Supplementary Information

Quantitative evidence for early metastatic seeding in colorectal cancer

Zheng Hu[1,2,3], Jie Ding [1,2,3§], Zhicheng Ma[1,2,3§], 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], Christina Curtis[1,2,3*]

Affiliations

[1] Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford, California, USA
[2] Department of Genetics, Stanford University School of Medicine, Stanford, California, USA
[3] Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA
[4] Department of Pathology, Stanford University School of Medicine, Stanford, California, 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] Azienda Ospedaliero-Universitaria Pisana and University 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, California, USA
[§] These authors contributed equally to this work

* Corresponding author: Christina Curtis, Stanford University School of Medicine, 265 Campus Drive, Lorry Lokey Building Suite G2120C, Stanford, CA 94305 Tel: 650-498-9943, Email: cncurtis@stanford.edu

Supplementary Figures (p. 2-50)

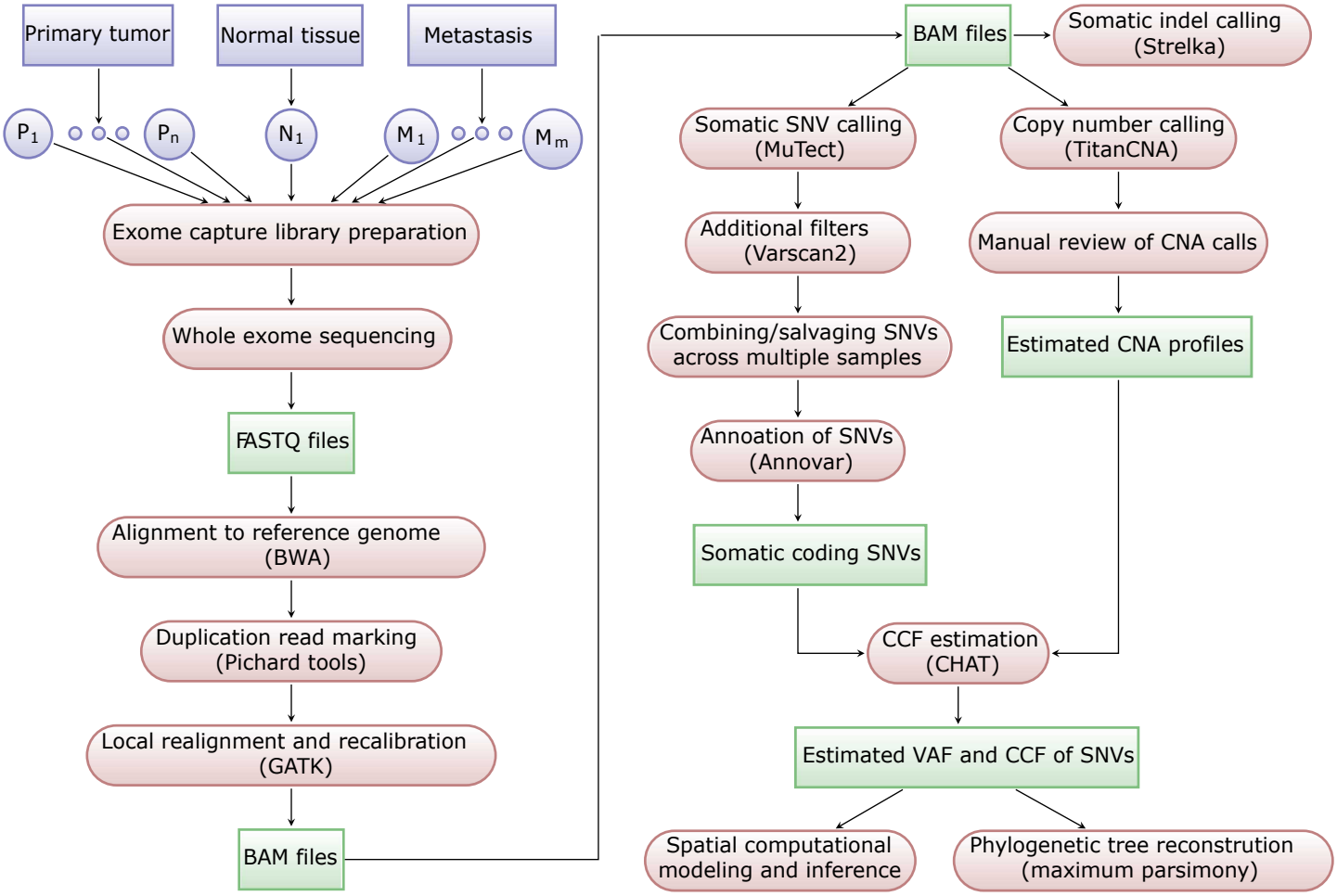
Supplementary Figures 1-26

Supplementary Tables (p.51-54)

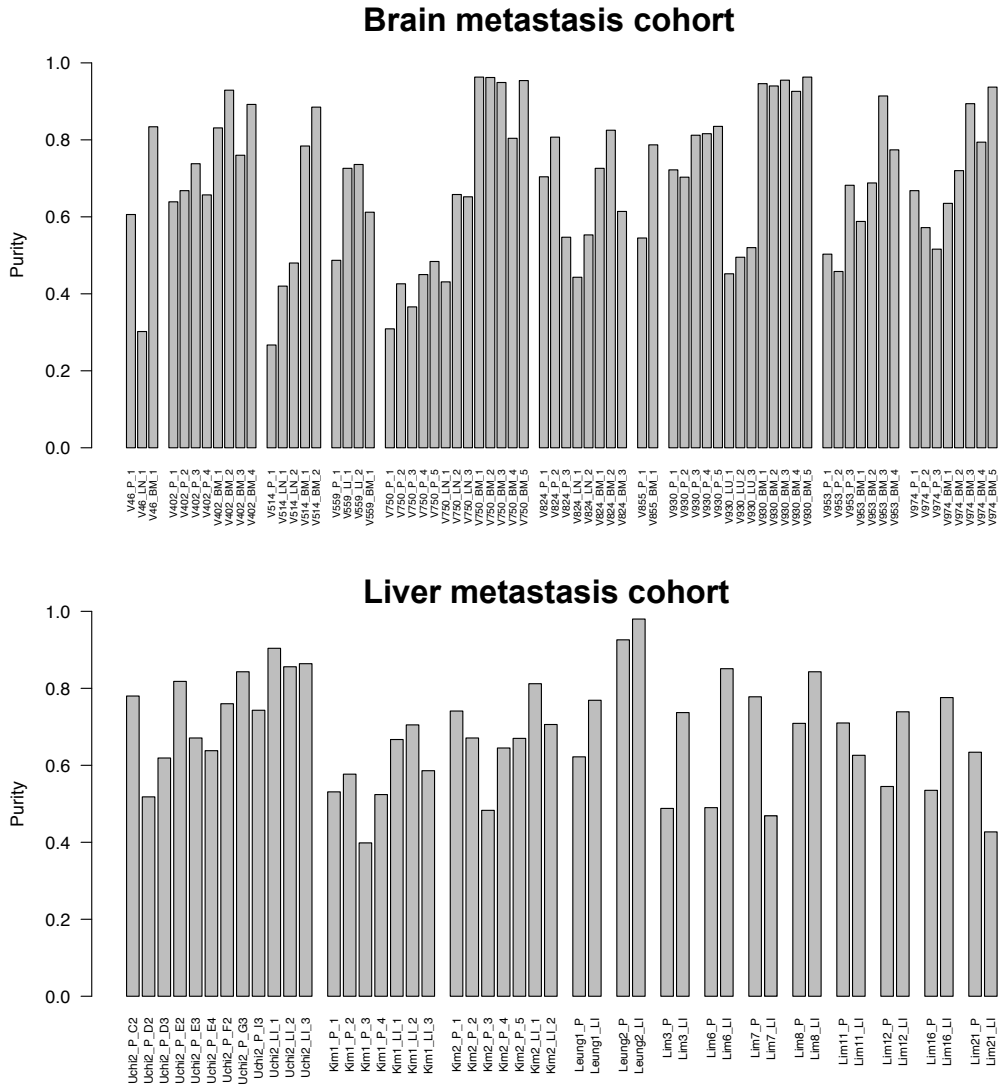
Supplementary Table 1. Clinical features of the metastatic colorectal cancer cohort
Supplementary Table 5. Spatial computational tumor model (SCIMET) parameters
Supplementary Table 6. Description of summary statistics for SCIMET
Supplementary Table 7. SCIMET Summary statistics for the metastatic colorectal cancer cohort

Supplementary Note (p.55-58)

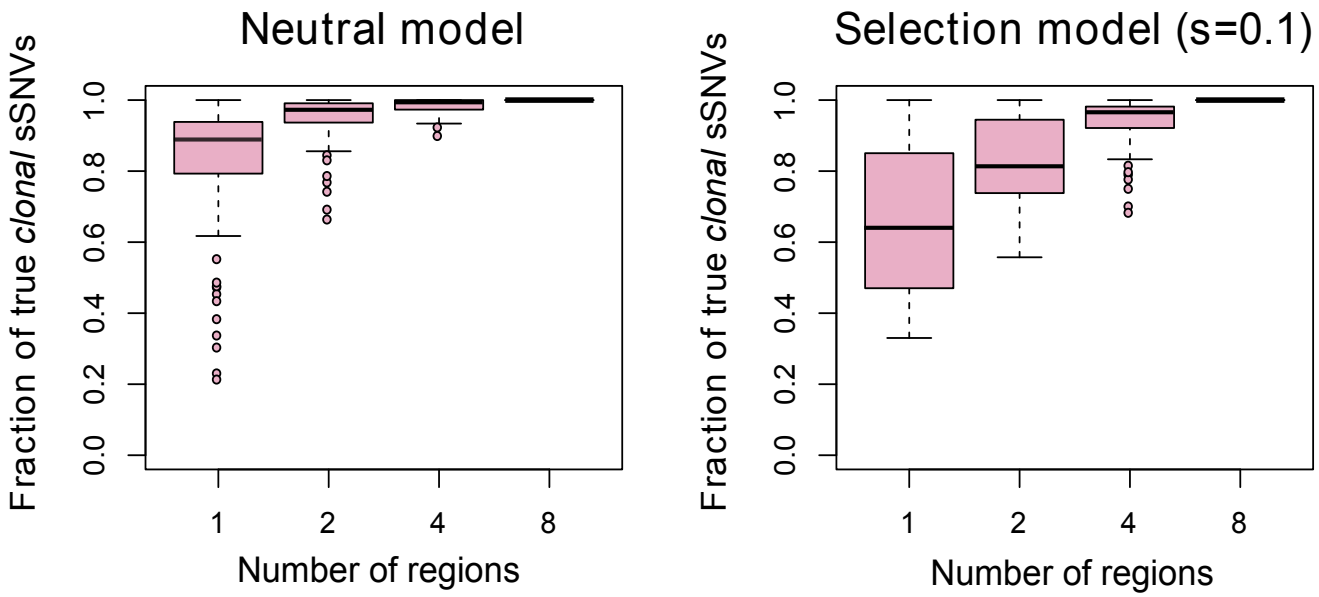
Supplementary Figures



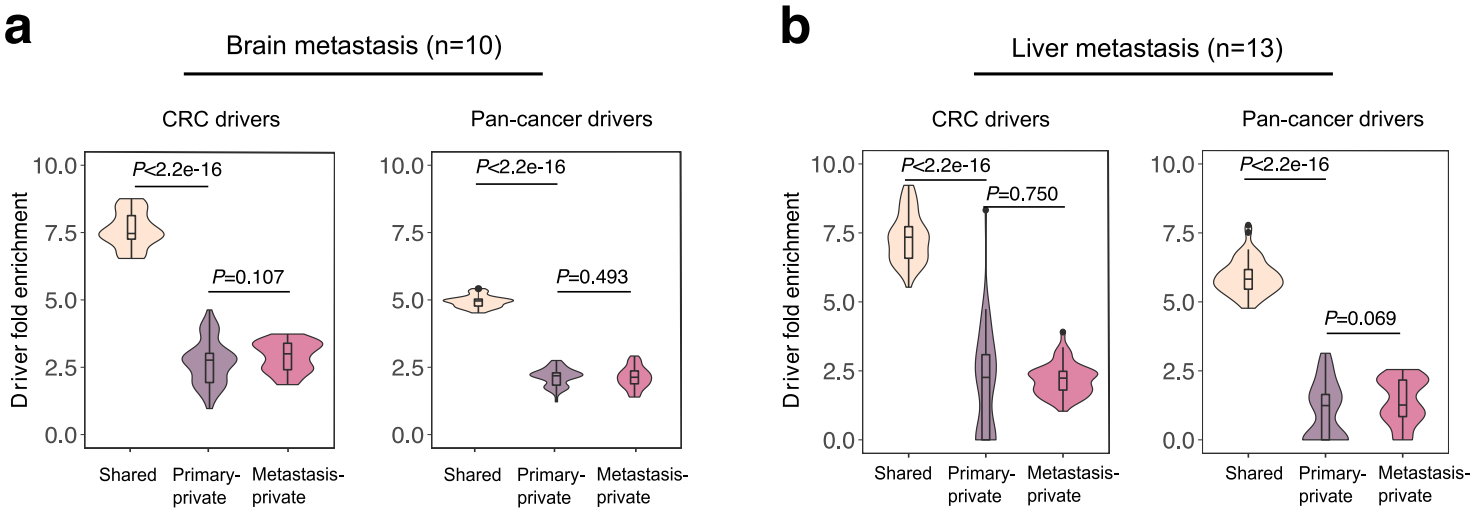
Supplementary Figure 1. Workflow summarizing the data input and analytical approach. Overview of analysis workflow based on whole-exome (multi-region) sequencing of paired primary colorectal cancers (CRCs), metastases, and normal controls. Blue modules represent the samples and data inputs. Red modules represent experiments, data analysis and computational modeling, where tools/methods are indicated by parentheses. Green modules represent the data output or results.



Supplementary Figure 2. Tumor cell purity estimates by TitanCNA in each tumor sample in the brain metastasis and liver metastasis cohorts. For the publicly available datasets (liver metastasis cohort), tumors with low tumor cell purity (<0.4) were excluded. The brain metastasis cohort consists of 72 tumor samples from 10 patients while the liver metastasis cohort consists of 46 tumor samples from 13 patients.

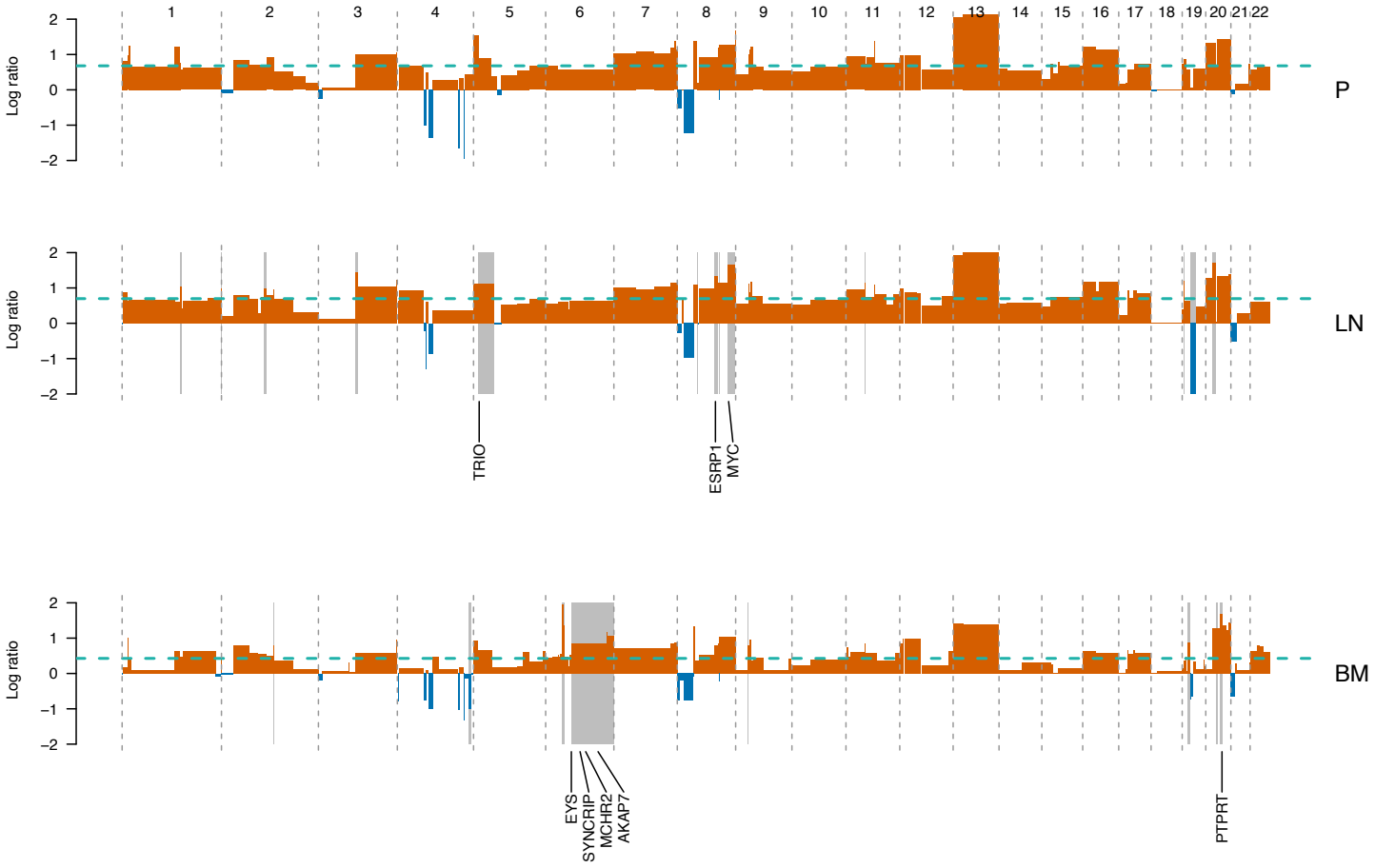


Supplementary Figure 3. Multi-region sequencing (MRS) is required to confidently identify *clonal* mutations. We employed our previously described spatial-agent based model of tumor growth [1] to evaluate the utility of sequencing increasing numbers (n=1, 2, 4, 8) of samples to correctly distinguish *clonal* and *subclonal* mutations under distinct evolutionary modes (neutral, s=0 and stringent selection, s=0.1). For each 'virtual' tumor, 8 regions were sampled and sequenced *in silico* (mean depth=80X) and clonal sSNVs were called based on a merged CCF cutoff of 60%. Similarly, 1, 2 or 4 regions were down sampled from the 8 regions to identify "clonal" sSNVs by this subset of regions. The fraction of true clonal sSNVs amongst all "clonal" sSNVs (merged CCF>60%) in the corresponding sampling scenario was evaluated. n=100 tumors for each model and number of regions/samples. Bar, median; box, 25th to 75th percentile (interquartile range, IQR); whiskers, data within 1.5 times the IQR.



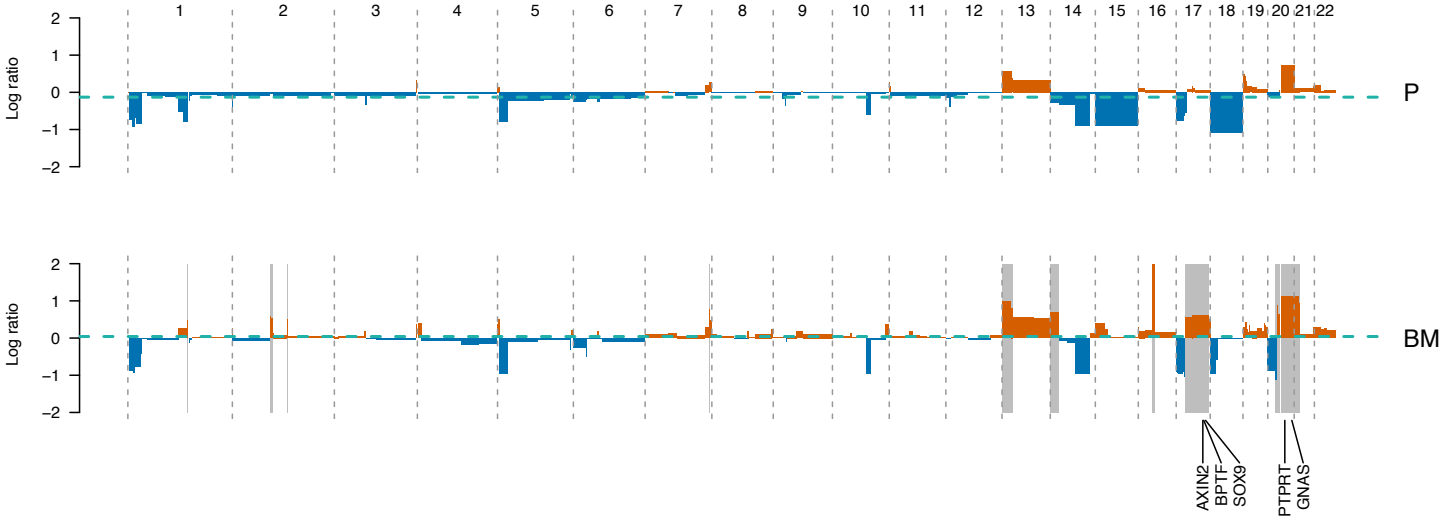
Supplementary Figure 4. Driver enrichment analysis. The enrichment of CRC or pan-cancer 'driver' genes amongst shared, primary-private, and metastasis-private *clonal* non-silent sSNVs and indels was evaluated separately in the (**a**) brain metastasis or (**b**) liver metastasis cohorts. A CCF value of 60% (or merged CCF=60% for tumors with multi-region sequencing data) was used to distinguish clonal and subclonal sSNVs. n=50 down-samplings of patients. *P*-value, Wilcoxon Rank-Sum Test (two-sided). Bar, median; box, 25th to 75th percentile (interquartile range, IQR); vertical line, data within 1.5 times the IQR.

V46

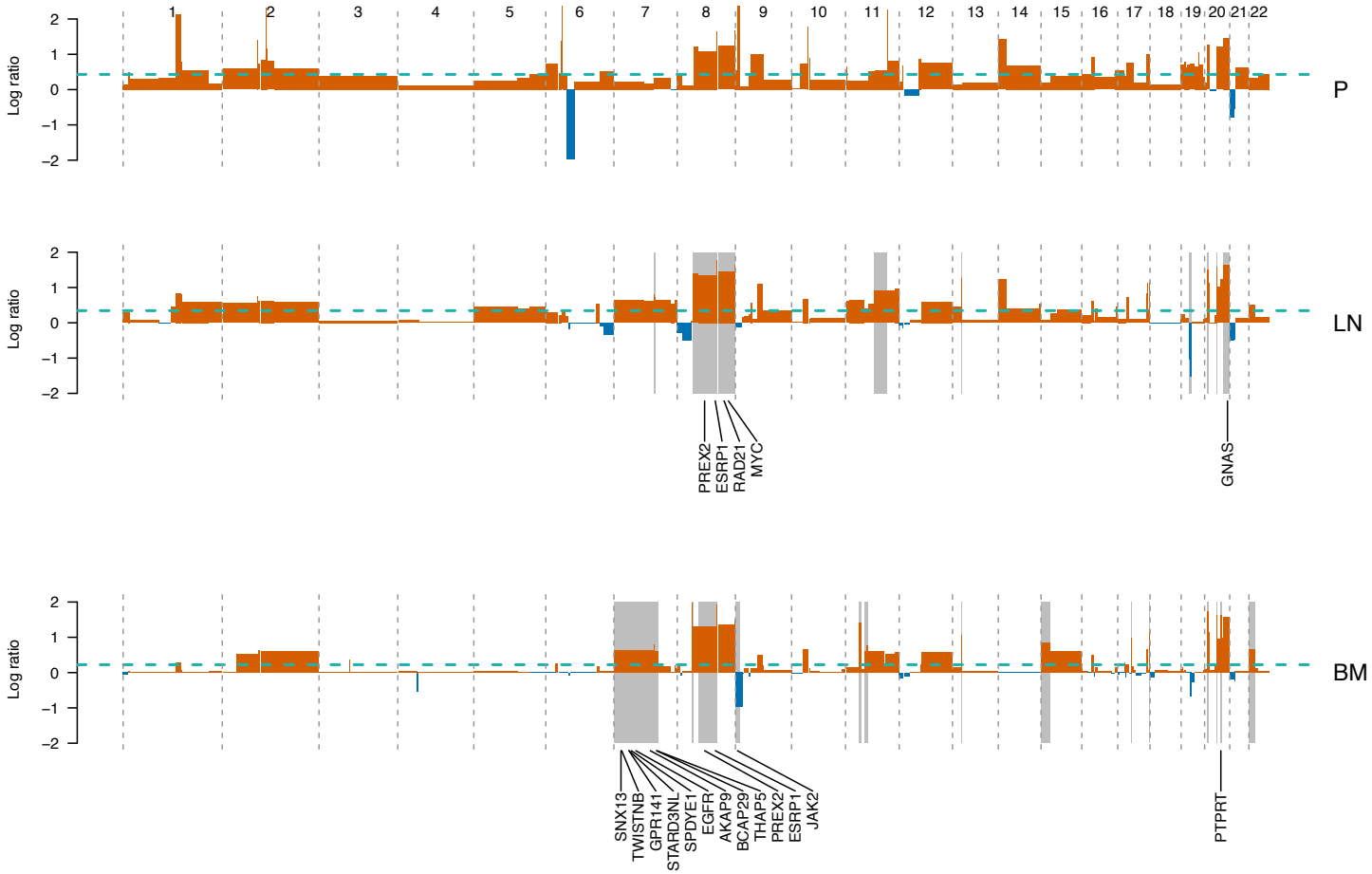


Supplementary Figure 5. Genetic concordance and heterogeneity amongst copy number alterations (CNAs) between paired primary CRCs and metastases. Segmented log depth ratios for each primary CRC and paired metastasis were adjusted for estimated tumor purity and ploidy in each sample and averaged across multiple-regions from the same tumor site. The green dashed line denotes mean log depth ratios. Grey boxes denote regions where CNAs were differentially altered in metastases relative to the primary CRC and satisfied the following criteria: absolute copy number is larger than 2.8 or less than 1.2; copy number relative to mean ploidy is larger than 0.8 or less than -0.8; and changes relative to the primary tumor in both absolute copy number and relative copy number are larger than 0.8 or less than -0.8. Putative CRC driver genes in the grey-boxed differential CNA regions are labeled.

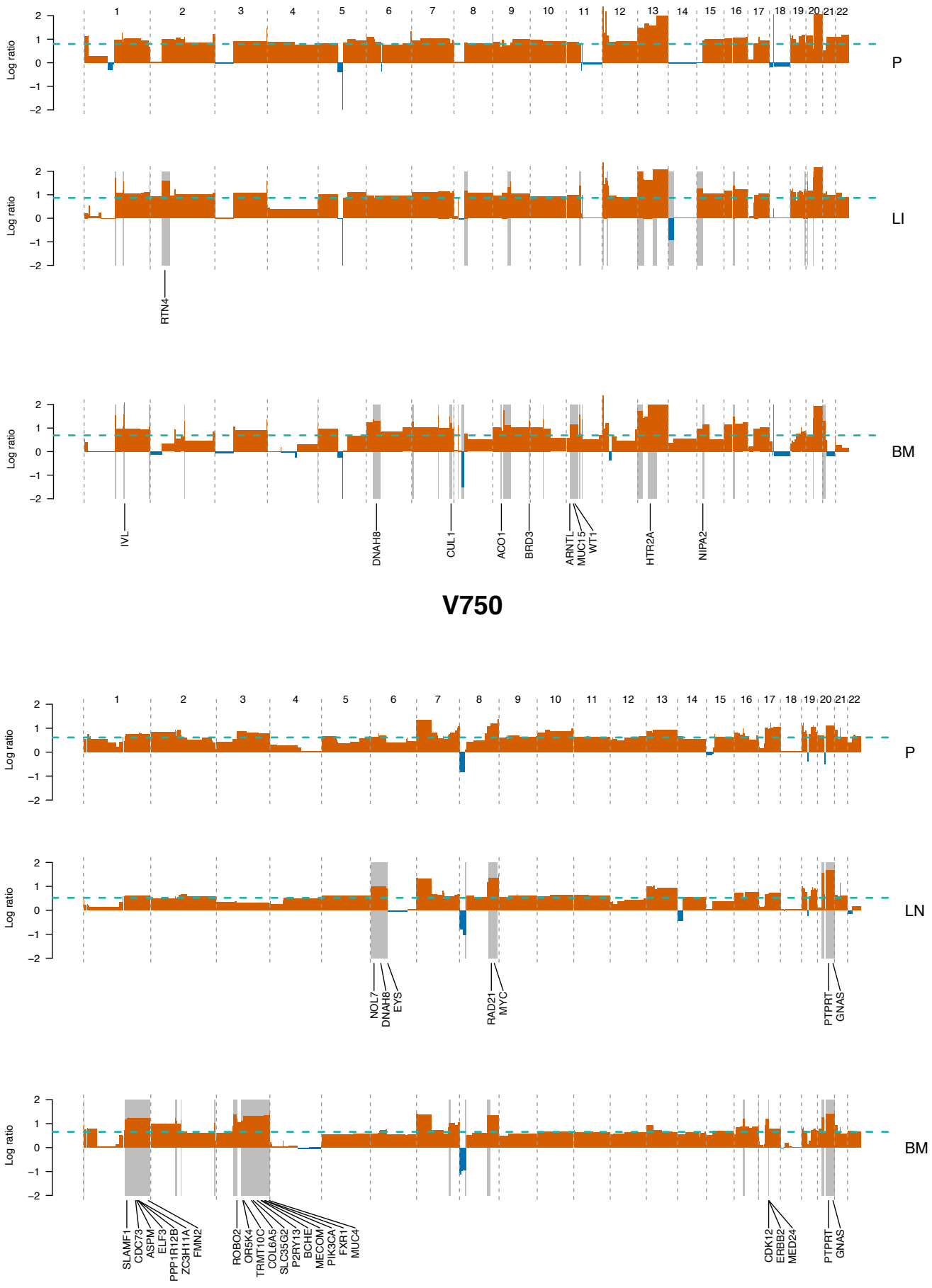
V402



V514

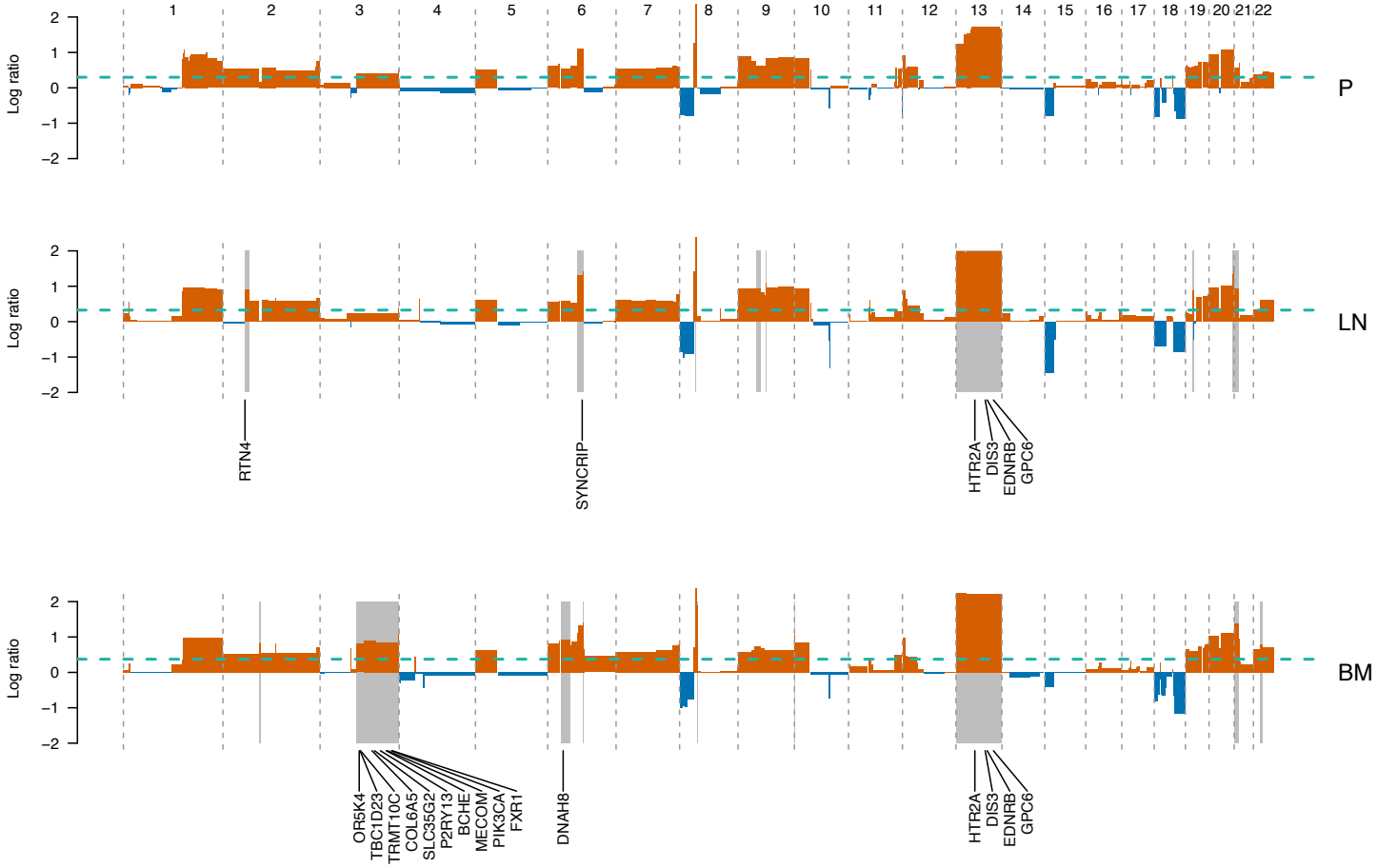


Supplementary Figure 5 - continued.

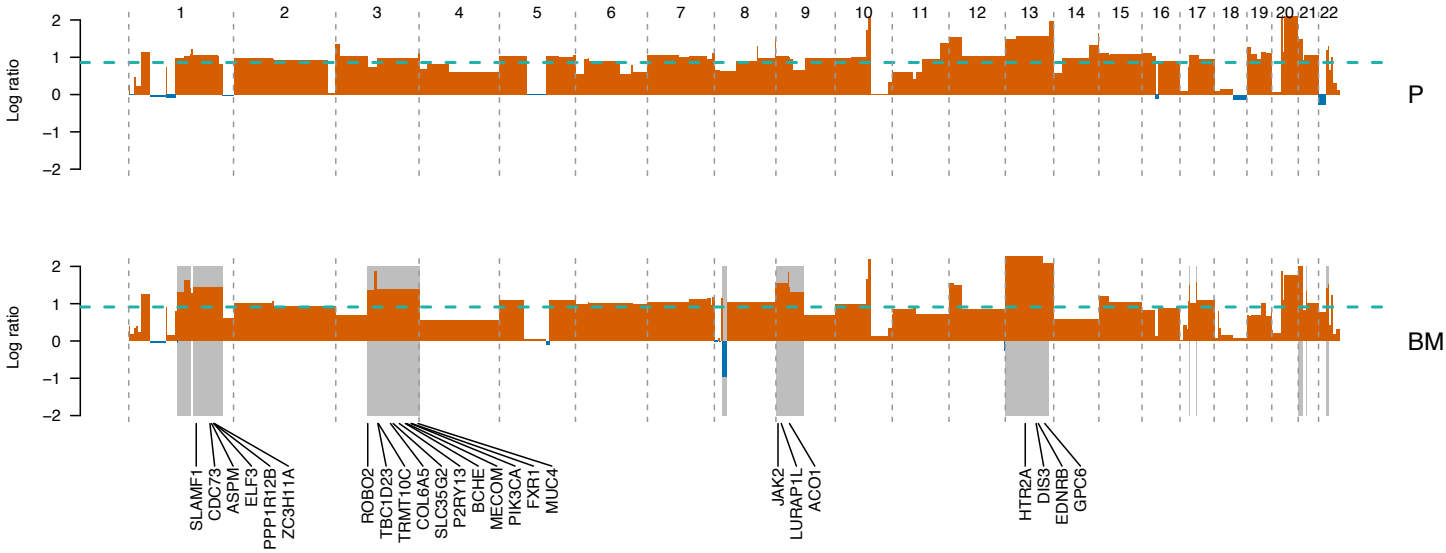


Supplementary Figure 5 - continued.

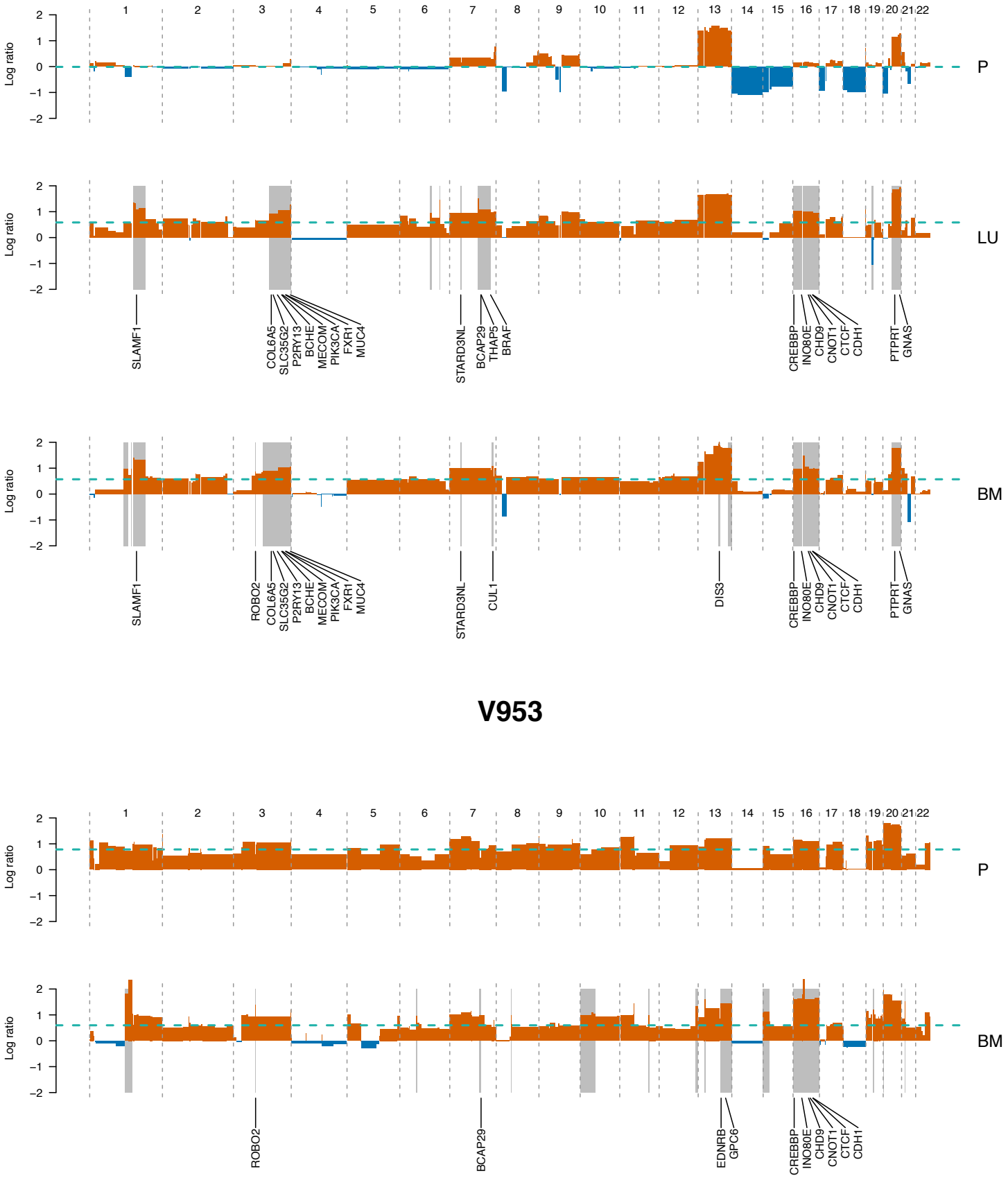
V824



V855

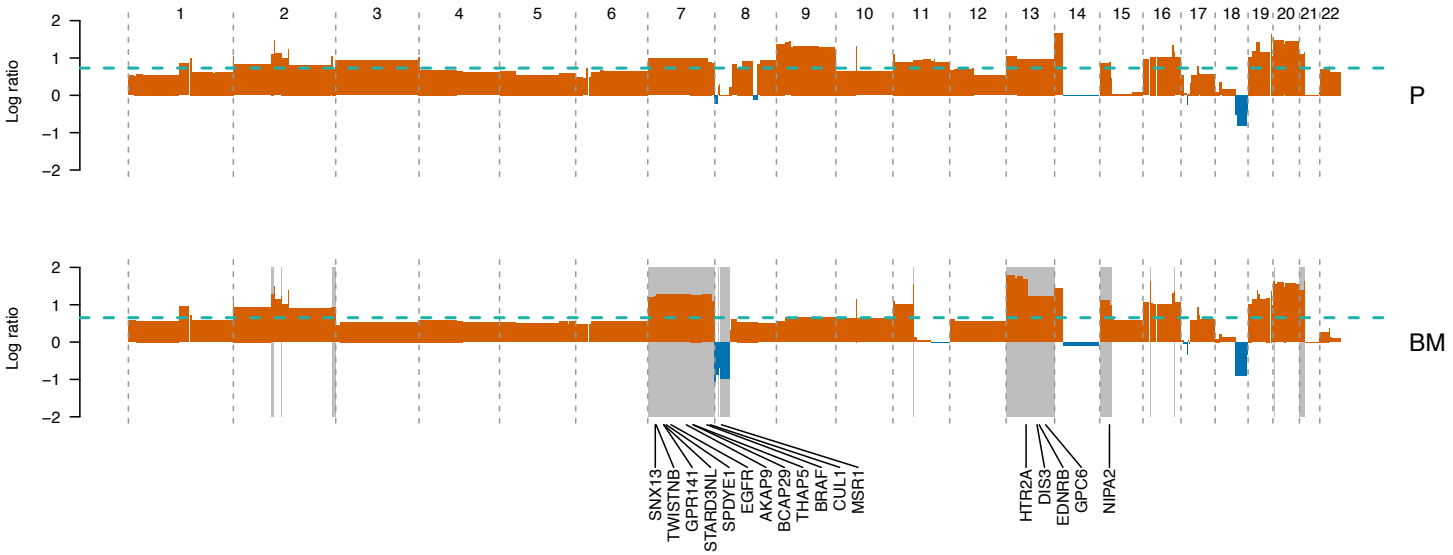


Supplementary Figure 5 - continued.

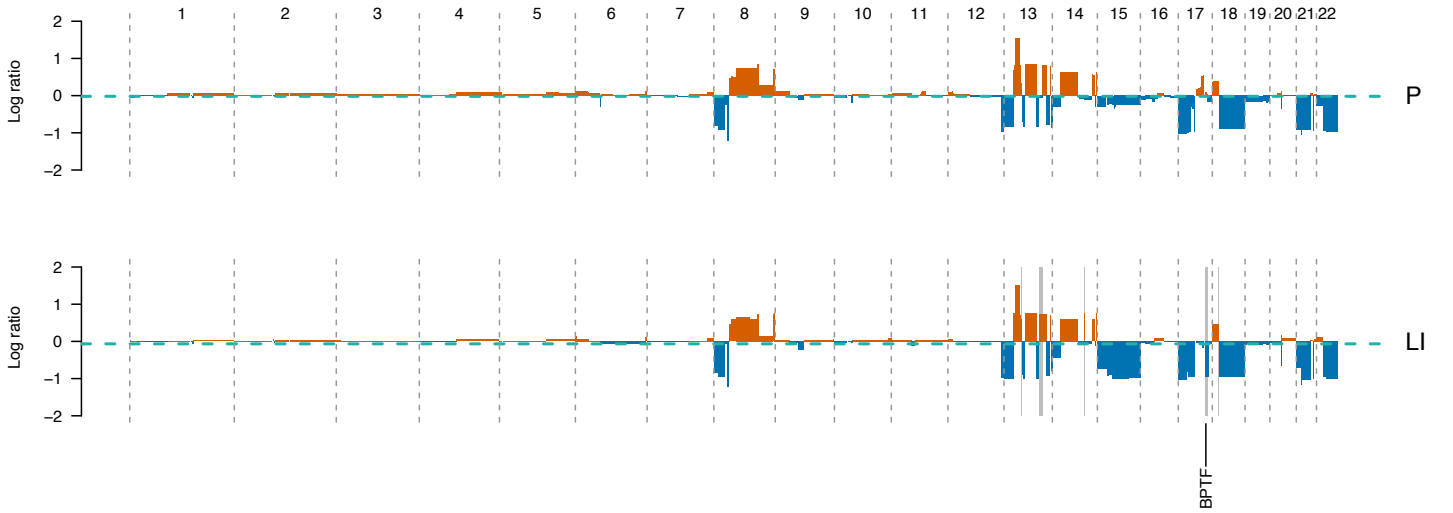


Supplementary Figure 5 - continued.

V974

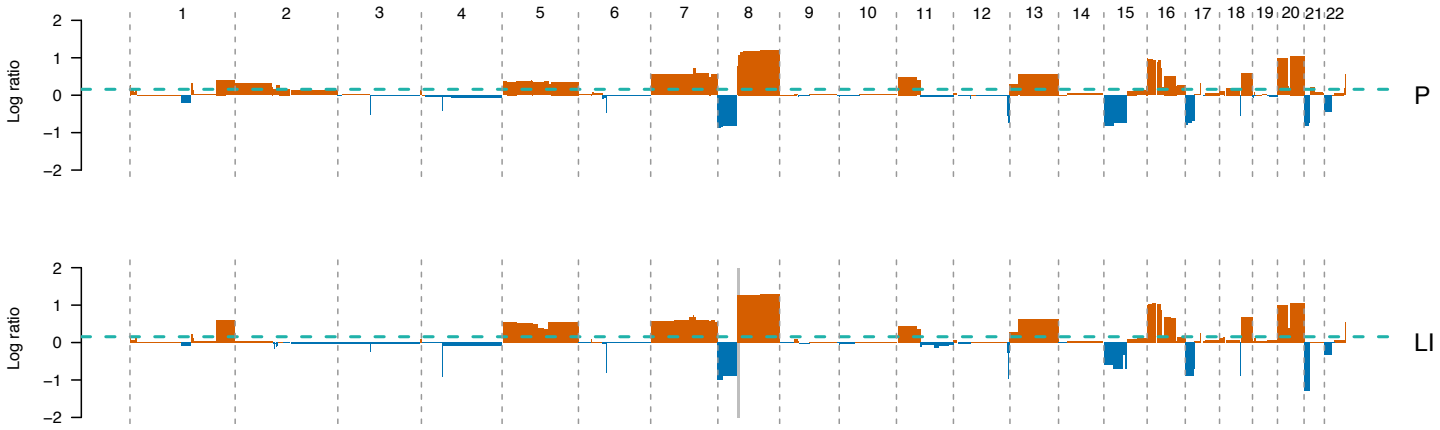


Uchi2

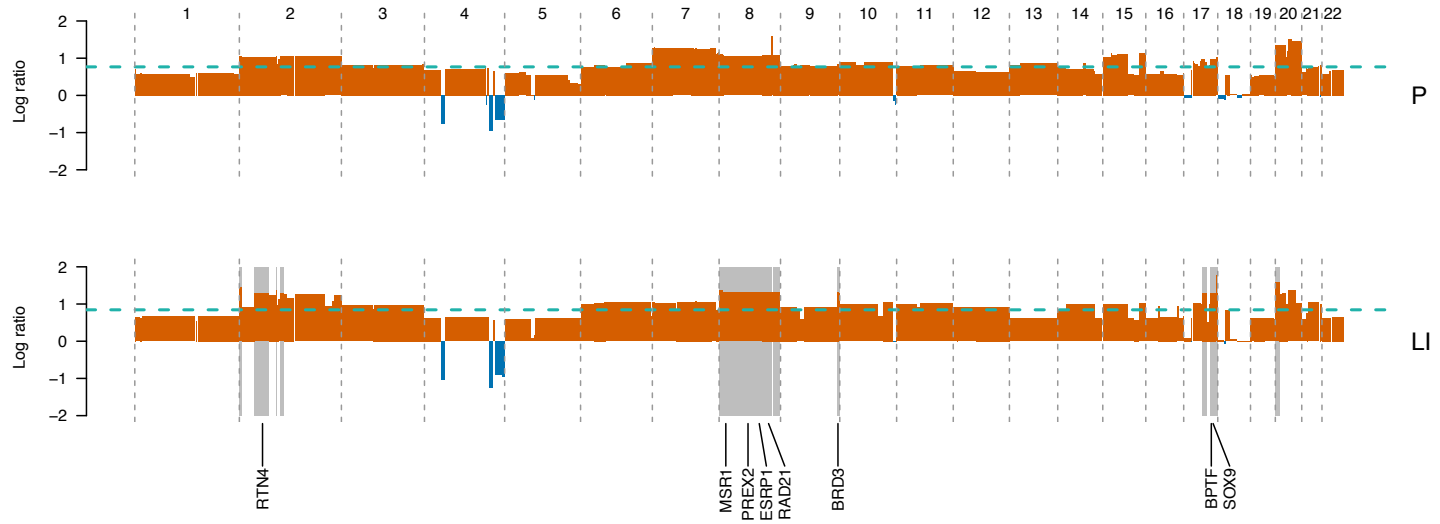


Supplementary Figure 5 - continued.

Kim1

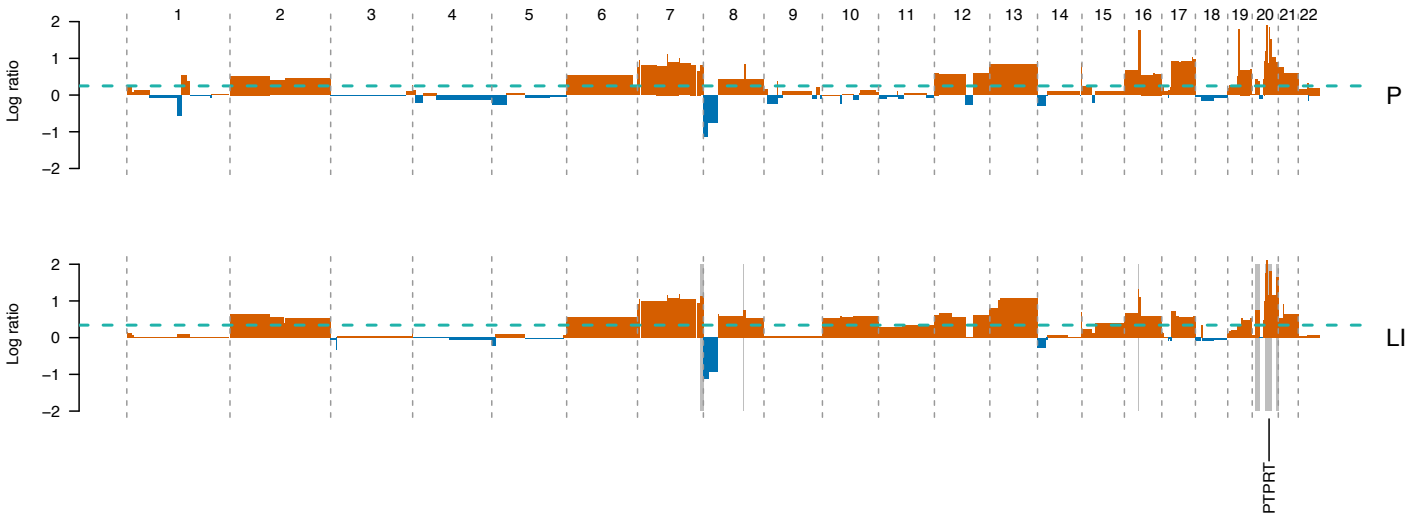


Kim2

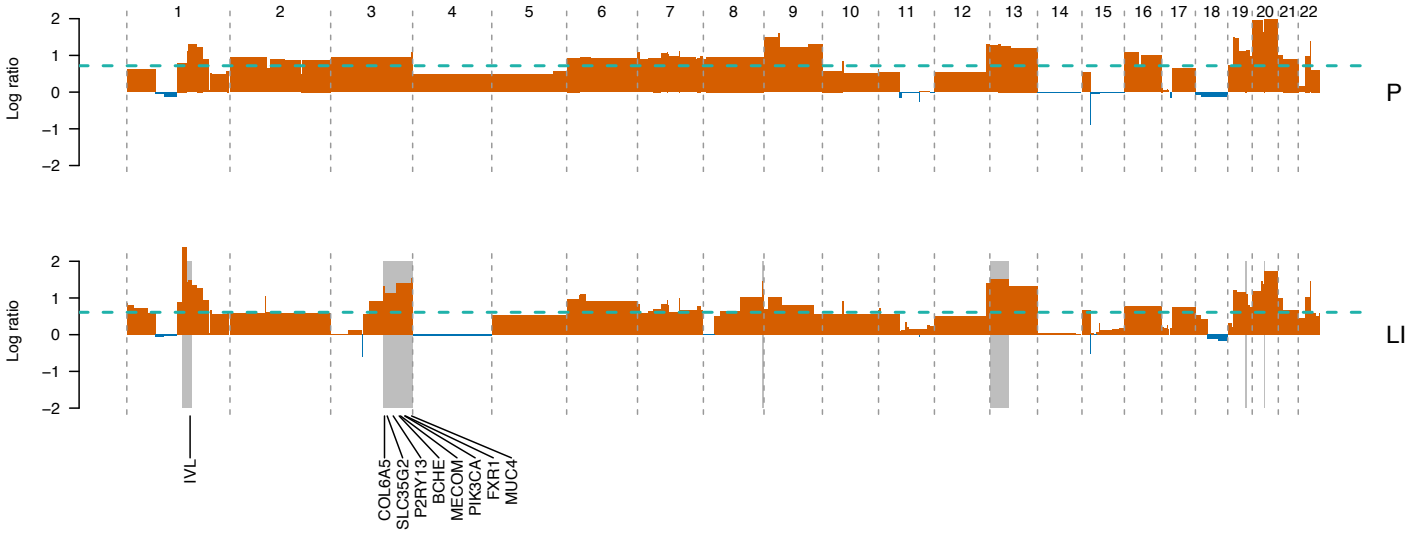


Supplementary Figure 5 - continued.

Leung1

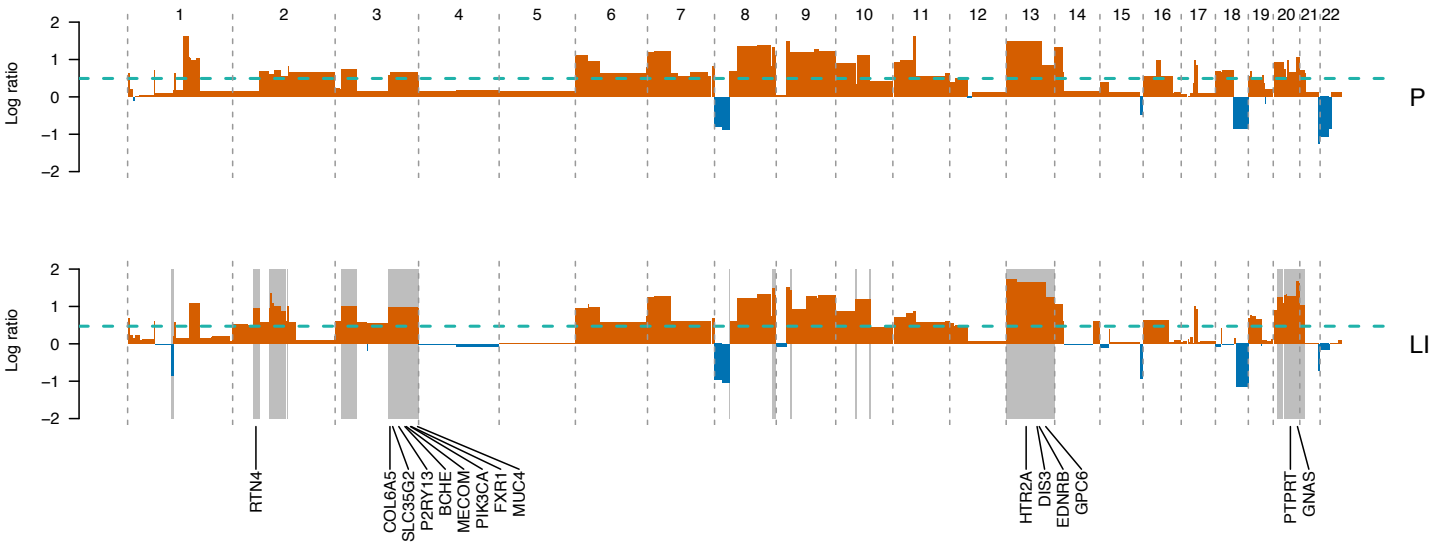


Leung2

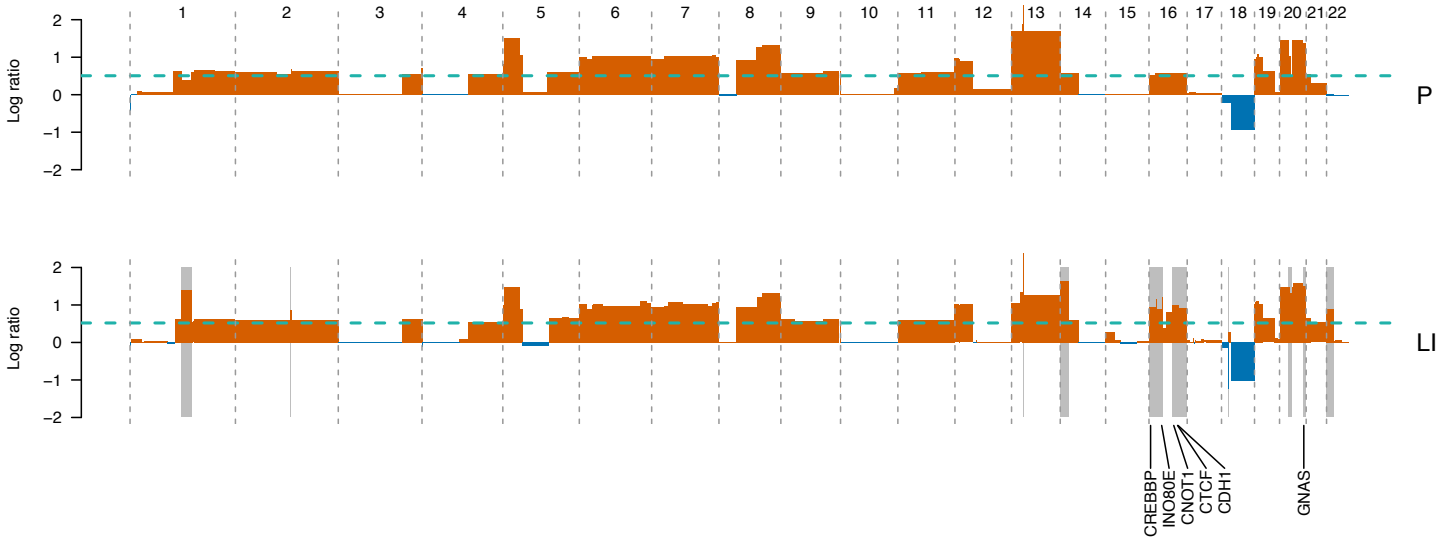


Supplementary Figure 5 - continued.

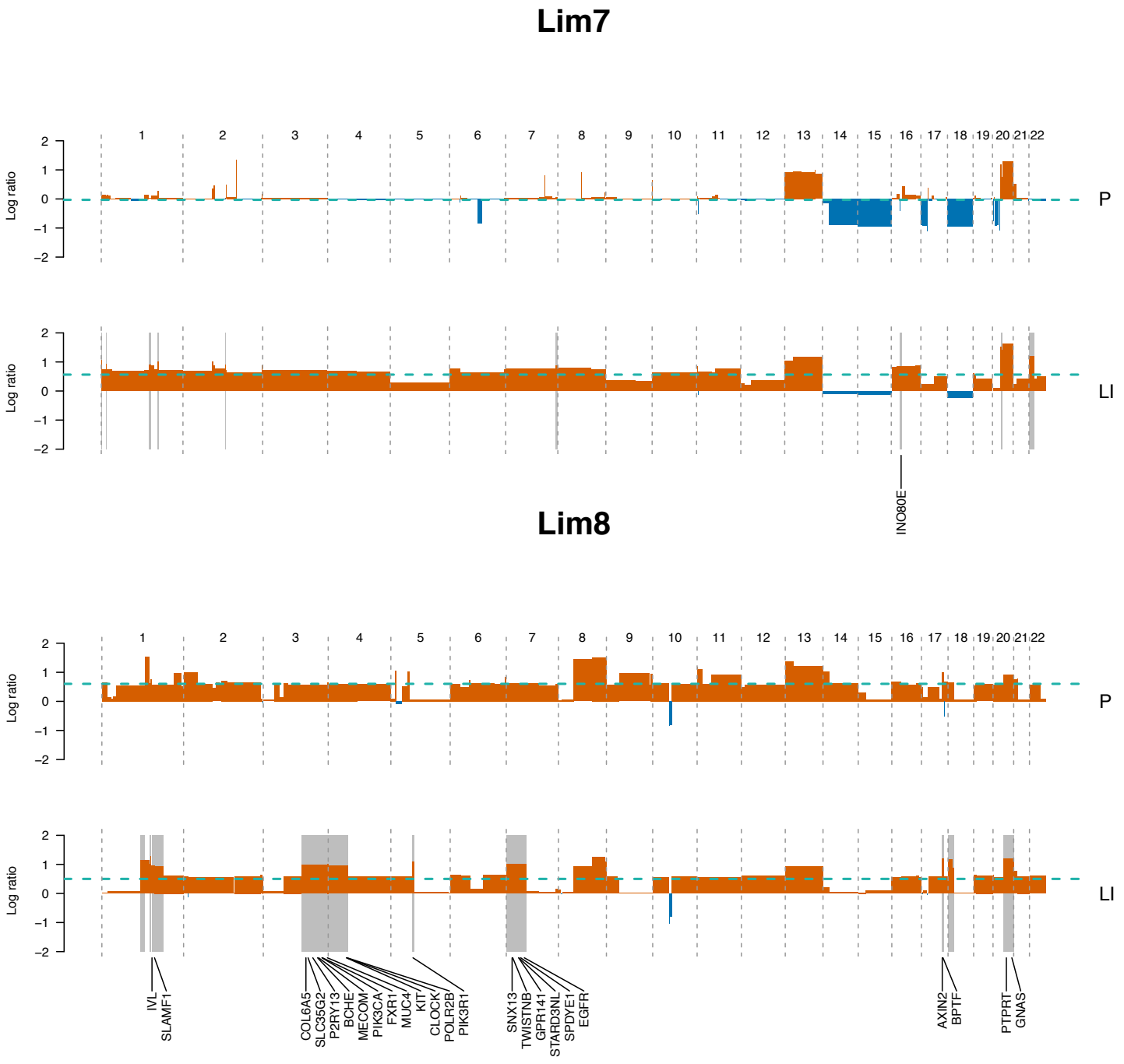
Lim3



Lim6

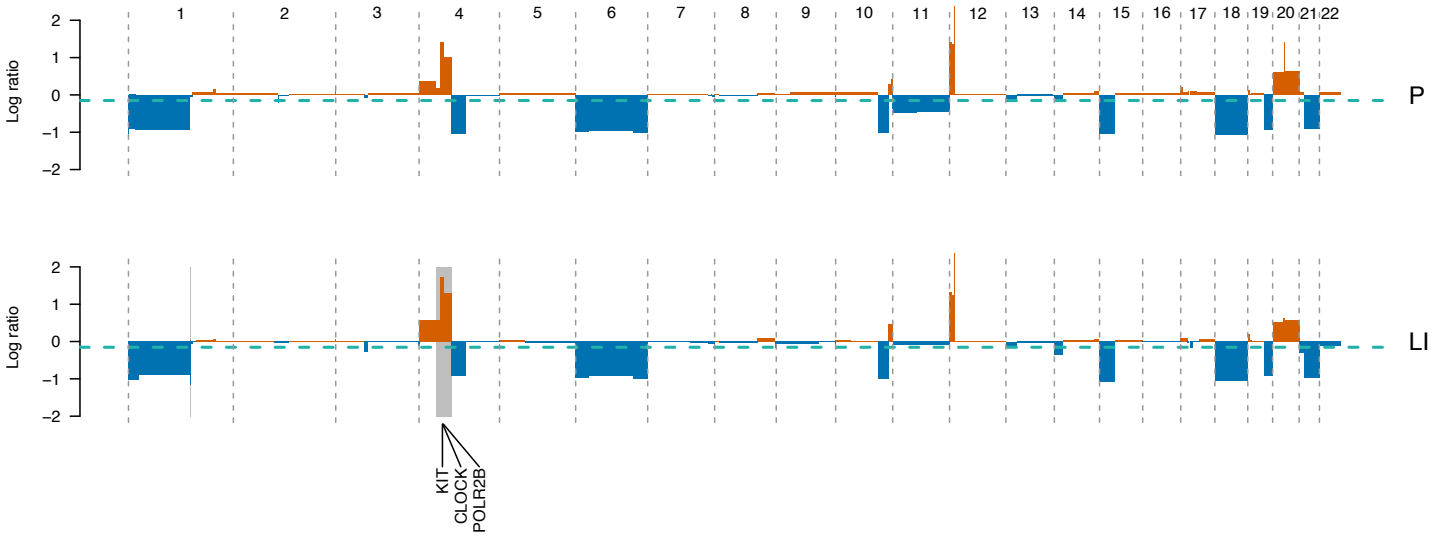


Supplementary Figure 5 - continued.

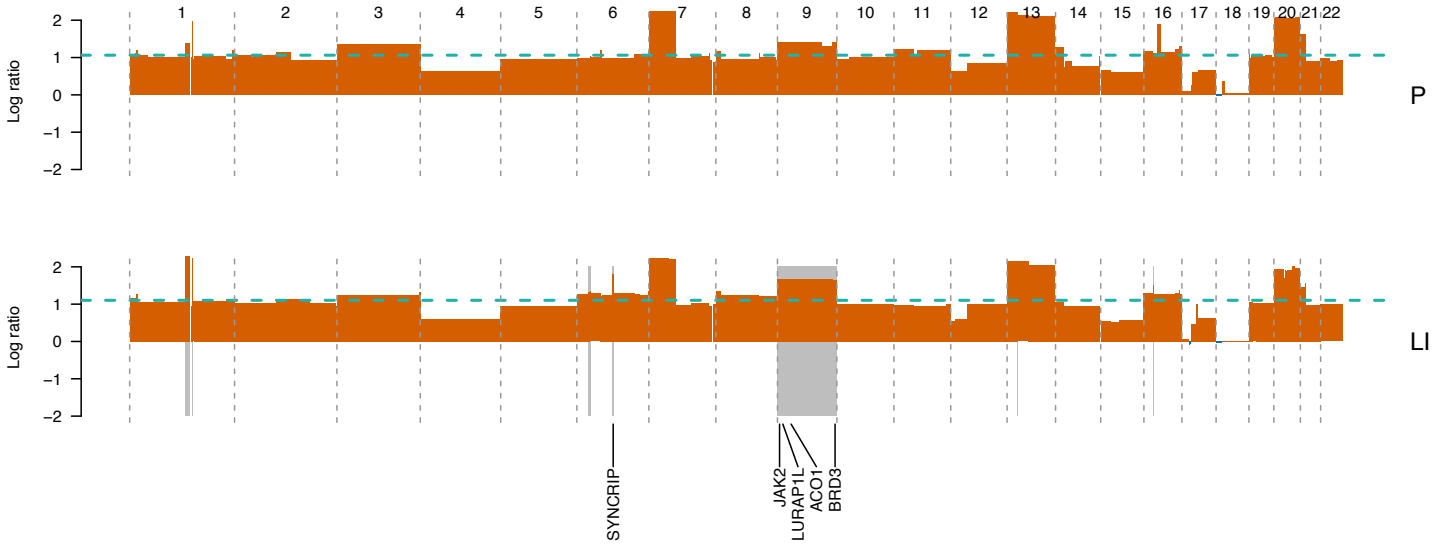


Supplementary Figure 5 - continued.

Lim11

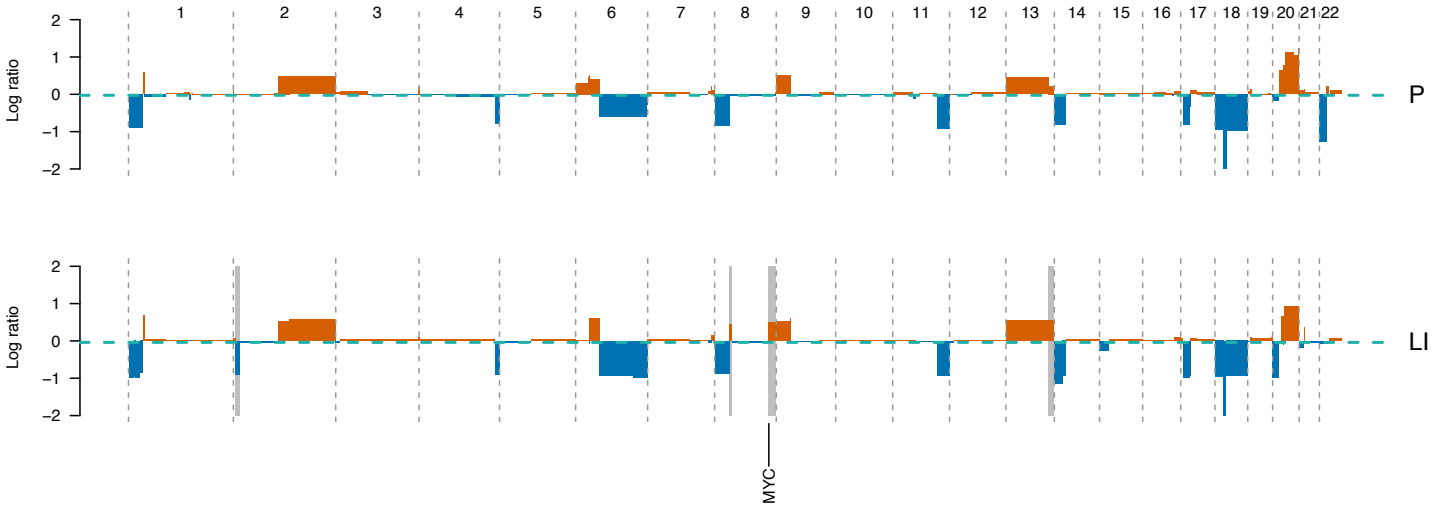


Lim12

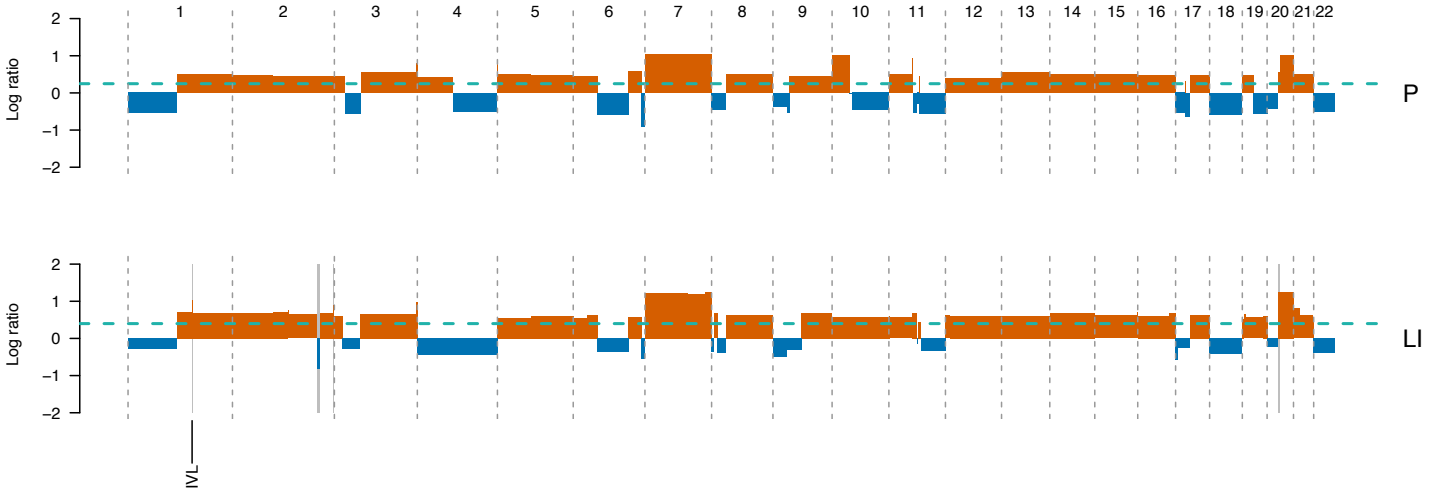


Supplementary Figure 5 - continued.

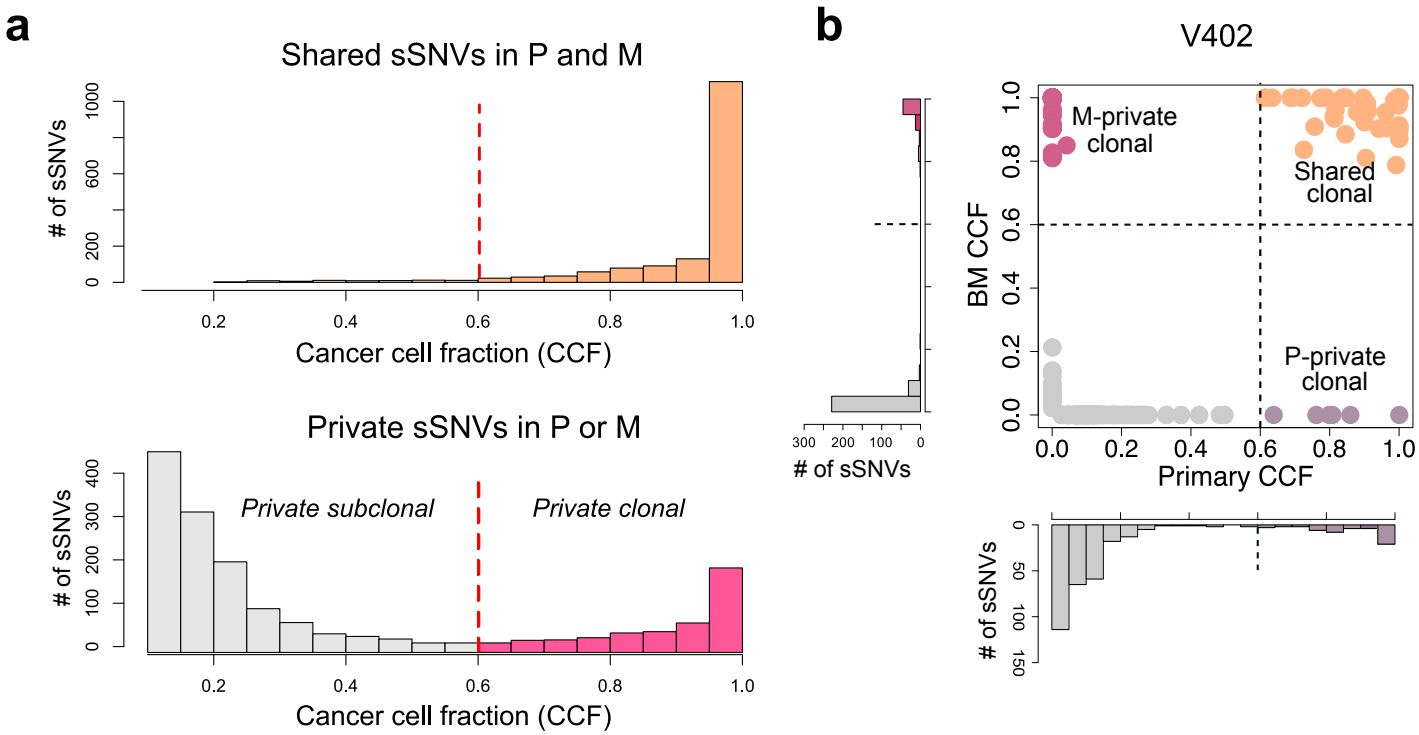
Lim16



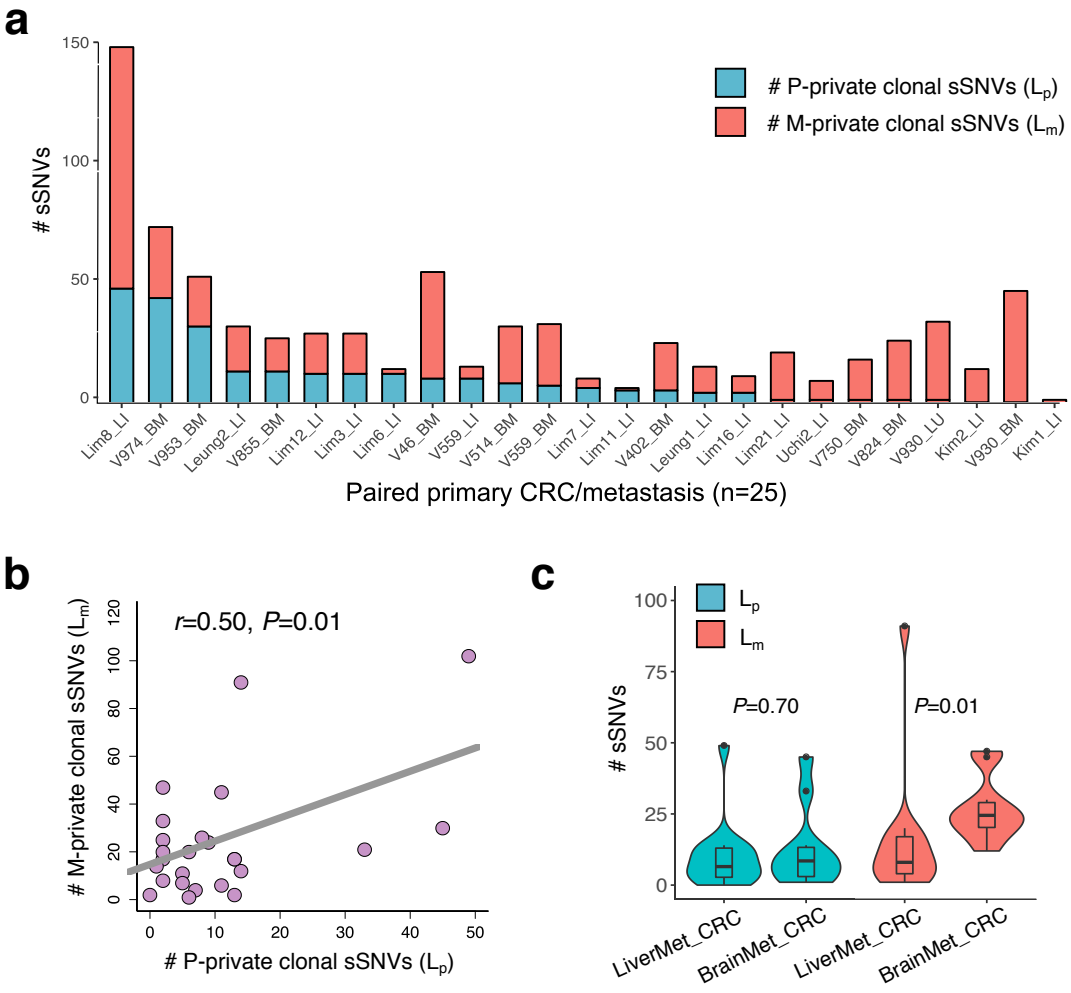
Lim21



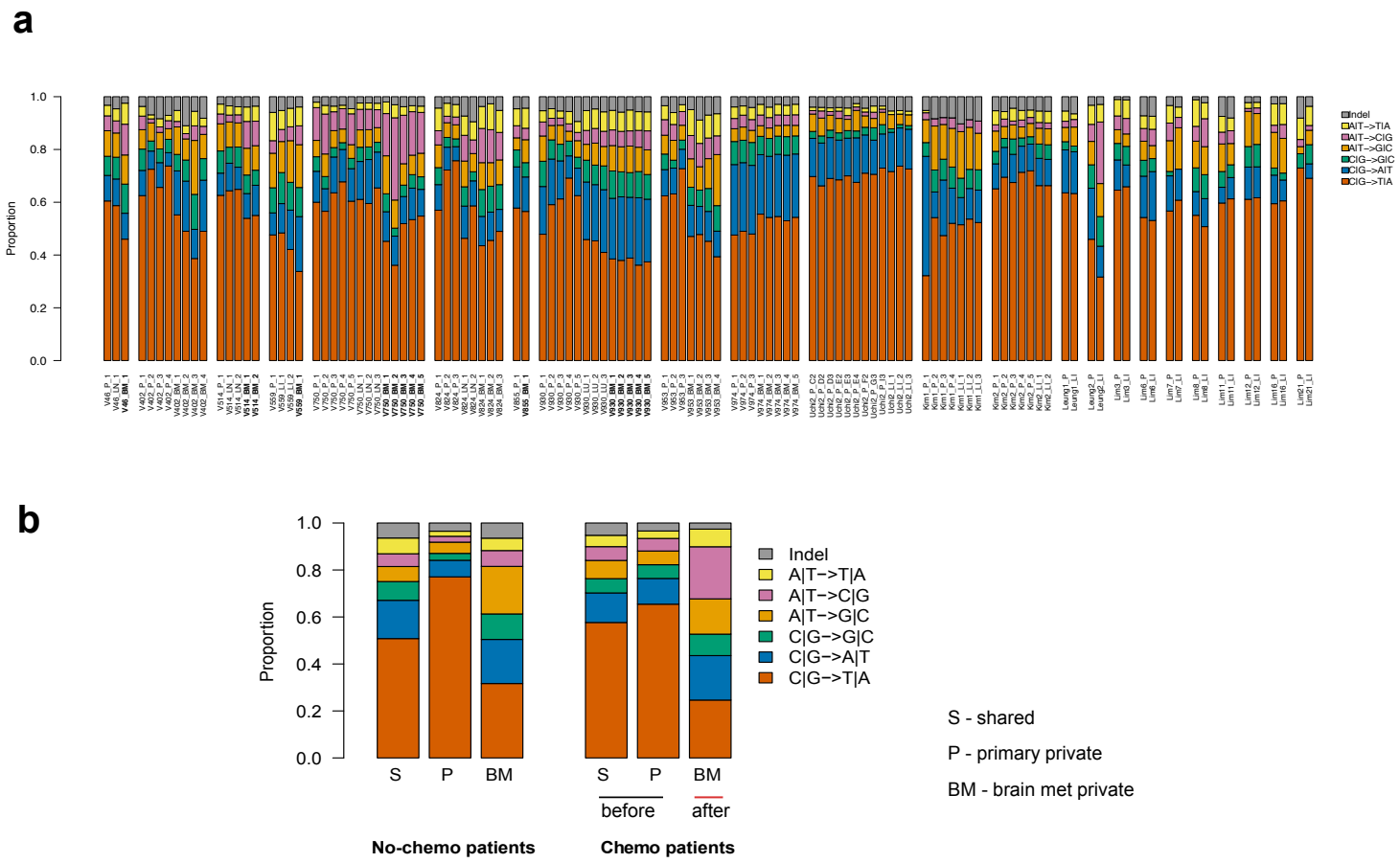
Supplementary Figure 5 - continued.



Supplementary Figure 6. Defining *clonal* and *subclonal* mutations based on the merged cancer cell fraction (CCF) from multi-region sequencing data. (**a**) The distribution of CCF values for P-M shared sSNVs (above) and P or M-private sSNVs (below). CCF values from cases with multi-region sequencing data were merged. For a given sSNV, the CCF estimates in different tumor sites are calculated separately by CHAT (Methods) and only CCF estimates ≥ 0.1 are counted. An sSNV is classified as a "shared" if it has a CCF≥0.2 (equivalent to VAF≥0.1) in both P and M for a given patient otherwise it is classified as "private" in P or M. The vast majority (99%) of P-M shared sSNVs have CCF>60%, a cutoff that also clearly distinguishes private clonal and subclonal sSNV clusters. (**b**) Scatterplot of merged CCF values in the primary versus brain metastasis (BM) for a representative patient (V402), where the cutoff of CCF=60% is indicated and classes of sSNVs are labeled. The CCF distribution of sSNVs in the primary tumor and metastasis are indicated on the corresponding histograms.

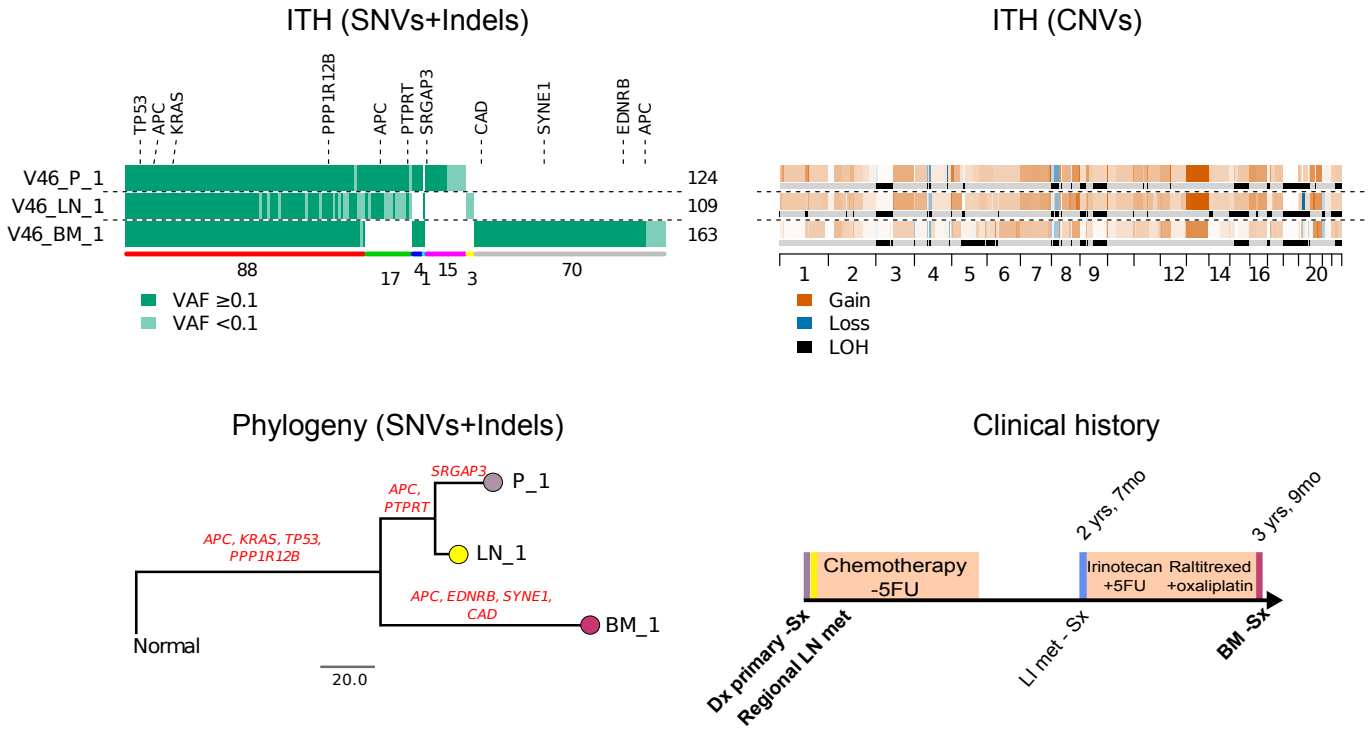


Supplementary Figure 7. The number of primary tumor-private *clonal* sSNVs (L_p) and metastasis-private *clonal* sSNVs (L_m) in mCRCs. A cutoff CCF=60% was used to identify clonal sSNVs. Merged CCF was used for tumors with multi-region sequencing data. (**a**) L_p and L_m values in 25 mCRC P/M pairs (10 brain metastases (BM), 14 liver metastases (LI) and 1 lung metastasis (LU)); (**b**) Pearson correlation (r) between L_m and L_p across patients in the mCRC cohort (n=25 P/M pairs). (**c**) L_p and L_m values in liver metastasis CRCs (n=14 P/M pairs) and brain metastasis CRCs (n=10 P/M pairs), respectively. *P*-value, Wilcoxon Rank-Sum Test (two-sided). Bar, median; box, 25th to 75th percentile (interquartile range, IQR); vertical line, data within 1.5 times the IQR.



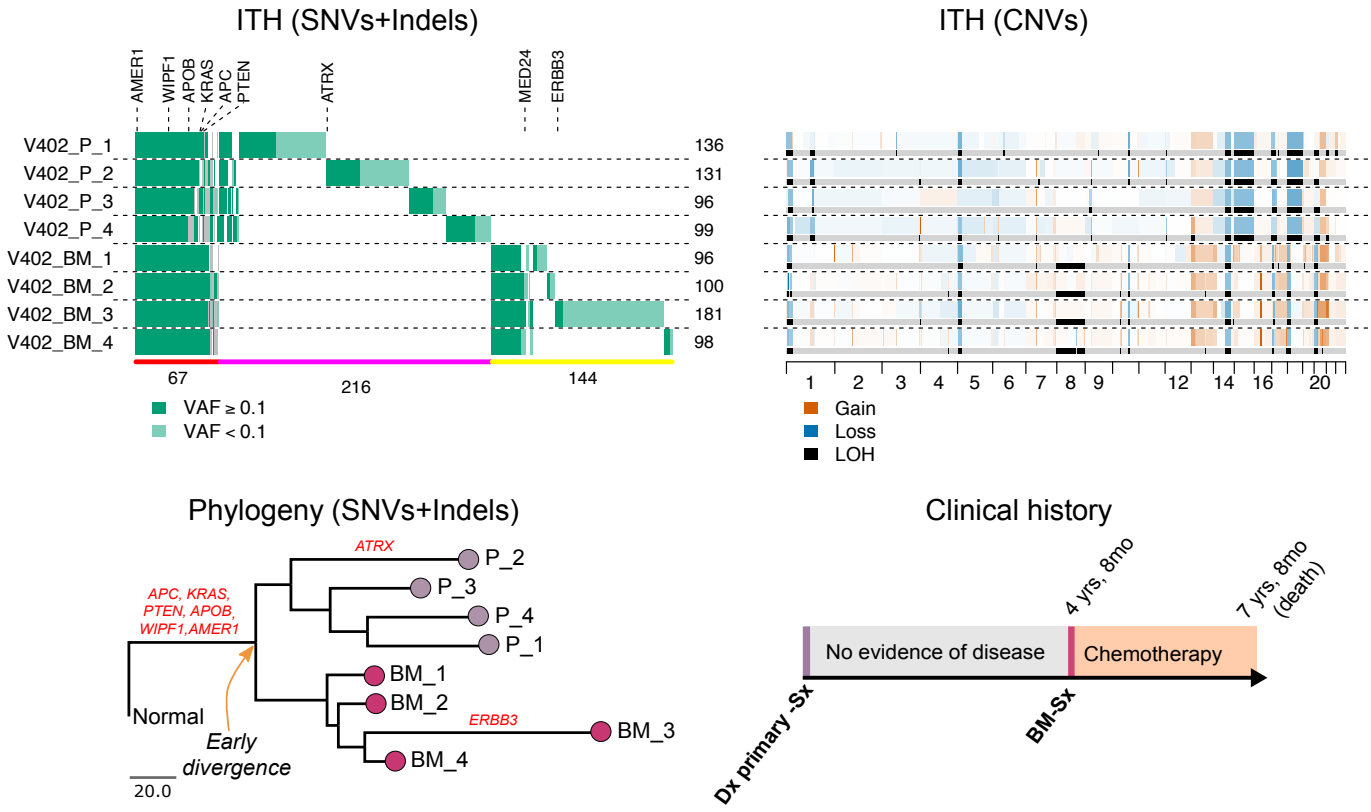
Supplementary Figure 8. Mutation spectra in whole-exome sequencing data from paired primary CRCs and metastases. (**a**) The proportion of seven mutation types (six sSNV types plus small indels) in each tumor region. Tumor samples acquired after chemotherapy are highlighted in bold. (**b**) The proportion of shared clonal (S), primary-private (P) or brain metastasis-private (BM) sSNVs and small indels in chemotherapy-naïve patients or patients who received chemotherapy (chemo-treated) prior to diagnosis of the brain metastasis are shown. Of note, all primary tumors were resected prior to therapy. Only coding sSNVs and small indels were included. A mutation is called as present in a given sample if there are at least 3 supporting variant reads and a VAF of at least 0.05 or a CCF of at least 0.1, regardless of the number of variant supporting reads.

V46

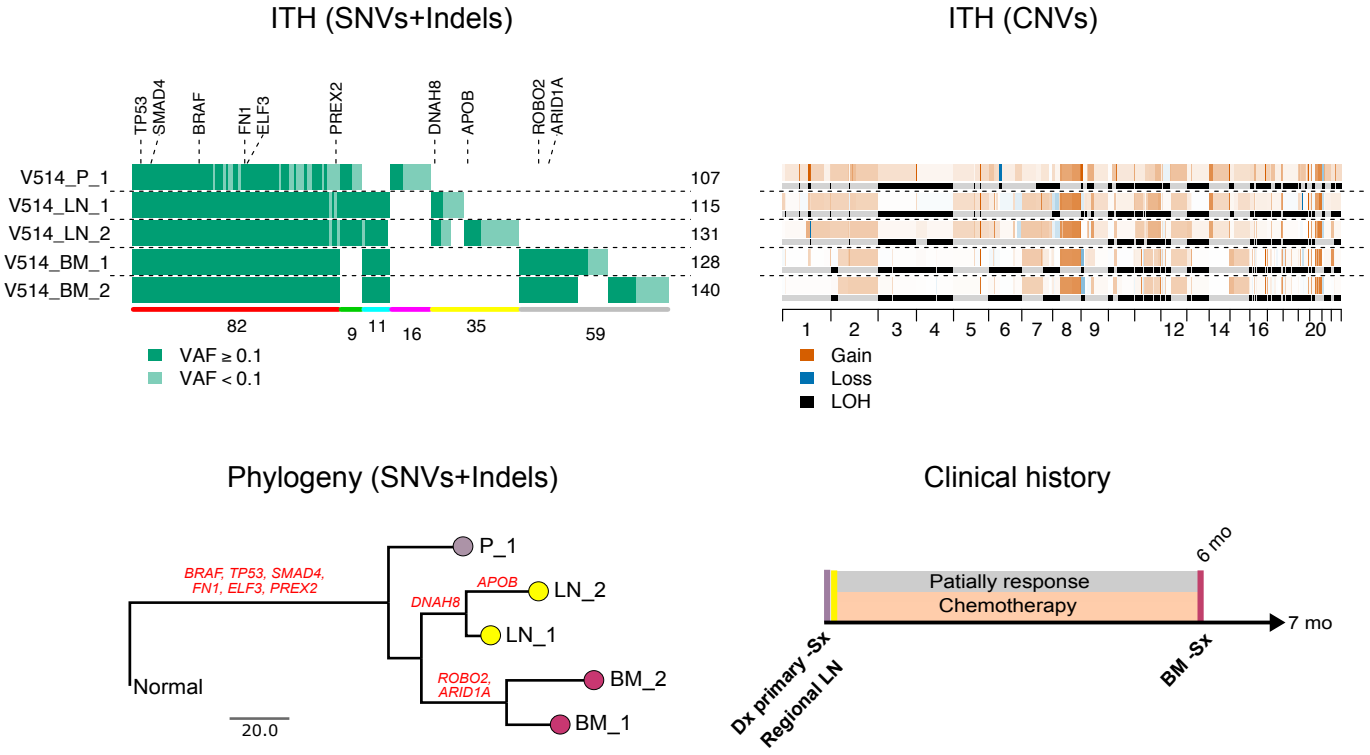


Supplementary Figure 9. Clinical history and intra-tumor heterogeneity (ITH) in paired primary CRCs and metastases. Patterns of within and between lesion heterogeneity amongst sSNVs and indels based on whole-exome sequencing of paired primary CRCs and metastases, where canonical CRC 'driver' genes are labeled. Dark and light green bars represent mutations with VAF≥0.1 and VAF<0.1. The number of mutations shared amongst different lesions is indicated below the corresponding colored horizontal bars (upper left): ubiquitously P-M shared *(*red*)*, partially P-M shared *(*green-M1 or blue-M2*)*, P-private *(*pink*)* or M-private (yellow-M1 or gray-M2 or cyan-M1&M2). M1 and M2 correspond to different metastatic sites in the same patient. The number of detected mutations in corresponding sample was labelled at the right of each row. Patterns of within and between lesion heterogeneity amongst CNAs (upper right). Phylogeny reconstruction via maximum parsimony (PHYLIP) based on mutational presence/absence (bottom left). Canonical CRC drivers are labeled. Clinical and treatment history (bottom right). Dx: diagnosis; Sx: surgical operation; BM: brain metastasis; LU: lung metastasis; LI: liver metastasis; LN: lymph node metastasis.

V402

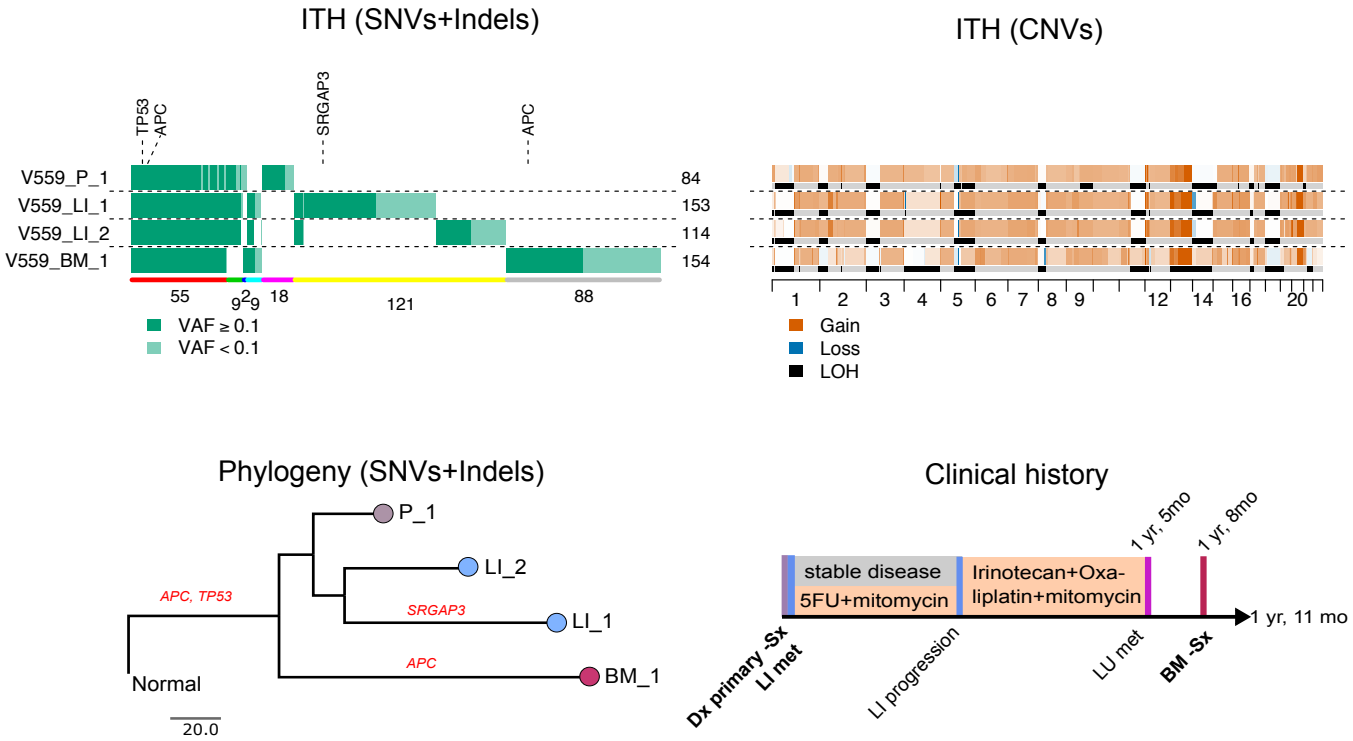


V514

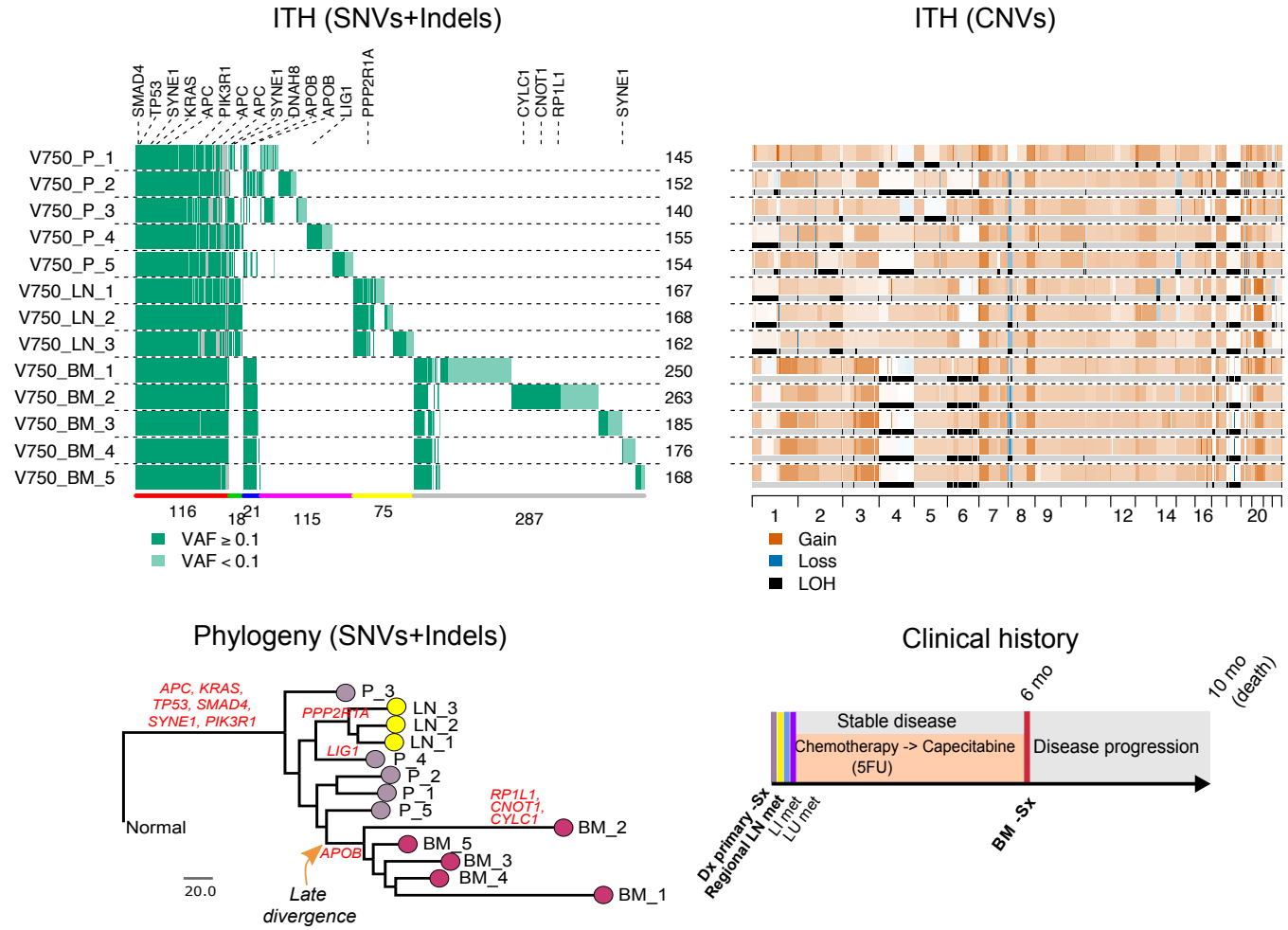


Supplementary Figure 9 - continued.

V559

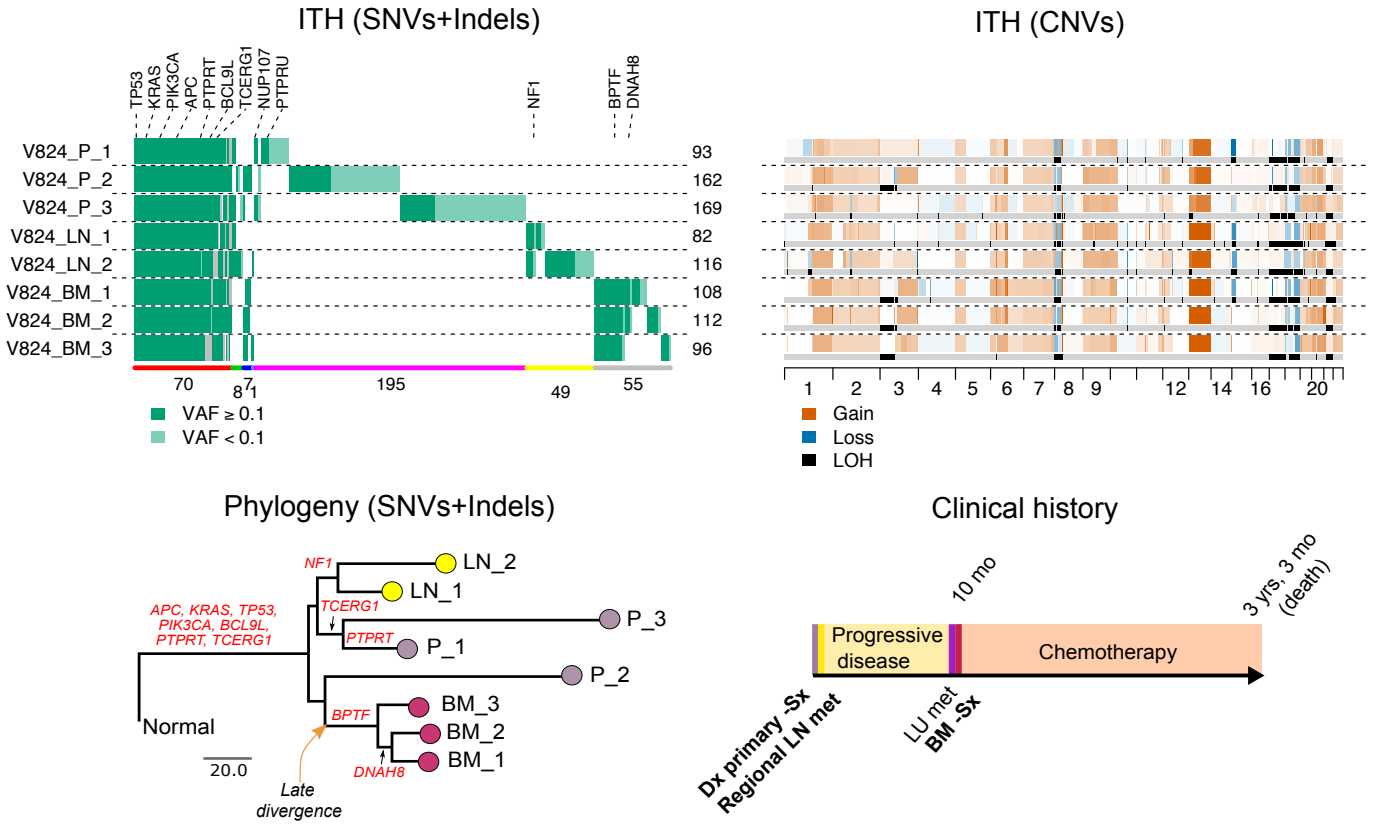


V750

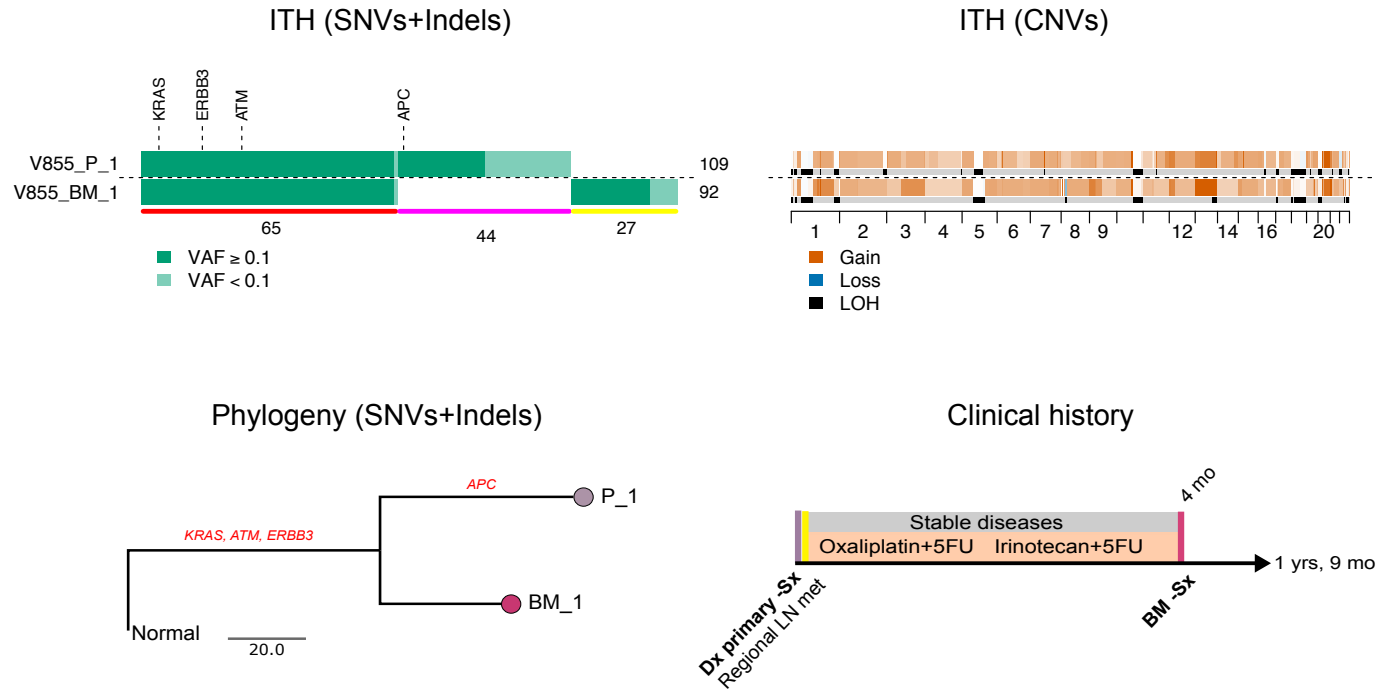


Supplementary Figure 9 - continued.

V824

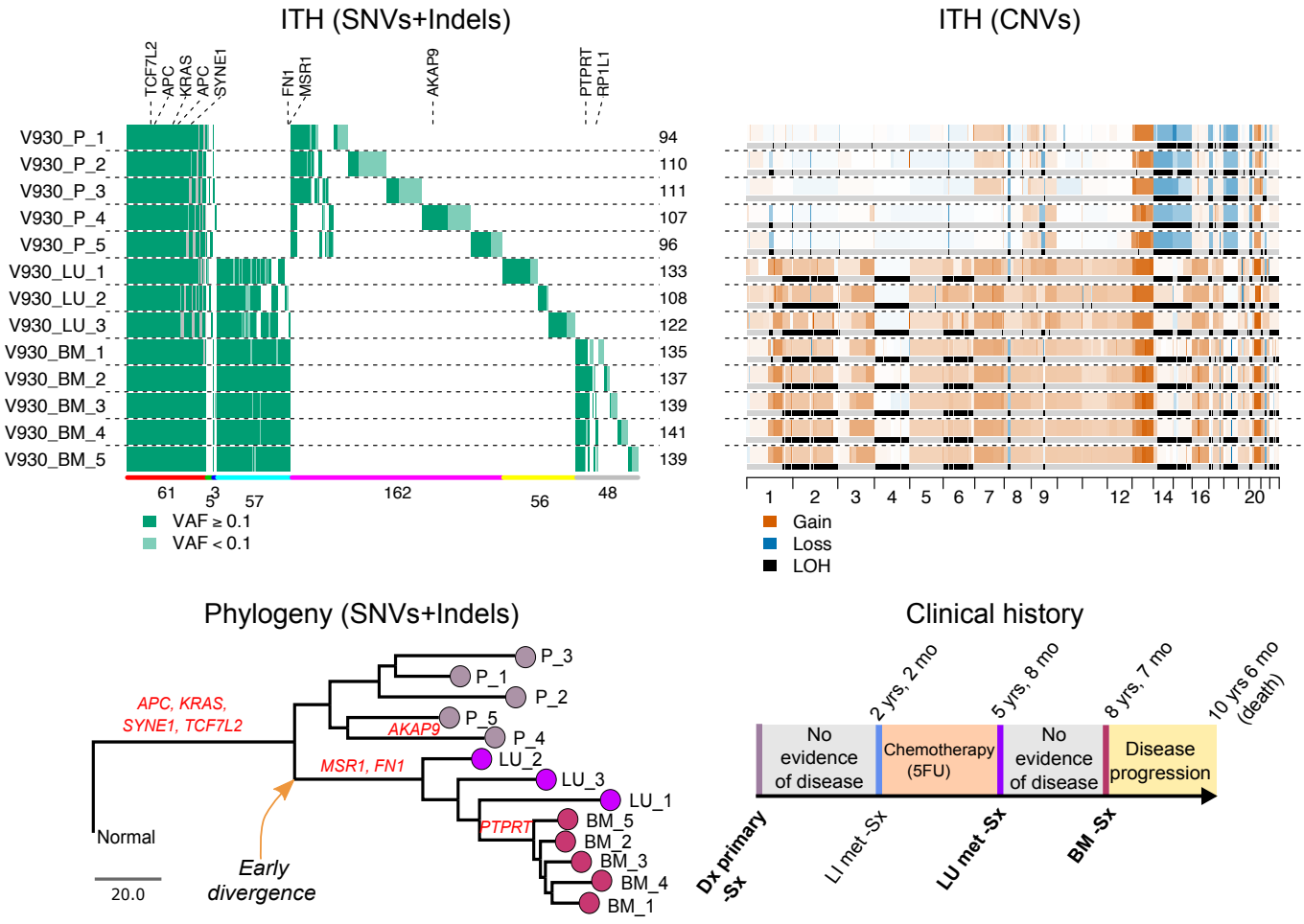


V855

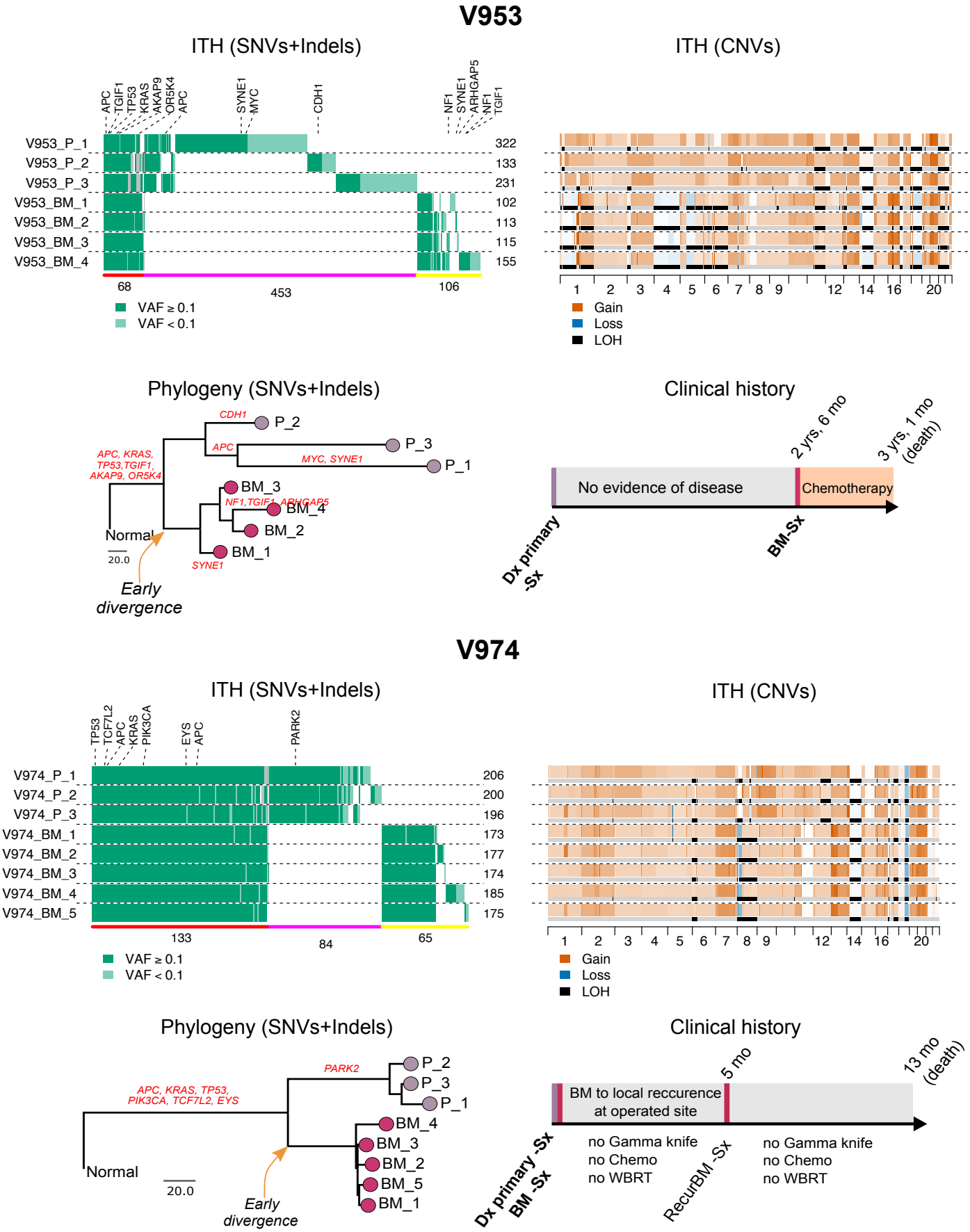


Supplementary Figure 9 - continued.

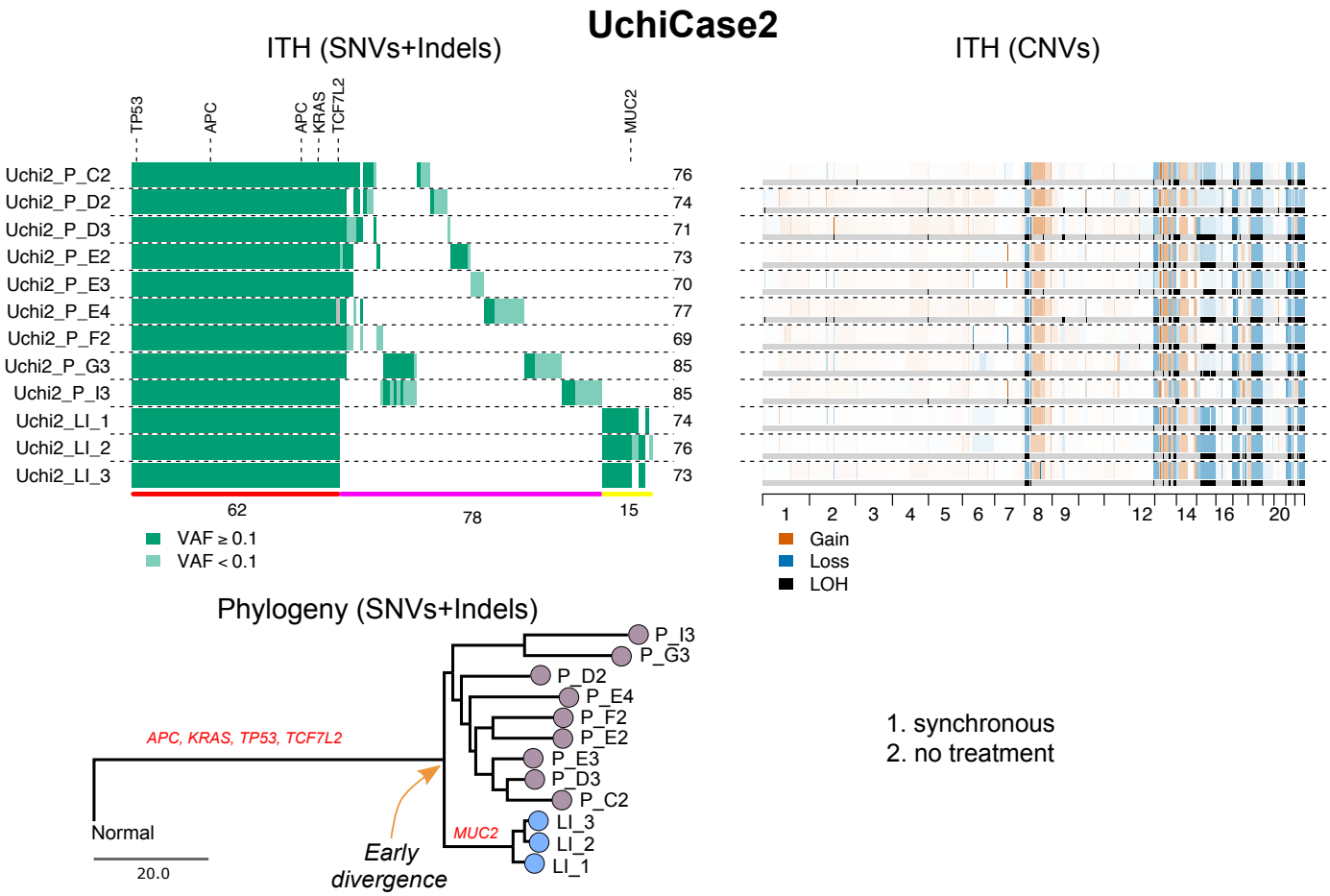
V930



Supplementary Figure 9 - continued.



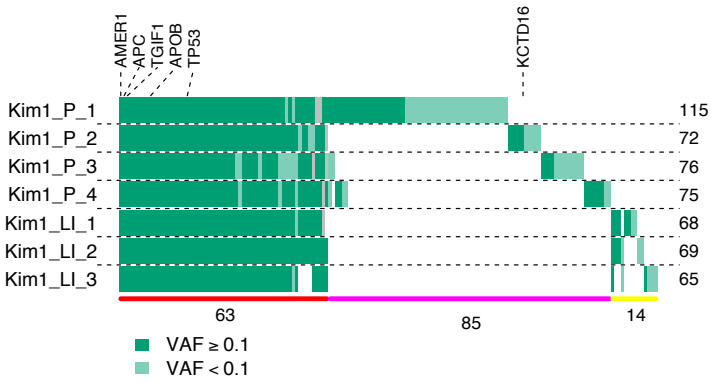
Supplementary Figure 9 - continued.



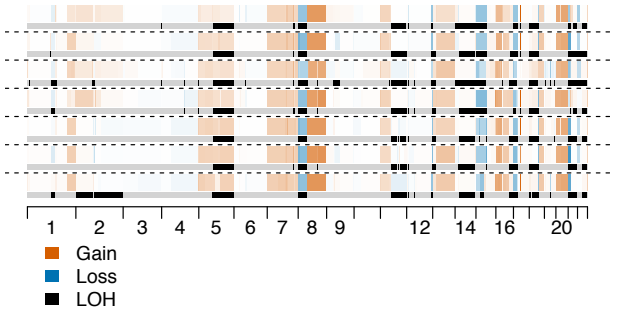
Supplementary Figure 9 - continued.

KimCRC1

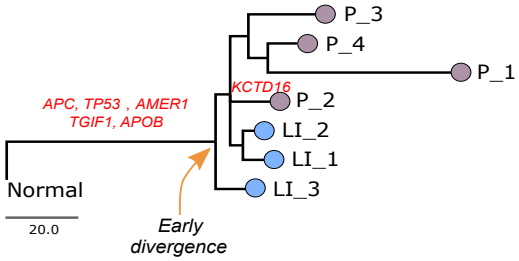
ITH (SNVs+Indels)



ITH (CNVs)



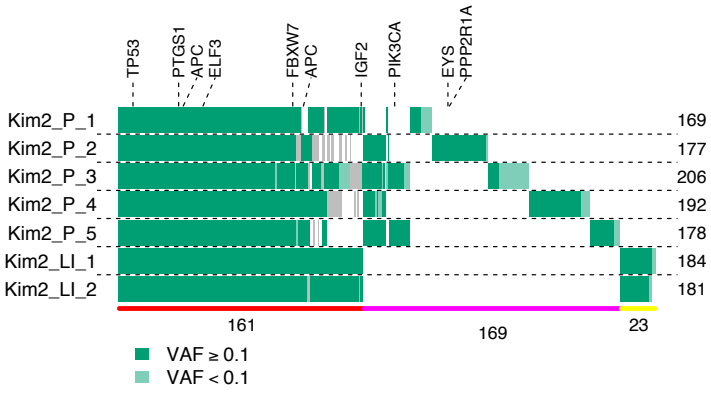
Phylogeny (SNVs+Indels)



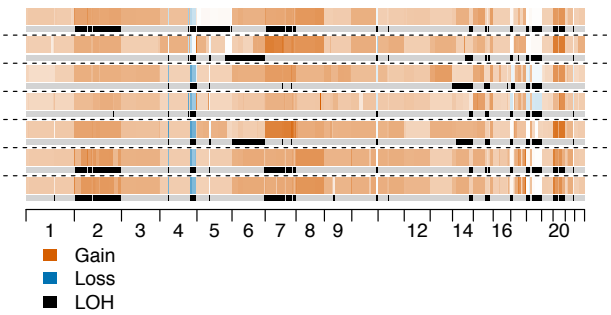
1. synchronous
2. no treatment

KimCRC2

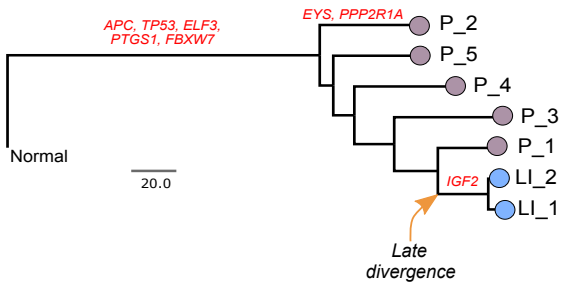
ITH (SNVs+Indels)



ITH (CNVs)

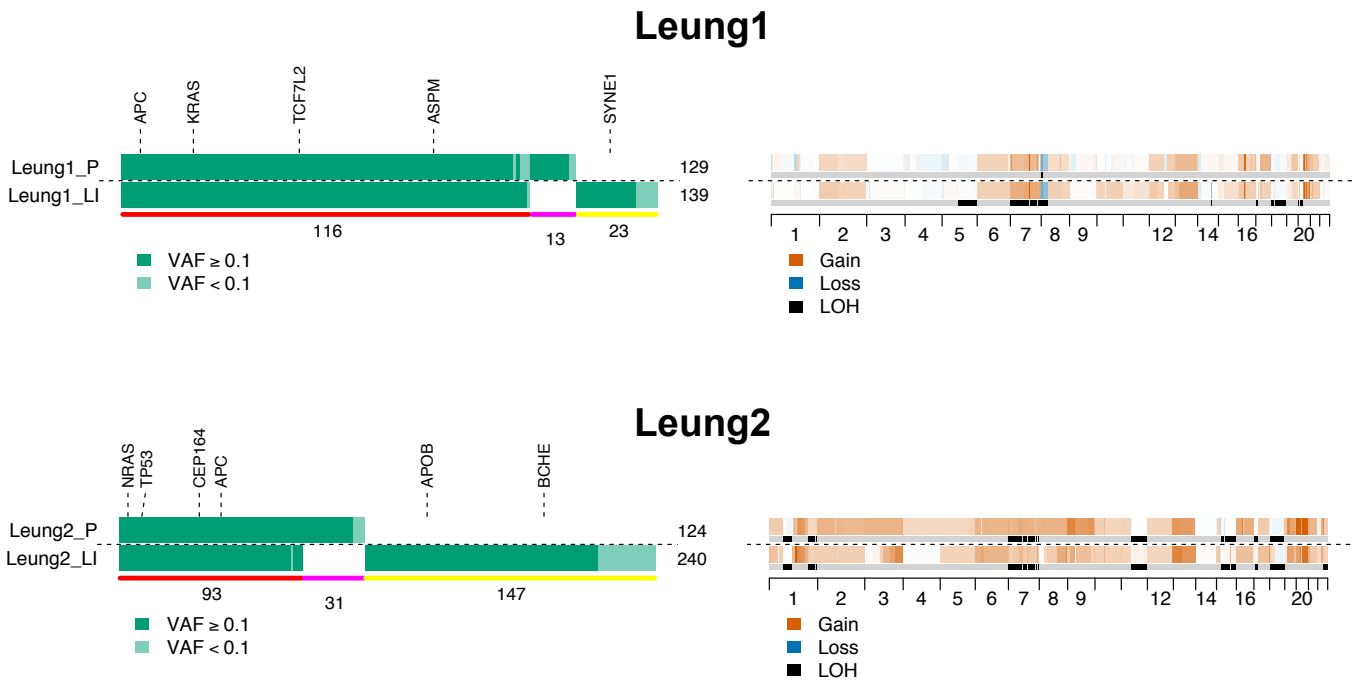


Phylogeny (SNVs+Indels)

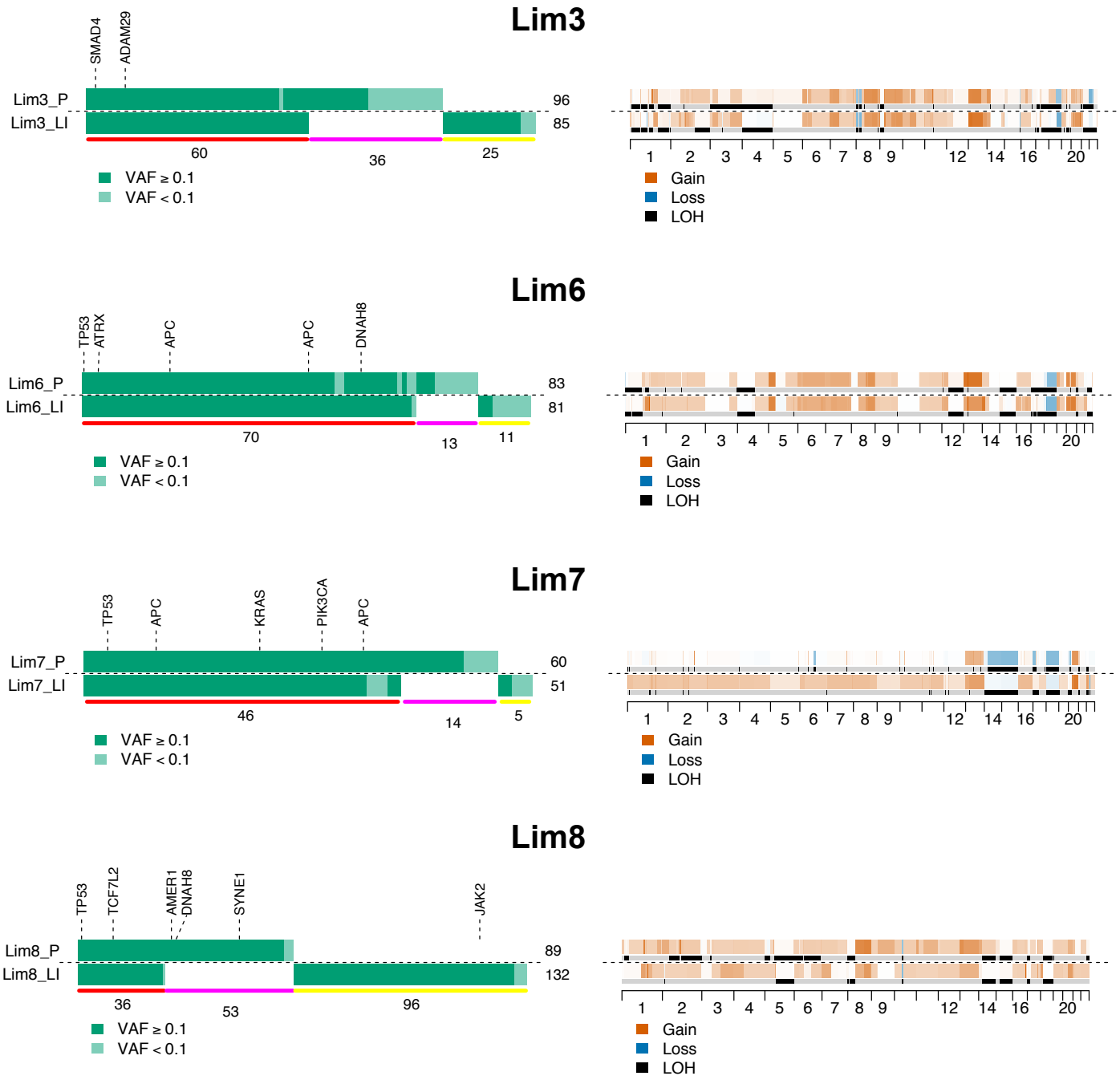


1. metsynchronous
2. no treatment

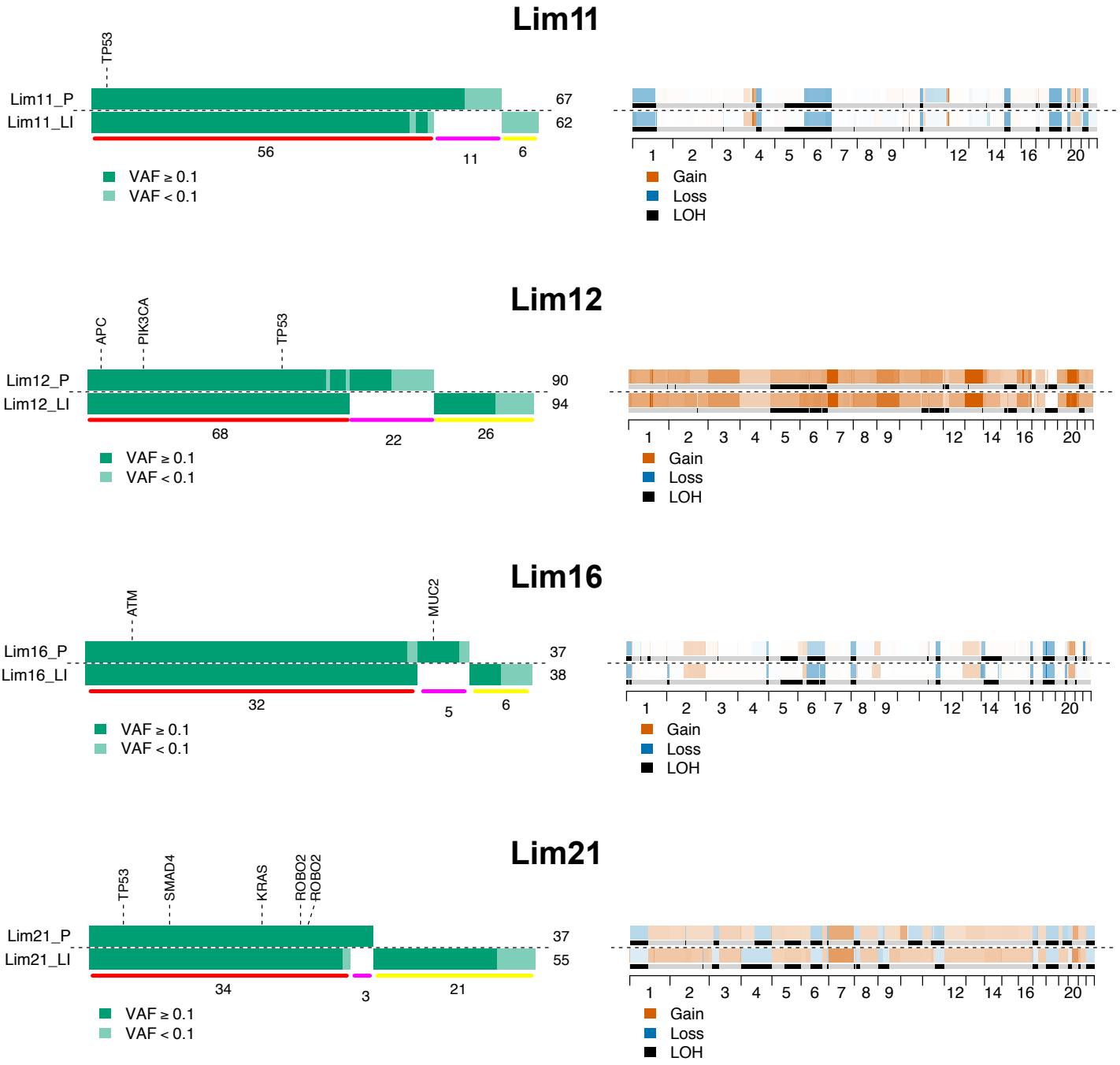
Supplementary Figure 9 - continued.



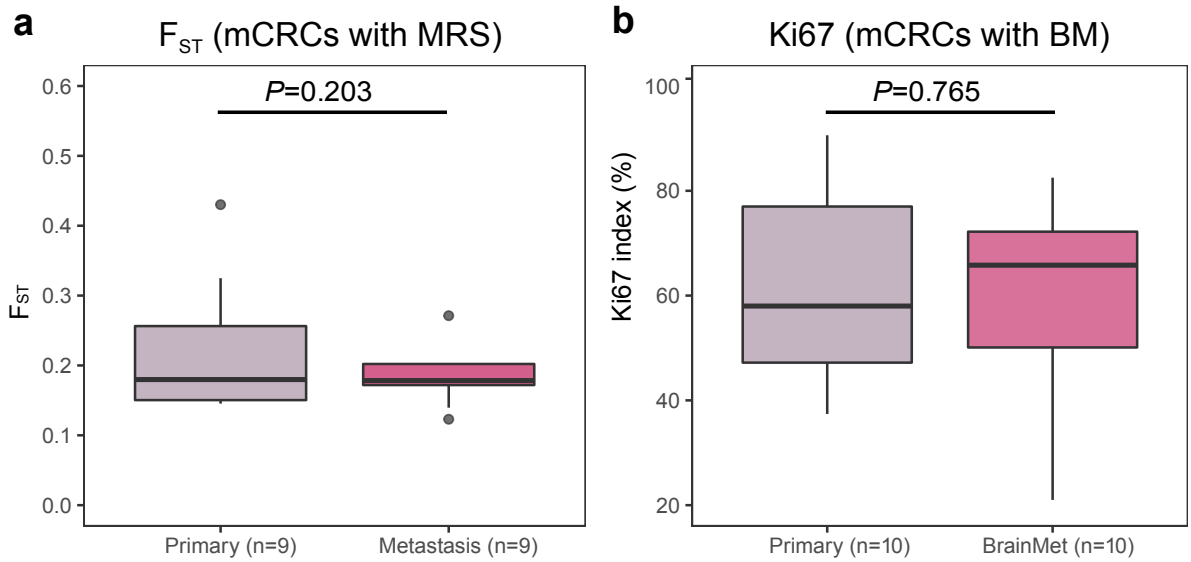
Supplementary Figure 9 - continued.



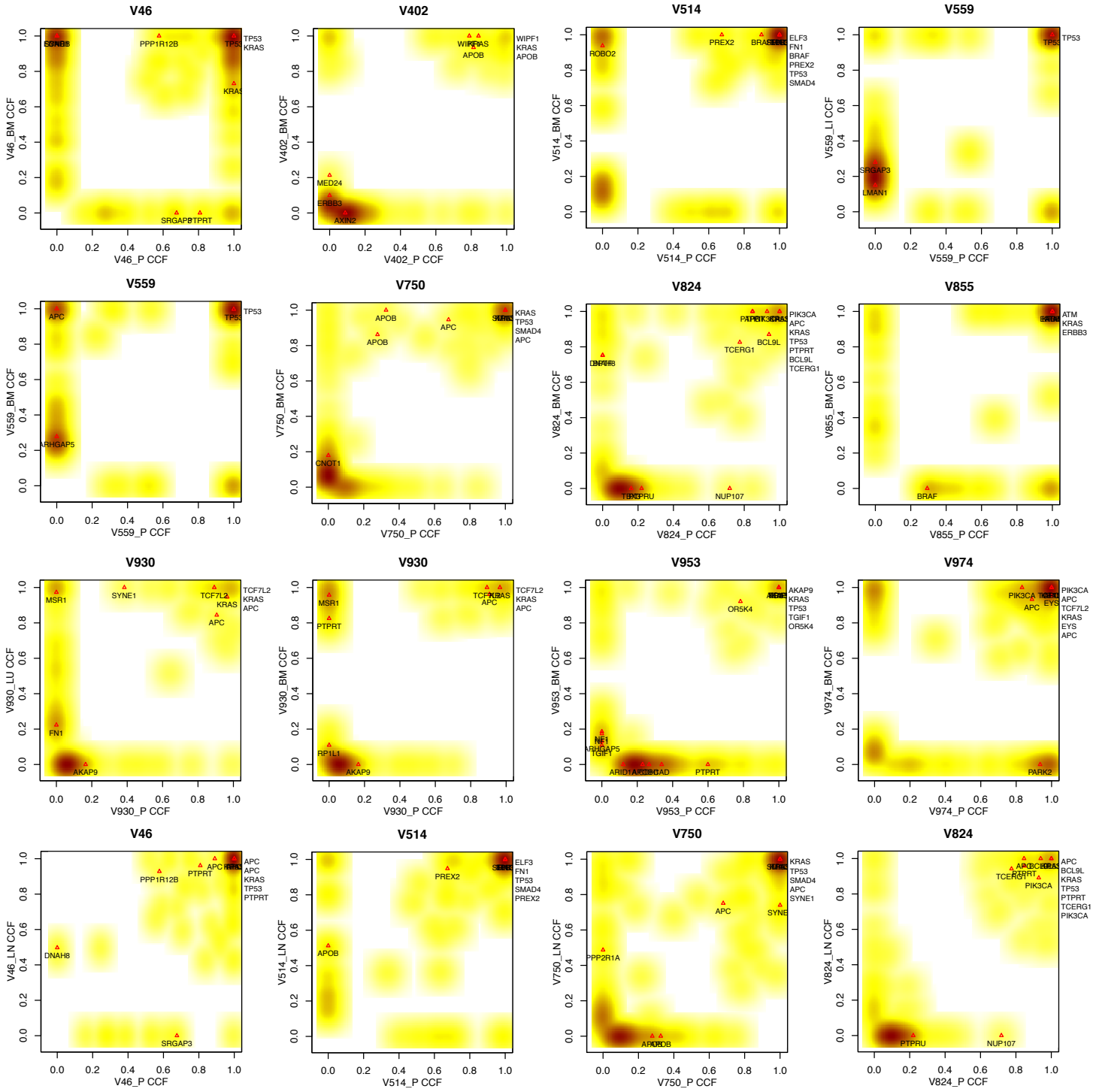
Supplementary Figure 9 - continued.



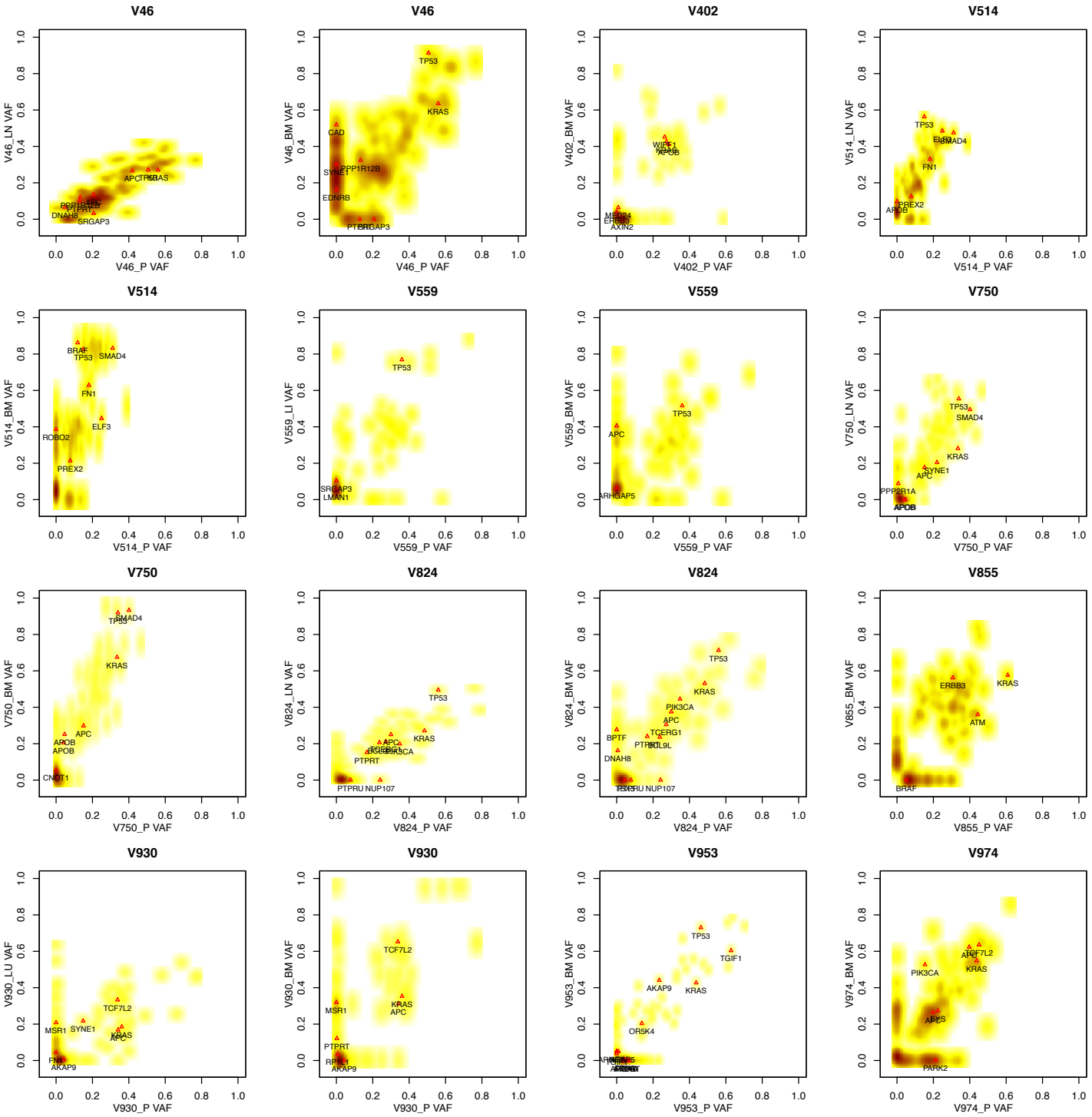
Supplementary Figure 9 - continued.



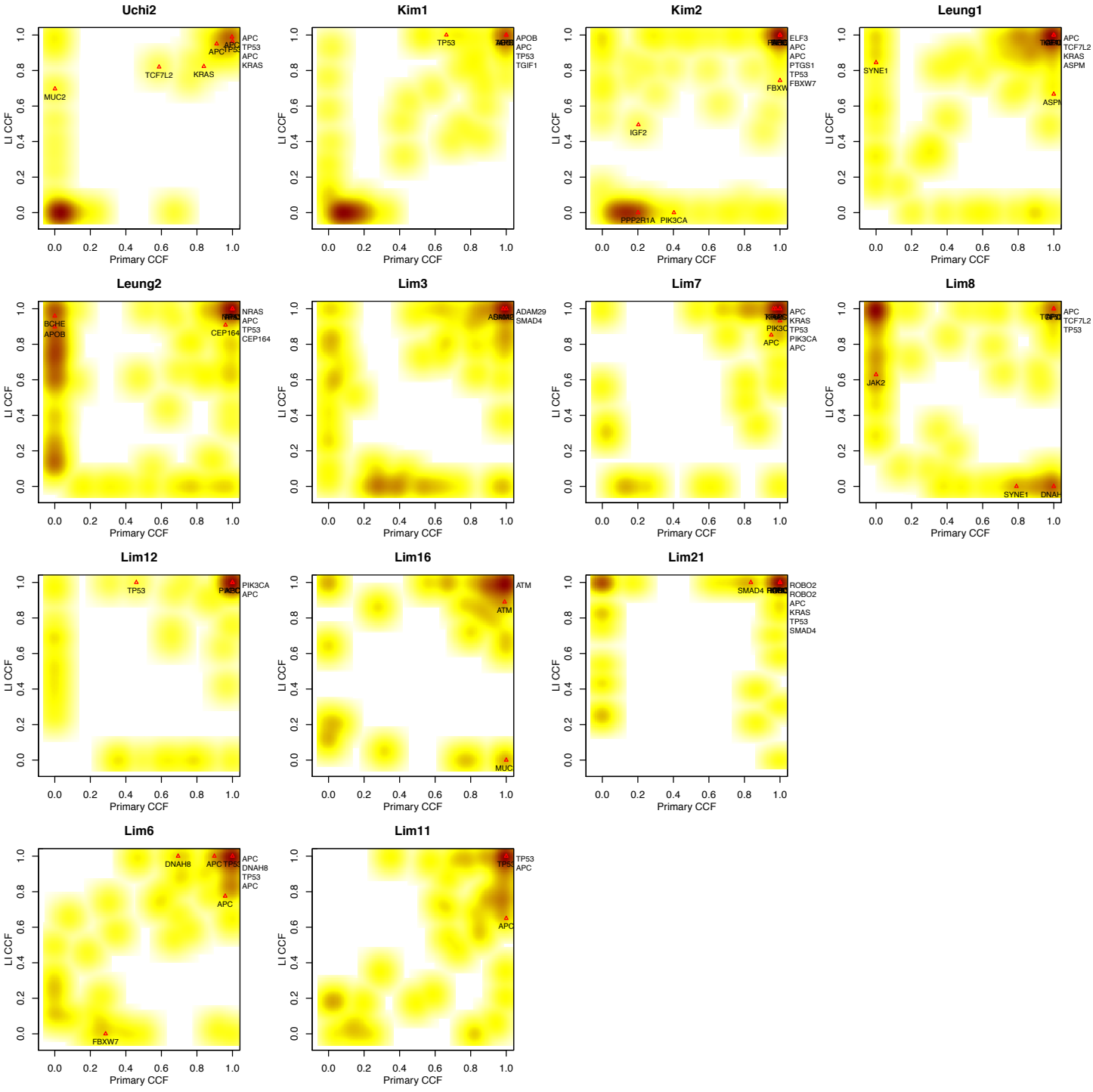
Supplementary Figure 10. F_{ST} based quantification of genetic divergence and Ki67 proliferative indices in metastatic CRCs. (**a**) F_{ST} based quantification of genetic divergence in paired primary CRCs and metastases computed based on subclonal sSNVs (merged CCF<60%) in cases with multi-region sequencing data (n=9 P/M pairs). (**b**) Ki67 proliferative indices in paired primary CRCs and brain metastases (n=10 P/M pairs). *P*-value, Wilcoxon Rank-Sum Test (two-sided). Bar, median; box, 25th to 75th percentile (interquartile range, IQR); vertical line, data within 1.5 times the IQR.



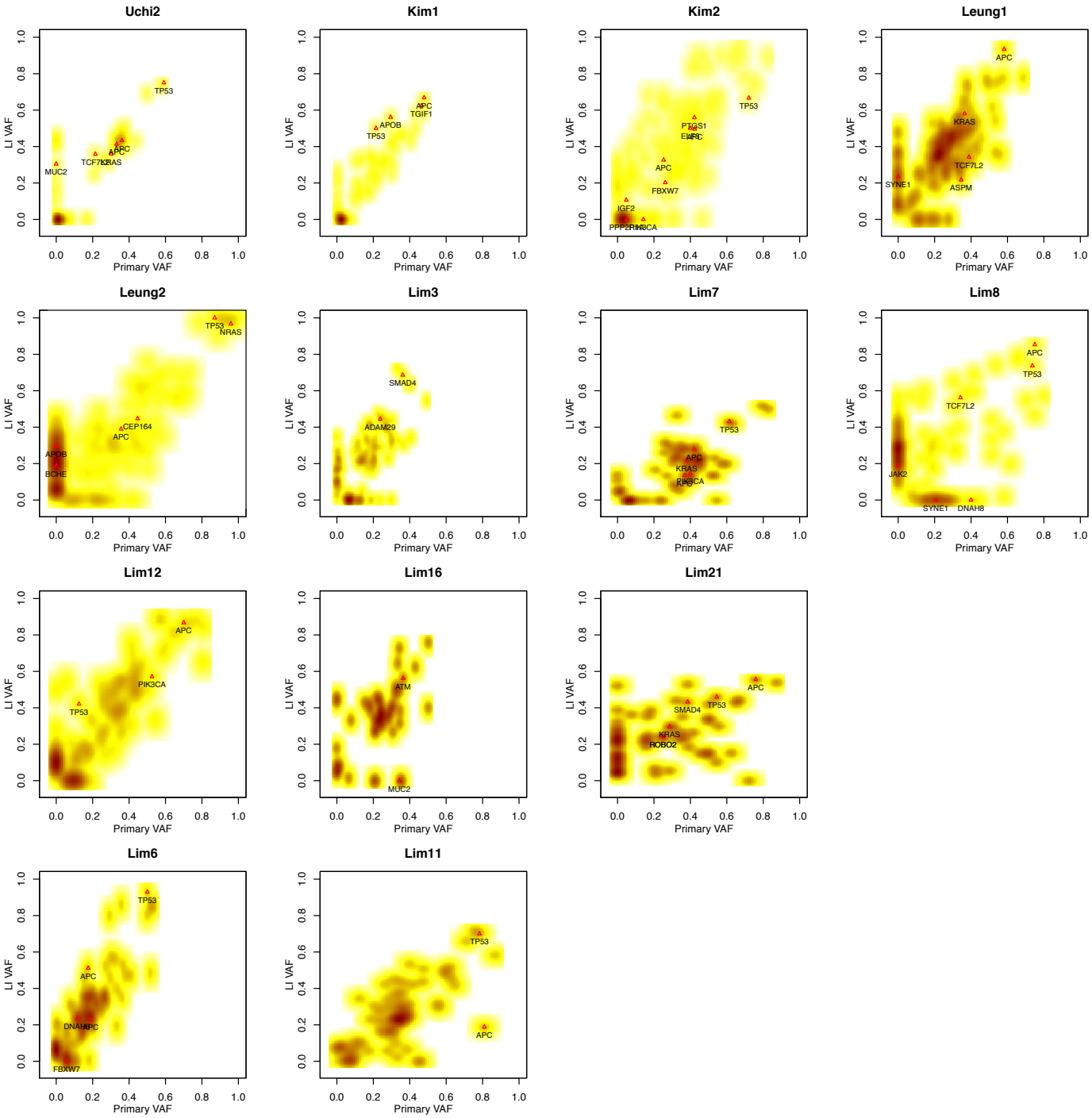
Supplementary Figure 11. Density plot of CCF estimates in paired primary CRCs and metastases in brain metastasis cohort. Merged cancer cell fraction (CCF) estimates are shown for tumors with multi-region sequencing (MRS) data. Putative CRC driver genes are labeled.



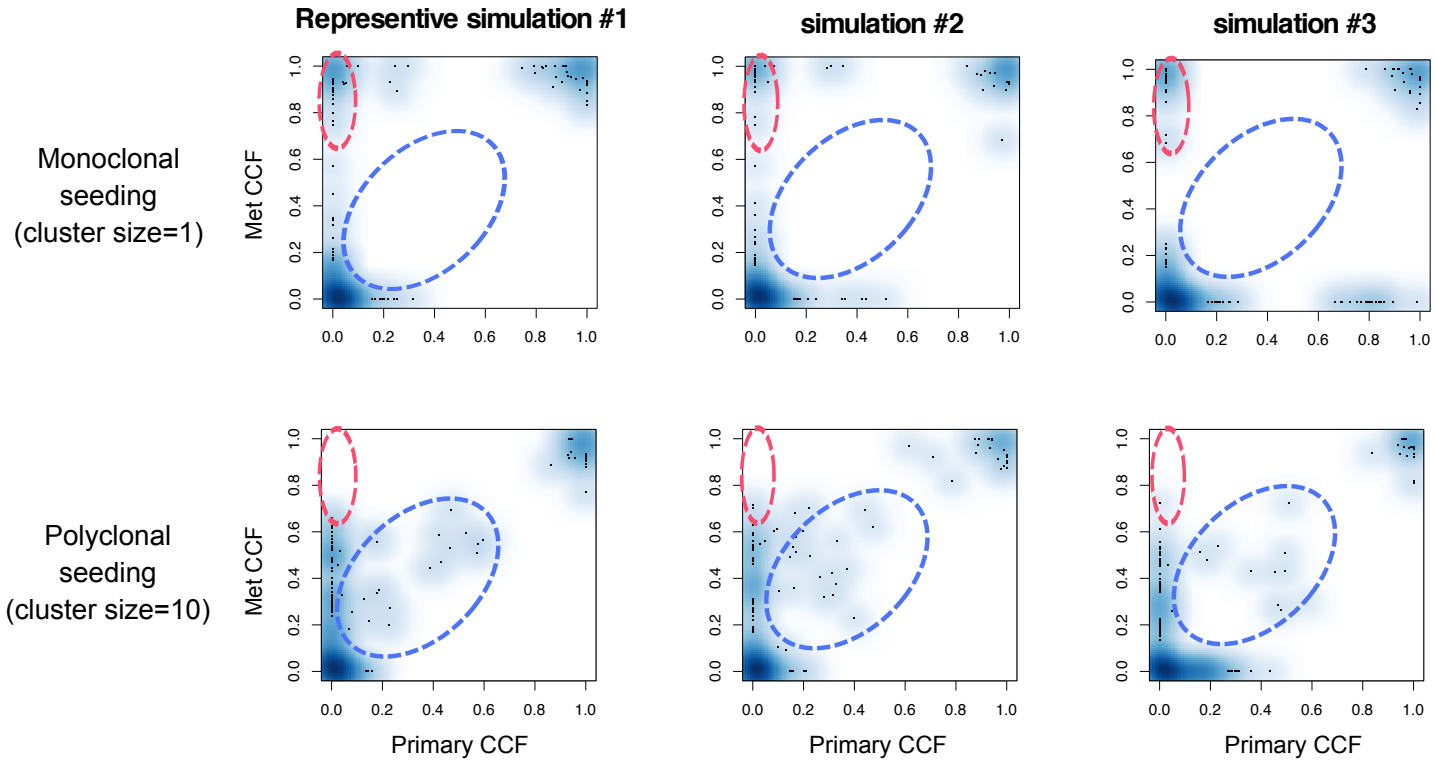
Supplementary Figure 11 - continued. Density plot of unadjusted VAF in paired primary CRCs and metastases in brain metastasis cohort. Merged VAFs are shown for tumors with multi-region sequencing (MRS) data. Putative CRC driver genes are labeled.



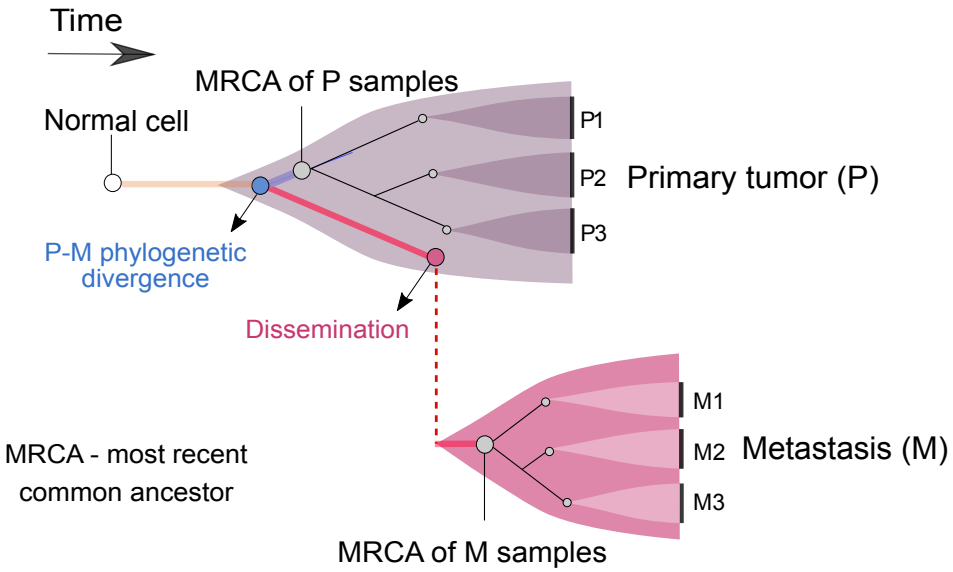
Supplementary Figure 12. Density plot of CCF estimates in paired primary CRCs and metastases in the liver metastasis cohort. Merged VAFs are shown for tumors with multi-region sequencing (MRS) data. Putative CRC driver genes are labeled.



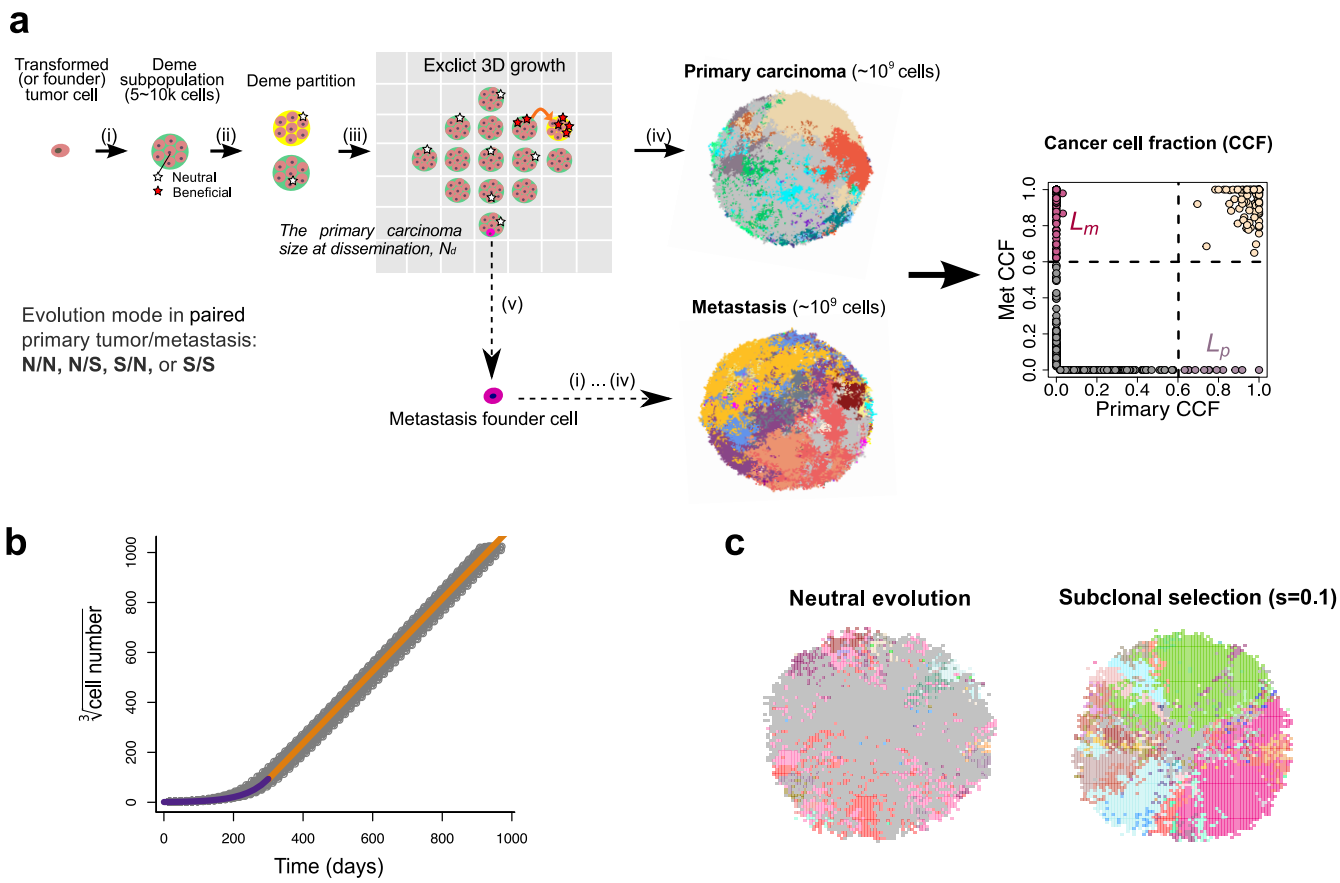
Supplementary Figure 12 - continued. Density plot of unadjusted VAF in paired primary CRCs and metastases in the liver metastasis cohort. Merged VAFs are shown for tumors with multi-region sequencing (MRS) data. Putative CRC driver genes are labeled.



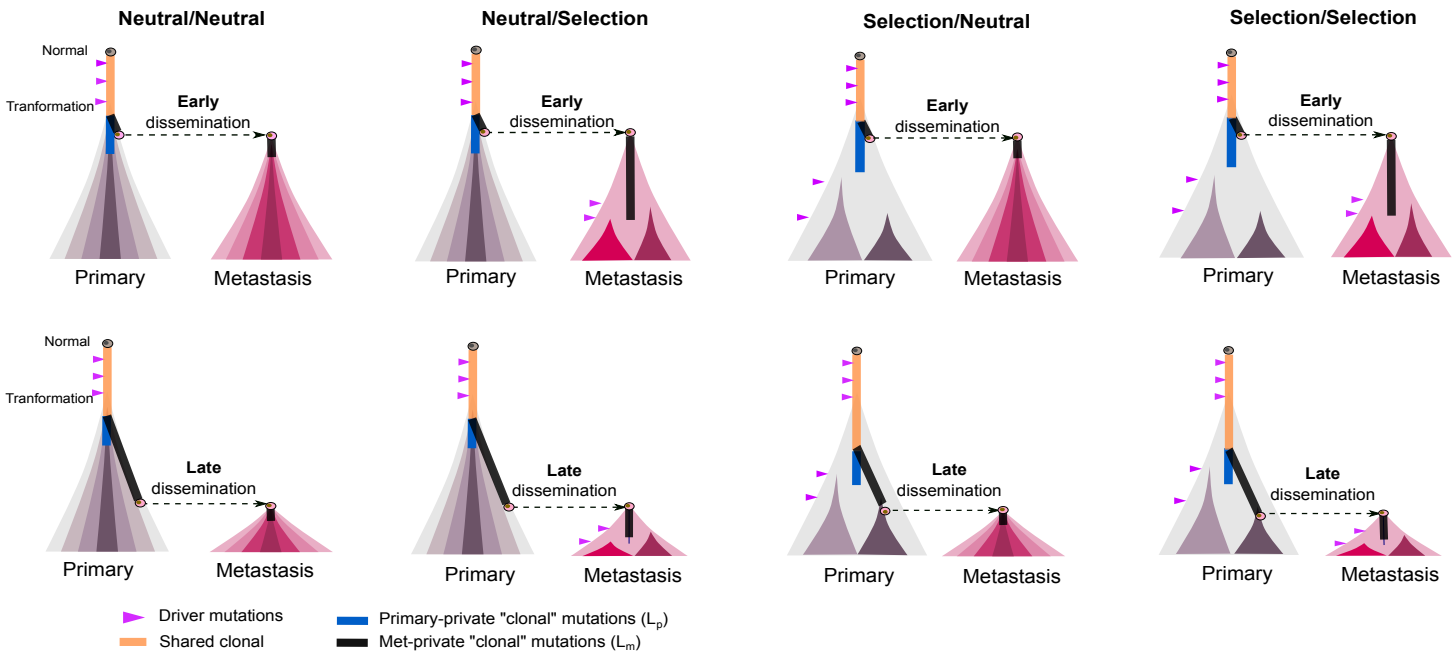
Supplementary Figure 13. Density plot of simulated CCFs in paired primary carcinoma and metastasis. The simulation framework and model are detailed in the Methods and **Supplementary Fig. 15**. During monoclonal seeding, a single cell was randomly chosen from the primary carcinoma and seeds the metastasis. Whereas during polyclonal seeding, 10 random cells sampled from whole primary tumor were chosen to seed a metastasis. The CCF density plot is shown where regions of metastasis (M)-private *clonal* mutations and primary (P)-M shared *subclonal* regions are indicated by red and blue ovals, respectively. Three exemplary CCF plots for P/M pairs are shown for monoclonal seeding (upper) and polyclonal seeding (lower), respectively.



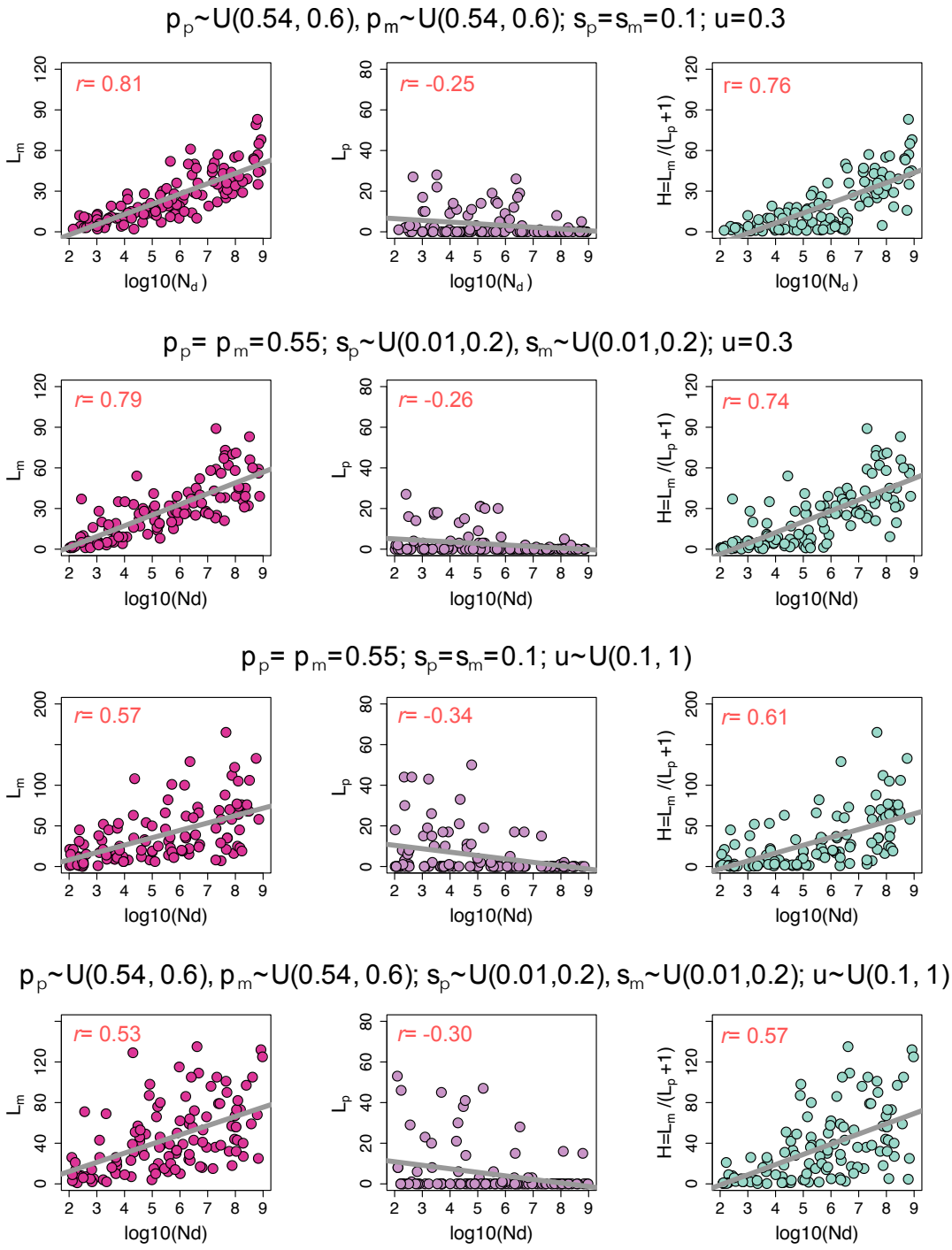
Supplementary Figure 14. Schematic illustration of the distinction between the time of P-M phylogenetic divergence and the actual time of dissemination. The time of metastatic dissemination may occur later than the time of phylogenetic divergence between primary tumor and metastasis. This is a well- known phenomenon in population genetics. Viewing the genealogical process backward, one can envision that some time is required for the metastasis-founding cell to coalesce to the common ancestor present in bulk primary tumor sequencing data.



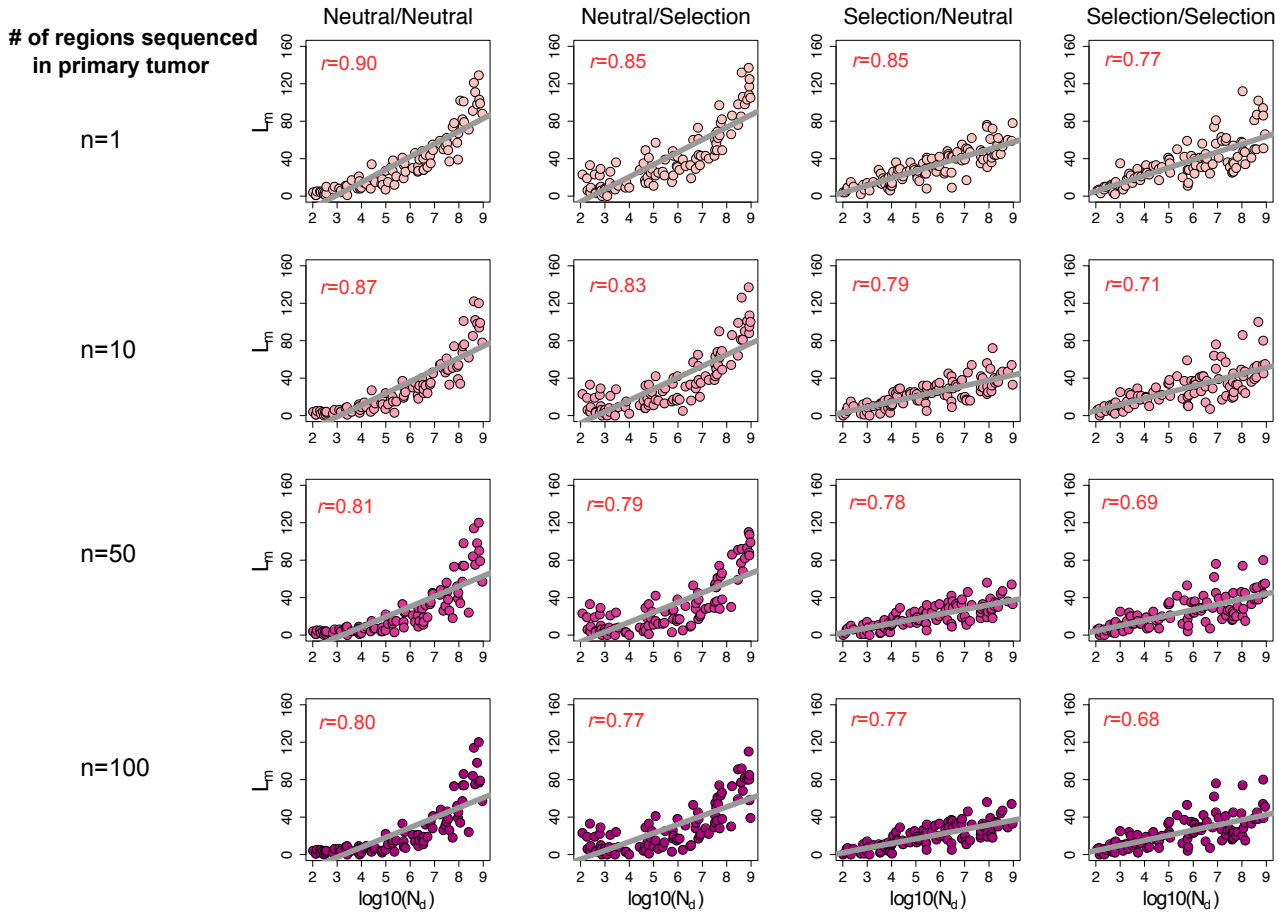
Supplementary Figure 15. A 3-D spatial-agent based model of tumor growth. (**a**) Tumor growth is simulated via the expansion of deme subpopulations (mimicking the glandular structure in primary colorectal cancers (CRCs) and metastases each containing 5-10K cells [3]) within a defined 3-D cubic lattice according to explicit rules dictated by spatial constraints, where cells within each deme are well-mixed and grow via a stochastic branching (birth-death) process, as previously described (Methods). The metastasis is assumed to be seeded by a single disseminated cell derived from the periphery of primary CRC when the tumor size is N_d. Metastatic growth follows the same spatial constraints and growth rate as in the primary CRC [5]. The final sizes of both the primary carcinoma and metastasis is ~10^9 cells (~2×10^5 demes). Different evolutionary scenarios in P/M pairs were simulated including Neutral/Neutral (or N/N), Neutral/Selection (or N/S), Selection/Neutral (or S/N) and Selection/Selection (or S/S). Here selection is modeled by assuming a constant beneficial mutation rate (denoted by u_b, ~10^{-5} per cell division [7]) that alters the cell birth/death probability according to the selection coefficient (denoted s). By simulating the acquisition of random mutations (neutral or beneficial), tracing the mutation genealogy of each cell as the tumor expands and subsequently spatially sampling and sequencing the 'final' virtual tumor as is done experimentally after resection or biopsy, we obtain the variant allele frequencies (VAF) and cancer cell fraction (CCF). Using this framework, we can evaluate the impact of the timing of dissemination, the bottleneck effect and subclonal selection on the number of primary-private clonal mutations (L_p), metastasis-private clonal mutations (L_m) and the metric $H=L_m/(L_p+1)$. (**b**) The deme subdivision model of peripheral growth results in exponential growth initially followed by power-law growth (cell number $N_t \sim t^3$) subsequently due to stringent spatial constraints in a relatively large tumor. Here the birth/death probability ratio, p/q=0.55/0.45≈1.2 at each cell generation was used, resulting in ~250 cell generations (or ~1000 days or ~3 years assuming 4 days for each cell cycle [8]) from recent founder cell to diagnosis of malignant primary tumor (~10^9 cells). (**c**) Representative 2D clone map simulated under neutral evolution (s=0) or stringent selection (s=0.1). Each dot indicates a deme and each color indicates a unique clone, where only mutations with VAF>0.4 in each deme are shown. Selection promotes subclonal expansions and genetic divergence between distant tumor regions.



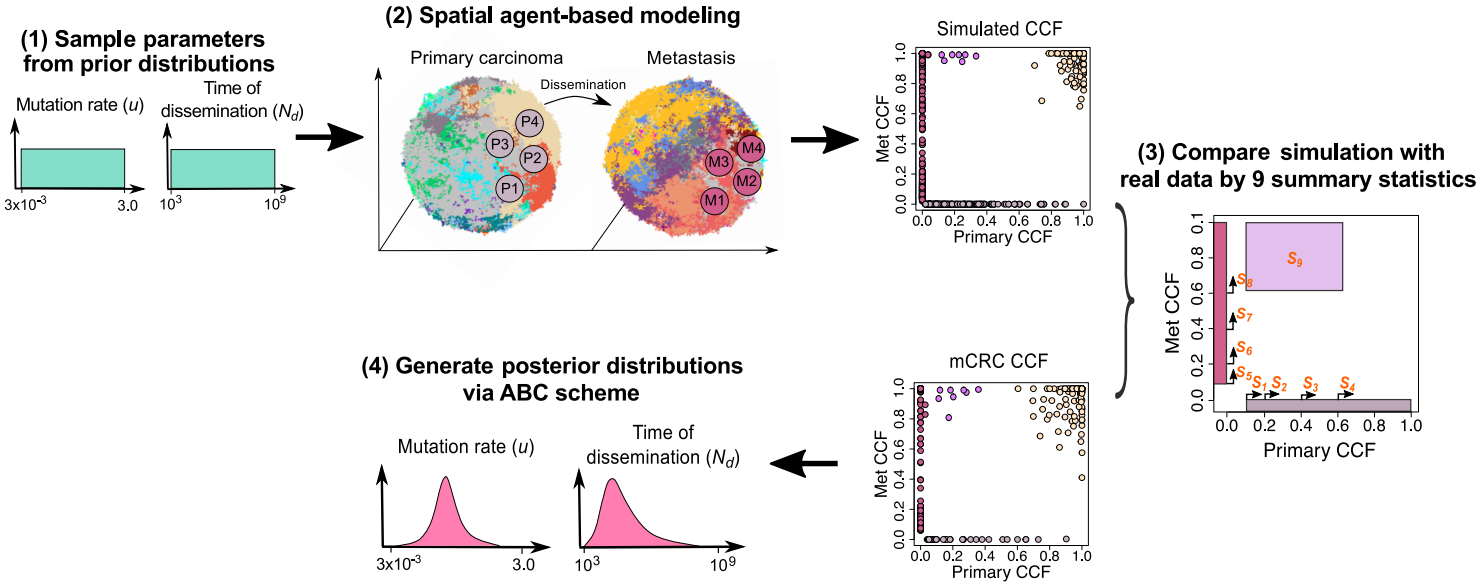
Supplementary Figure 16. The modes of tumor evolution in paired primary CRCs and metastases. Different modes of primary tumor and metastatic growth are considered, corresponding to four evolutionary scenarios, namely Neutral/Neutral, Neutral/Selection, Selection/Neutral and Selection/Selection. Stringent selection results in an increased number of P or M-private high-frequency mutations in the corresponding lesion. Some of these high-frequency mutations may be clonal in localized tumor regions. When the primary tumor grows in an effectively neutral fashion (Neutral/Neutral or Neutral/Selection), later dissemination gives rise to a larger number of metastasis-private high-frequency (or clonal) mutations (large L_m) due to the more stringent bottleneck effect. When the primary tumor is subject to subclonal selection, the primary carcinoma may possess many private high-frequency (even clonal) mutations (resulting in large L_p) if dissemination occurs early. We employ a spatial (3-D) agent-based tumor growth model to simulate these four scenarios and systematically investigate the contribution of selection and the timing of metastasis on the number of primary-private clonal mutations (L_p) and metastasis-private clonal mutations (L_m).



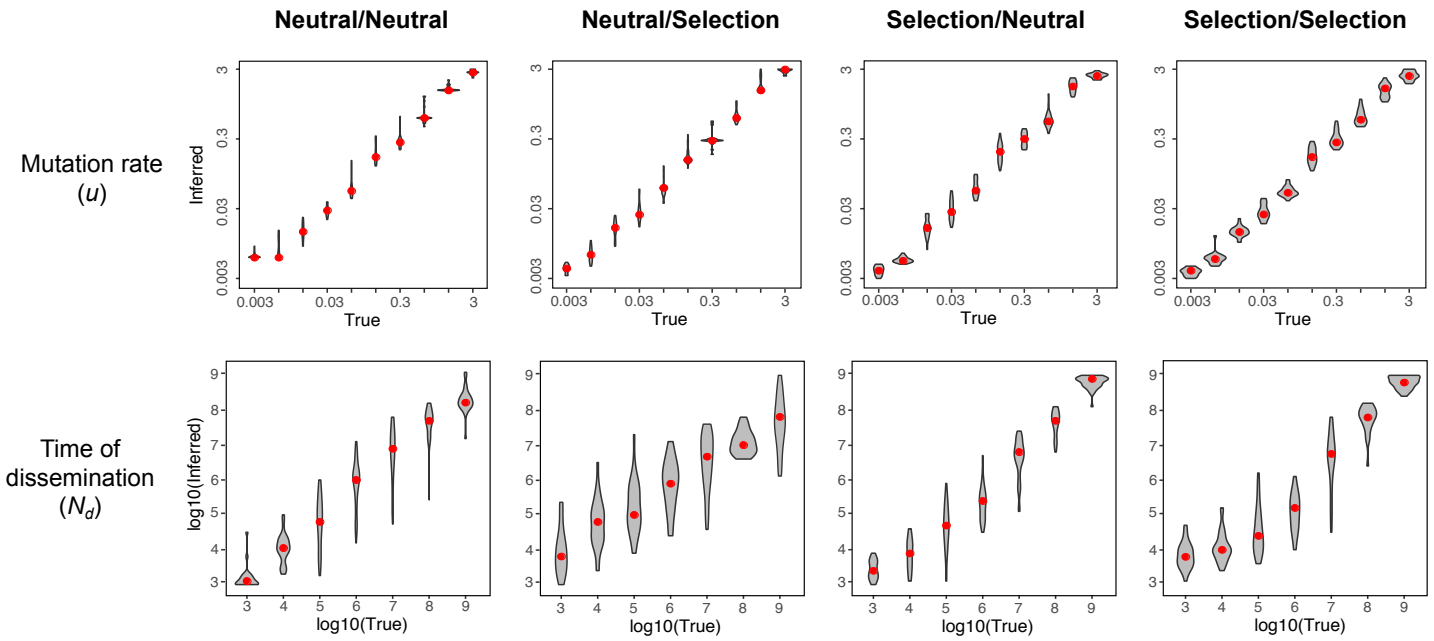
Supplementary Figure 17. Correlation between L_p, L_m and H and the timing of dissemination (N_d) based on spatial computational modeling of tumor growth. L_p and L_m correspond to the number of primary tumor-private and metastasis-private *clonal* sSNVs (CCF>60% in one site & CCF<1% in the other site), respectively. As in **Fig. 4b**, values for each parameter combination (row) are based on 100 paired primary tumors and metastases (n=100 P/M pairs) simulated within the spatial agent-based model and the timing of dissemination, N_d, was randomly sampled from a uniform distribution log10(N_d)~U(2,9). Here we evaluate the correlation (Pearson's r) between L_p, L_m and H with N_d while varying the cell birth probability p (death probability q=1-p), selection coefficient s and mutation rate u. p_p and p_m correspond to the cell birth probability during growth of the primary tumor and metastasis, respectively, while s_p and s_m are the selection coefficients. u is neutral passenger mutation rate per cell division at exonic regions. The parameter values used are as indicated.



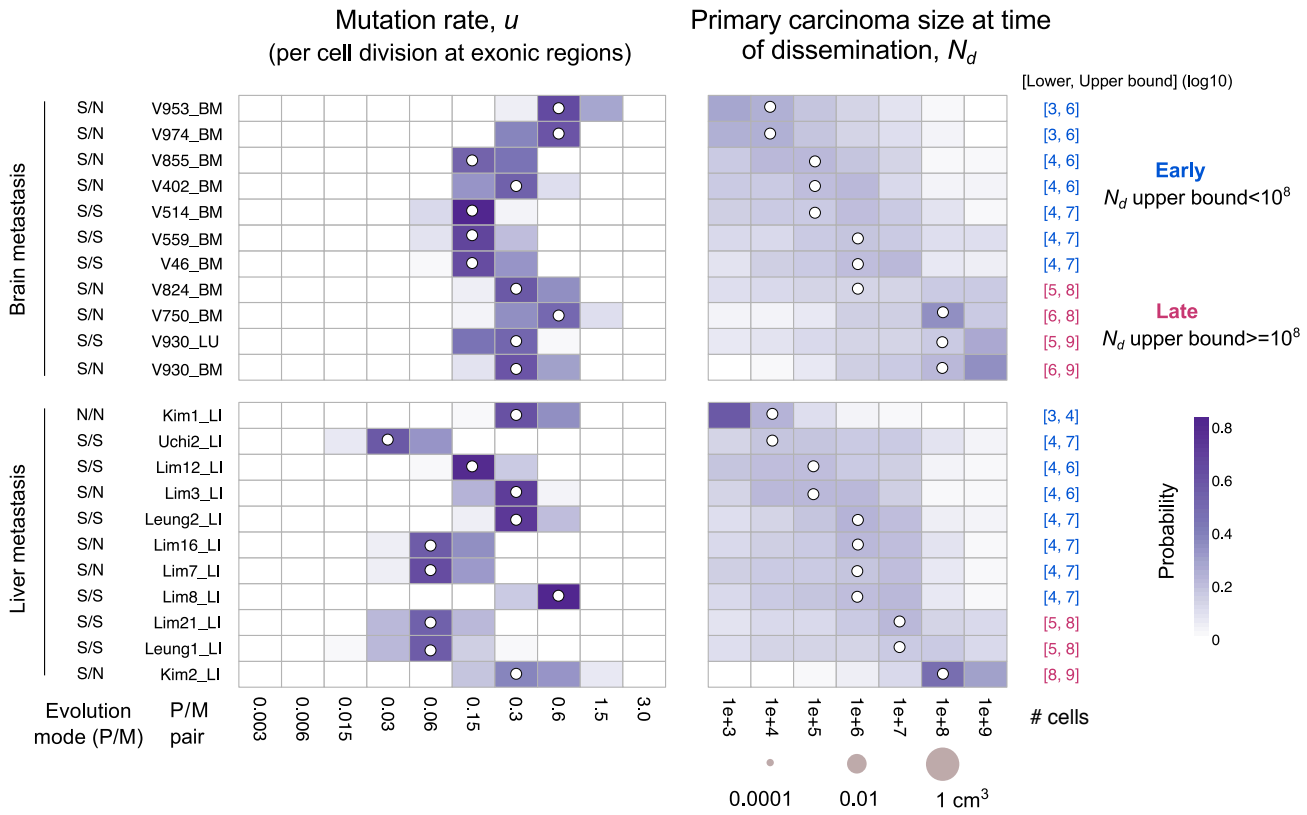
Supplementary Figure 18. Correlation between L_m and N_d under variable sampling intensities. Varied numbers of regions (n=1, 10, 50 and 100, each ~10^6 cells) from each primary tumor were randomly sampled and sequenced, and the sSNVs with CCF>1% in each region were considered. Here L_m is defined as the number of M-clonal sSNVs with CCF>60% in whole metastasis while absent in any of the sampled regions in primary tumor. As in **Fig. 4b**, values for each model and sampling scenario are based on 100 paired primary tumors and metastases (n=100 P/M pairs) simulated within the spatial agent-based model and the timing of dissemination, N_d, was randomly sampled from a uniform distribution log10(N_d)~U(2,9). Pearson's r is reported.



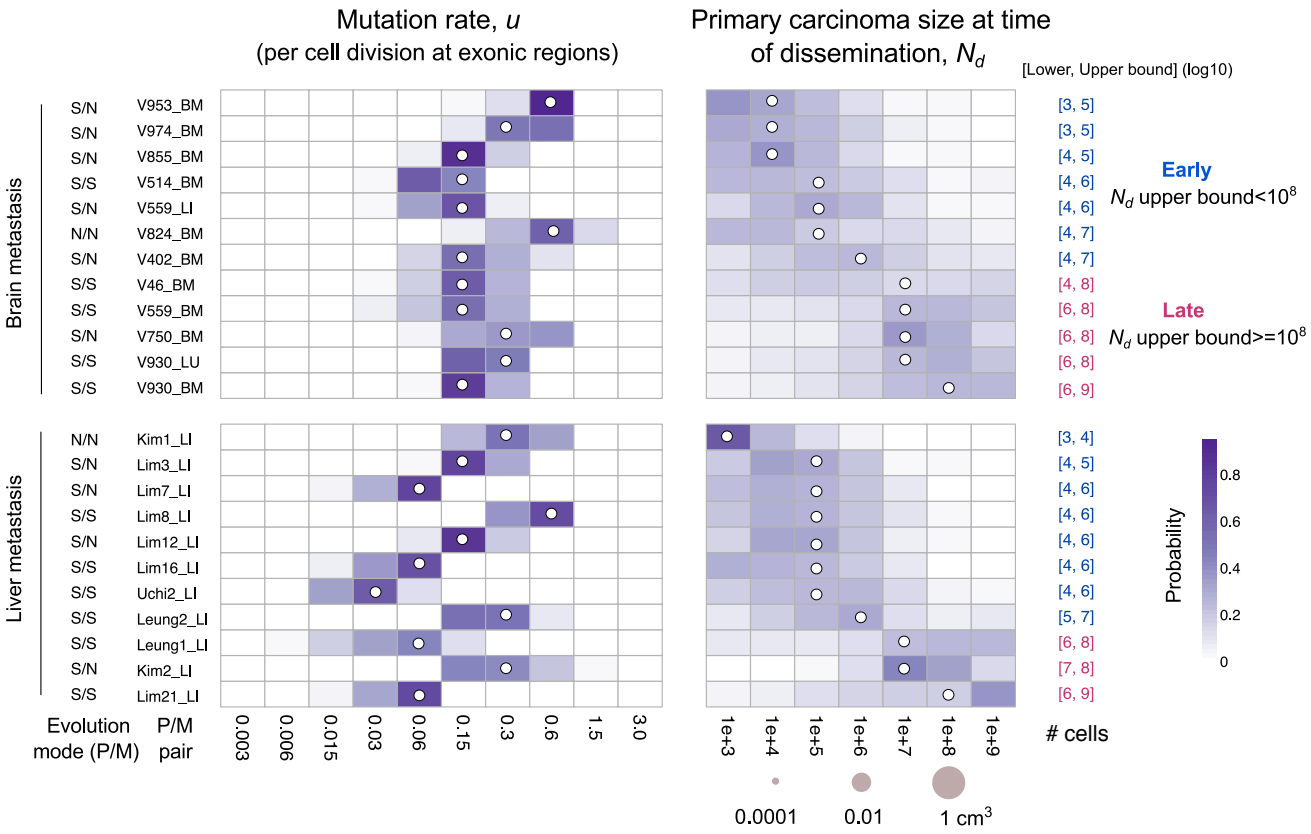
Supplementary Figure 19. Schematic of SCIMET (Spatial Computational Inference of MEtastatic Timing). The parameters we sought to infer via Approximate Bayesian Computation (ABC) [2] are $\boldsymbol{\theta}(u, N_d)$ where u is the mutation rate (per cell division in exonic regions) and N_d is the cell number in primary colorectal carcinoma at the time of dissemination. **Step (I):** Randomly sample u' and N_d' from two independent prior distributions of discrete values: $u \sim U\{0.003, 0.006, 0.015, 0.03, 0.06, 0.15, 0.3, 0.6, 1.5, 3\}$ and $N_d \sim U\{10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9\}$, respectively. **Step (II):** Simulate spatial tumor growth to obtain 'virtual' paired primary tumors and metastases (P/M pairs) (~10^9 cells each) under each of four evolutionary modes: Neutral/Neutral, Neutral/Selection, Selection/Neutral and Selection/Selection, where the selection coefficient is s=0.1 for selection and s=0 for neutral evolution. Four tumor regions (~10^6 cells) are sampled in each P/M, the sequencing process is simulated and nine summary statistics, $\boldsymbol{S'} = \{S_1, S_2, \ldots, S_9\}$ are computed based on the CCF (or merged CCF for multi-region sequencing data). **Step (III):** Compare the simulated summary statistics $\boldsymbol{S'}$ to the observed $\boldsymbol{S}$ for a given P/M pair by computing the distance between $\boldsymbol{S'}$ and $\boldsymbol{S}$, $d(\boldsymbol{S'}, \boldsymbol{S}) < \varepsilon$. **Step (IV):** If $d(\boldsymbol{S'}, \boldsymbol{S}) < \varepsilon$, accept the prior value of u' and N_d' and repeat steps (I)-(III). In our paper, we use a common variation of ABC [2,4]. Rather than using a fixed threshold, ε, we sort all distances calculated in by $d(\boldsymbol{S'}, \boldsymbol{S})$ (Step 3), and accept the $\boldsymbol{\theta'}$ that generated the smallest $100\times\eta$ percent distances. 70,000 P/M pairs were simulated under each of the four evolutionary scenarios. In total, 280,000 P/M pairs were simulated. η=0.01 was used thus the posterior is composed of 70000×0.01=700 data points. The ABC procedure is performed using the R package *abc* [6]. To select the tumor evolution model (N/N, N/S, S/N or S/S) in paired primary CRCs/metastases, 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 gives the probability of each model based on the posterior distribution. The model with the highest posterior probability was selected.



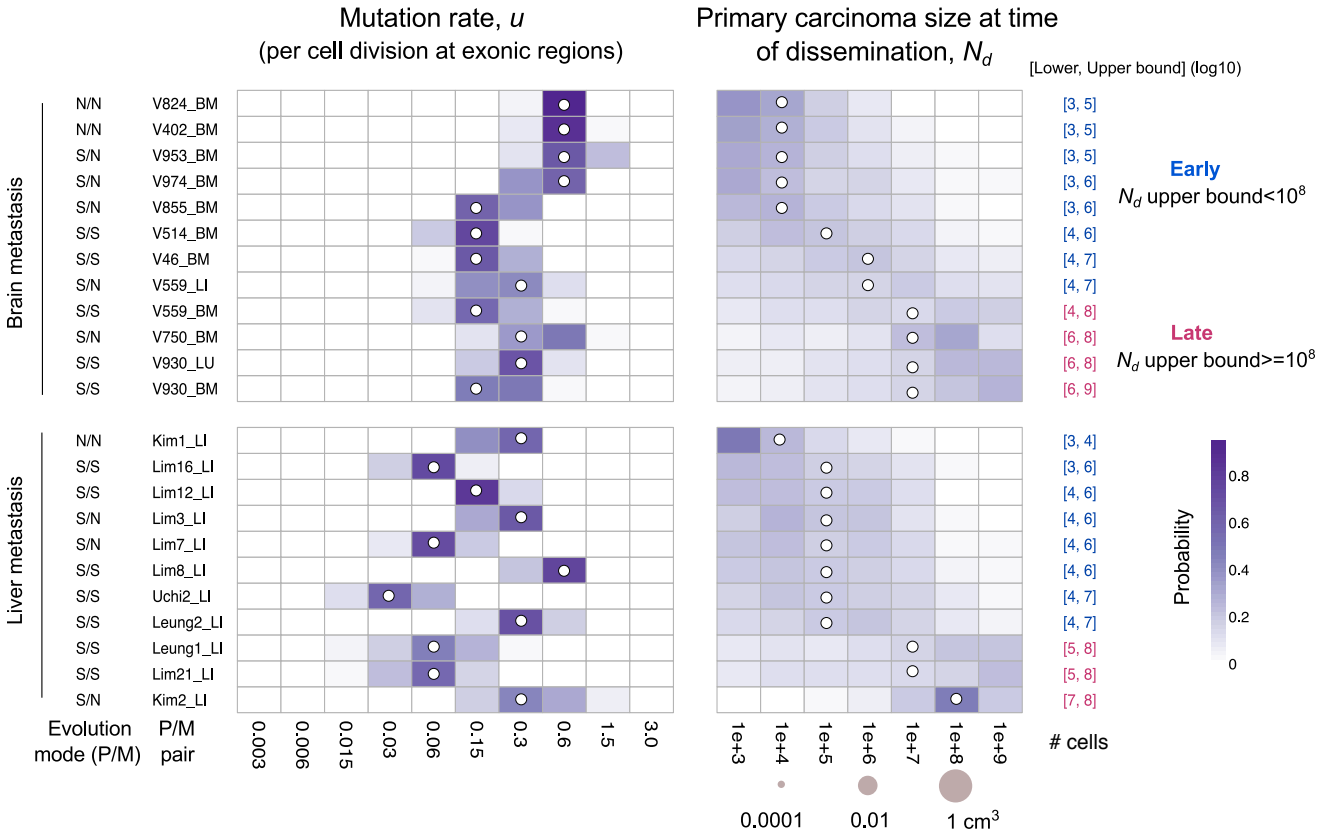
Supplementary Figure 20. Robust recovery of parameters via SCIMET on synthetic data. We evaluated the ability of SCIMET to recover the parameters *u'* and N_d*'* from synthetic data (virtual P/M pairs) via cross-validation. Scatterplots comparing the inferred (mean posterior value) versus true values are shown based on 200 Monte Carlo samplings under each combination of evolutionary modes. These results highlight the robustness of SCIMET across all evolutionary modes. Red circle, median of 200 simulations.



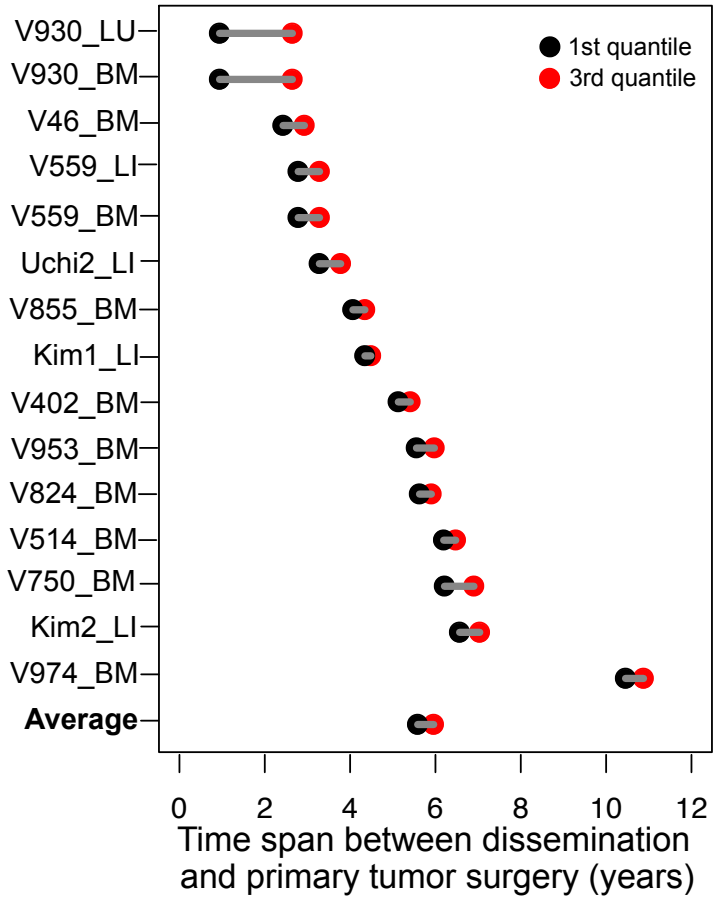
Supplementary Figure 21. Patient-specific inference of the mutation rate and timing of metastasis via SCIMET assuming a high birth/death probability ratio. Parameters are the same as **Fig. 5a** except that the birth/death probability ratio at each cell generation for the founding lineage in the primary carcinoma is p/q=0.6/0.4=1.5 (rather than p/q=0.55/0.45≈1.2), thus the tumor growth rate here is higher than assumed in **Fig. 5a**. All the metastases classified as *late* dissemination in **Fig. 5a** were also classified as *late* dissemination when assuming higher tumor growth rate as illustrated here. Two additional metastases (V824_BM and Lim21_LI) classified as *early* dissemination in **Fig. 5a** were classified as *late* dissemination here. These results highlight the robustness of SCIMET based inference to increases in the cell birth/death ratio.



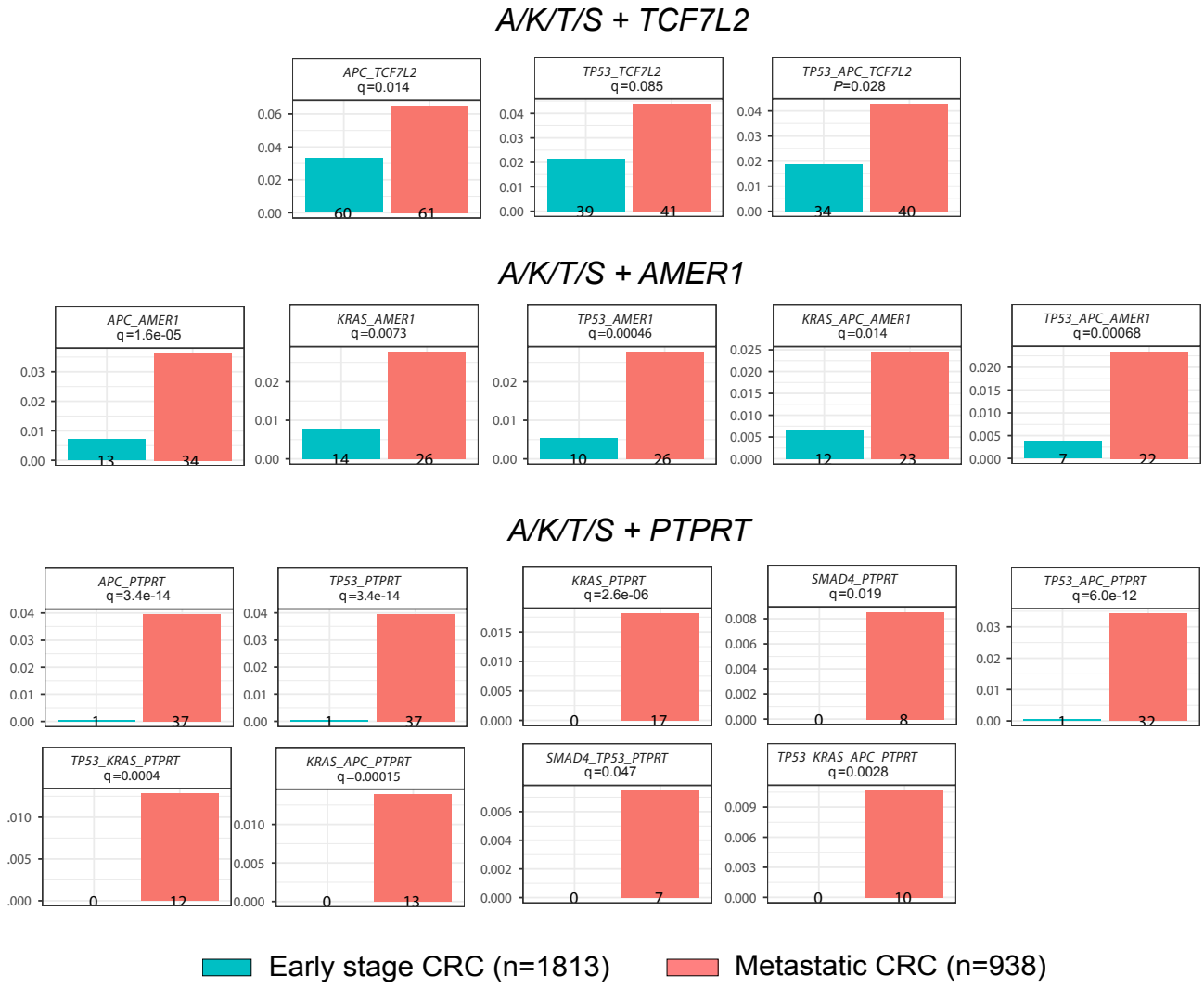
Supplementary Figure 22. Patient-specific inference of the mutation rate and timing of metastasis via SCIMET assuming collective dissemination (cluster size=10 cells). To model collective dissemination by localized cell cluster from primary tumor front, a cluster of 10 cells were randomly sampled as metastasis founder cells from a peripheral deme during the expansion of primary carcinoma. All other parameter values are the same as in **Fig. 5a**. All the metastases classified as *late* dissemination in **Fig. 5a** were also classified as *late* dissemination when assuming collective dissemination as illustrated here. Three additional metastases (V46_BM, V559_BM and Lim21_LI) classified as *early* dissemination in **Fig. 5a** were classified as *late* dissemination here. These results highlight the robustness of SCIMET based inference irrespective of whether collective cell dissemination or single cell seeding is assumed.



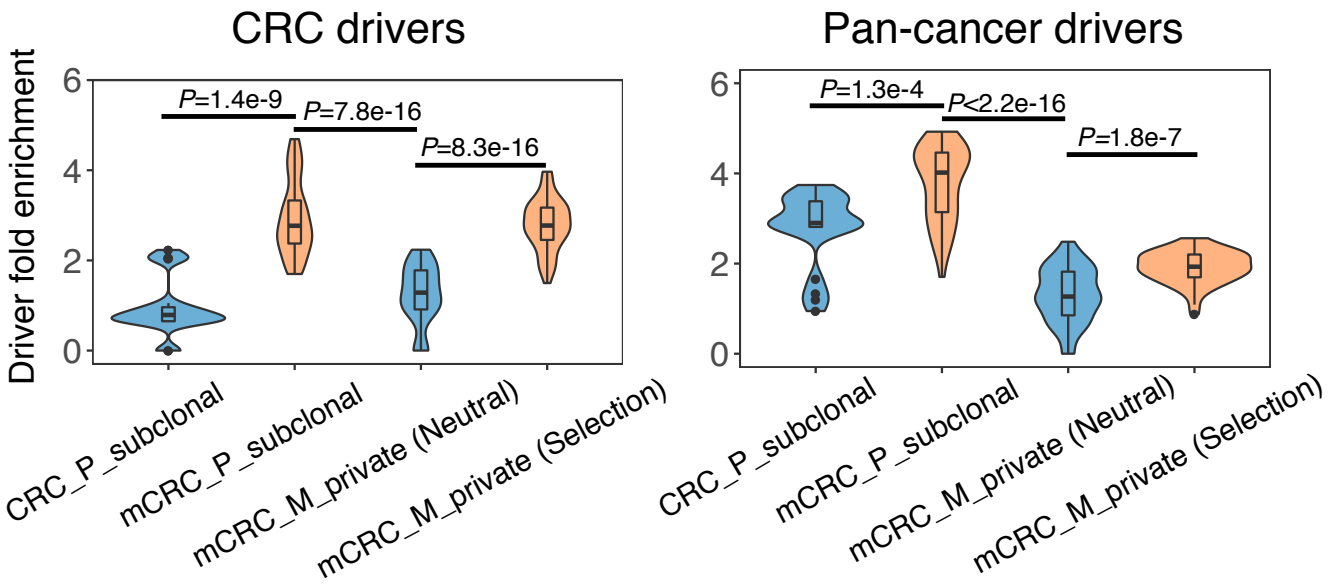
Supplementary Figure 23. Patient-specific inference of the mutation rate and timing of metastasis via SCIMET assuming single-region sequencing data. To evaluate the impact of limited spatial sampling, only one tumor region (~10^6 cells) was sampled and "sequenced" from each of paired virtual primary tumor and metastasis within SCIMET. All parameter values are the same as in **Fig. 5a**. All the metastases classified as *late* dissemination in **Fig. 5a** were similarly classified as *late* dissemination based on single-region simulation data as illustrated here. Two metastases (V559_BM and Lim21_LI) classified as *early* dissemination in **Fig. 5a** were classified as *late* dissemination here. This is anticipated since single-region sampling results in a larger number of metastasis-private *clonal* mutations (larger L_m and larger H) compared with multi-region sequencing, such that the timing of dissemination would be overestimated in accordance with the positive correlation between L_m or H and N_d. Overall, these results highlight the robustness of SCIMET and its utility for analyzing single region sequencing data.



Supplementary Figure 24. Estimation of the time span between dissemination and surgical resection of the primary tumor (years). We used the inferred primary tumor size at time of dissemination (N_d) by SCIMET and the recorded size of the primary tumor at the time of surgery to estimate the time elapsed between dissemination and surgery by employing an approximate growth function for our 3D spatial tumor growth model (Eq(S7) in Supplementary Note). The 1st and 3rd quantile of time span shown here correspond to the conversion from the 1st and 3rd quantile of N_d, respectively.



Supplementary Figure 25. Enrichment of canonical driver gene modules in metastatic versus early stage CRCs. We evaluated the enrichment of a subset of canonical 'core' drivers (*APC*, *KRAS*, *TP53* or *SMAD4; A,K,P,S*) plus recurrent mutations in candidate metastasis drivers (*AMER1, ATM, BRAF, PIK3CA, PIK3R1, PTPRT, TCF7L2*) identified in the mCRC cohort in an independent cohort of metastatic versus early stage CRC patients. Fisher's exact test was performed for all possible combinations of putative 'core' modules plus putative metastasis driver genes to determine the enrichment amongst metastatic versus early-stage CRCs (**Supplementary Table 8,** Methods). Significant (Two-sided Fisher's exact test with Benjamini–Hochberg adjustment for multiple testing; q-values<0.1) combinations based on the core canonical backbone plus one additional putative metastasis-associated driver gene are shown.



Supplementary Figure 26. Driver enrichment analysis stratified by the mode of tumor evolution. Violin plots illustrate the driver fold enrichment for shared, primary-private, and metastasis-private *clonal* non-silent sSNVs based on known CRC or pan-cancer 'drivers'. Analyses are similar to **Fig. 2c**, except that tumors were stratified by the mode of evolution namely neutral evolution or stringent selection (identified via ABC, **Fig. 5a**) (Methods). Amongst metastatic lesions (n=23), 10 were found to exhibit neutral evolution and 13 exhibited stringent selection. Amongst primary CRCs (n=21), the majority (19/21) evolved under stringent selection. We therefore compared primary tumors evolving under stringent selection (n=19) to an early stage CRC cohort (n=6) [1], where 4/6 exhibited patterns consistent with neutral evolution focusing on high-frequency non-silent mutations (CCF>20%) in primary tumors and metastases since high-frequency mutations are anticipated under stringent clonal selection. For metastatic tumors, we focused on all *metastasis-private* high-frequency non-silent mutations (CCF>20%), which include both private clonal (CCF>60%) and private subclonal (20%<CCF<60%) events. For primary tumors, we evaluated high-frequency private and shared subclonal mutations (20%<CCF<60%) to enable comparisons with an early stage CRC cohort. Bootstrapping was performed (Methods) where 60 percent of tumors were down-sampled at each iteration (n=50 down-samplings) and integrated to compute an enrichment score (n=20 down-samplings were repeated for the early stage primary CRC cohort). *P*-value, Wilcoxon Rank-Sum Test (two-sided). Bar, median; box, 25th to 75th percentile (interquartile range, IQR); vertical line, data within 1.5 times the IQR.

Supplementary Table 1. Clinical features of the metastatic colorectal cancer cohort

Patient ID	Sex	Age at diagnosis of primary tumor	Diagnosis history	Treatment before diagnosis of asynchronous met	Primary tumor size (cm)	Met tumor size (cm)	Total number of samples	Number of primary tumor samples	Number of met samples	MRS of both Primary and Met	Data source
V46	M	60	P&LN – LU(2y7m) – BM(3y9m)	yes	4	3 (BM)	3	1	1 (BM), 1(LN)	no	This study
V402	F	47	P – BM(4y8m)	no	7.5	5 (BM)	8	4	4 (BM)	yes	This study
V514	M	73	P&LN – BM(0y6m)	yes	9	2.5 (BM)	5	1	2 (BM), 2 (LN)	no	This study
V559	M	49	P&LI – LU(1y5m) – BM(1y8m)	yes	4.5	3.5 (BM)	4	1	1 (BM), 2 (LI)	no	This study
V750	M	65	P&LN&LI&LU – BM(0y6m)	yes	10	3 (BM)	13	5	5 (BM), 3 (LN)	yes	This study
V824	M	61	P&LN – BM&LU(0y10m)	no	8	5 (BM)	9	3	3 (BM), 3 (LN)	yes	This study
V855	M	57	P&LN – BM(0y4m)	yes	6	3 (BM)	2	1	1 (BM)	no	This study
V930	F	71	P – LI(2y2m) – LU(5y8m) – BM(8y7m)	yes	4	NA	13	5	5 (BM), 3 (LU)	yes	This study
V953	F	68	P – BM(2y6m)	no	8.1	5 (BM)	7	3	4 (BM)	yes	This study
V974	F	60	P&BM – RecBM(0y5m)	no	10	5 (BM)	8	3	5 (BM)	yes	This study
Uchi2	M	81	P&LI	no	5.2	NA	12	9	3 (LI)	yes	Uchi et al. 2016
Kim1	M	69	P&LI	no	6	NA	7	4	3 (LI)	yes	Kim et al. 2015
Kim2	M	79	P – LI(0y7m)	no	10.5	NA	7	5	2 (LI)	yes	Kim et al. 2015
Leung1	M	77	P&LI	no	NA	NA	2	1	1 (LI)	no	Leung et al. 2017
Leung2	M	64	P&LI	no	NA	NA	2	1	1 (LI)	no	Leung et al. 2017
Lim3	M	46	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim6	M	59	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim7	F	54	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim8	M	57	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim11	M	57	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim12	M	71	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim16	M	77	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Lim21	M	52	P&LI	no	NA	NA	2	1	1 (LI)	no	Lim et al. 2015
Total							118	55	63		

Abbreviations: P – primary tumor; BM – brain metastasis; LN – lymph node metastasis; LI – liver metastasis; LU – lung metastasis; met - metastasis; MRS – multi-region sequencing; NA- not available; & – synchronous; m- month; y - year

Note: samples from primary tumor and synchronously diagnosed metastases were untreated

Supplementary Table 5. Spatial computational tumor model parameters

Parameters	Description	Default in basic model	Justifications/Remarks
N_T	Final tumor size	N_T~10^9 cells for both primary tumor and metastasis	There are ~10^9 or more cells in a typical solid tumor.
K	Deme size	K=5,000–10,000 cells	The demes recapitulate the glandular structure often found in colorectal cancer in which the gland size is approximated at 2,000-10,000 cells [3] .The deme size recapitulates the degree of spatial constraint and clone mixing during tumor growth. For instance, small deme size represents stringent spatial constraint and reduced subclone mixing, thereby hindering the efficacy of selection. In contrast, large deme size results in relaxed spatial constraint and enhanced subclone mixing.
p and q	The birth and death probability for each cell at each generation during deme expansion, respectively	p=0.55 and q= 0.45	It has been reported that there is no significant growth rate difference in paired primary tumors and metastases [5]. We therefore assume the same birth and death rates in primary tumor and metastasis. Given the choice of p and q values here, it takes about 3 years (assuming 4 days for each cell cycle) for the tumor to grow from founder cell to diagnosis (~10^9 cells) (Fig. 15b).
u	Passenger mutation rate per cell division in the ~60Mb exonic regions	u=0.3	Mutation rate in normal somatic cells is at the order of 10^{-9} per base pair per cell division [9] . Because of the genomic instability in many cancers, the per-cell division mutation rate for cancer is significantly higher than normal cells. We assume a mutation rate 5×10^{-9} /base pair/division (equivalent to u=0.3 per cell division for the 60M exonic region) in the simulations, giving rise to 20-200 subclonal SNVs (10%<CCF<60%) in each bulk sample in the simulations which is in consistent with the observed number in current study.
u_b	Mutation rate of beneficial driver mutations per cell division	u_b=10^{-5}	Bozic et al [7] estimated u_b to be at the order of 10^{-5} per cell division in the genome.
s	Selection coefficient	s=0.1	We use relatively high selection s=0.1, in order to robustly distinguish with the evolutionary dynamics of effectively neutral evolution [1].
N_d	The primary tumor size in cell number at the time of dissemination	log10(N_d)~ U(2, 9)	We randomly chose 100 dissemination time points, correponding to the primary tumor size at the time of dissemination from a uniform distribution log10(N_d)~U(2, 9), each giving rise to an independent paired primary tumor and metastasis.
c	The number of cells from primary tumor seeding a metastasis	c=1	We assume one single cell from a deme in tumor periphery seeds the metastasis based on the pattern of commonly monoclonal seeding in the mCRC cohort.

Supplementary Table 6. Description of summary statistics for SCIMET

Summary statistics	Descriptions
S_1, S_2, S_3 and S_4	The total number of primary-private sSNVs that are present at merged CCF>10%, 20%, 40% and 60%, respectively.
S_5, S_6, S_7 and S_8	The total number of metastasis-private sSNVs that are present at merged CCF>10%, 20%, 40% and 60%, respectively.
S_9	The total number of sSNVs that are metastasis-clonal (merged CCF>60%) while primary-subclonal (10%<merged CCF<60%).

Supplementary Table 7. SCIMET Summary statistics for the metastatic colorectal cancer cohort

PM_pair	S1	S2	S3	S4	S5	S6	S7	S8	S9
V402_BM	106	29	9	6	24	21	20	20	2
V824_BM	108	23	4	2	39	29	26	25	3
V953_BM	295	190	64	33	54	33	21	21	2
V974_BM	66	59	49	45	35	30	30	30	1
V930_LU	32	12	7	2	88	78	48	33	2
V930_BM	29	11	6	2	52	48	47	47	1
V750_BM	63	26	6	2	98	42	19	17	11
V46_BM	20	19	12	11	58	54	51	45	4
V514_BM	16	16	16	9	42	27	26	24	2
V559_LI	18	15	13	11	103	68	13	6	5
V559_BM	13	13	11	8	66	65	34	26	4
V855_BM	32	32	21	14	21	21	14	12	2
Uchi2_LI	11	5	2	2	12	12	10	8	0
Kim1_LI	42	8	0	0	8	5	3	2	4
Kim2_LI	79	34	6	1	16	15	15	14	22
Leung1_LI	8	8	7	5	16	16	13	11	3
Leung2_LI	24	21	17	14	138	118	103	91	3
Lim3_LI	42	41	23	13	30	28	23	17	0
Lim7_LI	17	11	8	7	13	13	5	4	0
Lim8_LI	65	65	59	49	123	122	114	102	0
Lim12_LI	24	24	19	13	40	40	32	17	0
Lim16_LI	14	14	7	5	17	12	7	7	0
Lim21_LI	2	2	2	2	28	28	25	20	0

Supplementary Note

An algorithm for efficiently simulating virtual P/M tumor pairs and multi-region sequencing data

We devised an algorithm to accelerate the simulation of *virtual* P/M tumor pairs and multi-region sequencing data. The strategy is first to simulate the *deme partition history* in paired P/M tumors without involving mutation occurrence. Multiple regions in each of P and M (n=4, each composed of 10-100 demes) were then sampled and tracked back to the very first deme. All the demes that are not in the partition history of the sampled demes were trimmed (**Part I**). Second, we simulate the *cell evolutionary dynamics* along the partition history of the sampled demes via random cell birth or death and mutation occurrence (neutral or beneficial) (**Part II**). The growth dynamics and parameters employed here are the same as were employed for the 3-D agent-based model described in the Methods. Using this strategy, we only simulate the mutation occurrence along the *partition history* in sampled demes rather in the whole tumor, thus allowing for large-scale simulations of P/M tumor pairs under different evolution modes (effective neutrality or subclonal selection) via SCIMET.

Algorithm of Part I – Simulation of deme-partition history

For i=1 to 70,000:

While tumor volume < 0.2 million demes (~10^9 cells) in both primary tumor and metastasis:

Do

1) Primary tumor grows via the 3D agent-based model;
2) Randomly choose a peripheral deme where the disseminated cell or cell cluster will be sampled from (denoted as DC deme), at a time when primary tumor size is N_d;
3) Metastatic tumor grows via the same 3D agent-based model started from the sampled disseminated cell or cell cluster;
4) Keep track of the parental and offspring relationship for each deme;

End while

Sample four regions, each composed of 10-100 demes, from both primary tumor and metastasis;

Trace back the deme ancestors to the very first deme and trim the non-sampled demes to obtain the partition history of sampled demes;

End for

After obtaining the *deme partition history* of multi-region samples for each P/M pairs, we next simulate the *cell evolutionary dynamics* along the *partition history*. We simulate different evolutionary scenarios in P/M pairs, namely Neutral/Neutral (N/N), Neutral/Selection (N/S), Selection/Neutral (S/N) or Selection/Selection (S/S). The growth dynamics under each of the scenarios is as described in the paper.

Algorithm of Part II – Simulation of cell evolutionary dynamics

For tumor evolution mode in [N/N, N/S, S/N, S/S]:

For i=1 to 70,000:

1) Load deme-partition history;
2) Simulate cell birth-death process, mutation occurrence and deme partition in primary tumor started from a single transformed founder cell;
3) Randomly choose a dissemination cell or cell cluster from the DC deme;
4) Simulate cell birth-death process, mutation occurrence and deme partition in metastasis started from the disseminated cell or cell cluster;

End for

Implement the sequencing and mutation calling processes in each bulk sample to obtain the variant allele frequency (VAF) and cancer cell fraction (CCF) for each sSNV

End for

Mathematical analysis for the special case of neutral evolution and exponential growth

We consider a growing tumor cell population that starts from a single founder cell at transformation and expands via a stochastic continuous-time birth-death process. Each cell divides at rate *b* and dies at rate *d* with neutral point mutation rate *u* per cell division in the exonic region of the genome. Hence, the population growth rate is $\lambda=b-d$, the death-birth rate ratio is $\delta=d/b$ and the expected population size at time *t* is $N(t) = e^{\lambda t}$. Bozic et al. [10] have reported that the expected number of *clonal* mutations (m_c) and *subclonal* mutations present at a cancer cell fraction (CCF) larger than α (denoted by $m_s(\alpha)$) are:

$$m_c = \frac{\delta u}{1-\delta} \quad \text{(S1)} \quad \text{and} \quad m_s(\alpha) = \frac{u(1-\alpha)}{(1-\delta)\alpha} \quad \text{(S2), respectively.}$$

We denote m_i as the number of somatic mutations that occurred *prior to primary tumor founder cell,* including mutations that occurred during development, tissue self-renewal and multi-step carcinogenesis, which are typically shared by all descendent tumor cells including metastasis. The total number of somatic mutations present in the primary tumor with CCF larger than α is:

$$M_p(\alpha) = m_i + m_c + m_s(\alpha) = m_i + \frac{\delta u_p}{1-\delta_p} + \frac{u_p(1-\alpha)}{(1-\delta_p)\alpha} = m_i + \frac{u_p}{1-\delta_p}\left(\frac{1}{\alpha} + \delta_p - 1\right)$$

where δ_p is the death-birth rate ratio and u_p is the mutation rate during primary tumor growth.

We denote t_d as the time of dissemination (time 0 is the beginning of malignant growth from primary tumor founder cell) and N_d as the expected primary tumor size at the time of dissemination. Thus, $N_d = e^{\lambda t_d}$. We denote m_d as the number of somatic mutations that occurred on the metastasis-founder cell during the growth of primary tumor. Here the mutation rate is specified as per cell division and time is continuous calendar time with units

as days. We denote T_d as the expected number of cell generations at the time of dissemination, so $m_d = u_p T_d$. The cell generation time $g = 1/b$, hence, T_d is given by:

$$T_d = \frac{t_d}{g} = \frac{\ln(N_d)}{\lambda_p / b_p} = \frac{\ln(N_d)}{(b_p - d_p)/b_p} = \frac{\ln(N_d)}{1-\delta_p}$$

where δ_m is the death-birth rate ratio and u_m the mutation rate during metastatic growth.

Denoting $u_m=\gamma u_p$, the total number of somatic mutations present in the metastasis at CCF > α is:

$$M_m(\alpha) = m_i + u_p T_d + \frac{u_m}{1-\delta_m}(\frac{1}{\alpha} + \delta_m - 1) = m_i + \frac{u_p \ln(N_d)}{1-\delta_p} + \frac{\gamma u_p}{1-\delta_m}(\frac{1}{\alpha} + \delta_m - 1)$$

Setting $I(\alpha) = M_m(\alpha) - M_p(\alpha)$, thus:

$$I(\alpha) = M_m(\alpha) - M_p(\alpha) = \frac{u_p \ln(N_d)}{1-\delta_p} + \frac{u_p}{\alpha}(\frac{\gamma}{1-\delta_m} - \frac{1}{1-\delta_p}) - u_p(\gamma - 1)$$

When there is no significant change for the point mutation rate in metastasis, namely $\gamma \approx 1$ and $u_m \approx u_p = u$, thus:

$$I(\alpha) = \frac{u \ln(N_d)}{1-\delta_p} + \frac{u}{\alpha}(\frac{1}{1-\delta_m} - \frac{1}{1-\delta_p}) \quad \text{(S3)}$$

CRC metastases usually grow at rates comparable to or slightly faster than primary CRCs [5,11,12], thus $1 - \delta_p \approx 1 - \delta_m$. Therefore:

$$I(\alpha) = \frac{u \ln(N_d)}{1-\delta_p} \quad \text{(S4)}$$

Let $L_p(\alpha)$ and $L_m(\alpha)$ be the number of primary tumor-private and metastasis-private mutations with CCF>α in primary tumor and metastasis, respectively, then:

$$I(\alpha) = M_m(\alpha) - M_p(\alpha) \approx L_m(\alpha) - L_p(\alpha) \quad \text{(S5)}$$

According to our spatial model simulations (**Fig. 4b, Supplementary Fig. 18-19**), $L_p(\alpha) \approx 0$ when tumor is under neutral evolution and α is large (e.g. 0.6). Let L_m be the number of metastasis-private *clonal* mutations, thus:

$$L_m \approx \frac{u \ln(N_d)}{1-\delta_p} \quad \text{(S6)}$$

Therefore, in the special case of neutral evolution and exponential growth, the number of metastasis-private clonal sSNVs (L_m) is linearly related with the log scale of primary tumor size at time of dissemination ($\log(N_d)$). Of note, the positive relationship between L_m and $\log(N_d)$ is non-linear under spatial growth model suggesting the linearity between L_m and $\log(N_d)$ is model dependent (**Fig. 4b, Supplementary Fig. 18-19**).

Estimating the time span between dissemination and primary tumor resection

We converted the estimated primary tumor size at time of dissemination (N_d) to the time span between dissemination and surgical resection of the primary tumor. The conversion was based on the 3D spatial tumor growth model (**Supplementary Fig. 15b**) and the recorded size of the primary tumor at time of surgery (**Supplementary Table 1**) as well as the N_d estimations by SCIMET. The net tumor growth under the spatial agent-based model can be fitted by an initial exponential model followed by a power-law model as:

$$n(t) = \begin{cases} n_0 e^{rt}, & t < t_c \\ (w + mt)^3, & t \geq t_c \end{cases} \quad \text{(S7)}$$

where n(t) is the cell number t days after the primary tumor founder cell is established (n_0=1). By fitting the tumor growth simulation data and assuming a 4-day cell division time for colorectal cancer [8], the parameters in Eq. (S7) were estimated as: r=0.0454, w=-341.7, m=1.445 and $t_c \approx 300$ days (**Supplementary Fig. 15b**). Given a tumor with diameter, *D* (cm), the number of cells was estimated as: $\frac{4}{3}\pi(\frac{D}{2})^3 \times C = \frac{1}{6}\pi C D^3$, where *C* is the number of cells in a 1cm^3 tumor (equivalent to ~10^8 cells). Using Eq. (S7), the estimated N_d and primary tumor size at surgery, we can compute the time span (in years) between dissemination and primary tumor surgery for each patient (**Supplementary Fig. 24**).

References

1. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat Genet* **49**, 1015-1024 (2017).
2. Beaumont, M.A., Zhang, W. & Balding, D.J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-35 (2002).
3. Siegmund, K.D., Marjoram, P., Woo, Y.J., Tavare, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc Natl Acad Sci U S A* **106**, 4828-33 (2009).
4. Zhao, J., Siegmund, K.D., Shibata, D. & Marjoram, P. Ancestral inference in tumors: how much can we know? *J Theor Biol* **359**, 136-45 (2014).
5. Finlay, I.G., Meek, D., Brunton, F. & McArdle, C.S. Growth rate of hepatic metastases in colorectal carcinoma. *Br J Surg* **75**, 641-4 (1988).
6. Csilléry, K., François, O. & Blum, M.G. abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution* **3**, 475-479 (2012).
7. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A* **107**, 18545-50 (2010).
8. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci U S A* **105**, 4283-8 (2008).
9. Lynch, M. Evolution of the mutation rate. *Trends Genet* **26**, 345-52 (2010).
10. Bozic, I., Gerold, J.M. & Nowak, M.A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLoS Comput Biol* **12**, e1004731 (2016).
11. Bolin, S., Nilsson, E. & Sjodahl, R. Carcinoma of the colon and rectum--growth rate. *Ann Surg* **198**, 151-8 (1983).
12. Havelaar, I.J., Sugarbaker, P.H., Vermess, M. & Miller, D.L. Rate of growth of intraabdominal metastases from colorectal cancer. *Cancer* **54**, 163-71 (1984).