Hypermutable DNA chronicles the evolution of human colon cancer

Kamila Naxerova[a,1], Elena Brachtel[b], Jesse J. Salk[c], Aaron M. Seese[d], Karen Power[d], Bardia Abbasi[e], Matija Snuderl[f], Sarah Chiang[g], Simon Kasif[h,i], and Rakesh K. Jain[a]

[a]Edwin L. Steele Laboratory for Tumor Biology, Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114; [b]Department of Pathology, Massachusetts General Hospital, Boston, MA 02114; [c]Department of Medicine, University of Washington, Seattle, WA 98195; [d]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139; [e]School of Medicine, Boston University, Boston, MA 02118; [f]Department of Pathology, New York University Langone Medical Center and Medical School, New York, NY 10016; [g]Department of Pathology, Memorial Sloan–Kettering Cancer Center, New York, NY 10065; [h]Department of Biomedical Engineering, Boston University, Boston, MA 02215; and [i]Children's Hospital Informatics Program, Harvard-MIT Division of Health Sciences and Technology, Boston, MA 02215

Edited* by Stanley M. Gartler, University of Washington, Seattle, WA, and approved March 28, 2014 (received for review January 6, 2014)

Intratumor genetic heterogeneity reflects the evolutionary history of a cancer and is thought to influence treatment outcomes. Here we report that a simple PCR-based assay interrogating somatic variation in hypermutable polyguanine (poly-G) repeats can provide a rapid and reliable assessment of mitotic history and clonal architecture in human cancer. We use poly-G repeat genotyping to study the evolution of colon carcinoma. In a cohort of 22 patients, we detect poly-G variants in 91% of tumors. Patient age is positively correlated with somatic mutation frequency, suggesting that some poly-G variants accumulate before the onset of carcinogenesis during normal division in colonic stem cells. Poorly differentiated tumors have fewer mutations than well-differentiated tumors, possibly indicating a shorter mitotic history of the founder cell in these cancers. We generate poly-G mutation profiles of spatially separated samples from primary carcinomas and matched metastases to build well-supported phylogenetic trees that illuminate individual patients' path of metastatic progression. Our results show varying degrees of intratumor heterogeneity among patients. Finally, we show that poly-G mutations can be found in other cancers than colon carcinoma. Our approach can generate reliable maps of intratumor heterogeneity in large numbers of patients with minimal time and cost expenditure.

lineage tracing | microsatellites | tumor phylogenetics

Human cancers are composed of a continually evolving population of genetically and phenotypically divergent cells (1). This reservoir of diversity feeds the natural selection process that fundamentally drives disease progression through acquisition of metastatic properties and emergence of therapy-resistant clones (2–4). In recent years, characterization of intratumor heterogeneity has received increased attention as advanced sequencing technologies have enabled more detailed analysis of tumor cell populations (5–8).

Depending on the context, the term "intratumor heterogeneity" refers either to differences between cells that coexist in one localized tumor region or to variation in clonal composition between spatially separated parts, most notably between a primary tumor and its metastases (in the latter case, "intracancer heterogeneity" is a more appropriate terminology). The extent of genetic divergence between primary and metastatic tumors (and the history of dissemination encoded therein) is beginning to be investigated, but relatively few patient data are currently available. The canonical "linear progression" model of metastasis states that a genetically advanced cell metastasizes late in primary tumor development (9–11). This aggressive clone generates new metastases in a so-called "metastasis shower" (12). Linear progression predicts that metastases will be genetically similar to the primary tumor and to each other. The alternative "parallel progression" model (9) posits that metastasis occurs early in tumor evolution and consequently expects metastases to be substantially different from one another, and from the primary tumor, because they evolve separately over long periods of time. As more data become available, both scenarios can likely be corroborated. Importantly, different modes of metastasis may be prevalent in different cancer types. For example, studies of pancreatic adenocarcinoma (7) and triple-negative breast cancer (8) demonstrated that the primary tumor and its metastases share a majority of mutations, thereby indicating late dissemination. A recent comparative sequencing study in renal cell carcinoma, on the other hand, found substantial genetic divergence among primary and metastatic tumors (5). Notably, however, two metastases in distinct anatomical locations were almost identical to one another, suggesting a common founder clone related to a spatially discrete portion of the primary tumor. This example highlights how studying intratumor heterogeneity and mitotic history can reveal the evolution of systemic disease. Many clinically relevant questions in this area remain unanswered. What role does heterogeneity play at different progression stages? Clonal diversity in early, preneoplastic lesions increases the risk of malignancy (13); the final step of disease advancement, metastasis, on the other hand, appears to go hand in hand with a steep drop in intracancer heterogeneity (14). Does heterogeneity increase resistance to therapy (15), or is homogeneity created by the late expansion of a particularly aggressive clone associated with resistance?

Addressing these and other questions about the evolution of metastatic cancer will require analyzing large numbers of

Significance

Genetic heterogeneity in systemic cancer is of great clinical interest because it impacts therapeutic response and reflects how tumor cells grow and spread. We present a methodology that enables efficient evaluation of intratumor heterogeneity in patients through analysis of neutral somatic variation hotspots. Using only 20 genomic markers, we demonstrate a unique pattern of clonal diversity in each patient. Some tumors are significantly more diversified than others. Our data suggest that distinct clones can give rise to lymphatic and distant metastases. Our methodology is applicable to other human cancer types and facilitates high-throughput investigation of tumor evolution.

Author contributions: K.N. and R.K.J. designed research; K.N., A.M.S., K.P., and B.A. performed research; E.B., M.S., and S.C. contributed new reagents/analytic tools; E.B. provided human tissue samples; J.J.S. contributed analysis and interpretation; M.S. and S.C. contributed human tissue samples.; R.K.J. supervised and guided the research; K.N., J.J.S., and S.K. analyzed data; and K.N. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: naxerova.kamila@mgh.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1400179111/-/DCSupplemental.

patients with different types of tumors. Ideally, whole genome or exome sequencing would be performed on multiple specimens from each patient. With sequencing capacities continually rising, this approach will likely become feasible in the future. Presently, although, only large genome centers can regularly generate and process datasets of this magnitude. A further complication is that broad DNA sequencing of most archival clinical specimens is precluded due to a lack of patient consent. To study intratumor heterogeneity more efficiently, and therefore more widely, it would be expedient to target selected regions of the tumor genome that are enriched for somatic variation. Genes frequently altered in cancer are an option, but because driver mutations affect competitive advantage, their distribution may not reflect the correct phylogenetic relationships among tumor cell populations. Accurate reconstruction of cell division and migration events that occurred during tumor evolution can also be achieved with neutral genetic markers. Short repeats (microsatellites) in noncoding regions are especially suited for this purpose. Due to replication slippage (16), mutations are introduced frequently but presumably have no effect on fitness. In patients with DNA mismatch repair (MMR) defects and resulting microsatellite instability (MSI), variation in dinucleotide repeats has been used to study several aspects of tumor progression (17–19), but mutation rates in tumors with intact MMR are too low to make this approach widely applicable (20).

Recent research identified a particularly mutable class of polyguanine (poly-G) repeats as a hotspot of somatic variation even in normal cells (21). Analysis of poly-G repeats has successfully been used to study phylogenetic relationships between single cells in mouse development (22–24) and has been adapted for detecting preneoplastic clonal expansions in ulcerative colitis patients (25).

Here we show that analysis of poly-G repeats can determine lineage relationships in human cancer. We analyze a cohort of 22 colon cancer patients and find that most tumors contain an abundance of poly-G variants. We use poly-G mutation profiles to build well-supported phylogenetic trees that show ancestral relationships between primary tumors and their metastases. Our work demonstrates how a simple and highly scaleable assay can be used to generate reliable maps of clonal architecture in formalin-fixed and paraffin-embedded (FFPE) tumor samples.

Insertions/deletions of one or more base pairs (bps) in poly-G runs are a byproduct of normal replication. Human DNA polymerase replicates unique sequences with high fidelity, but replication accuracy significantly decreases in short tandem repeats (26, 27). Guanine homopolymers are particularly prone to replication slippage errors and can have mutation frequencies as high as 10^{-4} per base per cell division (28). Fig. 1 illustrates schematically how poly-G variants accumulate in genetic lineages as the zygote divides to give rise to the trillions of cells that constitute the adult human. A given poly-G tract has a certain probability of undergoing an insertion or deletion mutation during each division. This probability depends on a variety of factors, including the composition of the sequence surrounding the poly-G tract (26), and generally increases with repeat length (29). Because mutations are inherited by all daughter cells, each cell's unique mutational profile encodes its cell division history and its location in the organism's "cell lineage tree" (30–32). If single cells were isolated and their genomes individually analyzed, it would be possible to reconstruct the phylogenetic relationships between them, as has been demonstrated in murine development (22) and cell culture (21) using poly-G tracts, other microsatellites (30, 32), or random genomic regions (33) as lineage markers.

The primary drawback of this approach is that it can be very challenging to expand single cells from normal tissue to generate sufficient material for sequence analysis, and whole genome amplification can introduce artifacts for which it is difficult to

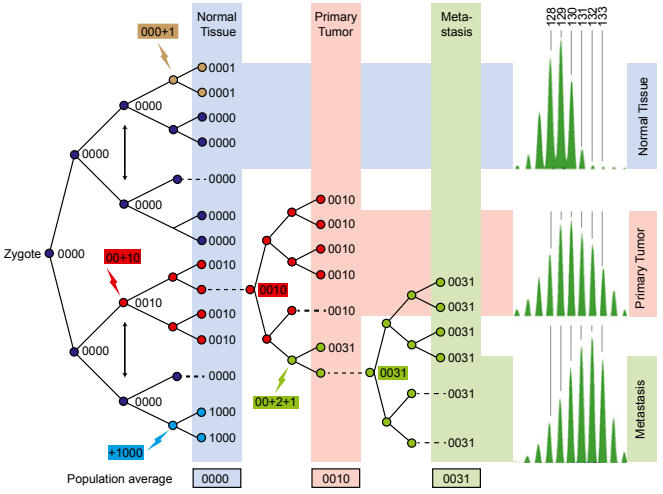

Fig. 1. Propagation of neutral poly-G mutations in normal and neoplastic somatic cell lineages—schematic representation. The vector (0000) represents the genotype of the zygote at four hypothetical poly-G alleles. During each cell division, an allele has a defined probability of undergoing a length alteration, noted as –1 for a deletion and +1 for an insertion. As cells divide and acquire mutations during development, extensive mixing occurs (black arrows between tree branches). As a result, mature tissues consist of cells that are derived from all branches of the tree, all harboring distinct mutational profiles. When a sample of normal tissue is analyzed, a majority of cells will not be mutated at any given locus, and the sample will have the zygote genotype (blue bar symbolizing cell composition of normal tissue sample). During tumorigenesis, the clonal expansion of one founder cell leads to a locally confined population of cells that all share its genotype (red bar) and can thus be differentiated from the zygote genotype. The founding of a monoclonal metastasis (green bar) is analogous. The right side shows examples of poly-G genotypes for marker Sal45 for normal tissue, a primary colon cancer, and a metastasis to the ovary. A family of fragments is generated during PCR due to the high mutability of poly-G tracts. The highest intensity peak (in this example, 129 bp in normal tissue, 130 bp in the primary tumor, and 132 bp in the metastasis) corresponds to the true length of the poly-G tract in the sample; adjacent peaks are created by slippage of Taq polymerase during amplification.

control. In bulk tissue analysis, on the other hand, the genomes of hundreds of thousands or millions of cells from divergent genetic lineages are combined in one sample and the mutational profile of any single cell is rendered undetectable. Even in relatively homogeneous tissues, such as the liver parenchyma, cells derive from many different branches of the cell lineage tree because extensive mixing occurs during development (22). The result is that at any given locus, most cells will not be mutated. Analyzing a bulk tissue sample therefore yields the genotype of the most recent common ancestor of all cells—that is, the zygote or "germline" genotype in the case of normal tissue (34).

A fundamentally different scenario arises during carcinogenesis, as one transformed cell begins to proliferate and create a locally confined population of daughter cells that are all closely related to each other. Sampling this population will reveal the genotype of the most recent common ancestor—the tumor founder cell. As the tumor grows, it accumulates new mutations that may become detectable if a clone becomes locally dominant or metastasizes to form a colony of homogeneous progeny at a distant site. Phylogenetic analysis relying on bulk tissue samples is therefore uniquely possible in cancer because clonal expansions unmask genetic variants that can be used to trace lineage. The right panel of Fig. 1 shows examples of poly-G tract genotypes in normal (polyclonal) human tissue, a primary tumor, and its metastasis. Because poly-G tracts are inherently hypermutable, Taq polymerase slippage during PCR generates a fragment distribution instead of a single product. This fragment distribution

or "PCR stutter distribution" can be precisely quantified, at single bp resolution, by capillary electrophoresis following PCR with fluorescent primers. The highest intensity peak represents the true genotype. If a tumor sample stutter pattern shifts from the normal reference derived from the same patient, then that sample contains new mutated alleles (*SI Appendix*, Fig. S1 contains representative examples of poly-G mutations and demonstrates PCR reproducibility). The presence of two primary peaks reflects heterozygosity in a sample. We sought to determine whether mutations in poly-G sequences could be found in human colon cancer patients.

Results

Poly-G Mutations Are Present in Most Colon Cancers. We began by screening a cohort of 22 human colon cancers for somatic mutations in a panel of 20 poly-G tracts. The cases in our cohort were consecutive patients who underwent colectomy and received a diagnosis of invasive carcinoma at Massachusetts General Hospital (see *Materials and Methods* for more details on patient selection). Anonymized patient information (pathological diagnosis, tumor size, histologic grade, stage, anatomic location of the tumor, neoadjuvant therapy, etc.) is presented in *SI Appendix*, Table S1. Because our ultimate goal was to study metastatic progression, we further subselected patients who had at least three lymph node metastases and/or distant metastases. Next, we screened matched pairs of primary tumor and normal tissue for poly-G variants at 20 genomic loci. DNA was extracted from FFPE tissue cores and subjected to poly-G tract profiling. Mutated alleles were found in 91% of patients (Fig. 2*A* and complete genotype information in *SI Appendix*, Table S2). As expected, colon cancers with MSI harbored the most alterations. Nevertheless, MSS tumors also contained abundant mutations. These mutations were qualitatively different from those observed in MSI cancers, indicating slippage errors during normal DNA replication rather than defective DNA MMR. Loss of DNA MMR proteins, such as MLH1 and PMS2, leads to frequent generation of new alleles in the growing tumor and results in a distinctively broadened stutter distribution. The changes that we observed in MSS tumors, on the other hand, typically consisted of a shift of the stutter pattern by 1 or 2 bp without broadening of the distribution, pointing to the presence of just one new allele that was shared by a large percentage of sampled cells (Fig. 2*B*). Characteristics of mutations found in MSS tumors are detailed in *SI Appendix*, Fig. S2. We found 66 (80%) deletions and 17 (20%) insertions. The same 4:1 ratio of deletions to additions was previously reported in a study of poly-G mutations in ulcerative colitis patients (25), implying that replication slippage in poly-G tracts preferentially leads to loss of repeat units. The preponderance of deletions also suggests that most of the observed changes were not caused by loss of heterozygosity (where longer and shorter alleles would have an equal chance of being affected). Eighty-three percent of alterations involved 1 bp, and 17% involved 2 or 3 bp. We do not know whether larger mutations (>1 bp) arose in a stepwise fashion or during a single larger replication slippage event. Alterations involving multiple bases do occur in poly-G repeats, albeit less frequently than single bp mutations (28). Therefore, for the overview in Fig. 2*A*, we counted every alteration, regardless of its magnitude, as one mutation. Given the relatively small percentage of larger changes, we do not expect misclassification of step-wise mutations as one-time events to be a significant source of bias. In contrast to MSS tumors, MSI tumors contained an abundance of large mutations between 4 and 13 bp (*SI Appendix*, Table S2). Ninety-six percent of these

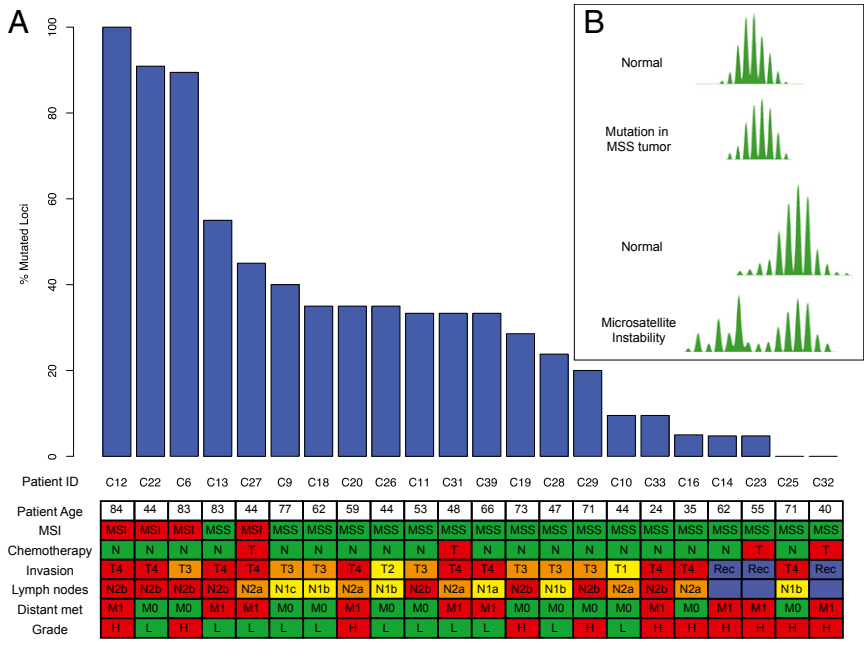

Fig. 2. Poly-G mutations in 22 human colon cancers. (*A*) Mutation frequency plotted as mutations per number of interrogated loci for each patient. Clinical characteristics are listed in the table below, including defects in DNA MMR (MSI, red; MSS, green), chemotherapy (N, green, therapy naïve; T, red, neoadjuvant therapy), extent of invasion (Rec, blue, recurrence, no primary tumor; T1, yellow, into submucosa; T2, yellow, through submucosa and extending into muscularis propria; T3, orange, through muscularis propria into pericolorectal tissues; T4, red, through serosa), lymph node status (N1a, yellow, metastasis in one lymph node; N1b, yellow, 2–3 lymph nodes; N1c, yellow, tumor deposits in the subserosa, mesentery, or nonperitonealized pericolic or perirectal tissues without regional lymph node metastasis; N2a, orange, metastasis in 4–6 lymph nodes; N2b, red, seven or more regional lymph nodes), distant metastasis (M0, green, absent; M1, red, present), and histologic grade (H, red, high grade; L, green, low grade). (*B*) Genotypes for marker Sal52 in a MSS cancer (*Upper*) and a cancer with MSI (*Lower*). The presence of a multitude of alleles in the MSI sample leads to a broadening of the distribution, whereas a simple shift indicates a single mutation in the MSS tumor.

were deletions, consistent with previous reports of an increased deletion-to-insertion ratio in DNA MMR-deficient cells (28).

The mutational fingerprint of each cancer is composed of two distinct types of alterations: "founder mutations" that were already present in the cell of origin at the time of transformation and "progressor mutations" that accumulated during tumor development. Colonic stem cells divide very frequently—every 30 h by some estimates (35)—and would therefore be expected to accumulate large numbers of founder mutations over the years, with total mutational burden increasing with age. Recent studies show a correlation between age at diagnosis and total number of somatic mutations in acute myeloid leukemia (36) and colorectal cancer (20, 37). We tested this correlation in our data after excluding MSI cases, as a distinct mutational mechanism is operational in these tumors. We found a significant positive correlation between patient age and mutation frequency (Fig. 3*A*), suggesting that mutations already present in the genomes of normal founder cells at the time of tumor initiation constitute an appreciable portion of the poly-G tract mutation profile. Tumor size, lymph node status, and presence of distant metastases were not significantly associated with mutation frequency, but because we specifically selected cases with lymphatic or distant metastasis, our cohort is biased for patients with advanced disease and not suited for rigorously testing this relationship. Exposure to neoadjuvant chemotherapy was also not associated with the number of poly-G variants per tumor.

However, we did find a highly significant inverse correlation between mutation frequency and histologic grade (Fig. 3*B*). Age and tumor grade were not associated. We excluded differences in tumor-derived DNA content as a confounding factor (Fig. 3*C*). Representative histological images visualizing tumor cell content in low- and high-grade tumors are provided in *SI Appendix*, Fig. S3. A majority of high-grade cancers had mutation frequencies close to zero—that is, they harbored neither progressor nor founder mutations, even in the upper quartile of the age distribution. For example, patient C25 had a mutation frequency of 0% at age 71. Progressor mutations accumulate during tumor development, and their number could conceivably be low if a tumor founder cell already contains a particularly advantageous set of oncogenic mutations that allow it to expand rapidly. However, the lack of a substantial number of founder mutations is surprising. One possible explanation is that poorly differentiated tumors derive from a cell population with relatively short mitotic history, such as a distinct, quiescent stem cell population that divides more rarely than the colonic stem cell whose proliferation constantly replenishes the epithelial compartment (38).

Poly-G Tract Profiles Generate a Map of Tumor Evolution. Founder mutations, by definition, are present in all tumor cells. Progressor mutations, on the other hand, may be differentially distributed across tumor regions and can be used for lineage tracing. To determine whether poly-G mutations could be used to reconstruct phylogenetic relationships between multiple tumor samples from the same patient, we selected four patients for deeper analysis. We collected between 8 and 15 spatially separated samples from different regions of the cancer (primary tumor mass, lymph node metastases, and distant metastases) and generated poly-G tract profiles for each sample using the same 20 markers used in our initial screen. (For tumors that were sampled repeatedly, we chose the primary tumor region with the greatest number of mutations for the overview in Fig. 2*A*.) To facilitate data analysis, we developed a semiautomated method for converting poly-G stutter distributions into genotypes (detailed in *Materials and Methods* and *SI Appendix*, Fig. S4). Finally, we created phylogenetic trees illustrating the lineage relationships between all sampled tumor parts. Every patient's tree provided unique insights into tumor evolution and metastatic progression. We have included phylogenetic analyses of four additional colorectal cancer cases in *SI Appendix*, Figs. S6–S9.

Poly-G tract profiling assigns metastases to their tumor of origin. We began by examining a case in which the phylogenetic relationships were at least partially known. Patient C39 was a 66-y-old male who underwent total colectomy without neoadjuvant chemotherapy and was found to have two spatially separated foci of invasive carcinoma, a 5.5 cm tumor in the cecum that arose within an adenoma and a 6 cm tumor in the sigmoid colon (Fig. 4*A*). Both cancers were low grade. One of the dissected lymph nodes near the inferior mesenteric artery revealed metastatic carcinoma in close proximity to the sigmoid tumor. We asked whether poly-G tract profiling could accurately link the lymph node metastasis to its tumor of origin in the sigmoid colon and moreover determine whether the two carcinomas had common or independent origins. We found seven variants in the most mutated parts of the cecal tumor and seven in the sigmoid lesion. That the tumors had the same number of mutations suggested similar mitotic ages, yet the mutations were largely mutually exclusive (Fig. 4*B* and full genotype data in *SI Appendix*, Table S3; because both tumors had similar numbers of mutations, only the sigmoid tumor is depicted in the overview in Fig. 2*A*). The phylogenetic tree constructed from these data located the two tumors in two independent evolutionary branches with high confidence values based on 1,000 bootstrap replicates (Fig. 4*C*). The lymph node metastasis was correctly assigned to the sigmoid tumor's branch.

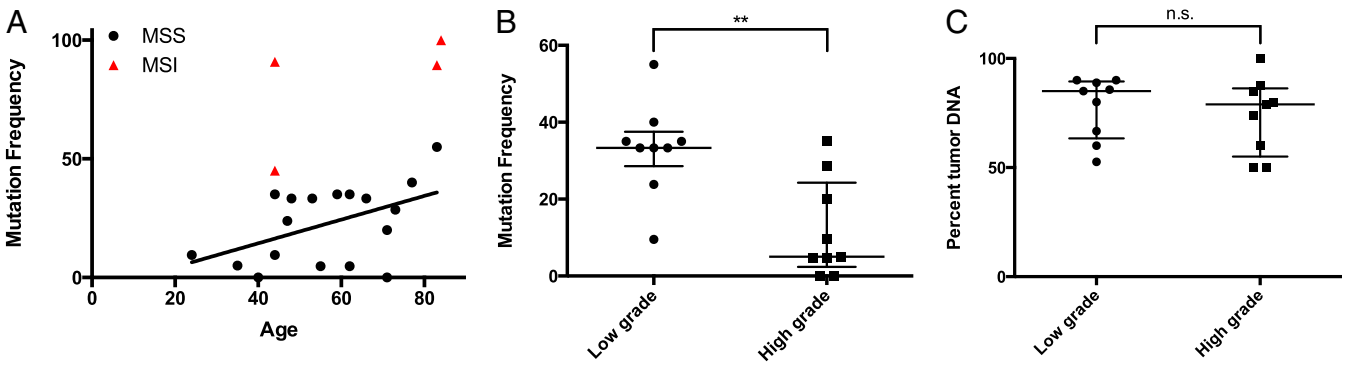

Fig. 3. Association between mutation frequency and age/histologic grade. (*A*) Age is positively correlated with mutation load in MSS tumors, $P = 0.0416$ (linear regression after exclusion of MSI cases, $R^2 = 0.23$) (MSI, red triangles; MSS, black dots). (*B*) Low-grade tumors contain more mutations than high-grade tumors, $P = 0.0041$ (two-tailed Mann–Whitney test). Only MSS tumors are included in this comparison. (*C*) Tumor DNA content in poorly differentiated and well-differentiated tumors is similar, $P = 0.4$ (two-tailed Mann–Whitney test). Individual values are plotted with their median and interquartile range in both *B* and *C*.

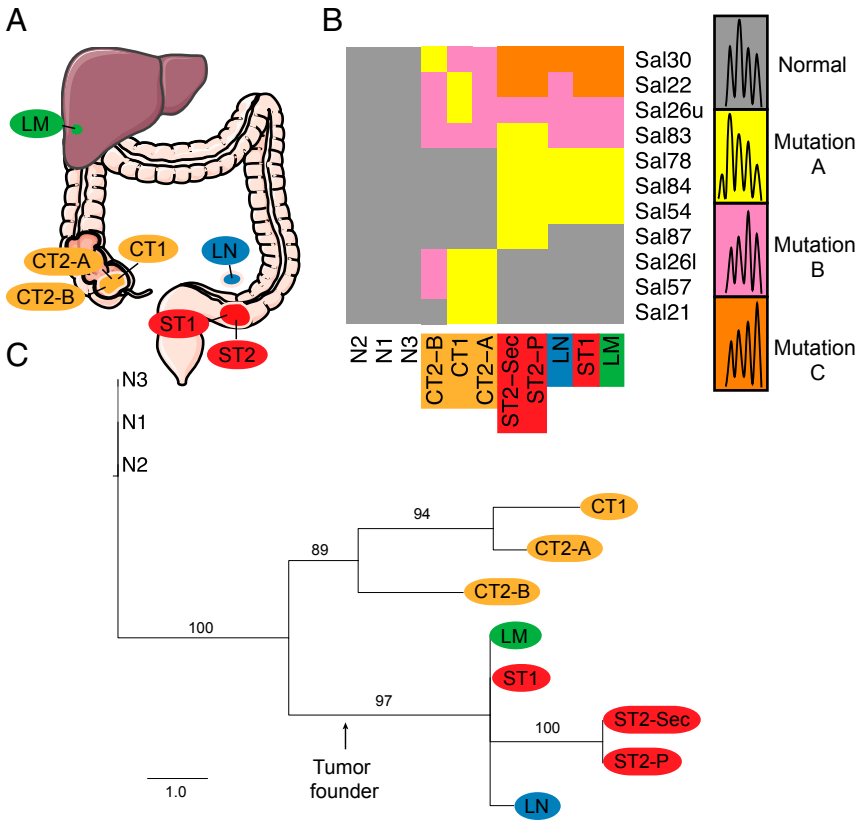

Fig. 4. Patient C39 with two synchronous adenocarcinomas of the colon. (*A*) Tumor location overview. CT, cecal tumor; LM, liver metastasis; LN, lymph node metastasis; N, normal; ST, sigmoid tumor. Tumor sizes are drawn to scale. Letters A and B indicate that two samples were taken from the same FFPE block. Additions "P" and "Sec" indicate that a block was analyzed twice, once via punch biopsy (P) and once via macrodissection of tissue sections (Sec). All other samples are derived from separate blocks. (*B*) Complete mutation heatmap, with poly-G markers in the rows and patient samples in the columns. Gray squares signify allele distributions that are indistinguishable from a normal reference sample (N1). Colored squares indicate a shift in allele distribution—that is, a poly-G mutation. If multiple different mutations exist per marker, they are indicated with additional colors. The right panel shows hypothetical examples of poly-G mutations. Because each marker harbors a distinct and unique set of mutations, we simply denote them with mutation A (yellow), B (pink), and C (orange) for the purposes of the heatmap. Detailed mutation information, including magnitude and direction, is provided in *SI Appendix*. (*C*) Phylogenetic tree constructed by neighbor-joining. Confidence values for each interior branch were calculated from 1,000 bootstrap replicates and are displayed adjacently. Branches with confidence values below 70% were collapsed into polytomies (i.e., nodes that give rise to more than two branches because the available mutation information was not sufficient to further resolve lineage relationships at the desired confidence level). The tree was rooted using a normal tissue sample as an outgroup. As expected, all normal samples have the same genotype.

One year after the initial surgery, and after six cycles of adjuvant chemotherapy with folinic acid, fluorouracil, oxaliplatin (FOLFOX), two liver metastases (1 cm and 0.5 cm) were resected. We genotyped the smaller lesion, and phylogenetic reconstruction connected it to the same evolutionary branch as the sigmoid tumor and excluded the cecal carcinoma as a source of metastasis. Notably, the liver lesion had the same mutational profile as sigmoid tumor area ST1, which was removed before administration of adjuvant chemotherapy, indicating that in this patient, cytotoxic therapy had not substantially changed the poly-G tract profile.

Extensively diversified primary tumor gives rise to homogeneous metastases. Patient C13 was an 83-y-old female with a 7.0 cm invasive colonic adenocarcinoma and metastases to the left and right ovaries (Fig. 5*A*). All lesions were removed in one surgery; the patient did not receive any prior chemotherapy. The tumor was moderately differentiated (low grade), was MSS, and involved the ileum, ileocecal valve, and cecum. We generated poly-G tract profiles for three normal tissue samples, eight primary tumor samples, three right ovary metastasis samples, and four left ovary metastasis samples. (A detailed description of specimens based on the surgical pathology report and the full genotype data are provided in *SI Appendix*, Table S4.) Fourteen loci were mutated in at least one sample, and each sample contained at least seven distinct mutations (Fig. 5*B*). As expected, all normal samples had the same genotype across all poly-G tracts. The primary tumor, by contrast, was highly diversified. Tumor regions PT5 and PT7 clustered in a distinct branch that had segregated from the rest of the tumor very early in its evolution (Fig. 5*C*). Neither region shared the majority of mutations found in other parts of the tumor, but instead harbored unique variants not found in any other sample. The ileal portion of the tumor (PT3-A and PT3-B) produced two samples that were identical to each other, yet distinct from the cecal part of the tumor. Tumor regions PT1 and PT6 shared a majority of mutations and were almost identical to samples from the ovarian metastases. All metastases clustered together on the branch with the greatest "depth" (39)—that is, the branch that contained the most mutated samples and was separated from the normal root by the greatest number of cell divisions. The tree allowed us to answer several important questions about this cancer's evolution. We observed extensive heterogeneity between different regions of the primary tumor, indicating that clonal populations had evolved locally for some time without intermixing. Some parts were so distinct from each other that we could not detect any shared mutations (e.g., PT5 vs. PT1). In contrast to the primary tumor, the metastases showed only minimal diversification. These results are consistent with

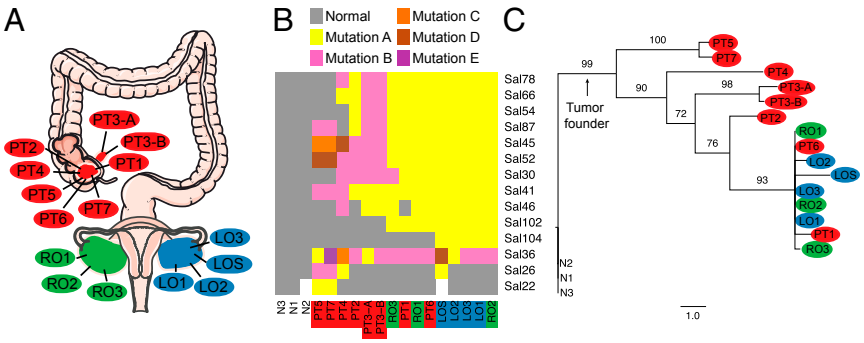

Fig. 5. Patient C13 with invasive adenocarcinoma of the colon and metastasis to the ovaries. (*A*) Approximate anatomical localization of all analyzed samples. LO, left ovary metastasis; N, normal; PT, primary tumor; RO, right ovary metastasis. Tumor sizes are drawn to scale. Letters A and B indicate that two samples were taken from the same FFPE block. All other samples are derived from separate blocks. The surgical pathology report provides a description of each tumor block, but the exact spatial orientation of each sample is not always known. For example, PT3-A and PT3-B are located in the ileum, and PT1 and PT2–7 are located in the cecum, but consecutive numbers do not necessarily imply that the tumor samples are adjacent to each other. (*B*) Complete mutation heatmap, with poly-G markers in the rows and patient samples in the columns. Gray squares signify allele distributions that are indistinguishable from a normal reference sample (N1). Colored squares indicate a shift in allele distribution—that is, a poly-G mutation. If multiple different mutations exist per marker, they are indicated with additional colors. White squares indicate missing data due to amplification failure. Because each marker harbors a distinct and unique set of mutations, we simply denote them with mutation A (yellow), B (pink), C (orange), and so forth for the purposes of the heatmap. Detailed mutation information, including magnitude and direction, is provided in *SI Appendix*. (*C*) Neighbor-joining tree with bootstrap values; branches with bootstrap values below 70% collapsed into polytomies.

metastasis occurring in late stages of primary tumor evolution. Specifically, they imply that a genetically advanced clone (residing in PT1 or PT6) gave rise to both metastases or that one ovary metastasis gave rise to the other in quick progression (i.e., without further diversification). Retrograde metastasis (40) of the ovarian lesion clone to tumor regions PT1 and PT6 is an alternative explanation consistent with the data.

Lymph node metastases can be phylogenetically distinct from distant metastases. Patient C13's left and right ovary metastases were similar to each other, but we also found genetically divergent metastases. Patient C31 was a 48-y-old female who received neoadjuvant FOLFOX chemotherapy and underwent surgery for a 3.2 cm MSS adenocarcinoma located at the hepatic flexure and a large 13 cm metastasis to the right ovary (Fig. 6*A*). The tumor had also metastasized to the mesenteric lymph nodes. We isolated four primary tumor samples, eight right ovary metastasis samples, and two tumor samples from the mesenteric lymph nodes. In the primary tumor, mutations were present in 33% of interrogated poly-G tracts (Fig. 6*B* and full genotype data in *SI Appendix*, Table S5). As in patient C13, patient C31's phylogenetic reconstruction showed that the ovarian tumor was distinct from the primary cancer and formed the deepest branch of the tree (Fig. 6*C*). The metastasis had a ∼40-fold larger volume than the primary tumor, implying that the metastatic clone must have been able to substantially increase its net growth rate (possibly this "growth spurt" happened in the early developmental stages of the metastasis, before it reached its large size). Because a relatively large number of mutations distinguished the primary tumor and the ovarian metastasis, they could have evolved separately for a substantial amount of time (consistent with parallel progression). However, we cannot exclude the possibility that we simply failed to sample the primary tumor region containing the subpopulation that gave rise to the metastasis. Interestingly, the ovarian clone did not spread to the lymph nodes: two independent samples from a large mass of matted lymph nodes were almost identical to the primary tumor in genetic composition across all markers. This finding shows that a primary tumor can contain multiple populations of clones with metastatic ability and raises the intriguing question of whether different routes of metastasis (lymphatic, hematogenous, intraperitoneal) are favored by genetically divergent cells.

A primary tumor and its widespread metastases are genetically homogeneous. Poly-G tract profiling of patient C27, a 44-y-old male with a mucinous adenocarcinoma that had spread extensively throughout the abdominal cavity, revealed a fundamentally different tumor evolution pattern than patients C13 and C31. Patient C27's descending colon harbored a small 1.5 cm tumor continuous with a 34.5 cm lesion that had essentially replaced the greater omentum (Fig. 7). In addition to this large mass, several serosal nodules and a splenic metastasis were resected after a course of neoadjuvant chemotherapy with FOLFOX and radiation treatment. The tumor had MSI, and the mutation rate was high with somatic alterations observed in 45% of interrogated loci (full genotype data provided in *SI Appendix*, Table S6 and mutation heatmap provided in *SI Appendix*, Fig. S5). In contrast to patients C13 and C31, whose samples revealed substantial variation, all specimens from patient C27 had similar poly-G tract profiles, and the topology of the resulting phylogenetic tree was flat (Fig. 7). Evidently, the tumor grew from a small lesion in the colon into a large omental mass and seeded a number of metastases while undergoing no significant spatial diversification. This is particularly surprising because this tumor was larger than the tumors in either patient C13 or C31, and its mutation rate was elevated due to MSI. Both these factors would be expected to lead to increased levels of diversity across different regions of the neoplasm (1). It therefore appears that one rapid clonal expansion that did not allow for regional "speciation" events created this cancer. Alternatively, patient C27's tumor cells may have had an exceptionally high motility, resulting in extensive mixing that rendered new clones generated during tumor growth undetectable. Both explanations, which are not mutually exclusive, point to an exceptionally aggressive phenotype. Future studies will determine whether spatial homogeneity is an adverse prognostic factor in colon cancer.

Poly-G Mutations Are Present in a Variety of Other Human Cancers. By testing a small panel of human tumors at 12 or more poly-G loci, we found poly-G mutations in several cancer types in addition to colon cancer, including renal cell carcinoma, glioblastoma, cholangiocarcinoma, esophageal carcinoma, pancreatic islet cell tumor, breast cancer, and lung carcinoid tumor (*SI Appendix*, Tables S7 and S8). Our dataset is not comprehensive enough to determine average tumor mutation frequency in cancers other

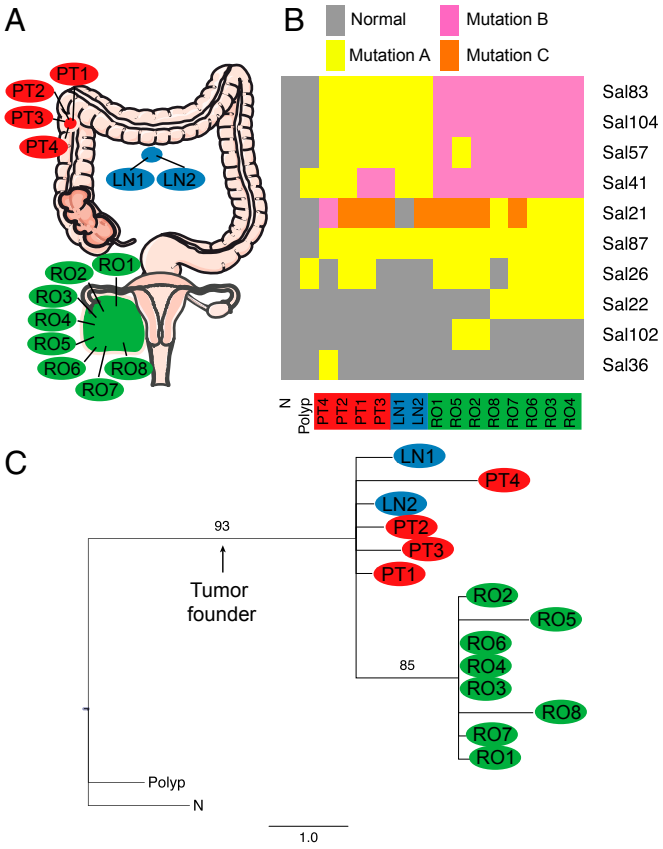

Fig. 6. Patient C31 with adenocarcinoma of the colon, lymph node, and ovarian metastases. (*A*) Tumor location overview. LN, lymph node metastasis; N, normal; PT, primary tumor; RO, right ovary metastasis. Tumor sizes are drawn to scale. (*B*) Complete mutation heatmap, with poly-G markers in the rows and patient samples in the columns. Gray squares signify allele distributions that are indistinguishable from a normal reference sample (N). Colored squares indicate a shift in allele distribution—that is, a poly-G mutation. If multiple different mutations exist per marker, they are indicated with additional colors. Because each marker harbors a distinct and unique set of mutations, we simply denote them with mutation A (yellow), B (pink), C (orange), for the purposes of the heatmap. Detailed mutation information, including magnitude and direction, is provided in *SI Appendix*. (*C*) Neighbor-joining tree with bootstrap values; branches with bootstrap values below 70% collapsed into polytomies.

than colon, although ongoing investigation of a breast carcinoma cohort indicates that variants are less frequent in this cancer type, presumably because breast epithelial cells do not divide as frequently as colonic cells.

Initial results suggest that the observed distinction between spatially heterogeneous and homogeneous tumors in colon cancer will also apply to other cancers. For example, one renal cell carcinoma showed a 33% mutation frequency, but most mutations were only detectable in select tumor portions (*SI Appendix*, Table S8). Analysis of a breast cancer (patient B1, Fig. 8*A*) comprising two lymph node metastases and four tumor nodules separated by several centimeters indicated that all lesions had a common origin because they shared some variants. However, we also found heterogeneously distributed mutations that allowed us to deduce that tumor focus TF1 had seeded the larger lymph node metastasis LN2, whereas tumor focus TF4 contained a distinct mutational profile and had segregated early on in its evolution. By contrast, patient O1's (Fig. 8*B*) malignant peripheral nerve sheath tumor showed homogeneity similar to patient C27's colon cancer. Patient O1 had a 14 cm calf tumor and a histologically similar 1.7 cm cancer on his left hand resected, and 1 y later, he underwent excision of a 6.5 cm lung metastasis. Poly-G tract profiling revealed identical mutations in all eight calf tumor and two lung metastasis samples, which suggests that the calf tumor was the source of the lung metastasis, whereas the tumor in his left hand showed no alterations and likely represented an independent transformation.

Discussion

We have shown that somatic mutations in noncoding poly-G repeats can be used to build maps of clonal architecture in human cancers. Poly-G tract profiling is sensitive enough to detect many distinct clonal populations within a tumor and produces reliable phylogenies that elucidate each patient's individual path of progression. The technique is widely useful in outlining clonal expansions that occurred during carcinogenesis.

In two patients with clear genetic divergence between primary and distant lesions, the metastases shared some alterations with the primary tumor but had also acquired private mutations. These data are consistent with previous findings in colorectal cancer (41) and pancreatic cancer (7). Patient C13's cancer supports the late metastasis paradigm. Patient C31 could potentially represent a case of parallel progression because relatively few mutations were shared between the distant metastasis and the primary tumor, with the caveat that sampling of the primary tumor might have missed the region harboring the precursor of the ovarian metastasis. In two other patients (C39 and C27), primary tumors and metastases shared a majority of mutations and were phylogenetically indistinguishable at the given resolution.

In two instances, we had the opportunity to compare distant and lymphatic metastases. In one patient (C31), we found that cancer cells that had disseminated to the lymph nodes had the same genotype as the primary tumor, whereas a distant ovarian metastasis had a distinct mutational profile and contained many private alterations. Two plausible explanations exist for this re-

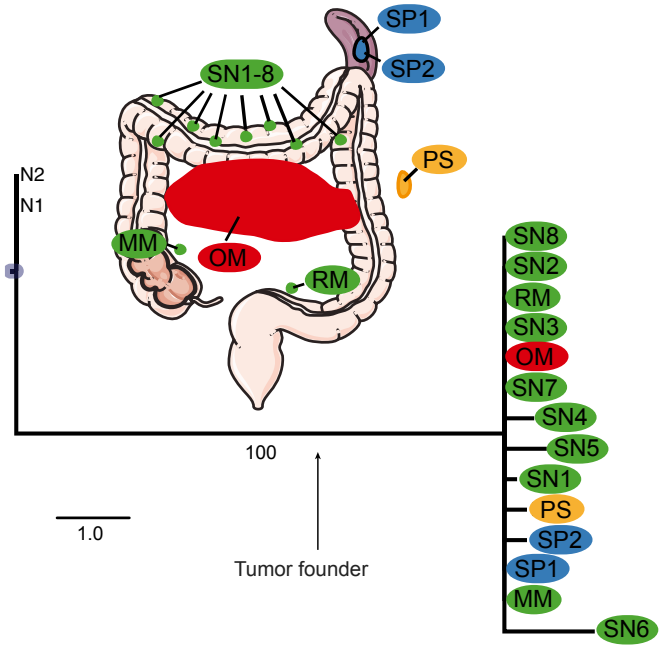

Fig. 7. Patient C27 with mucinous adenocarcinoma of the colon. MM, mesenteric margin; N, normal; OM, omentum; PS, peritoneal side wall metastasis; RM, retroperitoneal margin; SN, serosal nodule; SP, spleen metastasis. Neighbor-joining tree with bootstrap values; branches with bootstrap values below 70% collapsed into polytomies.

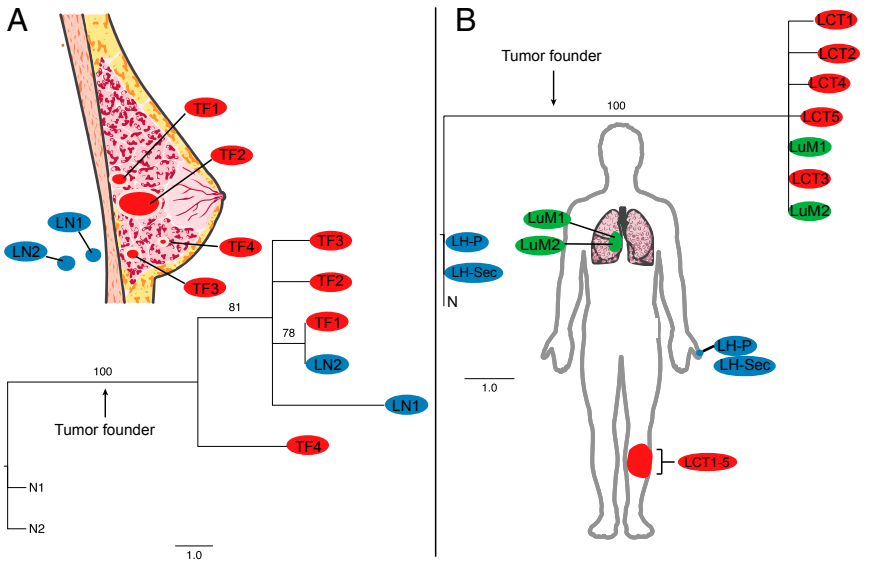

Fig. 8. Patient B1 with multifocal breast cancer and patient O1 with malignant peripheral nerve sheath tumor. (*A*) Patient B1. LN, lymph node metastasis; N, normal; TF, tumor focus. Neighbor-joining tree with bootstrap values; branches with bootstrap values below 70% collapsed into polytomies. (*B*) Patient O1. LCT, left calf tumor; LH, left hand nodule; LuM, lung metastasis; N, normal. Additions "Sec" and "P" indicate that the sample was analyzed twice, once using a biopsy punch (P) and then by retrieving tumor tissue from sections (Sec). Neighbor-joining tree with bootstrap values; branches with bootstrap values below 70% collapsed into polytomies.

sult. It is possible that after the ovarian metastasis had already formed, a sweeping clonal expansion occurred in the primary tumor and gave rise to the lymph node metastasis. However, this hypothesis does not account for the larger mutational load in the ovarian metastasis, which suggests that its founder clone had undergone a larger number of divisions than the clone dominating the primary tumor and the lymph node metastasis. An alternate explanation more consistent with our data is that large numbers of tumor cells continuously drain from the original site to the lymph node, which contains a polyclonal sample of cells from the primary tumor and is therefore indistinguishable from it. Future studies will determine, in a larger cohort of patients, whether genetic divergence between lymph node and distant metastases is a more general phenomenon. It would be of significant clinical and biological interest to evaluate whether lymphatic metastases might be formed through a distinctive migration mechanism.

Clonal diversity varies substantially between patients. Some tumors were diversified (C13, C31, B1), whereas others shared the same genotype across all primary and metastatic tumor samples (O1), in one case despite an elevated mutation rate caused by MSI (C27). We did not find any obvious connection between administration of chemotherapy before surgery and intratumor heterogeneity. For example, both patients C31 (diversified) and C27 (homogeneous) received neoadjuvant chemotherapy with FOLFOX. Although the "flat" clonal expansions (42) clearly represent younger entities than the diversified cancers (17), we currently do not know whether these differences in population structure are mirrored in divergent clinical behavior. Clonal diversity in the premalignant lesion of Barrett's esophagus represents a risk factor for future cancer development (13), which suggests that heterogeneity promotes malignancy, but the situation may be different in established cancers and/or differ by cancer type. In breast cancer, intratumor heterogeneity, as defined by cell surface marker expression, correlates with the histopathological stage (43), but how phenotypic heterogeneity relates to genetic diversity is not known. Determining whether genetic heterogeneity, or lack thereof, is associated with important clinical variables will be important in future studies. One limitation of our approach in this regard is that it relies on spatially distinct clonal expansions. Genetic heterogeneity within a sample cannot be detected if an allele is present at a frequency below 40–60% (25). Subclonal diversity below this threshold would therefore have to be evaluated with complementary techniques such as fluorescence in situ hybridization (44) or deep sequencing (8).

Our data further show a positive correlation between age at diagnosis and mutation frequency. Laiho et al. found a similar association when analyzing CA-dinucleotide repeats in colorectal cancers (20). These results accord with growing evidence that a large proportion of mutations [more than 50% by some estimates (37)] found in human cancers are not acquired during tumor development but are already present in the tumor founder cell. Mutations accumulate in normal cells but typically remain undetectable because no clonal expansion takes place. Recent work shows that after expansion of single normal human hematopoietic stem cells, comparable numbers of mutations can be observed as in acute myeloid leukemia (36). Because cells in different human tissues proliferate at varying rates, the mitotic history of a tumor founder cell is likely a significant factor in the variation among cancer mutation rates (45).

Intriguing in this context is that mutation frequency inversely correlates with tumor grade. Poorly differentiated tumors have significantly fewer poly-G mutations. Extending the argument that mutation frequency is codetermined by the mitotic history of the tumor founder cell, this observation suggests that less differentiated tumors might derive from a rarely dividing cell. In the colon, two distinct progenitor populations have been identified: one that is located among Paneth cells, expresses Lgr5, and displays the characteristics of an actively dividing tissue stem cell and another located at the +4 position, showing signs of quiescence (38). It may be that poorly differentiated tumors arise from the latter population.

Exome-wide analysis of somatic alterations in colorectal cancers by The Cancer Genome Atlas (TCGA) shows a distribution of mutation frequencies that is remarkably similar to our findings (46). Of 224 sequenced tumors, 16% were hypermutated with more than 12 mutations per megabase; in the remaining non-hypermutated tumors, mutation frequencies varied by approximately one order of magnitude. In our analysis, 18% of tumors

had MSI and very high mutation rates. Mutation frequencies in MSS cancers, excluding cases with no mutations, ranged from 5% to 55%. It remains to be determined whether the association between mutation frequency and histologic grade exists for exonic point mutations, as tumor grade was not part of the standardized dataset for the TCGA study. Many mutagenic factors (such as exposure to carcinogens or oxidative stress) might have quantitatively different effects on single base exonic substitution rate compared with replication slippage-mediated intragenic/intronic microsatellite mutation rate; we therefore do not necessarily expect to see the same effect in exome data.

In summary, we have shown that a highly scaleable PCR assay of endogenous mutational hotspots can generate reliable lineage information in human cancer with low time and cost expenditures. We have used this assay to generate biological insights into the origin and progression of metastatic colon cancer. In many cases, analysis of 20 poly-G markers yielded sufficient information to build robust phylogenetic trees. It is unlikely that interrogation of additional loci would contribute substantial new information in those instances. However, a larger number of poly-G markers might be able to resolve lineage relationships in cases like patient C27, whose carcinoma showed limited intratumor heterogeneity through the lens of our standard 20 marker panel. Our methodology can be used with FFPE specimens, which are collected in hospitals around the world on a daily basis. Our study only used tissues that were also available to the pathologist at the time of diagnosis. It is conceivable that lineage testing could be quickly performed for individual patients to improve clinical decision processes, for example distinguishing multicentric lung cancer from intrapulmonary metastasis (47). Because detecting mutated alleles in poly-G tracts does not require sequencing, patient privacy would be protected.

Compared with deep whole genome or exome sequencing, the resolution of poly-G tract profiling is relatively low, and the assay does not provide information on actionable mutations. Therefore, it is primarily useful for applications that focus on questions of lineage and clonality (as opposed to the study of causal variants) and require a large number of samples. Poly-G tract profiling could also be used as an efficient screening technology for selecting samples of interest for deeper analysis by next-generation sequencing.

Materials and Methods

Patient Selection and Tissue Collection. This study was approved by the Institutional Review Board of Massachusetts General Hospital, Boston. We searched the pathology database of Massachusetts General Hospital for patients who underwent surgery between 2010 and 2012 and whose diagnosis contained ICD9 code 153, "Malignant neoplasm of the colon." We reviewed the search results and selected 22 consecutive patients who underwent resection of a primary colon carcinoma along with at least three lymph node metastases and/or distant metastases. Eighteen patients were treatment naïve, three had received neoadjuvant chemotherapy, and one patient received neoadjuvant chemotherapy and radiation. Detailed patient information is provided in *SI Appendix*, Table S1. Histologic grade and other tumor characteristics were copied from the "final pathological diagnosis" section of the official surgical pathology report (i.e., all classifications were made by a pathologist according to Massachusetts General Hospital standards). For each patient, we then reviewed all available histology slides and FFPE tissue blocks and selected areas of homogeneous tumor for sampling. Tumors with a predominant stromal component were excluded. By default, we used a 1.5 or 2 mm biopsy punch to extract cores of tumor and normal tissue directly from the block. For small tumor samples, we cut 10 μm tissue sections and macrodissected tumor cells after staining slides with a PCR-compatible stain (Histogene, Life Technologies). Samples were de-paraffinized with xylene, washed with 100% ethanol, air-dried, and incubated with Proteinase K overnight as previously described (48). DNA was extracted with phenol-chloroform and precipitated with ethanol and sodium acetate. We estimate that the average tissue sample had a volume of 3 mm^3 and contained 9×10^6 cells.

Genotyping. A panel of primers flanking 35 poly-G tracts in the human genome was previously published (25). We used a randomly selected subset of primers from this panel (20 loci were sufficient to generate reliable phylogenies in most of our patients). Marker identification numbers are provided alongside full genotype data in all *SI Appendix* tables. Forward primers incorporated a fluorescent dye (HEX or 6-FAM) on their 5′ end. Reverse primers contained a 5′ GTTTCTT "pigtail" sequence (49). Because our DNA was derived from FFPE tissue and heavily fragmented, we included 90 ng of DNA (as determined by spectrophotometry) in each reaction to ensure the reproducibility of stutter patterns. Every PCR was performed in triplicate in a 10 μL volume with 1 μM forward and reverse primers, 200 μM of each dNTP, 2.5 units Taq Polymerase, 1× PCR buffer, and 1× Q-solution (Qiagen) to facilitate amplification of GC-rich templates. After 42 amplification cycles, PCR products were resolved by capillary electrophoresis using an ABI Genetic Analyzer 3130xl. MSI was tested using the Bethesda Markers as described in ref. 50. We did not distinguish between MSI-low and MSS tumors. Electropherograms were viewed with GeneMapper 4.0. The 22 tumor-normal pairs in our cohort were scored for the presence of mutations by visual comparison of the stutter distributions for each marker. If the tumor sample showed a consistent shift in the stutter pattern that was reproducible across all three replicates, we recorded a mutant genotype, denoting a repeat contraction with m[number of deleted bases] and an expansion with p[number of added bases]. If two distinct alleles were discernible and at least 6 bp apart, we scored them separately. Instances of loss of heterozygosity were not counted as mutations, but they were used as data points in the phylogenetic reconstruction. To facilitate analysis of multiple tumor regions from the same patient, we developed an automated approach that allowed us to compare stutter patterns across many samples in an objective manner. *SI Appendix*, Fig. S4 provides an overview of our algorithm. We exported peak information (size, height) from GeneMapper and fed it into an analysis pipeline within the R environment for statistical computing (www.R-project.org). For each patient and marker, we calculated pairwise correlation coefficients among all stutter distributions and used these as inputs to a hierarchical clustering algorithm. The resulting dendrogram divided all samples into categories that corresponded to different mutations. We examined the branches of the dendrograms and determined at which height to cut the tree based on three criteria: (*i*) Normal samples had to cluster separately from mutated tumor samples, (*ii*) replicates had to cluster within the same clade (allowing for some variation due to PCR failure), and (*iii*) all mutation categories could be verified by manual review of electropherograms. Genotype assignments were recorded in a matrix that contained the mutational status of every sample at 20 poly-G loci. Because we did not want to make assumptions about the likelihood of a particular allele distribution occurring, we treated mutations as unordered characters. This dataset was used for phylogenetic analysis.

Phylogenetic Reconstruction. We reconstructed phylogenies using two independent approaches. First, we calculated a distance matrix for each patient using an "equal or not" distance (31). This method increases the distances between two samples if they have unequal genotypes, regardless of the magnitude of the difference. We then used neighbor-joining (51) in R to infer the phylogenetic relationships between samples. In the very rare case of missing values, we imputed them using the nearest neighbor. We used bootstrapping with 1,000 replicates to test the reliability of the resulting trees (52) and collapsed all interior branches with bootstrap values below 70% into polytomies. Next, we used Bayesian inference of phylogeny—a methodology that relies on a fundamentally different set of principles than neighbor-joining—to construct the phylogenies. The results were almost identical in all cases, confirming the robustness of our approach. Bayesian phylogenies and posterior probability values for all clades are presented in *SI Appendix*, Fig. S10. We used the software MrBayes (53) with the same model parameters that were previously used for the analysis of poly-G tract mutation profiles (21).

Other Statistical Analyses. Statistical analysis was performed in Prism (Graphpad) and R. We used linear regression to test the association between mutation frequency in MSS tumors and four variables of interest (tumor size, lymph node status, presence of distant metastasis at diagnosis, and age).We used a two-tailed Mann–Whitney test to compare mutation frequencies in low- and high-grade tumors ($n = 9$ for each group after excluding MSI cases). We did not correct for multiple testing, as the number of tests was small and our sample size ($n = 18$) limited (with correction, the P value for the association between mutation frequency and grade would still be significant, but the association with age would not).

ACKNOWLEDGMENTS. We thank Marshall Horwitz for comprehensive advice and extensive technical training; Constance Cepko, Stephen Elledge, Igor Garkavtsev, Raju Kucherlapati, Robert Morris, Kornelia Polyak, Dennis Sgroi, Shamil Sunyaev, and Sridhar Ramaswamy for fruitful discussions; and Jessica Fessler for help with experimentation. This work was supported by the Department of Defense, Research Innovator Award W81XWH-10-1-0016 (to R.K.J.) and Predoctoral Traineeship Award W81XWH-11-1-0146 (to K.N.).

1. Marusyk A, Almendro V, Polyak K (2012) Intra-tumour heterogeneity: A looking glass for cancer? *Nat Rev Cancer* 12(5):323–334.
2. Fidler IJ (2003) The pathogenesis of cancer metastasis: The 'seed and soil' hypothesis revisited. *Nat Rev Cancer* 3(6):453–458.
3. Greaves M, Maley CC (2012) Clonal evolution in cancer. *Nature* 481(7381):306–313.
4. Merlo LMF, Pepper JW, Reid BJ, Maley CC (2006) Cancer as an evolutionary and ecological process. *Nat Rev Cancer* 6(12):924–935.
5. Gerlinger M, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366(10):883–892.
6. Anderson K, et al. (2011) Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* 469(7330):356–361.
7. Yachida S, et al. (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* 467(7319):1114–1117.
8. Ding L, et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464(7291):999–1005.
9. Klein CA (2009) Parallel progression of primary tumours and metastases. *Nat Rev Cancer* 9(4):302–312.
10. Klein CA (2008) Cancer. The metastasis cascade. *Science* 321(5897):1785–1787.
11. Weinberg RA (2007) *The Biology of Cancer* (Garland, New York), 1st Ed.
12. Weinberg RA (2008) Mechanisms of malignant progression. *Carcinogenesis* 29(6): 1092–1095.
13. Maley CC, et al. (2006) Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet* 38(4):468–473.
14. Liu W, et al. (2009) Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* 15(5):559–565.
15. Shibata D (2012) Cancer. Heterogeneity and tumor history. *Science* 336(6079): 304–305.
16. Viguera E, Canceill D, Ehrlich SD (2001) Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J* 20(10):2587–2595.
17. Shibata D, Navidi W, Salovaara R, Li ZH, Aaltonen LA (1996) Somatic microsatellite mutations as molecular tumor clocks. *Nat Med* 2(6):676–681, www.nature.com/nm/journal/v2/n6/abs/nm0696-676.html.
18. Tsao JL, et al. (1998) Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: Evidence of adenomas with stem cell architecture. *Am J Pathol* 153(4):1189–1200.
19. Tsao JL, et al. (2000) Genetic reconstruction of individual colorectal tumor histories. *Proc Natl Acad Sci USA* 97(3):1236–1241.
20. Laiho P, et al. (2002) Low-level microsatellite instability in most colorectal carcinomas. *Cancer Res* 62(4):1166–1170.
21. Salipante SJ, Horwitz MS (2006) Phylogenetic fate mapping. *Proc Natl Acad Sci USA* 103(14):5448–5453.
22. Salipante SJ, Kas A, McMonagle E, Horwitz MS (2010) Phylogenetic analysis of developmental and postnatal mouse cell lineages. *Evol Dev* 12(1):84–94.
23. Salipante SJ, Thompson JM, Horwitz MS (2008) Phylogenetic fate mapping: Theoretical and experimental studies applied to the development of mouse fibroblasts. *Genetics* 178(2):967–977.
24. Zhou W, et al. (2013) Use of somatic mutations to quantify random contributions to mouse development. *BMC Genomics* 14:39.
25. Salk JJ, et al. (2009) Clonal expansions in ulcerative colitis identify patients with neoplasia. *Proc Natl Acad Sci USA* 106(49):20871–20876.
26. Ellegren H (2004) Microsatellites: Simple sequences with complex evolution. *Nat Rev Genet* 5(6):435–445.
27. Weber JL, Wong C (1993) Mutation of human short tandem repeats. *Hum Mol Genet* 2(8):1123–1128.
28. Boyer JC, et al. (2002) Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. *Hum Mol Genet* 11(6):707–713.
29. Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *Am J Hum Genet* 62(6):1408–1415.
30. Wasserstrom A, et al. (2008) Reconstruction of cell lineage trees in mice. *PloS One* 3(4):e1939.
31. Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E (2005) Genomic variability within an organism exposes its cell lineage tree. *PLOS Comput Biol* 1(5):e50.
32. Frumkin D, et al. (2008) Cell lineage analysis of a mouse tumor. *Cancer Res* 68(14): 5924–5931.
33. Carlson CA, et al. (2012) Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nat Methods* 9(1):78–80.
34. Salk JJ, Horwitz MS (2010) Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. *Semin Cancer Biol* 20(5):294–303.
35. Potten CS, Kellett M, Roberts SA, Rew DA, Wilson GD (1992) Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut* 33(1):71–78.
36. Welch JS, et al. (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell* 150(2):264–278.
37. Tomasetti C, Vogelstein B, Parmigiani G (2013) Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci USA* 110(6):1999–2004.
38. Li L, Clevers H (2010) Coexistence of quiescent and active adult stem cells in mammals. *Science* 327(5965):542–545.
39. Wasserstrom A, et al. (2008) Estimating cell depth from somatic mutations. *PLOS Comput Biol* 4(4):e1000058.
40. Kim MY, et al. (2009) Tumor self-seeding by circulating cancer cells. *Cell* 139(7): 1315–1326.
41. Jones S, et al. (2008) Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci USA* 105(11):4283–4288.
42. Siegmund KD, Marjoram P, Tavaré S, Shibata D (2009) Many colorectal cancers are "flat" clonal expansions. *Cell Cycle* 8(14):2187–2193.
43. Park SY, et al. (2010) Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer. *Clin Cancer Res* 16(3):876–887.
44. Snuderl M, et al. (2011) Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* 20(6):810–817.
45. Lawrence MS, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214–218.
46. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330–337.
47. Shimizu S, et al. (2000) High frequency of clonally related tumors in cases of multiple synchronous lung cancers as revealed by molecular diagnosis. *Clin Cancer Res* 6(10): 3994–3999.
48. Shibata D (1994) Extraction of DNA from paraffin-embedded tissue for analysis by polymerase chain reaction: New tricks from an old friend. *Hum Pathol* 25(6):561–563.
49. Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: Primer modifications that facilitate genotyping. *Biotechniques* 20(6):1004–1006, 1008–1010.
50. Loukola A, et al. (2001) Microsatellite marker analysis in screening for hereditary nonpolyposis colorectal cancer (HNPCC). *Cancer Res* 61(11):4545–4549.
51. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
52. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39(4):783–791.
53. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.