Science. Author manuscript; available in PMC 2018 January 07.

Published in final edited form as:

Science. 2017 July 07; 357(6346): 55–60. doi:10.1126/science.aai8515.

Origins of lymphatic and distant metastases in human colorectal cancer

Kamila Naxerova[1,2,*], **Johannes G. Reiter**[3], **Elena Brachtel**[4], **Jochen Lennerz**[4], **Marc Van de Wetering**[5], **Andrew Rowan**[6], **Tianxi Cai**[7], **Hans Clevers**[5], **Charles Swanton**[6,8], **Martin A. Nowak**[3,9], **Stephen J. Elledge**[2,10], and **Rakesh K. Jain**[1]

[1]Edwin L. Steele Laboratories for Tumor Biology, Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA [2]Division of Genetics, Department of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA [3]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA [4]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA [5]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and UMC Utrecht, 3584CT Utrecht, the Netherlands; Cancer Genomics Netherlands, UMC Utrecht, 3584CG Utrecht, the Netherlands [6]The Francis Crick Institute, London, United Kingdom [7]Department of Biostatistics, Harvard University, Boston, MA 02115, USA [8]University College London Cancer Institute, London, United Kingdom [9]Department of Mathematics; Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA [10]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

Abstract

The spread of cancer cells from primary tumors to regional lymph nodes is often associated with reduced survival. One prevailing model to explain this association posits that fatal, distant metastases are seeded by lymph node metastases. This view provides a mechanistic basis for the TNM staging system and is the rationale for surgical resection of tumor-draining lymph nodes. Here, we examine the evolutionary relationship between primary tumor, lymph node and distant metastases in human colorectal cancer. Studying 213 archival biopsy samples from 17 patients, we used somatic variants in hypermutable DNA regions to reconstruct high-confidence phylogenetic trees. We found that in 65% of cases, lymphatic and distant metastases arose from independent subclones in the primary tumor, whereas in 35% of cases they shared common subclonal origin. Therefore, two different lineage relationships between lymphatic and distant metastases exist in colorectal cancer.

The spread of cancer cells from the primary tumor to regional lymph nodes is one of the most important factors predicting survival in patients with epithelial cancers (1). Lymph node metastasis uniformly associates with worse outcomes in breast (2), prostate (3), lung (4) and colorectal cancer (5), the most frequent cancers in the U.S. population. In colorectal

*Corresponding author. naxerova.kamila@mgh.harvard.edu.

carcinoma, the presence of cancer cells in tumor-draining lymph nodes defines stage III disease and triggers administration of adjuvant chemotherapy (6). The 5-year survival for patients with stage II (no lymph node metastases) is 82.5%, in contrast to 59.5% for patients with stage III disease (7).

In most patients, lymph node metastasis is not the cause of death, but is correlated with spread to vital organs (8). The association between lymphatic and distant metastasis has been known for at least 150 years (9) and, together with the observation that lymph node disease often precedes systemic disease, has engendered the view that affected lymph nodes may give rise to distant metastases (10–12). The concept of such a sequential progression or metastatic cascade (13), in which the primary tumor (T) seeds lymph node metastases (N), which in turn seed distant metastases (M), provides a mechanistic basis for the TNM staging system. A corollary of the sequential progression model is that surgical resection of positive lymph nodes will reduce recurrence rates. Indeed, resection of regional lymph nodes has been performed for more than 100 years (14). More recently, a number of clinical trials have shown that lymph node removal does not always improve patient survival (15). These findings have inspired the alternative view that lymph node metastases do not give rise to distant metastases (16) and suggest that treatment strategies may need to be re-evaluated (17).

Given its potential impact on patient care, a better understanding of the evolutionary relationship between lymph node and distant metastases is critical. We still do not know whether a single metastatic subclone evolves in the primary tumor, subsequently spreading to lymph nodes and distant sites (18–21), or whether multiple subclones in the primary tumor independently seed lymphatic and distant metastases (22–24). Here, we begin to examine these questions by studying the evolutionary history of colorectal cancer metastases.

Indels in hypermutable DNA enable reconstruction of tumor phylogenies

We conducted a systematic review of 1373 patient records and diagnostic materials at Massachusetts General Hospital (Fig. S1) and initially identified 19 colorectal cancer patients for whom formalin-fixed and paraffin-embedded samples from primary tumors, lymph node and distant metastases were available (Fig. 1A). We collected multiple tumor regions for each patient (mean 12.6, range 7-29) for a total of 239 samples (92 primary tumor biopsies, 59 lymph node metastases, 52 distant metastases, 36 normal tissue/germline samples) (Table S1). Of the 19 patients, 17 had liver metastases, one had an ovary metastasis, and one had multiple metastases in the omentum.

To trace the evolution of these cancers, we used a methodology (25) that leverages insertion/deletion (indel) mutations in hypermutable, non-coding polyguanine repeats (Fig. S2). The mutation rate of polyguanine repeats is several orders of magnitude higher than the mutation rate of non-repetitive DNA (26), making these sequences a rich reservoir of neutral somatic variation. Previous work has shown that indels in polyguanine repeats (27) as well as other microsatellites (28) accurately reconstruct evolutionary events modeled in cell culture. Furthermore, *in silico* models of polyguanine tract evolution have demonstrated that

revertant or parallel mutations do not significantly affect phylogenetic reconstruction accuracy when the number of interrogated markers is larger than 10 (29). These properties make polyguanine tracts attractive tools for phylogenetic analyses. Here, the mutation information from 20-43 polyguanine markers was generally sufficient to resolve lineages at our chosen confidence threshold (clade confidence > 70%, Fig. S3 and S4). Polyguanine markers were distributed across many chromosomes (Table S2). Therefore, any individual chromosomal alteration (gain or loss) would not be expected to substantially influence phylogenetic reconstruction. In total, our data set consisted of 19,541 individual genotypes. We developed a fully automated pipeline for data filtering, noise reduction and phylogenetic reconstruction (supplementary methods). First, to avoid artifacts created by contamination with normal cells, we implemented rigorous purity criteria, eliminating 11% of specimens from our study (Fig. S5 and Table S1). For all specimens belonging to the same patient, we then calculated a pairwise distance measure, the Jensen-Shannon distance (JSD) (30), over all polyguanine repeats. This distance reflected how much the samples had genetically diverged and formed the basis of our phylogenetic reconstruction with the neighbor-joining method (31).

Indel mutation patterns in the colorectal cancer cohort differed among patients. 81% of all observed alterations were deletions and 19% were insertions. A 4:1 ratio of deletions to insertions has previously been described by us and others (25, 32), indicating that it is an inherent property of polyguanine repeats. However, among individual cancers, we observed a relatively wide range of deletion frequencies, ranging from 45% in patient C69 to 100% in patient C77 (Fig. S6). This suggests that additional determinants, such as alterations in specific DNA repair proteins, may contribute to a skewing of mutation patterns in individual tumors. For example, a cancer showing microsatellite instability due to loss of *MLH1* protein expression almost exclusively harbored deletions (Fig. 1B) that also were of a significantly larger size than deletions in microsatellite stable tumors (Fig. 1C, Fig. S6).

Next, we explored whether accumulation of polyguanine indels is a cancer-related process or whether these mutations can be found in age-matched normal intestinal stem cells (ISCs). We analyzed DNA from 18 clonal expansions of human ISCs. Stem cell donors for 12 of these expansions were children (ages 4-14) and 6 expansions were from a 66-year old adult (Fig. S7). Adult ISCs had diverged significantly farther from a polyclonal germline reference than ISCs from children (Fig. 2A), suggesting that polyguanine indels accumulate in normal ISCs. We also observed a significant correlation between clonal mutation frequency and patient age in our colorectal cancer cohort (Fig. S8). The mean clonal mutation frequency in normal ISCs from the 66-year-old donor was lower than the mean frequency in cancers from age-matched (50-69 year old) patients (Fig. 2B), but with significant overlap of the two distributions. The baseline polyguanine mutation burden of a colorectal cancer therefore partially consists of alterations that are present in all intestinal cells.

Two distinct patterns of metastatic dissemination exist in colorectal cancer

The main goal of our study was to illuminate the evolutionary relationship between lymphatic and distant metastases. We aimed to sample lymph nodes as comprehensively as possible and included 91.3% of resected positive nodes in our analysis (Fig. S9,

Europe PMC Funders Author Manuscripts

supplementary methods). We first investigated the genetic distances among lymph node metastases, primary tumor biopsies, and distant metastases. For 33 of 45 (73%) lymph node metastases, the distance to the primary tumor (Fig. 2C) was shorter than the distance to distant metastases, and 31 of 45 (69%) distant metastases had a shorter genetic distance to the primary tumor than to any lymph node metastasis (Fig. 2D). This indicates that both types of metastatic lesions likely originated from distinct subclones in the primary tumor in most cases. To test this hypothesis, we examined all phylogenetic trees according to formal criteria. Patients were classified into two categories based on tree topology (Fig. 2E). We reasoned that lymph node and distant metastases had a common origin if a patient's tree contained a clade that contained at least one lymph node and at least one distant metastasis, but no primary tumor samples. Existence of such a branch indicates that both types of metastases were seeded from the same subclone, or that lymph node metastases gave rise to distant metastases. Formally, the reverse – seeding of lymphatic metastases from distant metastases – is also possible. A patient was classified as having distinct origins of lymphatic and distant metastases if no such clade existed. In all distinct origin cases, lymphatic and distant metastases were each more closely related to a primary tumor region than to each other.

To assess the robustness of each tumor's origin classification, we employed a bootstrapping strategy. We performed repeated random sampling (n=1000) of a tumor's mutation data to determine whether our origin classification was sensitive to changes in a limited number of polyguanine markers (supplementary methods). 14 of 17 tumors (82%) were classified with a bootstrap value above 80% (Fig. 2F), confirming the robustness of our phylogenetic data and our classification scheme. We also evaluated our classification by utilizing the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (33). (The nature of polyguanine genotyping data suggests the use of distance-based phylogenetic methods; see supplementary methods). Origin classification outcomes did not change for any of our 17 patients, further demonstrating the reliability of our results (all phylogenetic trees can by downloaded from datadryad.org, ID xxx).

Common origin of lymphatic and distant metastases

In 6 out of 17 tumors (35%), we found common origin of lymphatic and distant metastases. Selected phylogenetic trees with a classification confidence score above 80% for common origin are displayed in Fig. 3, along with pertinent clinical information.

Patient C38's cancer (Fig. 3A and anatomical sketch in Figure 1B) spread to the omentum and to the mesenteric lymph nodes. Furthermore, several satellite nodules had formed within the colonic epithelium, spatially separated from the primary tumor. We also investigated a piece of tumor that had invaded a vein. Notably, phylogenetic reconstruction showed that all lesions whose formation had depended on cell migration (that is, the satellite nodules, the distant and lymph node metastases, and the tumor within the vein) shared common ancestry, while the primary tumor had a divergent genetic profile.

Patient C69's cancer (Fig. 3B) showed a similar pattern, with one metastatic subclone giving rise to several liver metastases and a lymph node metastasis. Polyguanine indels clearly

attributed all metastases to the same evolutionary branch, even though M1, M2 and L1 were resected several months earlier than M3, and the patient received chemotherapy and bevacizumab between surgeries.

Patient C58 (Fig. 3C and anatomical sketch in Fig. 1C) had widespread metastases to the mesenteric lymph nodes and the liver. Most lymph node metastases were closely related to the primary tumor. However, a distinct subclone had formed in several lymph nodes that were located in close anatomical proximity (L3, L5, L6). The liver metastasis derived from the same subclone found in this lymph node group.

Further examples of patients with common origin of lymphatic and distant metastases are shown in Fig. 4D and Fig. S3. In all common origin cases, tree topologies are consistent with one of two modes of dissemination: the primary tumor seeded lymph node metastases, which in turn seeded distant metastases, with the formal possibility of the reverse. Or one genetically distinct ancestor evolved within the primary tumor and subsequently colonized lymph nodes and distant sites. In both scenarios, lymphatic and distant metastases share a common origin.

The common origin category is compatible with the idea of sequential progression and can explain important clinical observations, such as the well-established correlation between lymphatic and distant disease. In a majority of patients, however, tree topologies indicated independent seeding of lymphatic and distant metastasis from the primary tumor.

Distinct origins of lymphatic and distant metastases

Cancers in the distinct origins group, which encompassed 11 out of 17 patients (65%), contained multiple, genetically distinct metastasis ancestors. Fig. 4 shows selected phylogenetic trees with a distinct origin classification confidence score above 80% (the complete set, along with confidence values for each clade, is provided in Fig. S4).

Patient C66's tumor (Fig. 4A) is a representative example of the distinct origins category. The cancer harbored multiple subclones at different stages of evolution that had seeded genetically distinct metastases. Area P2, for example, was most closely related to lymph node metastasis L3, while area P1 was the origin of liver metastases M1 and M2. The tree shows that lymph node metastases were seeded continuously throughout the development of the tumor, but did not metastasize further. The liver metastases, on the other hand, arose in later evolution stages from the genetically most advanced clone. They constitute the terminal, most mutation-rich branch of the tree and, as a group, are more homogeneous than the lymph node metastases.

Patient C12's tumor (Fig. 4B) partially resembled that of patient C58. Its phylogenetic tree also showed a group of lymph node metastases (L2, L3, L4) that either derived from the same ancestral clone or gave rise to each other, while other lymphatic lesions (L1) were seeded independently. Notably, as for C58, the closely related nodes also were in anatomical proximity. However, the patient's liver metastasis (M1) did not arise from this subclone, but instead had distinct origins in primary tumor area P8.

Another noteworthy case from the distinct origins group is patient C53 (Fig. 4C) who underwent resection of two metastases located in the right liver lobe and one metastasis in the left liver lobe. Phylogenetic reconstruction showed that the metastases in the right liver diverged relatively early. After their divergence, the primary tumor evolved further and independently gave rise to lymph node metastasis L1 and the left liver metastasis, M3.

Further examples of cancers in the distinct origins category are shown in Fig. 4D-F and Fig. S4. In all these cases, the phylogenetic data indicate that lymph node metastases were not the source of distant metastases (also see explanatory schematic in Fig. S10).

Common clinicopathological variables do not correlate with origin classification

Next, we examined whether our origin classification was correlated with (and thus potentially influenced by) any clinicopathologic variables. We did not observe any significant differences in the number of positive nodes, the ratio of positive to examined nodes, the number of lymph nodes included in the final data set, the number of excluded nodes (Fig. S11A-D), the number of sampled primary tumor regions, the percentage of T3 vs. T4 stage patients (no T1/2 stage tumors were part of this cohort), the distribution of primary tumor sizes, the presence of vascular invasion, or the fraction of patients with synchronous vs. metachronous distant metastasis (Fig. S12A-E) between origin categories.

Most importantly, we found no association between origin and treatment history. Only one patient (C77) had neo-adjuvant chemotherapy (Table S3). In six patients, distant metastases were resected after the primary tumor and the lymph node metastases had already been removed, and all received treatment in the intervening time interval. Three of these patients (C69, C65, C36) fell into the common, and three (C66, C39, C63) into the distinct origins category (p=0.6).

Discussion

The presence of lymph node metastases is an important prognostic factor for most cancers, but the underlying reason has been unclear. One prevailing model posits that lymph node metastases are precursors of distant metastases and their surgical resection is necessary to attain a “cancer-free” state (34). An alternative model posits that distant metastases arise independently of lymph node metastases (16).

Our data show that lymph node metastases and distant metastases indeed often do have a common origin. While our phylogenies do not allow us to distinguish between sequential progression and common subclonal origin, many phylogenies in the common origin category are compatible with seeding of distant metastases from lymph nodes.

However, in a majority of patients, we find strong evidence of independent origins of lymph node and distant lesions. If independent seeding is prevalent, what is the reason for the association of lymphatic and distant metastasis? It could be that the association is driven by the common origin subset of patients. An alternative possibility is that most cells in tumors

Europe PMC Funders Author Manuscripts

belonging to the distinct origins category have the ability to metastasize. In such tumors, all cells that disseminate would have an increased likelihood of colonizing distant sites (35). Establishing lymph node metastases may be a more efficient process than establishing distant metastases and may therefore happen earlier and more frequently. This model would also be compatible with clinical observations, including the correlation between lymphatic and distant metastasis, the sometimes modest benefits of lymphadenectomy, and the advantage of early primary resection (assuming that even in such highly metastatic cancers, the survival rate of disseminated cells is relatively low, so that the tumor needs to grow to a certain size in order to metastasize efficiently).

Most metastases in our cohort were resected from the liver, which is the most frequent distant site of colorectal metastasis (36). Since venous blood from the intestines reaches the liver directly through the portal vein, it is possible that liver metastases are preferentially seeded hematogenously. Cancer cells that migrate through lymph nodes enter the venous circulation in the subclavian vein. The first capillary bed such cells encounter is the lung. It is therefore possible that lung metastases are more frequently seeded through the lymph nodes.

All cancers in our study were retrospectively collected specimens. Archival samples are mostly not suitable for whole genome or exome sequencing because patient consent for such comprehensive genetic profiling was not obtained at the time of surgery. Polyguanine repeat genotyping, on the other hand, is a limited analysis of length polymorphisms in non-coding DNA. It does not produce any information about functional or disease-related genes. Raw data produced by our method contain the lengths of PCR amplicons in arbitrary units, allowing for complete disclosure of mutation information while making patient identification impossible. Therefore, polyguanine repeat analysis represents a safe and effective method for studying tumor evolution in a patient population that would otherwise be inaccessible.

We conclude that the evolutionary relationship between lymphatic and distant metastases can take on two different forms in colorectal cancer. In the future, it will be important to determine whether cancers in the common and distinct origin categories exhibit different clinical behaviors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Matthias Nahrendorf, Filip Swirski, Jeff Gerold and Tim Padera for helpful comments and careful review of the manuscript. This work was supported by the Department of Defense W81XWH-10-0016 (R.K.J), W81XWH-12-1-0362 (S.J.E.), W81XWH-15-1-0579 (K.N.), National Human Genome Research Institute U54 HG007963 (T.C.), National Cancer Institute P01-CA080124 (R.K.J), R35-CA197743 (R.K.J.), Francis Crick Institute FC001169 (C.S.), Austrian Science Fund J-3996 (J.R.), Ludwig Institute for Cancer Research (S.J.E. and R.K.J). The Program for Evolutionary Dynamics is supported in part by a gift from B Wu and Eric Larson. K.N. conceived and designed the study and performed experiments. K.N. and J.G.R. analyzed data. E.B. and J.L. reviewed tissue specimens and clinical records. M.v.d.W., A.R., H.C. and C.S. provided DNA samples. K.N., J.G.R., E.B., J.L., T.C., C.S., M.A.N., S.J.E. and R.K.J. discussed results and strategy. R.K.J. supervised the study. K.N. wrote the manuscript, which was revised and approved by all authors. Raw polyguanine profiling data,

distance matrices and phylogenetic trees can be downloaded from https://steelelabs.mgh.harvard.edu/lymphmet and from datadryad.org (Accession number xxx).

References

1. Nathanson SD. Insights into the mechanisms of lymph node metastasis. Cancer. 2003; 98:413–423. [PubMed: 12872364]
2. McGuire WL. Prognostic factors for recurrence and survival in human breast cancer. Breast cancer research and treatment. 1987; 10:5–9. [PubMed: 3689982]
3. Gervasi LA, et al. Prognostic significance of lymph nodal metastases in prostate cancer. J Urol. 1989; 142:332–336. [PubMed: 2501518]
4. Naruke T, Suemasu K, Ishikawa S. Lymph node mapping and curability at various levels of metastasis in resected lung cancer. J Thorac Cardiovasc Surg. 1978; 76:832–839. [PubMed: 713589]
5. Chang GJ, Rodriguez-Bigas MA, Skibber JM, Moyer VA. Lymph node evaluation and survival after curative resection of colon cancer: systematic review. J Natl Cancer Inst. 2007; 99:433–441. [PubMed: 17374833]
6. National Comprehensive Cancer Network. [Accessed August 18, 2016] Colon Cancer (Version 2.2016). www.nccn.org. (available at https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf)
7. O'Connell JB, Maggard MA, Ko CY. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. J Natl Cancer Inst. 2004; 96:1420–1425. [PubMed: 15467030]
8. Wong SY, Hynes RO. Lymphatic or hematogenous dissemination: how does a metastatic tumor cell decide? Cell Cycle. 2006; 5:812–817. [PubMed: 16627996]
9. Virchow R. Die krankhaften Geschwülste: Dreissig Vorlesungen gehalten währen des Wintersemesters an der Universität zu Berlin. 1863
10. McBride CM. The surgeon as oncologist. South Med J. 1978; 71:1331–1333. [PubMed: 715481]
11. Halsted WS. I. The Results of Radical Operations for the Cure of Carcinoma of the Breast. Ann Surg. 1907; 46:1–19.
12. Sleeman J, Schmid A, Thiele W. Tumor lymphatics. Semin Cancer Biol. 2009; 19:285–297. [PubMed: 19482087]
13. Weinberg RA. Mechanisms of malignant progression. Carcinogenesis. 2008; 29:1092–1095. [PubMed: 18453542]
14. Moynihan B. The surgical treatment of cancer of the sigmoid flexure and rectum. Surg Gynecol Obstet. 1908; 6:463–6.
15. Gervasoni JE, Sbayi S, Cady B. Role of lymphadenectomy in surgical treatment of solid tumors: an update on the clinical data. Annals of surgical oncology. 2007; 14:2443–2462. [PubMed: 17597349]
16. Cady B. Lymph node metastases. Indicators, but not governors of survival. Arch Surg. 1984; 119:1067–1072. [PubMed: 6383272]
17. Engel J, Emeny RT, Hölzel D. Positive lymph nodes do not metastasize. Cancer metastasis reviews. 2012; 31:235–246. [PubMed: 22198520]
18. Liu W, et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat Med. 2009; 15:559–565. [PubMed: 19363497]
19. McPherson A, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nat Genet. 2016; 48:758–767. [PubMed: 27182968]
20. Gibson WJ, et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. Nat Genet. 2016; 48:848–855. [PubMed: 27348297]
21. Gerlinger M, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892. [PubMed: 22397650]
22. Yachida S, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467:1114–1117. [PubMed: 20981102]

Europe PMC Funders Author Manuscripts

23. Reiter JG, et al. Reconstructing metastatic seeding patterns of human cancers. Nat Commun. 2017; 8 14114.
24. Haffner MC, et al. Tracking the clonal origin of lethal prostate cancer. J Clin Invest. 2013; 123:4918–4922. [PubMed: 24135135]
25. Naxerova K, et al. Hypermutable DNA chronicles the evolution of human colon cancer. Proceedings of the National Academy of Sciences. 2014; 111:E1889–98.
26. Boyer JC, et al. Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet. 2002; 11:707–713. [PubMed: 11912186]
27. Salipante SJ, Horwitz MS. Phylogenetic fate mapping. Proceedings of the National Academy of Sciences. 2006; 103:5448–5453.
28. Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. Genomic variability within an organism exposes its cell lineage tree. PLoS computational biology. 2005; 1:e50. [PubMed: 16261192]
29. Salipante SJ, Thompson JM, Horwitz MS. Phylogenetic fate mapping: theoretical and experimental studies applied to the development of mouse fibroblasts. Genetics. 2008; 178:967–977. [PubMed: 18245843]
30. Endres DM, Schindelin JE. A new metric for probability distributions. IEEE Transactions on Information theory. 2003; 49:1858–1860.
31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4:406–425. [PubMed: 3447015]
32. Salk JJ, et al. Clonal expansions in ulcerative colitis identify patients with neoplasia. Proceedings of the National Academy of Sciences. 2009; 106:20871–20876.
33. SOKAL RR. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. 1958; 38:1409–1438.
34. Sleeman JP, Cady B, Pantel K. The connectivity of lymphogenous and hematogenous tumor cell dissemination: biological insights and clinical implications. Clin Exp Metastasis. 2012; 29:737–746. [PubMed: 22669542]
35. Vogelstein B, Kinzler KW. The Path to Cancer --Three Strikes and You're Out. N Engl J Med. 2015; 373:1895–1898. [PubMed: 26559569]
36. Patanaphan V, Salazar OM. Colorectal cancer: metastatic patterns and prognosis. South Med J. 1993; 86:38–41. [PubMed: 8420014]

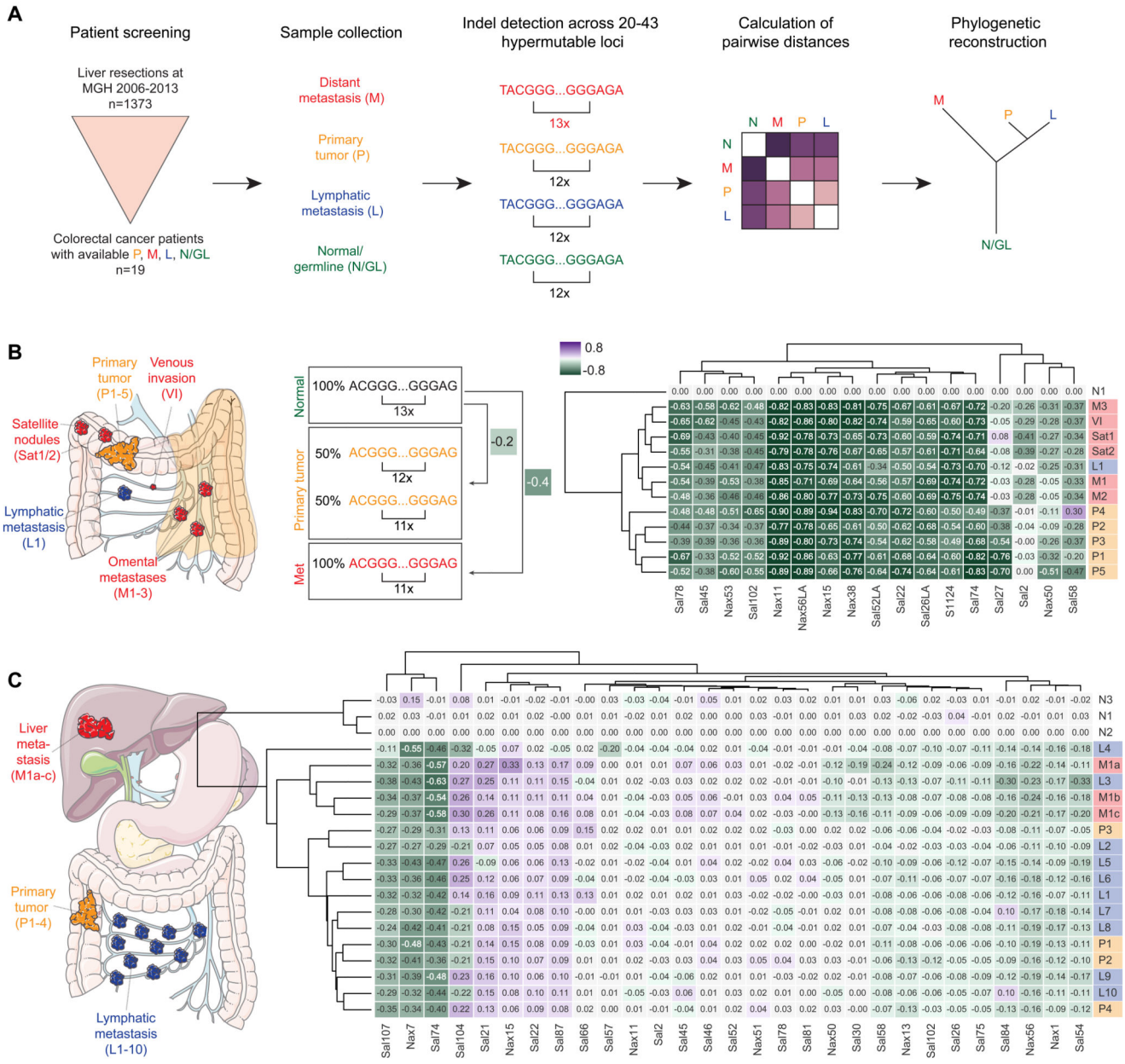

Fig. 1. Tracing tumor evolution through indels in hypermutable DNA.

(A) Study design schematic. DNA samples from primary tumor (P), distant metastases (M), lymph node metastases (L) and normal tissue/germ line (N/GL) from 19 colorectal cancer patients were genotyped across 20-43 hypermutable polyguanine repeats. The genetic divergence between two samples is the average distance across all markers. Pairwise distances between all samples from a patient were used as input for phylogenetic reconstruction with the neighbor-joining algorithm. **(B)** Anatomical sketch and raw data example for a cancer (C38) with microsatellite instability. Mutations can be present in varying percentages of cells within a sample. Therefore, the distance of a tumor sample to the normal reference is a continuous value. The heatmap shows tumor-normal distances

Europe PMC Funders Author Manuscripts

across all samples and markers. Dark green, deletions. Purple, insertions. Note that the heatmap only shows a small part of the full data set for a patient. Pairwise distances between all samples are used for phylogenetic reconstruction. **(C)** Anatomical sketch and raw data example for a microsatellite stable cancer (C58). Heatmaps for all patients are provided in Fig. S13.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

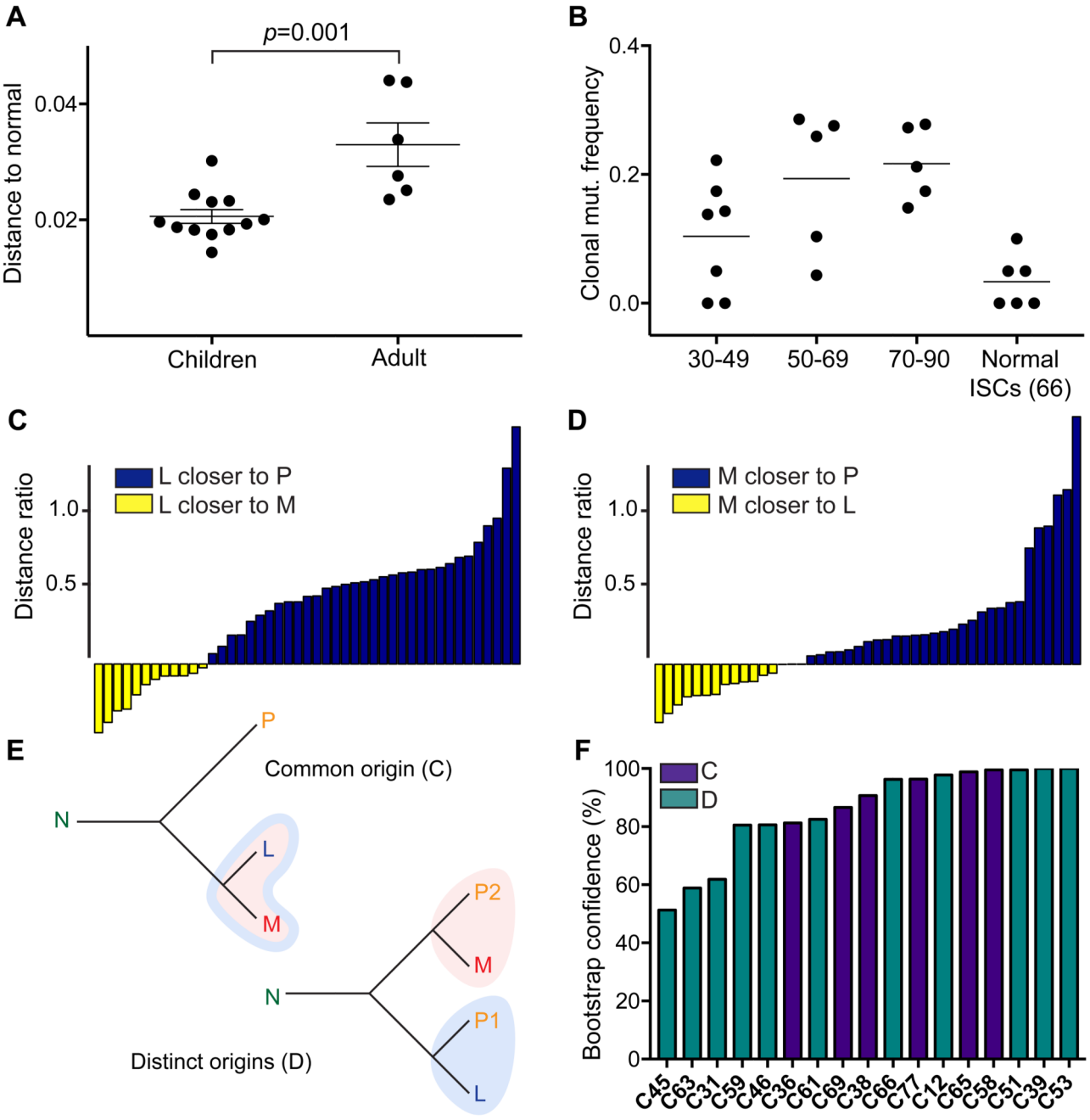

Fig. 2. Common versus distinct origins of lymph node and distant metastases.

(A) Clonal expansions of single intestinal stem cells (ISCs) from children (age<15, n=12) have fewer polyguanine indels than clonal expansions from a 66-year old adult (n=6). Data are mean +/- standard error of the mean, two-tailed t-test. **(B)** Clonal mutation frequency in cancers (defined as JSD => 0.11 in 95% of tumor biopsies) is correlated with patient age at diagnosis. Normal ISCs on average have fewer polyguanine indels than age-matched cancers, but the two distributions overlap. Lines indicate the mean. **(C)** Most lymph node metastases are more closely related to the primary tumor than to distant metastases. The plot

shows d(L-M)/d(L-P)-1, the distance of each lymph node metastasis (L) to its closest distant metastasis (M), divided by its distance to its closest primary tumor sample (P), minus one. Yellow, closest neighbor is a metastasis. Dark blue, closest neighbor is a primary tumor sample. **(D)** Analogous plot for distant metastases, showing d(M-L)/d(M-P)-1 **(E)** Classification of patients into cases with common or distinct origins of lymphatic and distant metastases. **(F)** Bootstrap values reflecting origin classification confidence for each patient. (Bootstrap n=1000).

Europe PMC Funders Author Manuscripts

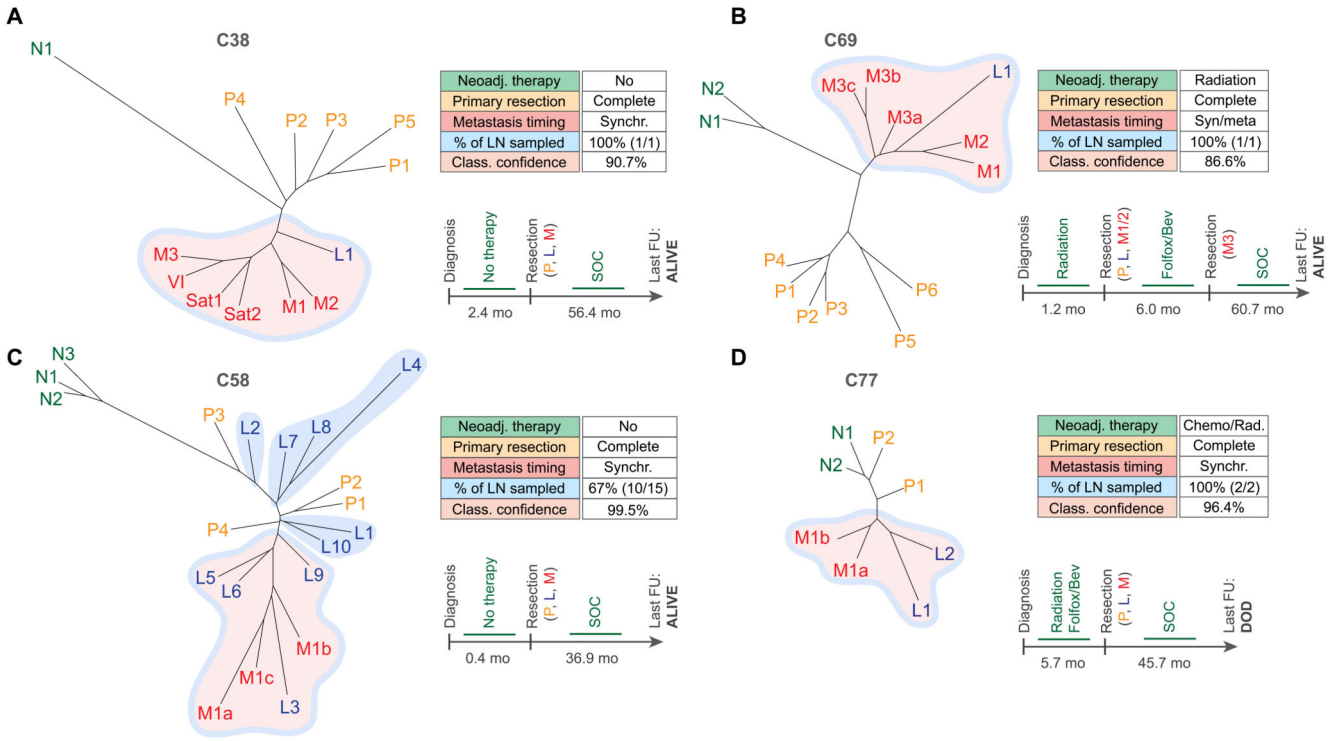

Fig. 3. Phylogenetic trees of cancers with common origin of lymphatic and distant metastases.

All trees except C38 (MSI case) are drawn to scale and were constructed with the neighbor-joining method. Seeding events (internal node/common ancestor and branches) that gave rise to distant metastases are shaded in red; events that gave rise to lymph node metastases are shaded in blue. Clinical information boxes show whether a patient received neoadjuvant therapy, whether primary tumor resection was complete (all margins unaffected), whether distant metastases occurred synchronously or metachronously, what percentage of suitable lymph nodes (i.e. those that were large and pure enough) was sampled, and the origin classification bootstrap value. Timelines summarize treatment and known lifespan for each patient. VI, venous invasion. Sat, satellite nodule. Lower case letters after sample numbers (a,b,c) indicate multiple biopsies from the same tumor mass. SOC, standard of care.

Europe PMC Funders Author Manuscripts

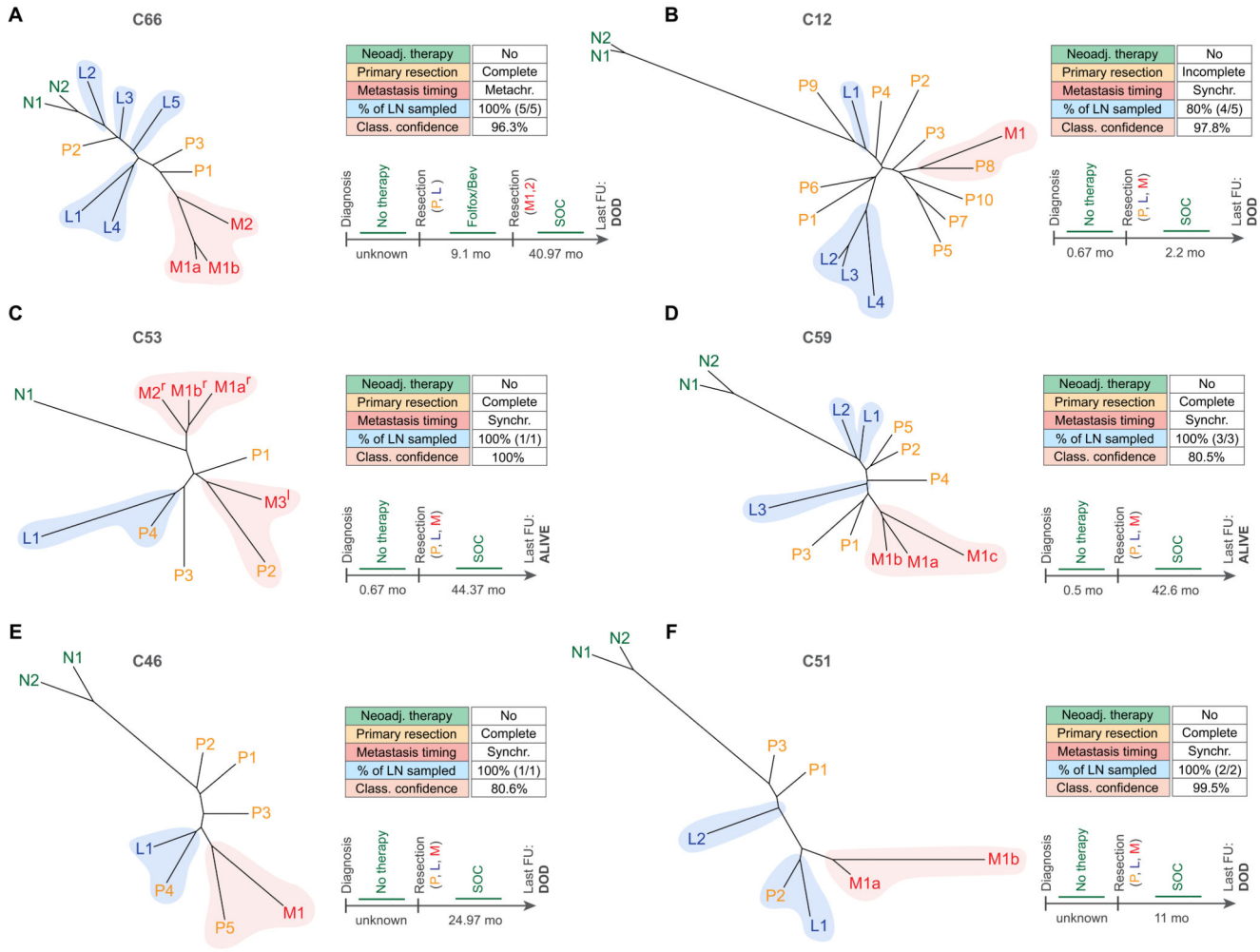

Fig. 4. Phylogenetic trees of cancers with distinct origins of lymphatic and distant metastases.
All trees except C12 (MSI case) are drawn to scale and were constructed with the neighbor-joining method. Shading and clinical information as in Figure 3.

Europe PMC Funders Author Manuscripts