Rapid evolution and biogeographic spread in a colorectal cancer

Joao M Alves[1,2,3], Sonia Prado-Lopez[1,2,3], Jose Manuel Cameselle-Teijeiro[4,5], David Posada[*,1,2,3]

1. Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain.
2. Biomedical Research Center (CINBIO), University of Vigo, Spain.
3. Galicia Sur Health Research Institute, Vigo, Spain.
4. Department of Pathology, Clinical University Hospital, Galician Healthcare Service (SERGAS), Santiago de Compostela, Spain.
5. Medical Faculty, University of Santiago de Compostela, Santiago de Compostela, Spain

* Corresponding author: dposada@uvigo.es

Keywords: tumor phylogenetics, cancer evolution, biogeography, colorectal cancer, metastatic dissemination

ABSTRACT

How and when tumoral clones start spreading to surrounding and distant tissues is currently unclear. Here, we leveraged a model-based evolutionary framework to investigate the demographic and biogeographic history of a colorectal cancer. Our analyses strongly support an early monoclonal metastatic colonization, followed by a rapid population expansion at both primary and secondary sites. Moreover, we infer a hematogenous metastatic spread seemingly under positive selection, plus the return of some tumoral cells from the liver back to the colon lymph nodes. This study illustrates how sophisticated techniques typical of organismal evolution can provide a detailed picture of the complex tumoral dynamics over time and space.

Cancer has long been recognized as a somatic evolutionary process mainly driven by continuous Darwinian natural selection, in which cells compete for space and resources[1]. With the increasing availability of high-throughput genomic data, several studies have started to explore the evolutionary relationships of tumor clones in order to identify the key molecular changes driving cancer progression[2], to better understand the subclonal architecture of tumors[3,4], and to determine the origins of metastases[5]. While sophisticated inferential methods have been put forward that make use of sequencing data to investigate the timing and the patterns of geographical dispersal of organismal lineages[6,7], their application in cancer research has only recently started[8,9].

In metastatic colorectal cancer (mCRC) many aspects underlying the dissemination of cancer cells to tissues beyond primary lesions have been difficult to determine. Although earlier models of mCRC progression have proposed a sequential metastatic cascade, with cells from the primary

tumor first escaping to local lymph nodes from where they seed distant tissues[10], conflicting evidence has recently emerged, as some genomic datasets seem to favor an independent origin of distant and lymph node metastases[5]. Here, to better understand the tempo and mode of diversification of the tumoral cells within the human body, we sampled and analyzed whole-exome sequencing data from 18 different locations of a mCRC (Fig. 1A) under a powerful Bayesian framework, typical of organismal phylogenetics, phylodynamics and biogeography.

After filtering out germline polymorphisms and single nucleotide variants (SNVs) in non-diploid regions, we detected 475 somatic SNVs with high confidence (Supplementary Table 1). A principal component analysis (PCA) of their allele frequencies showed a clear distinction between primary tumor and metastatic samples (Fig. 1B). Concordantly, we found a significant correlation between genetic and physical distances among these two groups, but not within (Supplementary Fig. 1). Albeit the extensive intratumor heterogeneity, we identified several clonal alterations in known CRC drivers[11], including two copy neutral loss of heterozygosity events in *APC* and *TP53,* plus a non-synonymous mutation in *KRAS* (Fig. 1C-D). Moreover, we also observed a clonal non-synonymous mutation in *MSLN*, a plasma membrane differentiation antigen which is emerging as an attractive target for cancer immunotherapy due to its potential involvement in the epithelial-to-mesenchymal transition, a cellular process thought to be required for metastatic dissemination[12].

We obtained a Bayesian estimate of the phylogeny, under a relaxed clock model with exponential growth, of the 21 tumor clones identified (Fig. 2A). All the metastatic lineages grouped together with high support, suggesting a monoclonal origin. The age of the tumor was estimated to be 6.94 – 6.45 years (95% Highest Posterior Density (HPD): 9.98/9.16 - 4.43/4.36) prior to clinical diagnosis (PCD). Also, the results imply an early origin of the metastatic ancestor, 4.20 years PCD (95% HPD: 6.30 - 2.46) (Supplementary Fig. 2), diverging within a short period of evolutionary time (posterior median divergence time = 2.58 years) from the ancestor of the tumor sample (tMRCA) (Fig. 2B). Despite the lack of a significant overall departure from neutrality across branches, evidence of positive selection (i.e., ratio of substitution rates at non-synonymous and synonymous sites (dN/dS) > 1) was found for four specific branches in the phylogeny, including the ancestral lineage that gave rise to all the metastatic clones, pointing out to changes potentially relevant for the acquisition of metastatic capabilities (Fig. 2A). The most notable mutation in this branch was a non-synonymous mutation in *ANGPT4*, an angiogenic gene known to promote cancer progression in multiple cancer types[13,14].

Furthermore, the Bayesian skyline plot (Fig. 2C) shows that the tumor underwent a very rapid demographic expansion coincident with the diversification of both primary tumor and metastatic clades, before eventually becoming stationary. Interestingly, the expansion of the metastatic

clade seems to slightly precede the one associated with the primary tumor. The posterior median estimate of the population growth rate per generation was 0.014 (95% HPD: 0.006 - 0.03), implying an average population doubling time of 193 days.

The colonization history of this tumor appears to have been quite complex. A dispersal-extinction biogeographic analysis placed the origin of sampled lineages around the geographical center of the primary tumor (Fig. 3A), subsequently radiating outwards in multiple directions. Additionally, we inferred with high confidence that the ancestral metastatic clone experienced an early long-distance dispersal to the liver (Fig. 3B), followed by a proliferation towards the nearby hepatic lymph nodes before eventually spreading "back" to the colonic lymph nodes. The number of implied migrations and movements was surprisingly high (Fig. 3C). Importantly, a distance-dependent model was heavily favored over a distance-independent model (Fig. 3D), suggesting an overall negative correlation between geographical distance and the dispersal ability of the tumoral clones at the whole patient level.

Collectively, our analyses provide a detailed picture of the evolutionary history of this tumor. While we are not the first ones applying Bayesian phylogenetics for cancer dating[8,9,15], previous attempts used sample trees and absence/presence mutational profiles instead of clonal phylogenies and clonal sequences, and therefore are subject to potential biases[16,17]. Besides, the evolutionary framework presented here has several advantages over previous approaches. For example, it is based on Bayesian estimates obtained only after contrasting competing evolutionary and demographic models under a rigorous model selection framework. Also, our biogeographic approach allows for the presence of the same ancestral clone at more than one location, and is able to consider the spatial distance among samples, unlike the approach of El-Kebir *et al.*[17]. On the other hand, our analyses imply a series of assumptions. In particular, it presumes that the clonal genotypes were appropriately reconstructed. Indeed, clonal deconvolution remains a very hard problem[18], and we cannot rule out some degree of uncertainty in the precise combination of mutations assigned to any given clone. Nevertheless, we were reassured to some extent by the fact that comparable clonal genotypes were obtained when using a different deconvolution approach[19] (Supplementary Fig. 3). Moreover, our biogeographic model assumes that the geographical distances among samples more or less reflect the true "migration likelihood" of the tumoral clones. While we cannot prove that the distances used are realistic in this regard, different sets of distance matrices resulted in similar biogeographic solutions (Supplementary Fig. 4).

Importantly, early metastases, such as the one described here, have already been proposed in mCRC[8,9,15]. Although Leung *et al.*[20] recently inferred a late-dissemination model in mCRC, they failed to provide quantitative measurements, and their timing of metastatic dissemination was

simply determined by visual inspection of mutational trees, making their results difficult to interpret and compare with. Reinforcing the idea of an early cell dissemination, our results suggest a fairly rapid population increase during the parallel phylogenetic diversification of the metastatic and primary tumor clades. Although these analyses revealed a similar individual contribution of each clade to the overall variation in effective population size, the observed demographic trends are compatible with an early geographical expansion, and subsequent establishment, of the metastatic lineages into new anatomical sites, together with the expansion of primary tumor populations to nearby areas.

Our biogeographic reconstruction revealed a pattern of metastatic dissemination in which the primary tumor directly seeded liver metastases without an apparent early involvement of the lymphatic system. Previous studies have argued that metastatic spread in mCRC can potentially occur *via* the hepatic portal vein - a direct blood supply between the colon and the liver[5,21]. On this basis, metastatic dissemination in this patient seems to have started hematogenously, with a single episode of long-range dispersal across the hepatic portal vein into the liver, followed by a sequence of short-range migration episodes to nearby anatomical areas before eventually spreading to colonic lymph nodes. While the latter colonization has not yet been described in mCRC patients, it might represent some type of *self-seeding* mechanism, as previously observed in mCRC in mice[22]. Interestingly, we observed a similar migration pattern, albeit less detailed (Supplementary Fig. 5), using a different approach[17].

In conclusion, we believe that this study demonstrates the utility of a sound evolutionary framework for exploring the spatio-temporal dynamics of cancer cell populations from multi-regional sequencing data. By integrating concepts from population genetics, phylogenetics and biogeography, we were able to resolve the spatial architecture of this cancer, temporally connect phylogenetic events at time scales compatible with clinical observations, and recover past demographic changes shaping the spatial distribution of malignant clones. As more data continues to accumulate, future studies could extend these type of evolutionary analyses to other patients and cancer types, including polyclonal metastatic tumors[5], in order to obtain a more comprehensive and meaningful understanding of the cancer spread, which could ultimately be used to predict clinical outcomes, and guide targeted treatments[23].

Methods

Sample collection. A 51-year-old man was admitted to the University Hospital of Santiago de Compostela (CHUS) with a one-month history of weakness and weight loss. The patient died five days after admission, and the pathological assessment revealed a low-grade, moderately differentiated, adenocarcinoma of the descending colon, with multiple metastatic lymph-nodes, liver metastases, a metastatic focus in the right diaphragmatic peritoneum and multiple

intravascular micrometastases in both lungs (pT4aN2bM1c)[24]. During the warm autopsy, performed by JMC, a total of 18 samples were collected, including eight from the primary tumor (C1-C8), two from colonic lymph-node metastases (CL1, CL2), two from hepatic lymph-node metastases (HL1, HL2), four from liver metastases (L1-L4), and two healthy samples from the colon (N1, N2) (Fig. 1A). Sample collection was approved by a local ethics committee (CAEI Galicia 2014/015), and written informed consent was provided by the patient's family.

Tumor disaggregation and sorting. Tumor samples and normal CRC tissues were frozen in liquid nitrogen, placed in dry ice and transported to the laboratory. Next, samples were minced in pieces of 1 mm^3 with a scalpel and digested by incubation in Accutase (LINUS) for 1h at 37ºC. Thereafter, the cell suspension was filtered with a 70 µm cell strainer (FALCON). The cell pellets were washed twice and suspended in ice-cold Phosphate Buffered Saline (PBS) and then stained for 30 min with the Anti-EpCAM (EBA1) antibody (BD). Following three successive washes in PBS buffer, flow cytometry analyses and sorting of EpCAM positive cells were performed with a FACSARIA III (BD Biosciences). Then, DRAQ5 and 7AAD dyes were added in order to select nucleated cells and exclude non-viable ones.

DNA extraction and exome sequencing. The DNA was extracted from the 18 samples using the QIAamp DNA Mini kit (QIAGEN), and whole-exome sequencing was carried out at 60X with the Ion Torrent PGM platform at the Fundación Pública Galega de Medicina Xenómica (FPGMX) at Santiago de Compostela, Spain.

Detection of somatic variants. Sequencing reads were aligned to the Genome Reference Consortium Human Build 37 (GRCh37) using the Torrent Mapping Alignment Program 5.0.7 (TMAP). After alignment, single nucleotide variants (SNVs) were called independently for all tumor and normal samples using a standalone version of the Torrent Variant Caller 5.6.0 (TVC). Following a similar approach to de Leng *et al.*[25], a set of high-stringency thresholds were used to retain high confidence bi-allelic calls, including a minimum coverage of 20X for both tumor and healthy samples, a minimum variant allele frequency (VAF) of 0.05, and a minimum nucleotide (Phred) quality score of 20. Germline polymorphisms were filtered by excluding variants present in the healthy samples. Copy number profiles, as well as tumor purity estimates and global ploidy status, were obtained using the Sequenza toolkit[26] under default settings (binning window of 1 Mb).

Population structure. To test the existence of population genetic structure in anatomical space, we assessed the correlation between genetic (measured *via* F_{ST} estimates) and geographical distance, using the Mantel test function in the adegenet R package[27] (Supplementary Fig. 1).

Deconvolution of clonal populations. Since the accuracy of the clonal deconvolution from mixed samples largely depends on the quality of the inferred VAFs, and copy-number variation is known to alter the allele frequency of somatic mutations in bulk tumor samples, somatic calls showing a VAF < 0.075, with a read depth < 20 in all tumor and healthy samples, and/or overlapping with copy-number events were filtered out prior to clonal deconvolution. The number of tumor clones, as well as their genotype sequences, were then inferred using the CloneFinder algorithm[18], which has been previously shown to outperform other methods in both simulated and empirical datasets (but see Supplementary Information).

Bayesian phylogenetic model fitting, reconstruction and dating. Bayesian phylogenetic analyses were performed using BEAST 2.4.7[28]. First, the most appropriate evolutionary model (i.e., demographics and substitution rates) for our data was identified using Bayes factors[29]. A detailed description of the models tested can be found in Supplementary Table 2. For each candidate model, marginal likelihoods were obtained through a path-sampling analysis implemented in BEAST, using 100 independent Markov Chain Monte Carlo (MCMC) chains with 500,000 steps each. As a prior for the relaxed clock rate mean, a value of 4.6e-10 substitutions per site per generation derived experimentally for CRC[15] was used. For conversion to real time, a generation time of four days was assumed[15,30]. Moreover, since the clonal genotypes obtained only comprise variable genomic positions, an SNV ascertainment bias correction[31] was performed by modifying the "*constantSiteWeights*" attribute in the input XML file for BEAST. Posterior distributions under the model with highest support (i.e., Clock Model: Relaxed clock exponential; Tree: Coalescent Exponential Population) for the parameters of interest were obtained by running an MCMC chain during 100 million generations, sampled every 2000. Convergence was assessed using Tracer v1.6[32]. After discarding the first 10% of the samples as burn-in, point estimates for the different parameters were obtained using posterior means, and a maximum clade credibility topology was constructed using the median heights.

Demographic analysis. Demographic changes in the cancer cell population were inferred from a Bayesian skyline plot (BSP) analysis carried out in BEAST 2.4.7. The same prior distributions described above were used, with the exception of the coalescent tree prior, which was set to "Coalescent Bayesian skyline". The final skyline reconstruction was obtained using Tracer v1.6, setting the number of bins to 100 and the age of the youngest tip to 0 (i.e., the time of collection looking backwards).

Estimation of positive selection. The coding clonal sequences were concatenated into a multiple sequence alignment and analyzed using PAML 4.8a[33] to obtain maximum likelihood estimates of the non-synonymous/synonymous rate ratio (*dN*/*dS*) for the different branches of the inferred clonal genealogy in BEAST. The significance of these estimates was tested using likelihood ratio

tests (LRTs) comparing a model assuming a single *dN/dS* for the whole genealogy (model M0) and models assuming that a specific branch has a different dN/dS than the rest (two-ratio model)[34].

Inference of ancestral clonal ranges and migration history. The ancestral spatial distribution of the clones was reconstructed using BayArea[6] upon the inferred BEAST genealogy, together with the observed "geographic ranges" of the tumor clones (i.e., presence/absence of each clone at each of the 16 sampled locations of the tumor) (see Supplementary Information). Posterior distributions for the parameters of interest were obtained by running an MCMC chain during 100 million steps, sampling every 2000 generations. BayArea implements a probabilistic dispersal-extinction biogeographic model that considers how different lineages colonize new regions or disappear from them through time. To examine whether two-dimensional geographical distances played a role in the dispersal ability of tumor clones, two candidate biogeographic models were compared in BayArea using Bayes factors (computed with the Savage-Dickey density ratio method): the mutual-independence (null) model, in which clonal dispersal is not conditioned by spatial distance (i.e., distance power parameter, $\beta = 0$), versus a distance-dependent dispersal model, where the probability of dispersal is affected by spatial distance (i.e., $\beta > 0$: dispersal to nearby areas is more likely than to distant locations, or $\beta < 0$: long-distance dispersal events are favored over short-distance movements). In order to define the spatial distances, different 2D coordinate matrices describing the geographical location of the samples were explored (see Supplementary Information).

References

1. Nowell, P. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
2. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
3. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
4. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
5. Naxerova, K. *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. *Science* **357**, 55–60 (2017).
6. Landis, M. J., Matzke, N. J., Moore, B. R. & Huelsenbeck, J. P. Bayesian analysis of biogeography when the number of areas is large. *Syst. Biol.* **62**, 789–804 (2013).
7. Höhna, S. *et al.* RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology* **65**, 726–736 (2016).
8. Lote, H. *et al.* Carbon dating cancer: defining the chronology of metastatic progression in colorectal cancer. *Ann. Oncol.* **28**, 1243–1249 (2017).
9. Zhao, Z.-M. *et al.* Early and multiple origins of metastatic lineages within primary tumors. *Proceedings of the National Academy of Sciences* **113**, 2140–2145 (2016).
10. Weinberg, R. A. Mechanisms of malignant progression. *Carcinogenesis* **29**, 1092–1095 (2008).
11. Vogelstein, B. & Kinzler, K. W. The Path to Cancer — Three Strikes and You're Out. *New England Journal of Medicine* **373**, 1895–1898 (2015).
12. He, X. *et al.* Mesothelin promotes epithelial-to-mesenchymal transition and tumorigenicity of human lung cancer and mesothelioma cells. *Mol. Cancer* **16**, 63 (2017).
13. Brunckhorst, M. K., Xu, Y., Lu, R. & Yu, Q. Angiopoietins Promote Ovarian Cancer Progression by

Establishing a Procancer Microenvironment. *The American Journal of Pathology* **184**, 2285–2296 (2014).

14. Lukas, R. V., Gondi, V., Kamson, D. O., Kumthekar, P. & Salgia, R. State-of-the-art considerations in small cell lung cancer brain metastases. *Oncotarget* **8**, 71223–71233 (2017).
15. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4283–4288 (2008).
16. Alves, J. M., Prieto, T. & Posada, D. Multiregional Tumor Trees Are Not Phylogenies. *Trends Cancer Res.* **3**, 546–550 (2017).
17. El-Kebir, M., Satas, G. & Raphael, B. J. Inferring parsimonious migration histories for metastatic cancers. *Nat. Genet.* **50**, 718–726 (2018).
18. Miura, S. *et al.* Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty469
19. Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* **16**, 91 (2015).
20. Leung, M. L. *et al.* Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.* **27**, 1287–1299 (2017).
21. Mizuno, N., Kato, Y., Izumi, Y., Irimura, T. & Sugiyama, Y. Importance of hepatic first-pass removal in metastasis of colon carcinoma cells. *J. Hepatol.* **28**, 865–877 (1998).
22. Kim, M.-Y. *et al.* Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315–1326 (2009).
23. Tabassum, D. P. & Polyak, K. Tumorigenesis: it takes a village. *Nat. Rev. Cancer* **15**, 473–483 (2015).
24. Amin, M. B. *et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging. *CA: A Cancer Journal for Clinicians* **67**, 93–99 (2017).
25. de Leng, W. W. J. *et al.* Targeted Next Generation Sequencing as a Reliable Diagnostic Assay for the

Detection of Somatic Mutations in Tumours Using Minimal DNA Amounts from Formalin Fixed Paraffin Embedded Material. *PLoS One* **11**, e0149405 (2016).

26. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
27. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
28. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
29. Kass, R. E. & Raftery, A. E. Bayes Factors. *Journal of the American Statistical Association* **90**, 773 (1995).
30. Rew, D. A., Wilson, G. D., Taylor, I. & Weaver, P. C. Proliferation characteristics of human colorectal carcinomas measured in vivo. *Br. J. Surg.* **78**, 60–66 (1991).
31. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**, 439–447 (2000).
32. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* **67**, 901–904 (2018).
33. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).
34. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).

Acknowledgements

This work was supported by the European Research Council (ERC-617457- PHYLOCANCER awarded to D.P.) and by the Spanish Ministry of Economy and Competitiveness - MINECO (BFU2015-63774-P awarded to D.P.). D.P. receives further support from Xunta de Galicia. J.M.A. is currently supported by an AXA Research Fund Postdoctoral Fellowship. We want to thank Diana Valverde for her help with the DNA extractions from several samples. We want to additionally thank Nuria Estévez-Gómez, Pilar Alvariño and people from the Fundación Pública Galega de Medicina Xenómica (FPGMX) for their help with some of the experiments, and Tamara Prieto, Harald Detering, Diego Mallo, Laura Tomás and Sara Rocha for discussions. We also thank the Supercomputation Center of Galicia (CESGA) for providing computational resources.

Author contributions

D.P. conceived and supervised the study. J.M.C.T. obtained the tumor samples. S.P.L. processed the samples. J.M.A. performed all the analyses. J.M.A. and D.P. wrote the manuscript with input from all other authors.

Competing interests

The authors declare no competing interests.

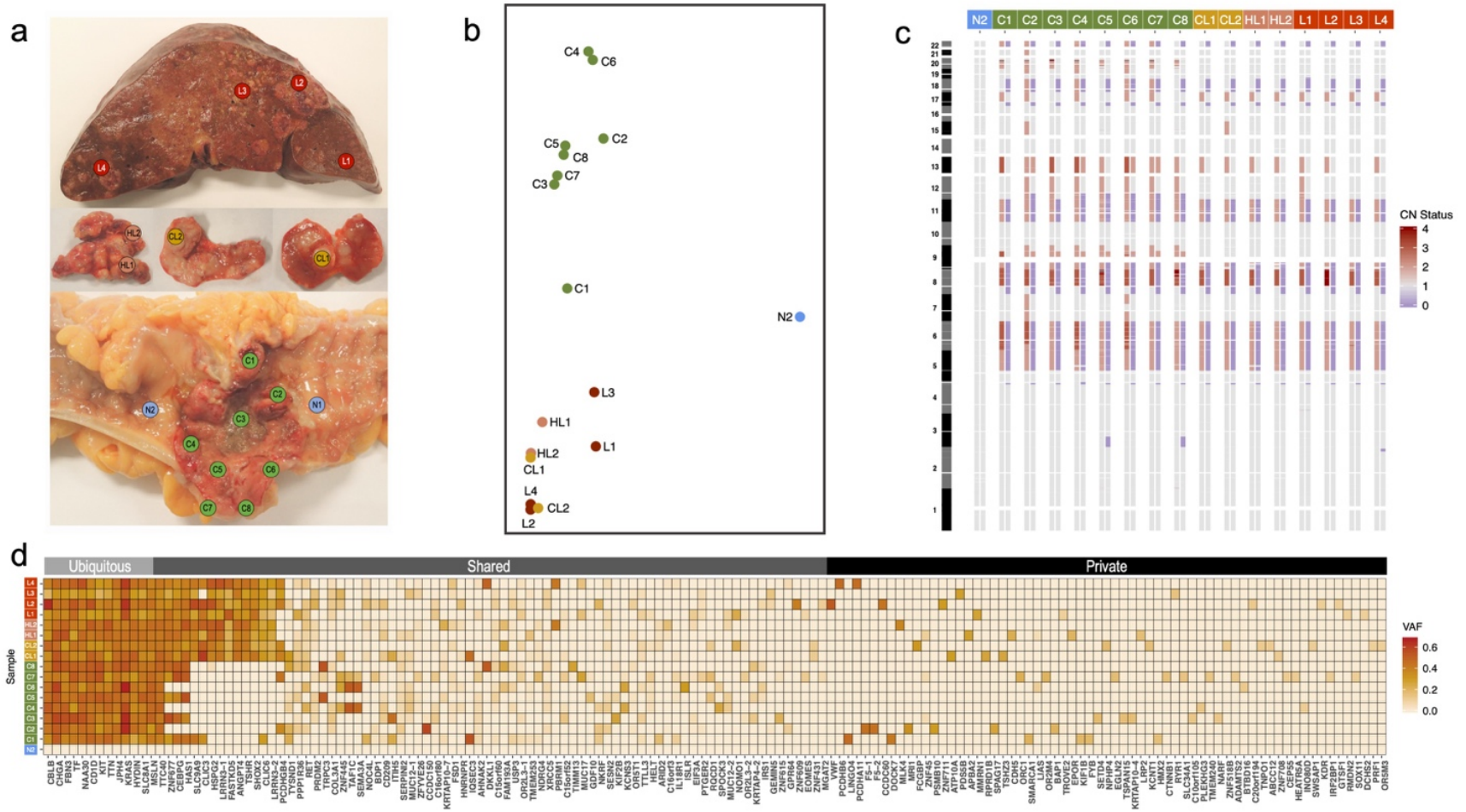

Figure 1. Genomic profiles of bulk tumor samples. a, Multiregional sampling scheme. A total of 18 samples were collected, including two samples from healthy tissue (in blue), eight from the primary tumor (green), two from proximal colonic lymph nodes (gold), two from distal hepatic lymph nodes (salmon), and four from liver metastasis (red). **b,** Principal component analysis (PCA) with variant allele frequencies (VAF) for all 475 somatic mutations detected. Each circle corresponds to a given sample, with colors highlighting the anatomical regions. **c,** Heatmap depicting genome-wide allele-specific copy number status (from 0 in blue to 4 in red) of healthy and tumor samples. Sample IDs are shown at the top. **d,** Heatmap with the observed allele frequencies (from 0 in white to 0.65 in red) of somatic mutations identified in the sequenced samples. Here only the non-synonymous mutations are shown (n = 156), sorted according to their mean VAF across all tumor samples. Gene names are displayed at the bottom of the map. Each row represents a single sample.

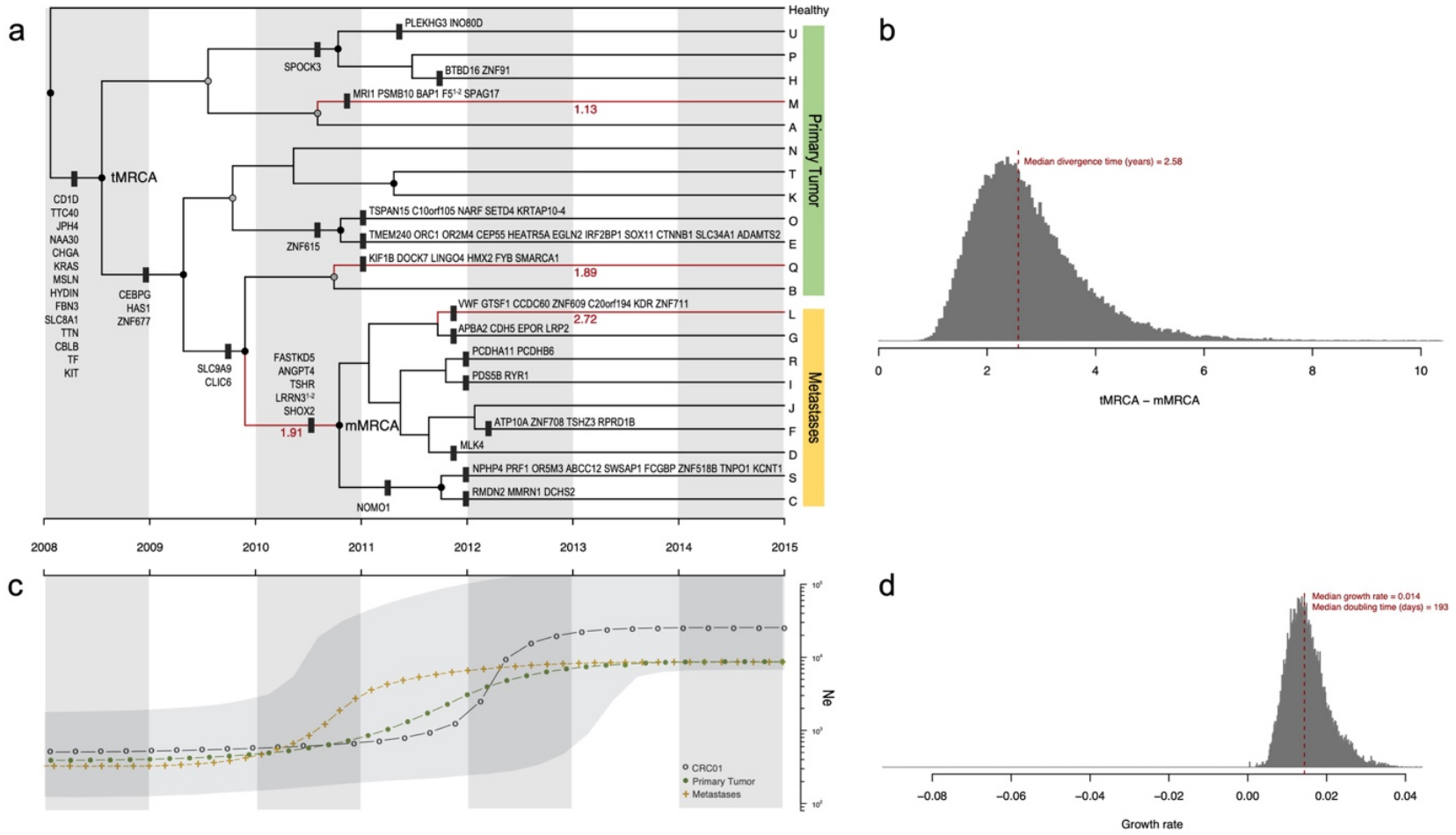

Figure 2. Phylogenetic and demographic reconstruction over time. a, Maximum clade credibility (MCC) tree resulting from the BEAST analyses using the CloneFinder-derived clones. Tree nodes with posterior probability values > 0.99 and > 0.50 are indicated with black and grey solid circles, respectively. Clone IDs (A-U) are shown at the tips of the tree. The x-axis is scaled to years (assuming one generation every four days; see Methods). Only non-synonymous mutations are shown. Tree branches showing a dN/dS ratio > 1 are highlighted in red together with the corresponding dN/dS value. **b,** Posterior probability distribution of the relative divergence time in years of mMRCA in relation to the tMRCA (tMRCA minus mMRCA). The dashed red line depicts the median age estimate of the mMRCA. **c,** Bayesian Skyline Plot (BSP) analysis. The y-axis is in log scale. The black dotted line represents the historical effective population size of the entire cancer cell population (Ne). The gray shading illustrates the 95% HPD interval. Green and golden dotted lines correspond to the effective population sizes of the primary and metastatic populations, respectively. **d,** Histogram illustrating the growth rate per generation of the tumor. The population doubling time is shown in days.

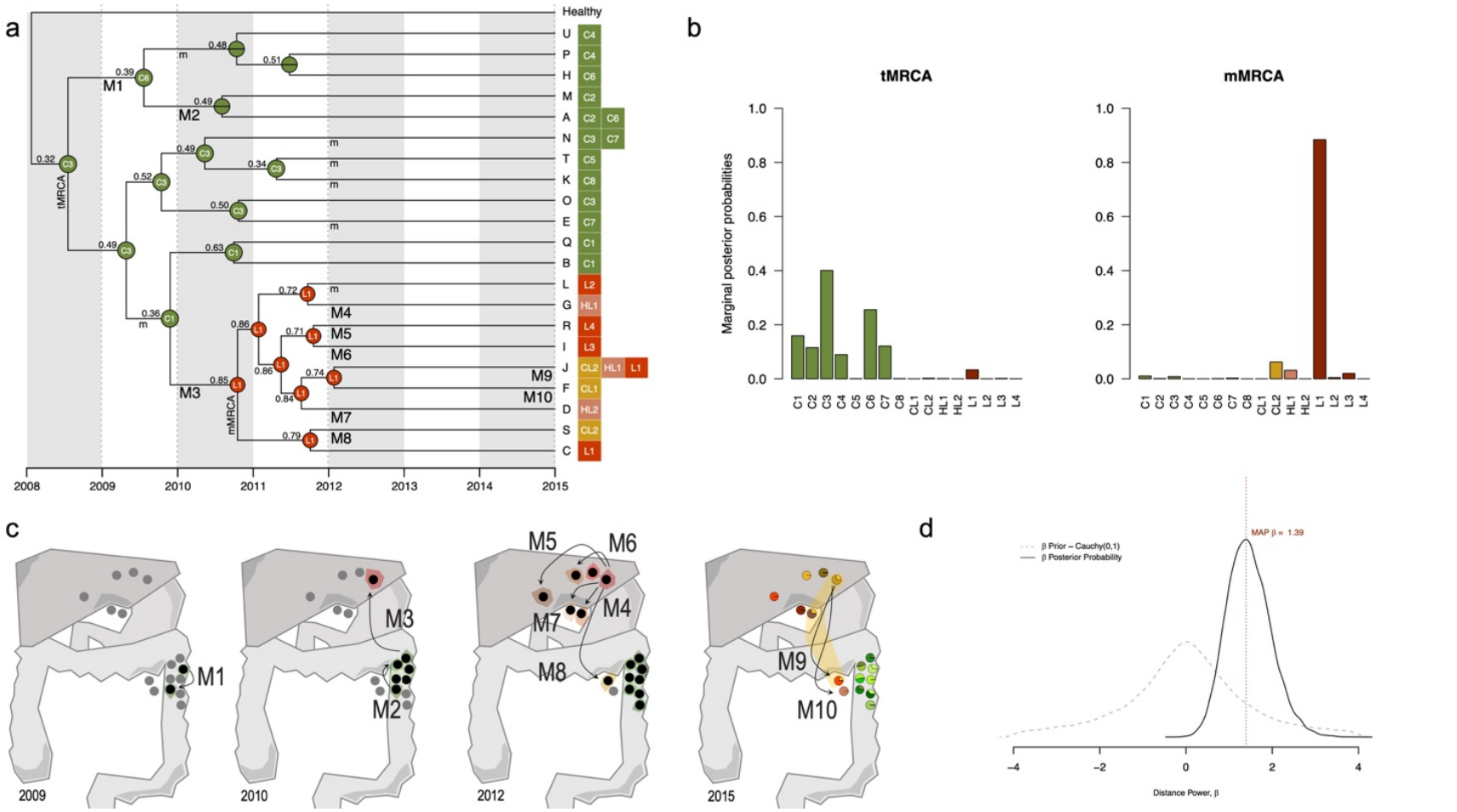

Figure 3. Inferred biogeographic history. a, Biogeographic reconstruction from BayArea, describing the geographical range (i.e., the set of occupied locations) of the ancestral clones. At each tree node, the range with the highest posterior probability is depicted. The sample ID is shown for those ancestral nodes whose inferred area ranges are restricted to a single location. The locations where the extant clones (A-U) were sampled are shown next to the tips. Migration events are depicted in the panel below represented by an uppercase "M" and numbered (M1-M10). A lowercase "m" indicates the remaining migrations inferred. **b,** Marginal posterior probabilities for the occupancy at single locations for the tumoral (tMRCA) and metastatic (mMRCA) ancestral clones. **c,** Schematic representation of the clonal dynamics in anatomical space over four time points. From 2009 to 2012, samples where BayArea inferred the presence of tumor clones are highlighted in black. Colored areas surrounding samples anatomical location represent the inferred spatial distribution of the clonal populations. Arrows highlight the inferred migration events. **d,** Comparison of the distance-dependent/independent dispersal models. The dashed grey line corresponds to the prior distribution for the distance power parameter, β ~ Cauchy(0,1). The solid black line indicates the posterior distribution obtained. The vertical dashed red line indicates the maximum *a posteriori* estimate of β.