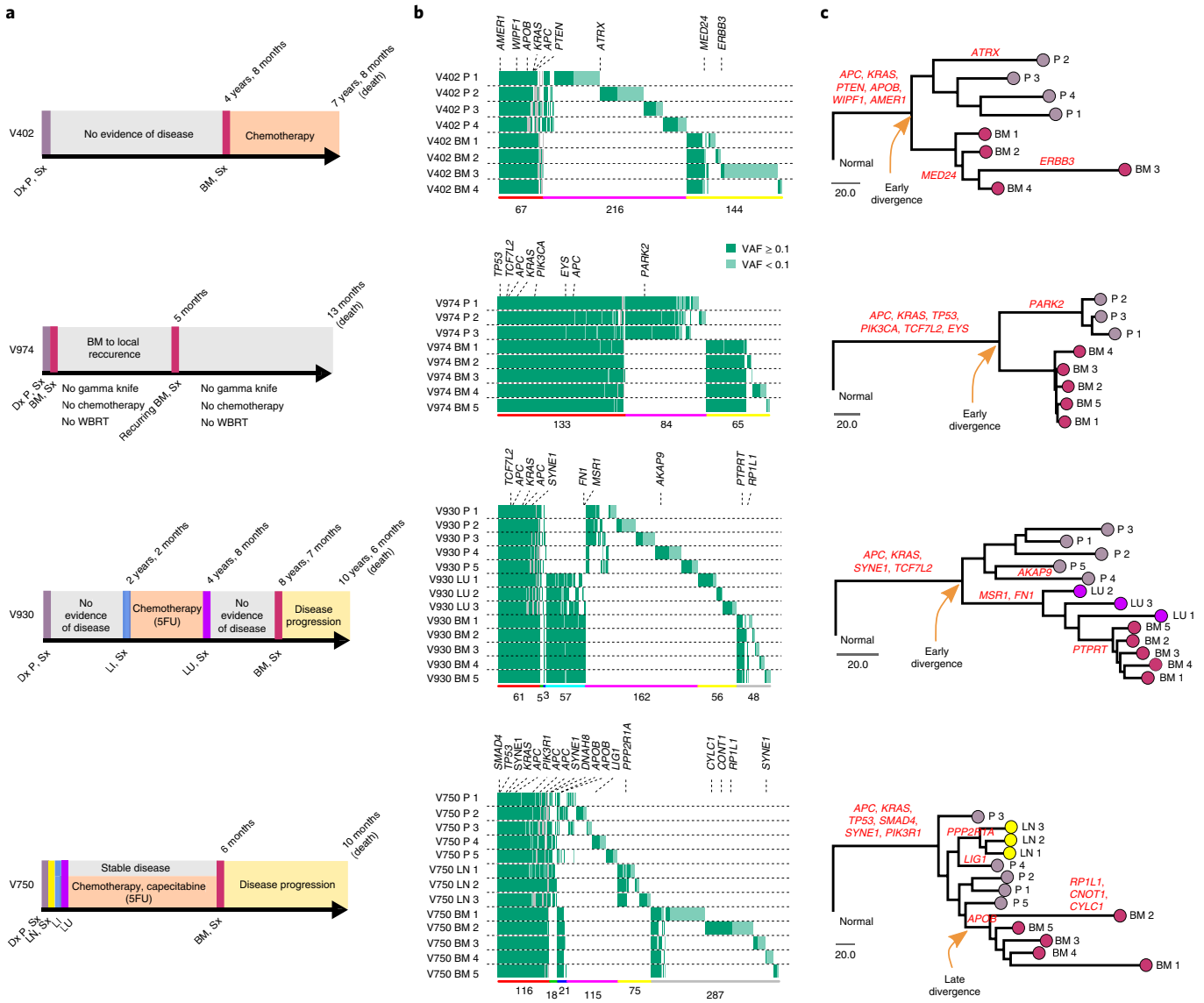

Fig. 3 | Within- and between-lesion heterogeneity in paired primary CRCs and metastases. a, Clinical and treatment history for four representative patients who had CRC with brain metastases. Dx, diagnosis; Sx, surgical resection. **b**, Patterns of within- and between-lesion heterogeneity among sSNVs and indels based on MRS of paired primary CRCs and metastases, for which canonical CRC driver genes are labeled. The number of mutations that are shared or private among different lesions is indicated below as the corresponding colored horizontal bars: ubiquitously P/M shared (red), partially P/M shared (green indicates M1; blue indicates M2), primary tumor-private (pink) or metastasis-private (yellow indicates M1 and gray indicates M2; or cyan indicates M1 and M2). P corresponds to primary tumor. M1 and M2 correspond to different metastatic sites in the same patient when multiple metastatic sites were sampled (V974: M1-LU, M2-BM; V750: M1-LN, M2-BM). **c**, Phylogeny reconstruction using maximum parsimony (PHYLIP) based on mutational presence or absence, for which canonical CRC drivers genes are labeled. VAF, variant allele frequency. WBRT, whole-brain radiation therapy. 5FU, 5-fluorouracil.

evolutionary scenarios (Fig. 4b). The positive relationship between L_m and N_d remains significant when accounting for variation in mutation rate, cell birth and death rate, and selection intensity during tumor growth (Supplementary Fig. 17). We next evaluated L_m by simulating sequencing reads from variable numbers of primary tumor regions ($n = 1$, 10, 50 or 100) while considering the whole metastasis as a bulk sample within our computational model. The positive correlation between L_m and N_d was highly significant under all sampling scenarios, pointing to the robustness of this observation (Supplementary Fig. 18). As expected, smaller L_m was observed when a greater number of primary tumor regions were sequenced because fewer mutations were private to the metastasis (Supplementary Fig. 18). Mathematical analysis of the special case of neutral evolution and exponential growth further demonstrates the positive relationship between L_m and N_d (Supplementary Note, Supplementary equation (6)). These data suggest that later dissemination results in more clonal mutations in the metastasis, many of which are at low frequency in the primary tumor and are often undetectable in bulk sequencing. Accordingly, later dissemination will give rise to more metastasis-private clonal mutations in real sequencing data, leading to higher PMGD. It should be noted that if sampling of the primary tumor was exhaustive or if the metastasis-founder clone could be traced—neither of which are generally practical for studies of tumors in human patients—one would expect very small L_m values and no correlation between L_m and N_d since all mutations in the metastasis-founder cell that accumulated during primary tumor growth would be captured. By contrast, the number of primary tumor-private clonal sSNVs (L_p) exhibited a

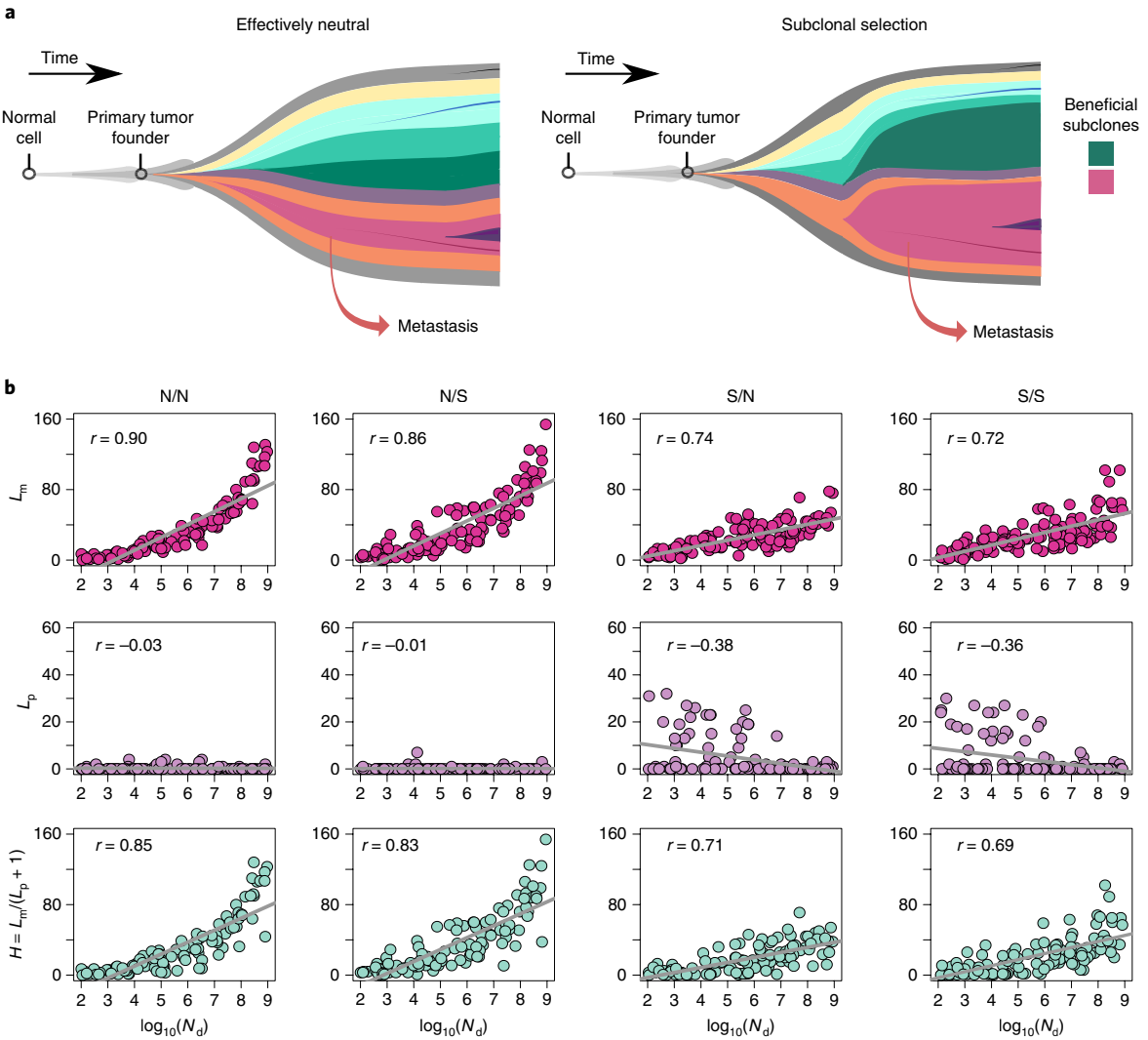

Fig. 4 | Correlation between the L_p, L_m and H and primary carcinoma size at the time of dissemination. a, Schematic illustration of effectively neutral (N) evolution and stringent subclonal selection (S), two distinct evolutionary modes that can occur during the growth of the primary tumor or metastasis. It is assumed that metastatic dissemination occurs during expansion of the primary CRC following malignant transformation, where N_d corresponds to the size of the malignant clone (carcinoma) at the time of dissemination. **b**, The correlation between the timing of dissemination and L_m, L_p or H, based on the spatial simulation of tumor growth ($n = 100$ tumors for each scenario; Pearson's r is reported). L_p and L_m correspond to the number of private clonal sSNVs (CCF > 60% in one site and CCF < 1% in the other site) in the whole primary carcinoma and metastasis, respectively, and $H = L_m/(L_p + 1)$.

slightly negative correlation with N_d when CRCs grew under stringent selection (S/N or S/S), whereas under neutral evolution (N/N or N/S) $L_p \approx 0$, regardless of the timing of dissemination (Fig. 4b and Supplementary Fig. 17).

We defined early dissemination as $N_d < 10^8$ cells (around 1 cm^3 in volume)—the size at which CRCs are generally clinically detectable—and late dissemination as $N_d \geq 10^8$ cells. To establish intuition for the relationship between PMGD and N_d, we defined $H = L_m/(L_p + 1)$. In the simulation studies, H was positively correlated with N_d (Fig. 4b and Supplementary Fig. 17), indicating that larger H values are associated with later dissemination. Indeed, late dissemination typically results in large H (>20) (Fig. 4b). The observation that most patients in the mCRC cohort exhibited small H values (median = 2.4, range = 0.5–23.5) suggests that early dissemination may be relatively common. Although H is strongly associated with the timing of dissemination, it does not capture all components of PMGD, including the mutation rate, as this is cancelled out in the division of L_m over L_p. Additionally, variation in L_p due to differences in the mode of evolution and sampling bias contribute to noise in H. To account for these sources of variability while estimating the timing of dissemination in individual patients, we turned to a powerful statistical inference framework grounded in population genetics theory.

Quantitative evidence for early metastatic seeding in CRC. In order to infer the timing of dissemination N_d, mutation rate u (per cell division in exonic regions) and mode of tumor evolution in P/M pairs, we developed SCIMET (spatial computational inference of metastatic timing), which couples our spatial (three-dimensional) agent-based model of tumor evolution with a statistical inference framework based on approximate Bayesian computation (ABC)[47,48] (Fig. 4a, Supplementary Figs. 15, 16, 19, Supplementary Tables 6, 7 and Methods). The use of ABC is well-established in population genetics and has been utilized to infer the parameters of tumor evolution[19,49]. As the patient genomic data were generally consistent with monoclonal seeding, we assumed that a single cell seeds the metastasis (Lim6 and Lim11 were therefore excluded from this analysis). Evaluation of SCIMET on virtual tumors demonstrates

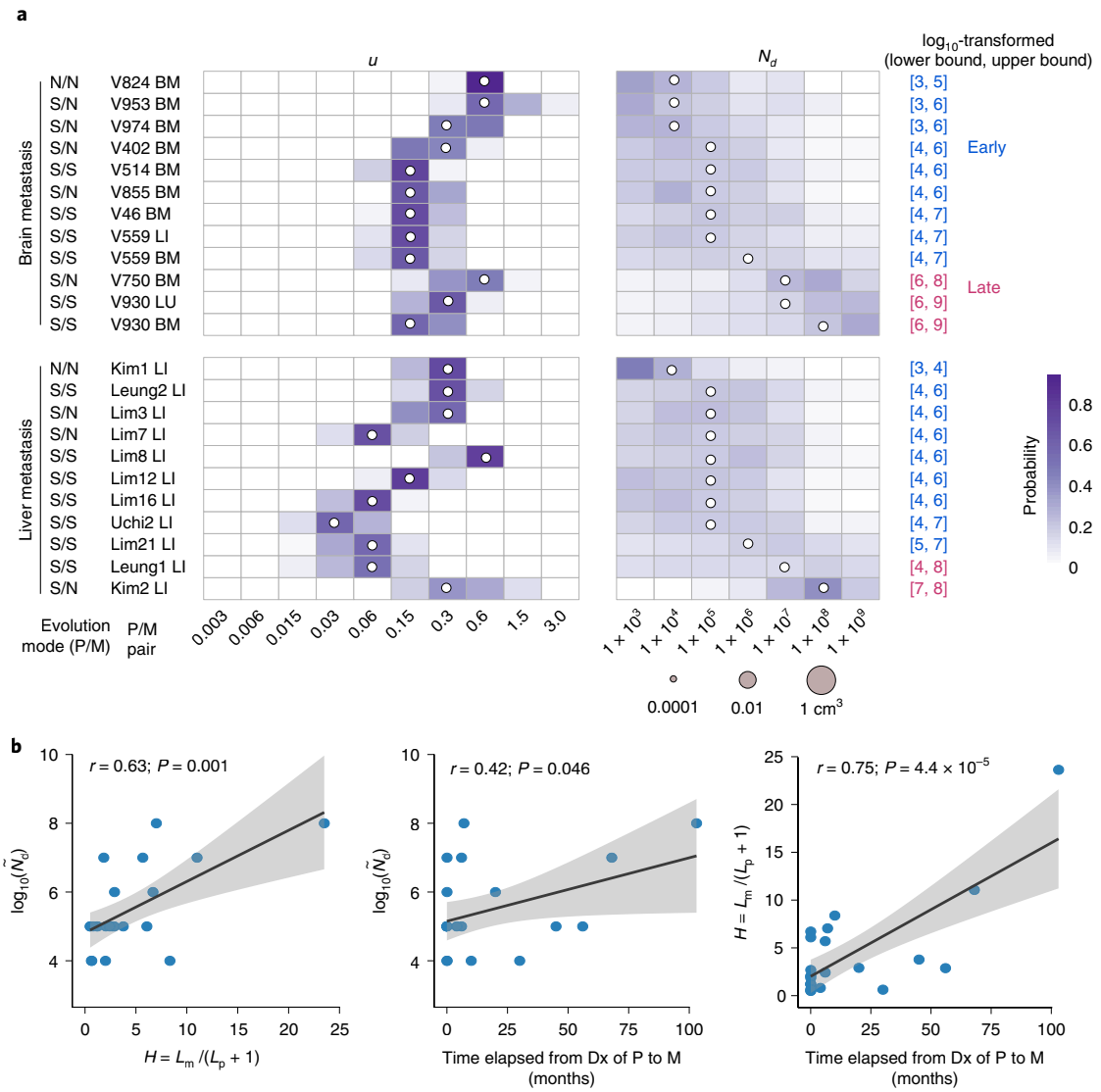

Fig. 5 | Patient-specific inference of the timing of metastasis in CRC. a, Heat map of the posterior probability distributions inferred by SCIMET for the mutation rate u (per cell division in exonic regions) and N_d (timing of metastatic dissemination relative to primary carcinoma size) in individual P/M pairs ($n = 23$) from 21 patients with mCRC. The median of the posterior distribution ($\widetilde{N}_d$) is indicated by a white circle at the corresponding value. For patients with more than one distant metastasis, each metastasis was analyzed independently. The mode of tumor evolution in each P/M pair was determined based on model selection within the statistical inference framework (Methods). We define early dissemination as N_d (upper bound) $< 10^8$ cells (approximately 1 cm^3 in volume) and use the third quartile of the posterior distribution as the upper bound to be conservative. Late dissemination is defined as N_d (upper bound) $\geq 10^8$. P/M pairs for which dissemination and seeding are inferred to have occurred early are denoted in blue, whereas those inferred to have disseminated late are denoted in magenta. **b**, Correlations between $\widetilde{N}_d$ based on SCIMET and the H metric as well as the time elapsed from diagnosis of the primary tumor to diagnosis of the metastasis ($n = 23$). The Pearson's r and P values are reported. Shading corresponds to the 95% confidence interval of the linear regression.

the accurate recovery of the mutation rate and timing of dissemination (Supplementary Fig. 20).

The majority (90%) of CRCs and metastases (57%) exhibited patterns consistent with subclonal selection (Fig. 5a). Inference of patient-specific mutation rates using SCIMET showed an order of magnitude variation across patients (inferred u or $\tilde{u} = 0.06$–0.6, corresponding to 10^{-9}–10^{-8} mutations per base pair per cell division). Notably, in 83% (19 out of 23) P/M pairs from 17 out of 21 patients, dissemination was estimated to occur early when the primary CRC was below the limits of clinical detection (inferred N_d or $\widetilde{N}_d < 10^8$ cells) and typically when the primary tumor was composed of fewer than 10^6 cells using conservative estimates (Fig. 5a and Methods). The $\widetilde{N}_d$ values were also significantly smaller than the tumor size documented at the time of diagnosis in this cohort (Supplementary Table 1). Of note, early dissemination was common irrespective of the site of distant metastasis (8 out of 10 brain and 10 out of 12 liver). Congruent results were obtained when accounting for higher ratios of cell birth and death rates in the primary CRC and metastasis (Supplementary Fig. 21), the collective dissemination of small clusters of cells ($n = 10$ cells; Supplementary Fig. 22) or single-region sampling (Supplementary Fig. 23). Among the four cases for which late dissemination was inferred, three had MRS data, enabling comparison with their phylogenies. For two patients (V750 brain metastasis and Kim2 liver metastasis), late dissemination was consistent with the tumor phylogeny (Fig. 3c and Supplementary Fig. 9). For patient V930, late dissemination was inferred for both the lung and brain metastases, consistent with the large H values (brain, $H = 23.5$; lung, $H = 11$).

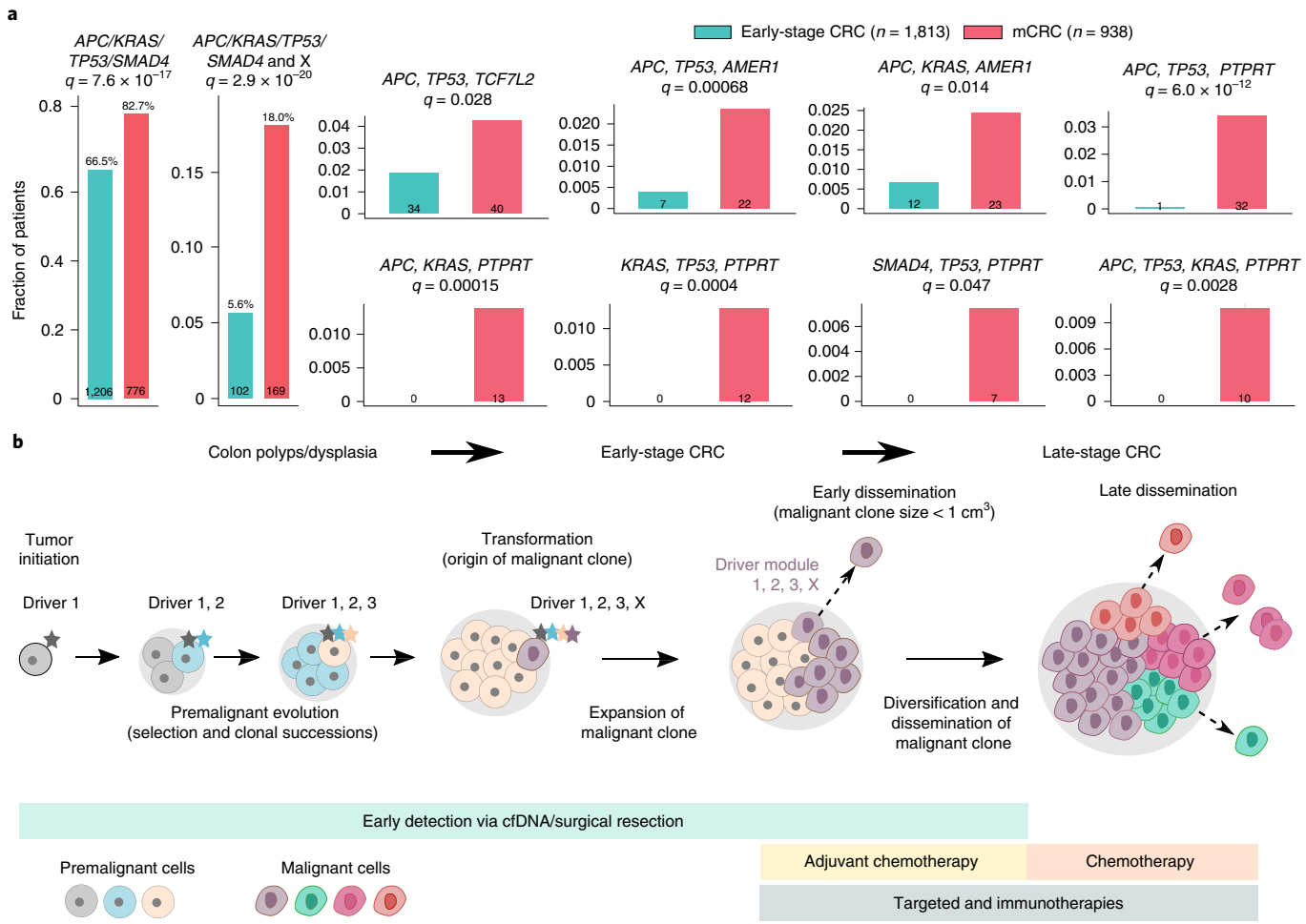

Fig. 6 | Enrichment of early driver gene modules in mCRC and clinical implications of early dissemination. a, The enrichment of canonical core CRC driver genes (*APC*, *KRAS*, *TP53* or *SMAD4*) plus recurrent mutations in candidate drivers (*AMER1*, *ATM*, *BRAF*, *PIK3CA*, *PTPRT* or *TCF7L2*) identified in the mCRC cohort was evaluated in an independent cohort of 2,751 patients with CRC. The combined bar plots (left) illustrate the overall frequency of the core module alone or with an additional candidate driver (X) in early-stage CRCs versus mCRCs. Individual bar plots indicate the frequency of specific modules. *q* values are based on two-sided Fisher's exact tests with Benjamini–Hochberg adjustment. **b**, Three stages of CRC progression are outlined: premalignancy (between initiation and transformation), early-stage CRC (between transformation and dissemination) and late-stage CRC (after dissemination). A set of potential interventions to prevent cancer mortality targets each stage could be implemented: for premalignant lesions, resection (after detection by colonoscopy or possibly cell-free DNA (cfDNA)); for early-stage CRC, surgical resection and possibly adjuvant chemotherapy; and for late-stage CRC, chemotherapy and/or targeted/immune therapies. Given the high rate (80% here) of early dissemination, before clinical detectability of the early-stage CRC, detection and resection of premalignant lesions will have the greatest impact on preventing cancer mortality. For tumors that undergo dissemination before clinical detectability, surgical resection alone, even of a small tumor, cannot prevent metastasis. Once the early-stage tumor is discovered, newly defined metastatic modules (**a**) may inform patient stratification to aid the directed use of adjuvant chemotherapy.

However, the tumor phylogeny indicates early divergence of the metastatic lineage (Fig. 3c). This case illustrates that phylogenetic divergence can occur before dissemination (Supplementary Fig. 14), emphasizing the need for a quantitative evolutionary framework to time metastasis.

The $\widetilde{N}_d$ values based on SCIMET were positively correlated with H (Pearson's $r = 0.63$, $P = 0.001$; Fig. 5b), consistent with the observation that the H metric reflects the timing of dissemination. Additionally, both $\widetilde{N}_d$ and H were positively correlated with the time elapsed between diagnosis of the primary CRC and distant metastasis (Fig. 5b), suggesting that metastases that are diagnosed later likely disseminated later. Furthermore, we estimated the time span between metastatic dissemination and surgical resection of the primary tumor using an approximate analytical function for our spatial tumor growth model and found that dissemination often occurred more than three years before surgery (Supplementary Fig. 24 and Supplementary Note).

Metastasis-associated early driver gene modules. As noted above, most canonical drivers were clonal and shared between paired primaries and metastases (Fig. 2), indicative of their early acquisition before transformation. Taken together with the finding that cancer cells seed metastases early in the majority of mCRCs in this cohort, specific combinations of early driver genes (modules) may confer metastatic competence. In support of this view, oncogene engineering of four canonical early driver genes (*APC*, *KRAS*, *TP53* and *SMAD4*) in wild-type primary colon organoids yielded metastases after xenotransplantation[50]. Similarly, in a mouse model of CRC, oncogenic *Kras* in combination with *Apc* and *Trp53* deficiency was sufficient to drive metastasis[51].

We therefore evaluated the association between the early driver modules defined in the mCRC cohort and metastatic proclivity by analyzing a collection of 2,751 patients with CRC, including 938 patients with metastatic disease (stage IV) and 1,813 patients with early-stage CRC (stages I–III) that were prospectively sequenced as

part of the MSK-Impact[32] and GENIE[33] studies (Methods). Notably, we find that numerous early driver gene modules were significantly enriched in metastatic relative to early-stage CRCs in this independent dataset after correction for multiple-hypothesis testing (Fig. 6a, Supplementary Fig. 25, Supplementary Table 8 and Methods). These modules consist of a backbone of canonical core CRC drivers (combinations of *APC*, *KRAS*, *TP53* or *SMAD4*) with one additional candidate metastasis driver (*TCF7L2*, *AMER1* or *PTPRT*). Collectively, the core modules plus an additional candidate metastasis driver show statistically significant enrichment in metastatic versus early stage CRCs (18% compared to 5.6%, respectively, $q = 2.9 \times 10^{-20}$). Examination of the prevalence and enrichment of individual modules indicates that *PTPRT* mutations in combination with canonical drivers were almost exclusively observed in patients with metastases (Fig. 6a and Supplementary Fig. 25). Thus, *PTPRT* appears to be a highly specific driver of metastasis. *PTPRT* mutations were previously reported in 26% of CRCs[52] and loss of *PTPRT* in CRC and in head and neck squamous cell cancers results in increased STAT3 activation and cellular survival[53,54]. It has therefore been proposed that *PTPRT* mutations may be predictive biomarkers for STAT3 pathway inhibitors, highlighting new therapeutic opportunities[54]. Other modules, which involved *AMER1* and *TCF7L2*, were also significantly enriched in metastatic cases, but were less specific; perhaps because an additional driver defines the module. We therefore identify a compendium of metastasis driver modules that can inform the stratification and therapeutic targeting of patients with aggressive disease.

Discussion

We describe a theoretical and analytical framework that yields quantitative in vivo measurements of the dynamics of metastasis in a patient-specific manner, while accounting for confounding factors, including the metastasis founder event, the mode of tumor evolution, mutation rate variation and tissue sampling bias. By analyzing genomic data from paired primary CRCs and distant metastases to the liver and brain from five patient cohorts within this evolutionary framework, we demonstrate that metastatic seeding often occurs early (17 out of 21 patients), when the carcinoma is clinically undetectable (~10^4–10^8 cells or 0.0001–1 cm^3) and years before diagnosis and surgery (Fig. 5 and Supplementary Figs. 21–24). The observation that early metastatic seeding was prevalent irrespective of the site of distant metastasis, indicates the generalizability of these results. Moreover, dissemination was early even when considering liver-exclusive and brain-exclusive metastases, which represent extremes in terms of their prevalence and prognosis. Collectively, these findings indicate that CRCs can be 'born to be bad', for which invasive and metastatic potential is specified early[19,20,55], illuminating the need to target the canonical drivers of tumorigenesis. However, not all tumors will metastasize and there is an urgent need to identify biomarkers that are associated with aggressive disease.

Towards this end, we validated metastasis-associated driver modules in an independent cohort, thus defining the molecular features of metastasizing clones. The overlap with drivers of initiation and combinatorial structure of these modules may explain why few drivers of metastasis have been identified to date. Although the canonical driver landscape is relatively sparse, there are nonetheless many possible combinations of mutations that collectively disrupt key signaling pathways (WNT, TP53, TGFB, EGFR and cellular adhesion) enabling niche independence and outgrowth at foreign sites[50].

Of note, the vast majority (90%) of primary tumors in the mCRC cohort exhibited subclonal selection consistent with the metastatic clone having a selective growth advantage (Fig. 5a). By contrast, a smaller proportion of early stage (I–III) CRCs (33%) exhibited patterns consistent with subclonal selection[23], suggesting that the mode of tumor evolution may correlate with disease stage or aggressiveness, although larger studies are needed to investigate this trend. Whereas drivers were not enriched in metastases when all cases were considered (Fig. 2c), stratifying by the mode of tumor evolution revealed the enrichment of private high-frequency (CCF > 20%) driver mutations in metastases evolving under stringent selection compared to those evolving neutrally (Supplementary Fig. 26), suggesting that further subclonal driver mutations may occur during the growth of some metastases. Nevertheless, a sizeable proportion (43%) of distant metastases evolved neutrally, potentially reflecting the high fitness of the metastatic clone, consistent with a fitness plateau[56].

The finding that early dissemination—which results in successful metastatic seeding—can occur before the primary tumor is clinically detectable in the majority (80%) of patients with mCRC in this cohort underscores the importance of detecting malignancy at the earliest possible stage (Fig. 6b). Such small tumors fall below the detection limits for current imaging modalities, but advances in profiling circulating cell-free tumor DNA may ultimately enable earlier non-invasive detection[57,58]. Importantly, a considerable number of patients with mCRC did not exhibit early systemic spread, suggesting that colonoscopy can be beneficial in this subgroup. Our data also raise the possibility that patients with early-stage disease with combinations of driver genes that confer a high risk of metastasis may particularly benefit from adjuvant chemotherapy to target micro-metastatic disease[59]. Although the clinical utility of this approach needs to be prospectively evaluated, our findings provide a rationale for patient stratification and therapeutic targeting.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at. https://doi.org/10.1038/s41588-019-0423-x.

Received: 21 December 2018; Accepted: 18 April 2019;
Published online: 17 June 2019

References

1. Vanharanta, S. & Massagué, J. Origins of metastatic traits. *Cancer Cell* **24**, 410–421 (2013).
2. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. *Science* **352**, 169–175 (2016).
3. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Emerging biological principles of metastasis. *Cell* **168**, 670–691 (2017).
4. Jones, S. et al. Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105**, 4283–4288 (2008).
5. Campbell, P. J. et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
6. Yachida, S. et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
7. Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184 (2017).
8. van de Vijver, M. J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med* **347**, 1999–2009 (2002).
9. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).
10. Sänger, N. et al. Disseminated tumor cells in the bone marrow of patients with ductal carcinoma in situ. *Int J. Cancer* **129**, 2522–2526 (2011).
11. Husemann, Y. et al. Systemic spread is an early step in breast cancer. *Cancer Cell* **13**, 58–68 (2008).
12. Rhim, A. D. et al. EMT and dissemination precede pancreatic tumor formation. *Cell* **148**, 349–361 (2012).
13. Hosseini, H. et al. Early dissemination seeds metastasis in breast cancer. *Nature* **540**, 552–558 (2016).
14. Siegel, R. L. et al. Colorectal cancer statistics, 2017. *CA Cancer J. Clin.* **67**, 177–193 (2017).
15. Andres, A. et al. Surgical management of patients with colorectal cancer and simultaneous liver and lung metastases. *Br. J. Surg.* **102**, 691–699 (2015).
16. Vatandoust, S., Price, T. J. & Karapetis, C. S. Colorectal cancer: metastases to a single organ. *World J. Gastroenterol.* **21**, 11767–11776 (2015).
17. Christensen, T. D., Spindler, K. L., Palshof, J. A. & Nielsen, D. L. Systematic review: brain metastases from colorectal cancer—incidence and patient characteristics. *BMC Cancer* **16**, 260 (2016).

18. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
19. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
20. Ryser, M. D., Min, B. H., Siegmund, K. D. & Shibata, D. Spatial mutation patterns as markers of early colorectal tumor cell mobility. *Proc. Natl Acad. Sci. USA* **115**, 5774–5779 (2018).
21. Uchi, R. et al. Integrated multiregional analysis proposing a new model of colorectal cancer evolution. *PLoS Genet.* **12**, e1005778 (2016).
22. Suzuki, Y. et al. Multiregion ultra-deep sequencing reveals early intermixing and variable levels of intratumoral heterogeneity in colorectal cancer. *Mol. Oncol.* **11**, 124–139 (2017).
23. Sun, R. et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**, 1015–1024 (2017).
24. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.* **12**, e1004731 (2016).
25. Hong, W. S., Shpak, M. & Townsend, J. P. Inferring the origin of metastases from cancer phylogenies. *Cancer Res.* **75**, 4021–4025 (2015).
26. Naxerova, K. & Jain, R. K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat. Rev. Clin. Oncol.* **12**, 258–272 (2015).
27. Zhao, Z. M. et al. Early and multiple origins of metastatic lineages within primary tumors. *Proc. Natl Acad. Sci. USA* **113**, 2140–2145 (2016).
28. Schwartz, R. & Schaffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**, 213–229 (2017).
29. Kim, T. M. et al. Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clin. Cancer Res.* **21**, 4461–4472 (2015).
30. Leung, M. L. et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.* **27**, 1287–1299 (2017).
31. Lim, B. et al. Genome-wide mutation profiles of colorectal tumors and associated liver metastases at the exome and transcriptome levels. *Oncotarget* **6**, 22179–22190 (2015).
32. Yaeger, R. et al. Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. *Cancer Cell* **33**, 125–136 (2018).
33. The AACR Project GENIE Consortium AACR Project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* **7**, 818–831 (2017).
34. Lee, S. Y. et al. Comparative genomic analysis of primary and synchronous metastatic colorectal cancers. *PLoS One* **9**, e90459 (2014).
35. Xie, T. et al. Patterns of somatic alterations between matched primary and metastatic colorectal tumors characterized by whole-genome sequencing. *Genomics* **104**, 234–241 (2014).
36. Brannon, A. R. et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol.* **15**, 454 (2014).
37. Tan, I. B. et al. High-depth sequencing of over 750 genes supports linear progression of primary tumors and metastases in most patients with liver-limited metastatic colorectal cancer. *Genome Biol.* **16**, 32 (2015).
38. Gonzalez-Perez, A. et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
39. The Cancer Genome Atlas Network Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
40. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
41. Mamlouk, S. et al. DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nat. Commun.* **8**, 14093 (2017).
42. Yano, J. M. et al. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell* **161**, 264–276 (2015).
43. Hayakawa, Y. et al. Nerve growth factor promotes gastric tumorigenesis through aberrant cholinergic signaling. *Cancer Cell* **31**, 21–34 (2017).
44. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population-structure. *Evolution* **38**, 1358–1370 (1984).
45. Fitch, W. M. Toward defining course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406 (1971).
46. Naxerova, K. et al. Origins of lymphatic and distant metastases in human colorectal cancer. *Science* **357**, 55–60 (2017).
47. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
48. Marjoram, P. & Tavaré, S. Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* **7**, 759–770 (2006).
49. Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C. & Tavare, S. Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. *Cancer Res.* **73**, 41–49 (2013).
50. Fumagalli, A. et al. Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. *Proc. Natl Acad. Sci. USA* **114**, E2357–E2364 (2017).
51. Boutin, A. T. et al. Oncogenic *Kras* drives invasion and maintains metastases in colorectal cancer. *Genes Dev.* **31**, 370–382 (2017).
52. Wang, Z. et al. Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* **304**, 1164–1166 (2004).
53. Zhang, X. et al. Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T. *Proc. Natl Acad. Sci. USA* **104**, 4060–4064 (2007).
54. Lui, V. W. et al. Frequent mutation of receptor protein tyrosine phosphatases provides a mechanism for STAT3 hyperactivation in head and neck cancer. *Proc. Natl Acad. Sci. USA* **111**, 1114–1119 (2014).
55. Turajlic, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx Renal. *Cell* **173**, 581–594 (2018).
56. Rogers, Z. N. et al. Mapping the in vivo fitness landscape of lung adenocarcinoma tumor suppression in mice. *Nat. Genet.* **50**, 483–486 (2018).
57. Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
58. Tie, J. et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci. Transl. Med.* **8**, 346ra92 (2016).
59. Casadaban, L. et al. Adjuvant chemotherapy is associated with improved survival in patients with stage II colon cancer. *Cancer* **122**, 3277–3287 (2016).

Acknowledgements

C. Curtis is supported by the National Institutes of Health through the NIH Director's Pioneer Award (DP1-CA238296) and NCI Cancer Target Discovery and Development Network (CA217851). This work was funded in part by grants from the American Cancer Society (IRG–58-007-54), the Emerson Collective Cancer Research Fund and a gift from the Wunderglo Foundation to C. Curtis. Z.H. is supported by an Innovative Genomics Initiative (IGI) Postdoctoral Fellowship. The project was supported in part by Cancer Center Support Grants from the National Cancer Institute to the Stanford Cancer Institute (P30CA124435) and the University of Southern California Norris Comprehensive Cancer Center (P30CA014089). We thank J. Caswell-Jin and A. Harpak for critical feedback on the manuscript. This study is dedicated to the memory of G. Borges, a tireless cancer warrior.

Author contributions

Z.H. implemented the computational and mathematical models, performed simulation studies and statistical analyses. J.D. implemented the genomic data analysis pipeline, analyzed and visualized genomic data and provided statistical advice. Z.M. processed clinical samples and generated the genomic data. Z.H., R.S. and J.A.S. analyzed the genomic data. Z.H., J.D., R.S., J.A.S. and C. Curtis interpreted the data. J.S.S. contributed to simulation studies. C.J.S., A.S.B. and P.B. performed pathology review. A.S.B. and M.P. performed immunohistochemistry experiments. M.P., P.B., F.L., C. Cremolini, A.F. and H.-J.L. contributed clinical samples and expertise. Z.H. and C. Curtis wrote the manuscript, which was reviewed by all authors. C. Curtis conceived and supervised the study.

Competing interests

A.S.B. has received research support from Daiichi Sankyo and honoraria for lectures, consultation or advisory board participation from Roche Bristol-Myers Squibb, Merck and Daiichi Sankyo as well as travel support from Roche, Amgen and AbbVie. M.P. has received honoraria for lectures, consultation or advisory board participation from the following for-profit companies: Bayer, Bristol-Myers Squibb, Novartis, Gerson Lehrman Group, CMC Contrast, GlaxoSmithKline, Mundipharma, Roche, Astra Zeneca, AbbVie, Lilly, Medahead, Daiichi Sankyo and Merck Sharp & Dome. P.B. has received travel support, honoraria for lectures, consultation or advisory board participation from the following for-profit companies: Biocartis, Novartis, Pfizer, Roche and Roche Diagnostics. C. Curtis is a scientific advisor to GRAIL and reports stock options as well as consulting for GRAIL and Genentech.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0423-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Clinical specimens, pathology review and sequencing studies. In brief, archived formalin-fixed paraffin-embedded (FFPE) tissue specimens from 10 patients with metastatic CRC, including primary tumor, matched metastases and adjacent normal colon tissue, were obtained from the Medical University of Vienna brain metastasis biobank, which was established in accordance with ethical guidelines (approval 078/2004). Tissue specimens were collected during the course of routine clinical care and clinical data were retrieved by retrospective chart review. All samples were de-identified and patients in the brain metastasis cohort were deceased prior to initiating this study. Brain metastases were available for all patients ($n=10$) and for several patients metastases to the liver ($n=1$), lung ($n=1$) and regional lymph nodes ($n=4$) were also available (Supplementary Table 1). For 6 of the 10 patients, multiple specimens ($n=3$–5) from both the primary tumor and metastasis were sampled and sequenced (Supplementary Table 1). Histological sections were independently reviewed by expert pathologists (A.S.B., P.B. and C.J.S.). The Ki-67 proliferative index was determined using immunohistochemical staining using the Ki-67 antibody, as previously described[60]. Consistent with the growth of CRC brain metastases in an expansive rather than infiltrating fashion[61], no normal brain parenchyma was observed within the main brain metastasis lesion.

For all patients, regions of high-cellularity (>60%) were selected for DNA isolation using the QIAamp DNA FFPE Tissue Kit (Qiagen). Libraries were prepared using the Agilent SureSelect Human All Exon kit or Ilumina Nextera Rapid Capture Exome kit for sequencing on the Illumina Hiseq 2000/2500 or Nextseq 500. Paired sequencing reads were aligned to the human reference genome build hg19 with BWA (v.0.7.10)[62]. Duplicate reads were flagged with Picard Tools (v.1.111). Aligned reads were further processed with GATK 3.4.0 for local re-alignment around insertions and deletions and base quality recalibration.

We also analyzed de-identified exome-sequencing data from patients with mCRC in four published datasets[21,29–31] using the same unified bioinformatics framework described below. After excluding tumors with low purity (<0.4), we retained 46 tumor specimens from 13 patients with mCRC who had paired samples of liver metastases and refer to this as the liver metastasis cohort.

Somatic SNV detection and filtering. sSNVs were called by MuTect (v.1.1.7)[63] with paired tumor and normal sequencing data. sSNVs that failed to pass the internal filters of MuTect, had fewer than 10 total reads or 3 variant reads in the tumor sample, fewer than 10 reads in the normal sample or mapped to paralogous genomic regions were removed. Additional Varscan (v.2.3.9)[64] filters were applied to remove sSNVs with low average variant base qualities, low average mapping qualities among variant-supporting reads, strand bias among variant-supporting reads and high average mismatch base quality sums among variant-supporting reads, either within each tumor sample or across all tumor samples from the same patient. Additional filtering removed sSNVs detected in a panel of normals by running MuTect in single-sample mode with less stringent filtering criteria (artefact detection mode). sSNVs called in at least two normal samples were included in the panel of normal sSNV list. For FFPE samples, sSNVs called in samples from one patient were checked against samples from all other patients to flag those that might be artefactual. The maximum observed VAFs across all samples from each patient were calculated based on raw output files from MuTect. sSNVs with maximum observed VAFs between 0.01 and 0.05 in at least two other patients were removed. Small indels were called with Strelka (v.1.0.14) and annotated by Annovar (v.20150617)[65]. sSNVs and small indels in protein-coding regions were retained for downstream analyses. Additional filters were applied to exclude possible artefactual sSNVs due to the processing of FFPE specimens. Specifically, artefacts among C>T/G>A sSNVs with bias in read pair orientation were filtered in each individual FFPE sample, similar to the previously described approach[66].

For patients with MRS data, we sought to exploit this information by retrieving read counts for sSNVs across samples from the same patient. To obtain depth and VAF information across all samples from the same patient, for each sSNV and in each tumor sample that an sSNV was not originally called in, the total reads and variant-supporting reads were counted using the mpileup command in SAMtools (v.1.2)[67]. Only reads with mapping quality ≥40 and base quality at the sSNV locus ≥20 were counted and used to calculate VAF values for that sSNV.

Copy-number analysis, tumor purity and CCF estimation. Copy-number analysis was performed using TitanCNA (v.1.5.7)[68]. In brief, TitanCNA uses depth ratio and B-allele frequency information to estimate allele-specific absolute copy numbers with a hidden Markov model, and estimates tumor purity and clonal frequencies. Only autosomes were used in copy-number analysis. First, for each patient, germline heterozygous SNPs based on dbSNP build 138 were identified using SAMtools and SnpEff (v.3.6) in the normal sample. HMMcopy (v.0.99.0)[69] was used to generate read counts for 1,000-bp bins across the genome for all tumor samples. Whole-exome sequences from multiple normal samples per patient were pooled separately for the purpose of calculating read counts in the bins and the pooled normal read depth data were used as controls only for the calculation of depth ratios. TitanCNA was used to calculate allelic ratios at the germline heterozygous SNP loci in the tumor sample and depth ratios between the tumor sample and the pooled normal data in bins that contained those SNP loci. Only SNP loci within whole-exome sequence-covered regions were then used to estimate allele-specific absolute copy-number profiles. TitanCNA was run with different numbers of clones ($n=1$–3). One run was chosen for each tumor sample based on visual inspection of fitted results, with preference given to the results with a single clone unless results with multiple clones had visibly better fit to the data. Results from tumor samples from the same patient were inspected together to ensure consistency. Overall ploidy and purity for each tumor sample was calculated from the TitanCNA results. For the public datasets including liver-exclusive mCRCs, cases with estimated purity >0.4 in both the primary tumor and paired metastases (Supplementary Fig. 2) were included since low purity hinders accurate SNV/CNA calling.

Mutational CCFs were estimated with CHAT (v.1.0)[70]. CHAT includes a function to estimate the CCF of each sSNV by adjusting its VAF based on local allele-specific copy numbers at the sSNV locus. sSNV frequencies and copy-number profiles estimated from previous steps were used to calculate CCFs for all sSNVs in autosomes (using a modified function). The CCFs were also adjusted for tumor purity. The merged CCF of each sSNV is computed by integrating CCFs from multiple regions when MRS data are available:

$$\mathrm{CCF}=\begin{cases}\dfrac{\sum_{i=1}^{k}\mathrm{CCF}_i\times d_i}{\sum_{i=1}^{k} d_i} & \mathrm{CCF}<1\\ 1 & \mathrm{CCF}\geq 1\end{cases} \quad (1)$$

where d_i and CCF_i are the sequencing depth and CCF estimation in region i, respectively. Of note, the vast majority (99%) of P/M shared sSNVs have CCF (or merged CCF) >60%, a cut-off that also optimally distinguishes the site-private clonal and subclonal sSNV clusters (Supplementary Fig. 6). We thus use 60% as the CCF cut-off to define clonal versus subclonal sSNVs in the PMGD analysis.

Data processing for downstream analyses. For each tumor site (primary or metastasis) in a patient, the average CCF estimate of a sSNV is set to 0 if neither of the following two criteria are met (1) VAF ≥ 0.03 and variant read count ≥3; (2) VAF ≥ 0.1 in any of the regions. The following additional filters were applied to summarize the MRS P/M data in a given patient. First, filter out sSNVs without VAF ≥ 0.05 and variant read count ≥3 or VAF ≥ 0.1 in any samples from this pair of sites. Second, filter out sSNVs with total read depth <20 from either of the two tumor sites. Third, filter out all sSNVs in chromosome regions with LOH in all specimens from one tumor site but not in all samples from the other tumor site. Fourth, for sSNVs not present in any specimens with LOH, filter out sSNVs that satisfy the following criteria in specimens from at least one of the two tumor sites: (1) absent in some samples with LOH; (2) present in any samples without LOH.

Driver enrichment analysis. Driver fold enrichment was determined based on colorectal adenocarcinoma driver genes (defined by combining IntOGen v.2016.5[38] and TCGA[39] including 221 genes, Supplementary Table 3) or all pan-cancer drivers, including 369 high-confidence genes[40] that had non-silent coding sSNVs/indels out of the total number of genes with non-silent coding sSNVs/indels. The resulting metric was normalized to the fraction of driver genes out of all genes in the human genome. Clonal mutations (CCF > 60% in primary tumor or metastasis; merged CCF was used for MRS data) were divided into three sets that represented shared, primary tumor-private and metastasis-private mutations, for which only distant metastases were considered. Driver gene fold enrichment was calculated for each set of mutations by randomly sampling 15 out of 25 P/M pairs from the whole cohort, aggregating them to calculate one driver enrichment score, and repeating this analysis 100 times ($n=100$ downsamplings) to derive a test statistic. For each downsampling, the driver enrichment score was calculated as:

$$\text{Enrichment fold score}=\frac{n(\text{driver non}-\text{silent clonal})/n(\text{all non}-\text{silent clonal})}{n(\text{driver genes})/n(\text{total genes})} \quad (2)$$

where n(all non-silent clonal) and n(driver non-silent clonal) correspond to the total number of non-silent clonal mutations and the number of non-silent clonal mutations in driver genes, respectively. Here n(driver genes) and n(total genes) correspond to the total number of drivers reported for CRC ($n=221$) or pan-cancer ($n=369$) and the number of coding genes in the genome ($n=22{,}000$), respectively.

Orthogonal validation of early metastasis driver gene modules. Clinical annotations and targeted sequencing data were obtained for the GENIE[33] (v.3.0) and MSK-Impact[32] CRC cohorts from Synapse (http://synapse.org/genie) and cBioPortal (http://www.cbioportal.org/study?id=crc_msk_2018), respectively. The MSK-Impact cohort includes early-stage primary CRCs, primary CRCs that are known to have metastasized and the metastatic lesion (predominantly liver) from 1,099 patients with mCRC and a total of 1,134 samples with available sequencing and clinical covariates, including stage, microsatellite status and time to metastasis. As the mCRC discovery cohort did not include any cases with microsatellite

unstable tumors, these were removed, as were cases with *POLE* mutations. Microsatellite stable samples were divided into early-stage non-metastatic samples ($n=57$), metastatic primary tumors ($n=440$) and metastatic samples ($n=498$).

The GENIE cohort is composed of 39,600 samples profiled with different targeted sequencing panels from which CRC samples were selected (oncotree codes: COADREAD, COAD, CAIS, MACR, READ and SRCCR). In order to avoid duplicated samples, all MSK-Impact samples from the GENIE cohort were removed, as were duplicated samples from the same patient, resulting in 2,666 samples, 1,756 of which were from primary tumors. As the GENIE cohort does not currently include stage or outcome information, all primaries are assumed to be non-metastatic, although some may be stage IV or diagnosed as metastatic in the future.

All possible combinations of recurrent putative metastasis driver genes (*APC*, *TP53*, *KRAS*, *SMAD4*, *PIK3R1*, *BRAF*, *AMER1*, *TCF7L2*, *PIK3CA*, *PTPRT* and *ATM*) identified in the mCRC cohort were evaluated in metastatic relative to early-stage cases using a two-sided Fisher's exact test (with Benjamini–Hochberg adjustment for multiple testing). The enrichment analysis was calculated for the combined MSK-Impact and GENIE primary CRC cohort, as well as for the MSK-Impact cohort alone (Supplementary Table 8). As the number of genes in a module increases, the specificity of the association with metastasis increases, whereas the frequency of the module and in turn power to detect an association decreases (Supplementary Fig. 25). Although combining datasets may potentially introduce some biases, because we assume that all GENIE primary samples are non-metastatic and microsatellite stable, this will render our analyses conservative. Indeed, it is worth noting that although these results are already highly significant, they are likely conservative for several reasons: (1) due to the short follow-up time, some cases with early-stage tumors may develop metastases; (2) imbalanced sample size with nearly twice as many patients with early-stage disease versus cases of metastatic disease; (3) several putative metastasis drivers that were identified in the mCRC cohort are not represented on the targeted sequencing panel and hence cannot be evaluated.

Phylogenetic tree reconstruction and F_{ST} computation. We ran PHYLIP[71] (http://www.trex.uqam.ca/index.php?action=phylip&app=dnapars) and applied the maximum parsimony method to reconstruct the phylogeny of multiple specimens from individual patients based on the presence or absence of SNVs and indels. When multiple maximum parsimony trees were reported, we chose the top ranked solution. FigTree (http://tree.bio.ed.ac.uk/software/Figtree/) was used to visualize the reconstructed trees. We computed the F_{ST} statistic for each primary tumor or metastasis using the Weir and Cockerham method[44] based on the adjusted frequency of subclonal sSNVs (merged CCF < 60%) identified in MRS data. Clonal mutations (merged CCF > 60%) did not contribute to intratumor heterogeneity and were excluded in F_{ST} calculations.

Spatial agent-based modeling of tumor progression. We extended our previously described three-dimensional agent-based tumor evolution framework[19,23] to model tumor growth, mutation accumulation and metastatic dissemination after malignant transformation under different evolutionary scenarios in P/M pairs. Pre-malignant clonal expansions before transformation do not alter the genetic heterogeneity within a tumor and were therefore were not modeled (Figs. 1c, 4a and Supplementary Fig. 15). We assume that dissemination occurs after malignant transformation of the founding carcinoma cell as invasion (a cardinal feature of carcinomas) is a requirement for metastasis. We have previously used this framework to model primary tumor evolution[23]. In this model, spatial tumor growth is simulated by the expansion of deme subpopulations (composed of approximately 5,000 cells with diploid genome), mimicking the glandular structures that are often found in CRCs and metastases and consistent with the number of cells found in individual CRC glands (around 2,000–10,000 cells)[72]. Model assumptions are detailed in Supplementary Table 5. The deme subpopulations expand within a defined three-dimensional cubic lattice (Moore neighborhood, 26 neighbors), through peripheral growth while cells within each deme are well-mixed without spatial constraints and grow by a random birth-and-death process (division probability P and death probability $Q=1-P$ at each generation). The notion of peripheral growth is supported by recent studies, which indicated that cancer cells at the periphery of the tumor proliferate much faster than those at the center[73]. Moreover, peripheral growth results in a power law model of net tumor growth (Supplementary Fig. 15b) and is supported by data on CRCs[74]. The first deme is generated via the same birth-and-death process, beginning with a single transformed founding tumor cell. Here we used the following parameters: $P=0.55$ and $Q=0.45$ for the deme expansion in both the primary tumor and metastasis. Thus the cell birth/death probability ratio for the founding lineage is $P/Q=0.55/0.45\approx1.2$. This is supported by the observation that there is no significant difference in proliferation rates based on Ki-67 staining of paired CRCs and brain metastases (Supplementary Fig. 10b), as previously reported in liver metastases[75]. Based on these values of P and Q, approximately 3 years are required from transformation to the diagnosis of primary carcinoma (approximately 10^9 cells; Supplementary Fig. 15b). Once a deme exceeds the maximum size (10,000 cells), it splits into two offspring demes via random sampling of cells from a binomial distribution (N_c, $P=0.5$), where N_c is the current deme size.

During the growth of the primary CRC, a single cell from a random deme at the tumor periphery is randomly chosen to seed the metastasis, which supported by mounting pathological evidence of invasive cells in the tumor front and the fact that blood vessels are also mostly distributed in the invasive front in CRC[76]. The total cell number at the time of metastatic dissemination is denoted by N_d. The metastasis grows via the same model as the primary tumor, starting from the disseminated tumor cell(s).

During each cell division, the number of neutral passenger mutations acquired in the coding portion of the genome follows a Poisson distribution with mean u. Thus, the probability that k mutations occurred in each cell division is as follows:

$$P(x=k)=\frac{u^k e^{-u}}{k!} \quad (3)$$

where an infinite sites model and constant mutation rate are assumed during tumor progression. For simplicity, we do not simulate CNAs, LOH or aneuploidy, and all mutations are heterozygous. Under the neutral model, all somatic mutations are assumed to be neutral passenger events and do not confer a fitness advantage, whereas in the subclonal selection model, beneficial mutations (or advantageous mutations) arise stochastically via a Poisson process with mean u_s during each cell division. We assume $u_s=10^{-5}$ per cell division in the genome[23,77]. We investigated a relatively strong positive selection coefficient ($s=0.1$), where s specifies the increase in cell division probability per cell division when a beneficial mutation occurs in the neutral cell lineage. The cell birth and death probabilities for a selectively beneficial clone are $P_s=P(1+s)$ and $Q_s=1-P_s=1-P(1+s)$, respectively, thus the selective advantage is defined as $s=P_s/P-1$. We selected $s=0.1$, as we have previously shown that the resultant patterns of between-region genetic divergence can be clearly distinguished from those arising under effectively neutral growth[23].

During simulation of primary and metastatic growth, each mutation is assigned a unique index that is recorded with respect to its genealogy and host cells, enabling analysis of the mutational frequency in a sample of tumor cells or the whole tumor during different stages of growth. We simulate growth until the primary and metastasis reach a size of approximately 10^9 cells (or around 10 cm^3) comparable to the size of the clinical samples studied here, which ranged from 4 to 15 cm in maximum diameter. To simulate each of the four scenarios of P/M growth, namely N/N, N/S, S/N or S/S, we used a mutation rate $u=0.3$ per cell division in the exonic region (corresponding to 5×10^{-9} per site per cell division in the 60 Mb diploid coding regions) and selection coefficients $s=0$ and $s=0.1$ when modeling neutral evolution and subclonal selection, respectively, during growth of the primary tumor or metastasis. Under each of the four scenarios of P/M growth, 100 time points (representing the primary tumor size at the time of dissemination, N_d) were sampled at random from a uniform distribution, $\log_{10}(N_d)\sim U(2,9)$, each giving rise to independent P/M pairs. The CCF from the whole tumor in both the primary tumor and metastatic lesions were obtained for each sSNV (site). CCFs > 60% in one site and CCFs < 1% in the other site were used to count the number of primary tumor-private and metastasis-private clonal sSNVs (L_p and L_m, respectively), consistent with the strategy used for patient samples.

SCIMET. We sought to infer two parameters that govern the dynamics of metastasis, namely u, the mutation rate per cell division in the exonic region, and N_d, the primary tumor size at the time of dissemination based on our spatial tumor simulation framework. The two parameters of interest (u and N_d) were randomly sampled from a prior discrete uniform distribution, namely 10 values from 0.003 to 3 for u; and 7 values from 10^3 to 10^9 cells (on $\log_{10}$ scale) for N_d (Supplementary Tables 6, 7 and Supplementary Fig. 19). Discrete prior distributions for u and N_d were used to estimate the order of magnitude rather than the precise values of these two parameters. We simulated 70,000 paired primary tumors and metastases (composed of 10^9 cells each) under each of the four evolutionary scenarios (N/N, N/S, S/N or S/S). After generating the virtual P/M tumors, multiple regions ($n=4$), each composed of approximately 10^6 cells, are sampled from an octant of tumor sphere, as was done for our clinical samples (Supplementary Fig. 19). The VAF of all sSNVs in the sampled bulk subpopulation is considered the true VAF (denoted by f_T), whereas the observed allele frequency is obtained via a statistical model that mimics the random sampling of alleles during sequencing. Specifically, we used a binomial distribution (n,f_T) to generate the observed VAF at each site given its true frequency f_T and number of covered reads n. The number of covered reads at each site is assumed to follow a negative-binomial distribution (negative binomial(size, depth)) where depth is the mean sequencing depth and size corresponds to the variation parameter[78]. We assume depth = 80 and size = 2 for the sequencing data in each tumor region. A mutation is called when the number of variant reads is ≥3, thereby applying the same criteria as for the patient tumors. The observed VAF for each mutation is converted to CCF and the merged CCF from four regions were computed (equation (1)) to mimic the patient genomic data. The nine summary statistics used to fit the CCF data are described in Supplementary Fig. 19 and Supplementary Table 6. The median values of the posterior probability distributions obtained from SCIMET are referred to as the inferred parameter values ($\tilde{u}$ and $\widetilde{N_d}$). To be conservative, we define early dissemination as N_d (upper bound) < 10^8 cells (around 1 cm^3 in volume) using the third quartile of the posterior distribution as the upper bound, whereas late dissemination is defined as N_d (upper

bound) $\geq 10^8$ cells (Fig. 5a). We also evaluated the robustness of SCIMET to a higher birth/death rate ratio (Supplementary Fig. 21), collective dissemination by a cell cluster ($n = 10$ cells; Supplementary Fig. 22) or single-region sequencing data (Supplementary Fig. 23). Of note, both a higher birth/death rate ratio and single-region sequencing data would result in overestimation of the timing of metastatic dissemination. A higher birth/death rate ratio yields a higher tumor growth rate thus the inferred primary tumor size at the time of dissemination would be larger than for a lower birth/death rate ratio. Single-region sampling results in a larger number of metastasis-private clonal mutations (larger L_m and larger H) compared with MRS, thus the timing of dissemination would be overestimated in accordance with the positive correlation between L_m or H and N_d. Overall, these comparisons demonstrate the robustness of SCIMET to different model assumptions.

We used a version of ABC based on the acceptance–rejection algorithm[79] to estimate posterior probability distributions for the parameters of interest $\boldsymbol{\theta}(u, N_d)$. The ABC version of rejection sampling is as follows:

For $i = 1$ to K under model $\mathbf{M}$ (N/N, N/S, S/N or S/S):

1. Sample parameters $\boldsymbol{\theta}'$ from the prior distribution $\pi(\boldsymbol{\theta})$.
2. Simulate data $\mathbf{D}'$ using model $\mathbf{M}$ with the sampled parameters $\boldsymbol{\theta}'$ and summarize $\mathbf{D}'$ as summary statistics $\mathbf{S}'$.
3. Accept $\boldsymbol{\theta}'$ if $d(\mathbf{S}', \mathbf{S}) < \varepsilon$, for a given tolerance rate ε, where $d(\mathbf{S}', \mathbf{S})$ is a measure of Euclidean distance between $\mathbf{S}'$ and $\mathbf{S}$.
4. Go to (1).

Using this scheme, we are able to approximate the posterior distribution by: $P(\boldsymbol{\theta}|d(\mathbf{S}', \mathbf{S}) < \varepsilon)$. We use a common variation of ABC[47,80] in which, rather than using a fixed threshold ε, we sort all K distances calculated by $d(\mathbf{S}', \mathbf{S})$ (step (3)), and accept the $\boldsymbol{\theta}'$ that generated the smallest $100 \times \eta$ percentage distances. We use $\eta = 0.01$ so that the posterior is composed of $70{,}000 \times 0.01 = 700$ data points. The ABC procedure is performed using the R package abc[81]. To determine the most probable model of tumor evolution (N/N, N/S, S/N or S/S) in P/M pairs, we run the postpr method implemented in the R package abc, which integrates all simulation data from the four models to run the ABC procedures (steps (1)–(4)) and outputs the probability of each model based on the posterior distribution. The model (N/N, N/S, S/N or S/S) with the highest probability was selected.

A Monte Carlo cross-validation scheme was performed to assess the performance of SCIMET. This procedure involves randomly sampling a combination of parameters u' and N_d' (true parameters) and sampling 10 simulations of the summary statistics $\mathbf{S}'$ under this parameter set to independently run the ABC scheme. The posterior parameters u and N_d with the maximum probability were used as parameter estimates for one simulation, namely $\tilde{u}$ and $\widetilde{N_d}$. The mean value of $\tilde{u}$ and $\widetilde{N_d}$ in 10 simulations was taken as the parameter estimate (inferred parameters) in the cross-validation. The process of Monte Carlo sampling and SCIMET inference was repeated 200 times under each of the four evolutionary scenarios (N/N, N/S, S/N and S/S). Comparison of the inferred versus true parameter values indicates the robustness of this approach (Supplementary Fig. 20).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data have been deposited at the European Genotype Phenotype Archive (EGA) under accession number EGAS00001003573. Data from previously published studies are available from the DDBJ (accession number JGAS00000000060)[21] and the SRA (accession numbers SRP052609, SRP074289 and SRP041725)[29–31].

Code availability

Code used for genomic data analysis and simulation studies are available at https://github.com/cancersysbio/mCRCs and https://github.com/cancersysbio/SCIMET.

References

60. Berghoff, A. S. et al. Differential role of angiogenesis and tumour cell proliferation in brain metastases according to primary tumour type: analysis of 639 cases. *Neuropathol. Appl. Neurobiol.* **41**, e41–e55 (2015).
61. Berghoff, A. S. et al. Invasion patterns in brain metastases of solid cancers. *Neuro-oncol.* **15**, 1664–1672 (2013).
62. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
63. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
64. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
65. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
66. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
67. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
68. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
69. Ha, G. et al. Integrative analysis of genome-wide loss of heterozygosity and mono-allelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**, 1995–2007 (2012).
70. Li, B. & Li, J. Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.* **15**, 473 (2014).
71. Felsenstein, J. Phylogeny inference package. *Cladistics* **5**, 164–166 (1989).
72. Siegmund, K. D., Marjoram, P., Woo, Y. J., Tavaré, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA* **106**, 4828–4833 (2009).
73. Lloyd, M. C. et al. Darwinian dynamics of intratumoral heterogeneity: not solely random mutations but also variable environmental selection forces. *Cancer Res.* **76**, 3136–3144 (2016).
74. Sarapata, E. A. & de Pillis, L. G. A comparison and catalog of intrinsic tumor growth models. *Bull. Math. Biol.* **76**, 2010–2024 (2014).
75. Finlay, I. G., Meek, D., Brunton, F. & McArdle, C. S. Growth rate of hepatic metastases in colorectal carcinoma. *Br. J. Surg.* **75**, 641–644 (1988).
76. Kather, J. N. et al. Identification of a characteristic vascular belt zone in human colorectal cancer. *PLoS One* **12**, e0171378 (2017).
77. Bozic, I. et al. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107**, 18545–18550 (2010).
78. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
79. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
80. Zhao, J., Siegmund, K. D., Shibata, D. & Marjoram, P. Ancestral inference in tumors: how much can we know? *J. Theor. Biol.* **359**, 136–145 (2014).
81. Csilléry, K., François, O. & Blum, M. G. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).

nature research

Corresponding author(s): Christina Curtis

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	BWA-MEM (0.7.10), Picard (1.111), GATK (3.4.0), MuTect (1.1.7), Varscan (2.3.9), Strelka (1.0.14), Annovar (20150617), DToxoG (1.14.4.0), TitanCNA (1.5.7), SAMtools (1.2), SnpEff (3.6), HMMcopy (0.99.0), CHAT (1.0), GENIE (3.0), cBioPortal (http://www.cbioportal.org/study?id=crc_msk_2018), IntOGen (2016.5)
Data analysis	R version 3.5.0, PHYLIP (http://www.trex.uqam.ca/index.php?action=phylip&app=dnapars), FigureTree (http://tree.bio.ed.ac.uk/software/Figuretree/), Custom software: mCRCs (https://github.com/cancersysbio/mCRCs), SCIMET (https://github.com/cancersysbio/SCIMET)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data have been deposited at the European Genotype Phenotype Archive (EGA) under accession number EGAS00001003573. Data from previously published studies are available at: DDBJ: JGAS00000000060 (Uchi et al.), and the SRA: SRP052609 (Kim et al.), SRP074289 (Leung et al.), SRP041725 (Lim et al.).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical methods were not used to predetermine sample size - rather all available samples that were suitable for inclusion were utilized. Exome sequencing data from 118 biopsies from 23 metastatic colorectal cancer (mCRC) patients with paired liver and brain metastases were analyzed. Additionally, targeted sequencing data from an independent cohort of 2,751 colorectal cancer patients was utilized to corroborate these findings.
Data exclusions	Samples with low tumor purity (<0.4) from publicly available metastatic colorectal cancer datasets were excluded since this adversely affects mutation detection, as noted in the Methods (page 12) and Figure S2.
Replication	Replication is not applicable for patient genomic data. For the simulation studies, the number of replicates are reported in the text. In particular, statistical inference of patient-specific parameters using SCIMET requires a large number of simulations, we performed 1000 simulations for each of the parameter combinations (n=70000 in total for each model).
Randomization	This was an observational study, no randomization was performed.
Blinding	Blinding was not considered appropriate for this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging