Tracing human cancer evolution with hypermutable DNA

Tracing human cancer evolution with hypermutable DNA

A dissertation presented

by

Kamila Naxerova

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Human Biology and Translational Medicine

Harvard University

Cambridge, Massachusetts

December 2013

Dissertation Advisor: Rakesh K. Jain Kamila Naxerova

Abstract

Tracing human cancer evolution with hypermutable DNA

Metastasis is the main cause of cancer morbidity and mortality. Despite its clinical significance, several fundamental questions about the metastatic process in humans remain unsolved. Does metastasis occur early or late in cancer progression? Do metastases emanate directly from the primary tumor or give rise to each other? How does heterogeneity in the primary tumor relate to the genetic composition of secondary lesions? Addressing these questions in representative patient populations is crucial, but has been difficult so far. Here we present a simple, scalable PCR assay that enables the tracing of tumor lineage in patient tissue specimens. Our methodology relies on somatic variation in highly mutable polyguanine (poly-G) repeats located in non-coding genomic regions. We show that poly-G mutations are present in a variety of human cancers. Using colon carcinoma as an example, we demonstrate an association between patient age at diagnosis and tumor mutational burden, suggesting that poly-G variants accumulate during normal division in colonic stem cells. We further show that poorly differentiated colon carcinomas have fewer mutations than well-differentiated tumors, possibly indicating a shorter mitotic history of the founder cell in these cancers. We collect multiple spatially separated samples from primary carcinomas and their metastases and use poly-G fingerprints to build well-supported phylogenetic trees that

illuminate each patient's path of progression. Our results imply that levels of intra-tumor heterogeneity vary significantly among patients.

Our approach can generate reliable lineage information in large numbers of patients with minimal time and cost expenditure. It can be used in its own right to study tumor evolution, or as an efficient screening tool to select samples for deeper analysis by next generation sequencing. Further development and successful application of targeted cancer therapies will vitally depend on an accurate understanding of clonal architecture in human tumors. The mitotic history of a neoplasm, as captured by neutral lineage markers, can provide an important backdrop on which to project the distribution of hundreds of therapeutically relevant mutations.

Table of Contents

Acknowledgements

First and foremost, I would like to thank my advisor Dr. Rakesh Jain for his generosity and support. Dr. Jain created a space of truly remarkable freedom in which I could pursue my own scientific interests, and even in (scientifically speaking) dark times he never lost his confidence in me.

I would also like to extend special thanks to Dr. Elena Brachtel without whom none of this work would have been possible. She is wise and funny and I learned a lot from her. Many heartfelt thanks also go to other members of the MGH Pathology department, in particular Matija Snuderl and Gregory Lauwers.

Jesse Salk from the University of Washington is an indispensible colleague. My many phone and email conversations with him over the years kept me sane and significantly contributed to the quality of the work.

I am very grateful to the members of my dissertation advisory committee, Raju Kucherlapati, Barbara Smith and Alex Toker, for many helpful discussions along the way. Two members of the Ragon Institute, Karen Power and Aaron Matthews, helped me, completely selflessly, to run immense numbers of fragment analyses. Finally, I am indebted to all the people who generously offered advice over the years: Simon Kasif in particular, Marshall Horwitz, Connie Cepko, Steve Elledge, Sridhar Ramaswamy and his group, and Shamil Sunyaev.

A Predoctoral Traineeship Award from the Department of Defense supported my work. Dr. Jain received funding through the Breast Cancer Innovator Award (DoD).

To my family: Dad, Mom, Matthias and Janna, thank you for everything.

Chapter 1: Using genetic tools to study metastatic progression in humans

Statement of contribution: Parts of this chapter correspond to a review that is currently in preparation. I wrote the manuscript in its entirety.

1.1 The importance of intra-tumor heterogeneity in systemic cancer

After virtual defenselessness in the face of metastatic disease for most of human history, a hopeful time has now begun in medicine. Metastasis causes 90% of human cancer deaths (Mehlen and Puisieux, 2006), but recent advances in molecularly targeted therapies have extended progression-free and overall survival for patients with some forms of metastatic cancer. These include BRAF-mutant melanoma (Chapman et al., 2011), EGFR-mutant (Maemondo et al., 2010) or ALK-rearranged (Shaw et al., 2011) lung cancer, EGFR-expressing colon cancer (Cunningham et al., 2004), and most gastrointestinal stromal tumors (Reichardt et al., 2012). Concurrent improvements in massively parallel sequencing technologies generate a steady stream of putative new targets for therapeutic intervention (Ciriello et al., 2013; Kandoth et al., 2013). It could be argued that the road toward the eradication of cancer will be straightforward from here: more targets will be discovered, novel therapeutics developed, resistance (Holohan et al., 2013) will be monitored dynamically and kept in check with precision medicine.

However, increased appreciation of genetic and phenotypic intra-tumor heterogeneity casts some doubt on the viability of this simple path forward. The emergence of lethal recurrent disease in all patients treated with targeted therapy, sometimes within a mere few months, suggests that cells that are inherently treatment resistant are invariably present in each tumor. Longitudinal monitoring of the clonal composition of tumors before and after treatment confirms the remarkable genetic plasticity of cancer (Landau et al., 2013). Even with advanced second-line therapies

(potentially consisting of complex drug combinations), it will likely be difficult to bridle the evolutionary potential of billions of cancer cells.

It also remains uncertain whether targeted therapies can successfully be transferred from the metastatic to the adjuvant setting (as recently reviewed in Polzer and Klein, 2013). While trastuzumab (an anti-HER2 antibody) and imatinib (a c-KIT and PDGFR tyrosine kinase inhibitor) have shown benefits as adjuvant therapies in breast cancer (Gianni et al., 2011) and gastrointestinal stromal tumors (Reichardt et al., 2012), respectively, recent adjuvant trials of EGFR inhibition in lung cancer (Goss et al., 2013) and colon cancer (Alberts et al., 2012) have not been encouraging. Genetic and/or phenotypic divergence between the primary tumor and minimal residual disease (MRD) is suspected to be the root cause of these difficulties (Aguirre-Ghiso et al., 2013).

Given these pressing clinical challenges, gaining a comprehensive understanding of intra-tumor heterogeneity is a central task of translational cancer research today. Marked diversity of cancer cells can be observed from a genetic and epigenetic perspective, with respect to gene and cell surface marker expression patterns, and with reference to differentiation state and propagation potential (Marusyk et al., 2012). The umbrella term “intra-tumor heterogeneity” is used 1) to describe diversity, as defined by any of the above measures, among cells intermingling in one localized tumor area, 2) to refer to differences between spatially distinct tumor regions, 3) or to denote heterogeneity among multiple non-contiguous lesions in metastatic disease. From the point of view of a clinician prescribing targeted therapies, genetic divergence between primary tumors, MRD and overt metastasis arguably is the most relevant form of intra-tumor diversity. In current practice, molecular analysis of small cell

populations from the primary tumor typically guides treatment strategies aimed at the eradication of both MRD and macroscopic metastasis. Progressively, however, we are becoming aware that genetic targets show discordance both within and between lesions (Gerlinger et al., 2012). Re-biopsy of metastases is therefore increasingly being advocated (Niikura et al., 2013). While repeat examination of surgically accessible metastases of advanced size sometimes is possible (notably always involving the risk of pain and infection for the patient), the genetic traits of MRD are far more difficult to test, mainly because the relevant cells or microscopic lesions are invisible and potentially widely dispersed.

Since the genetic composition of most metastatic (precursor) lesions cannot be analyzed due to these practical limitations, it is becoming increasingly expedient to gain a more principled understanding of how diversity arises within the primary tumor and how the bottleneck of metastatic dissemination modulates it. Surprisingly, the most basic steps of metastasis are still poorly understood. In particular, much uncertainty still surrounds the following questions: *When* do metastatic precursor cells leave the primary tumor? The time point of dissemination likely is a major determinant of genetic divergence. Also, by *what route* do metastatic cells spread to form widely disseminated disease? For example, do metastases give rise to other metastases or do they emanate independently from the primary tumor? An evidence-based model of metastatic progression that includes detailed knowledge of factors that shape the genetic relationships between primary tumor and metastases would likely improve treatment and prevention of advanced disease. Unfortunately, many decisive properties of metastasis in humans, among them latency and variable time to recurrence (TTM), are

not correctly recapitulated in many animal models (Klein, 2011). Obtaining data in the human setting therefore is highly desirable.

The following part of this chapter describes the theoretical models that currently dominate our thinking about metastatic progression and offers definitions of widely used concepts in this area. Comparative data of primary tumors and metastases in humans will be discussed in the next section, with an emphasis on how the available evidence corroborates these models. The chapter concludes with an outline of experimental techniques that can used to address the central questions of *when* and by *what route* systemic disease is established.

1.2 Models of metastatic progression in humans

The time course of metastasis varies significantly across different cancer types and patients. Metastases can be diagnosed at the same time as the primary tumor (synchronous) or after latency periods ranging from a few months to several decades (metachronous). The distribution of progression-free survival intervals in the population of patients that do relapse is tumor type specific and typically very broad. Currently there is no reliable way to predict TTM. While the size of the primary tumor at diagnosis correlates with overall risk of metastasis for many cancer types, it is not necessarily an indicator of how long the TTM will be. In colon cancer, for example, primary tumor characteristics in patients with synchronous vs. metachronous metastases are not significantly different (Tsai et al., 2007).

It appears that individual tumor biology not only controls whether recurrence will occur at all, but also determines TTM and how quickly metastatic disease will reach lethal dimensions once a cancer has relapsed. The biological mechanisms underlying

this large variation are virtually unknown, and many combinations of interacting causative elements are conceivable, among them the time point of metastasis relative to the evolutionary trajectory of the primary tumor, possible dormancy periods and growth rate characteristics at ectopic sites. Several conceptual frameworks exist that aim to integrate knowledge of these mechanisms and clinical observations to create a "unified theory of metastasis". As will become apparent below, drastically different models of metastatic progression in humans are still competing. This is a reflection of how challenging it is to study these processes *in vivo* and how much research remains to be done in this area.

Linear progression

The traditional model of metastatic progression is called "linear" because it postulates that primary tumor development and metastasis occupy sequential positions on a unidirectional timeline of events (Figure 1.1, blue bottom panel). A central assumption of linear progression is that only highly aberrant, genetically advanced cancer cells can effectively colonize distant organs. These cells are thought to arise in a step-wise fashion, through multiple clonal expansions, during the development of the primary tumor (Cairns, 1975). Since acquisition of metastasis-enabling mutations is a slow process, dissemination typically occurs in late stages of tumorigenesis, around the time or shortly before a primary tumor becomes clinically detectable. Consequently, the evolutionary divergence between primary and secondary neoplasms is relatively small, and the primary tumor is regarded as a good surrogate for the molecular and phenotypic properties of metastases.

Figure 1.1: Overview of human metastasis models. *The cell lineage tree of the primary tumor is depicted on the left side. Cells with different genetic alterations are indicated in different colors. During the first few divisions after transformation, all cells are still genetically similar to each other (red), but as the tumor continues to proliferate (dotted lines), diversity increases. Some cells die (X), and expansions of particularly fit clones can in turn locally decrease heterogeneity (light blue). Linear progression (blue background panel) assumes that late in tumor development, one of these clonal expansions gives rise to a metastasis (red connector), which can in turn spawn other metastases in a cascade. Parallel progression (green background panel), on the other hand, conjectures that metastasis occurs early on and that metastases evolve separately from each other and the primary tumor. Under the tumor self-seeding model (red background panel), metastatic cells return to their tumor of origin. Finally, metastatic precursors may lay dormant for variable time periods (yellow background panel).*

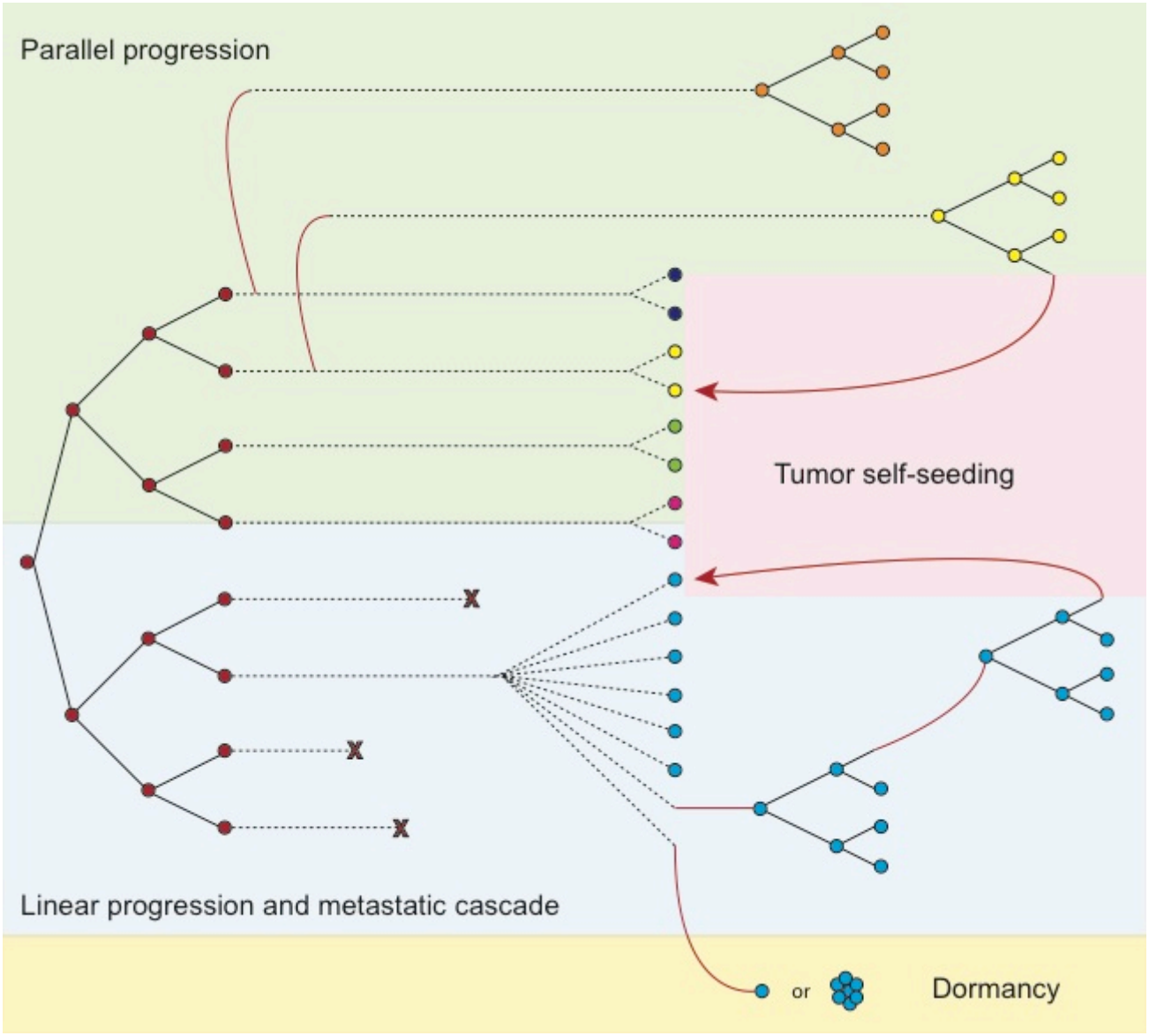

Figure 1.1 (Continued).

Metastatic cascades

Loosely associated with the linear progression model, by virtue of placing the development of widespread metastasis in the latest stages of carcinogenesis, is the conception that metastases, particularly those located in central organs with high blood flow such as the lung and liver, give rise to other metastases in a cascading manner (Bross et al., 1975). Cascades or "showers of metastases" (Weinberg, 2008) would be expected to lead to greater genetic homogeneity of systemic disease than direct descent from the primary tumor. A related, but somewhat distinct potential "cascade step" is the lymphatic system. Autopsy studies show that regional and distant lymph nodes are by far the most common site of metastasis: at time of death, lymphatic lesions are two-fold more frequent than metastases to the liver, which is the next-common site (Disibio and French, 2008). For many cancer types, the presence of cancer cells in regional lymph nodes is a negative prognostic indicator. Historically, it was therefore assumed that lymph node metastases are precursors of distant lesions. This belief motivated aggressive surgical interventions, such as the radical mastectomy and axillary lymph node dissection in breast cancer, to eradicate locoregional disease as thoroughly as possible (Fisher, 1980). Axillary lymph node dissection is used much less widely today because it does not seem to prolong survival (Giuliano et al., 2011), but it is still contentions where lymphatic disease can rise to distant metastases (Klein, 2013).

Parallel progression

Diametrically opposed to linear progression, though not strictly mutually exclusive, is the "parallel progression" model (Figure 1.1, green top panel), which posits that metastasis

occurs early in cancer development and that primary and secondary tumors evolve independently for a significant amount of time (Klein, 2009; 2013). Parallel progression assumes that successful dissemination and ectopic survival do not necessarily require a complex repertoire of acquired mutations, but can be accomplished by genetically inchoate cancer cells with few karyotypic abnormalities. The somatic evolution of these early disseminated tumor cells (DTCs) largely happens at distant organ sites and involves extensive adaption to local microenvironments. Therefore, substantial genetic disparity is expected between the primary tumor and its metastases, as well as between metastases in different anatomic locations. Parallel progression fundamentally doubts that molecular profiling of primary tumors is appropriate for selecting effective therapeutics against MRD and overtly metastatic disease.

Tumor self-seeding.

Both linear and parallel progression models regard metastasis as a unidirectional process that begins within the primary neoplasm and terminates at a distant site. “Tumor self-seeding” (Figure 1.1, red side panel) is a recently (Kim et al., 2009) articulated hypothesis stating that a bilateral, dynamic cell exchange exists between synchronous lesions. The primary tumor continually sheds cancer cells into the bloodstream, some of which will pass through the lung capillary network to enter the arterial circulation. A highly selected subset of these circulating tumor cells (CTCs) re-enters the original primary site to fuel local progression. Cells that extravasate from proliferating metastases could similarly return to the primary tumor. If tumor self-seeding indeed took place to a significant degree, it would create genetic links among primary

and secondary lesions that would be difficult (or impossible) to distinguish from late dissemination.

Dormancy

Among the more widely accepted (if unproven) concepts in metastatic progression in humans is dormancy (Figure 1.1, yellow bottom panel), a loosely defined term that is used to describe multiple distinct forms of tumor growth arrest. In the clinical setting, dormancy is invoked to explain “ultra-late” recurrences 10 or more years after resection of the primary tumor (Klein, 2011). In cell biological parlance, dormancy can either refer to a senescence-like state single disseminated tumor cells enter after they have become entrapped in foreign and potentially hostile tissue microenvironments (“tumor cell dormancy”) or to the indolent behavior of small subclinical neoplasms which exhibit no net growth (“tumor mass dormancy”). Dormancy is an important factor to consider in theoretical frameworks of metastasis because it is often impossible to judge whether a metastatic lesion appeared after a prolonged latency period because it disseminated late in cancer progression or because it underwent a period of dormancy at its new site.

Conclusion

The contemporaneous existence of these at times contradictory models illustrates how little we still know about metastatic progression in humans. Deepening our understanding of how systemic disease emerges has evident implications for future therapy development. Personalizing treatment according to the molecular profile of the primary tumor is a promising strategy for cancers that follow the linear progression model and metastasize in cascades. In the case of parallel progression and dormancy,

on the other hand, repeat biopsies or analysis of CTCs or circulating tumor DNA may be required to obtain updated information on the genetic constitution of target cells. It is possible that disparate modes of metastasis are prevalent in different tumor types, and that varying combinations of the "pure" models presented above can occur. The following section presents human-derived data on the relationship between primary tumors, MRD and overt metastases and discusses the compatibility of the results with different models of metastatic progression.

1.3 Evidence in humans

Growth kinetics

A strong line of evidence supporting the parallel progression model is the observation that growth rates of primary tumors and metastases are largely similar. It is commonly assumed that epithelial cancers develop over many years and even decades (Jones et al., 2008; Yachida et al., 2010). If a metastasis is discovered synchronously or just one or two years after primary tumor resection – a common event in breast cancer, for example (Demicheli et al., 1996) – and its growth rate is comparable to the primary tumor, it follows that it must also be many years old (i.e. disseminated early). Linear progression cannot explain the emergence of metastasis close to the time of primary tumor diagnosis without assuming dramatically elevated growth rates of secondary lesions. However, most imaging studies of metastases suggest that their doubling time is similar to the tumor of origin (Klein, 2009). One evident problem with this argumentation is that growth rates of subclinical neoplasms cannot be monitored. It could be that doubling times of *macroscopic* metastases are not significantly different

from primary tumors, but that their early growth stages are accelerated because tumor cells have already acquired proliferation-enabling mutations. Therefore, while tumor growth kinetics data seem to favor parallel progression, standing on their own they do not constitute proof.

Circulating and disseminated tumor cells

More direct evidence can be garnered from comparisons of primary tumors and precursors of overt metastases (i.e. MRD). CTCs and DTCs are defined as cytokeratin (Meng et al., 2004) or EpCAM-positive (Nagrath et al., 2007) cells found in the blood or bone marrow of cancer patients, respectively. Technically, cancer cells lodged in other tissues than the bone marrow are also considered DTCs, but only the bone marrow lends itself to sampling (and still at a high cost of discomfort to the patient). DTCs are significant prognostic biomarkers; their presence is associated with a higher risk of relapse in many common epithelial cancers (Riethdorf et al., 2008). CTC numbers, on the other hand, are predictive of survival in multiple forms of metastatic cancer (Cristofanilli et al., 2004; Danila et al., 2007; Krebs et al., 2011). In a simplified view, DTCs can be considered potential precursors of future metastases, while CTCs represent an aggregate liquid biopsy of proliferating lesions throughout the body.

Since they are targets of adjuvant therapy, there has been a long-standing interest in the genomics of DTCs. The consensus of many studies in this area is that DTCs have fewer genomic aberrations than the corresponding primary tumor (Schardt et al., 2005; Schmidt-Kittler et al., 2003; Weckermann et al., 2009). This result corroborates the model of parallel progression because it suggests that DTCs left the primary tumor in early stages, before more complex genomic aberrations were

acquired. Whether the genetic divergence between DTCs and primary tumors can be confirmed with high-resolution techniques remains to be determined. Genomic analyses of DTCs were often performed with comparative genomic hybridization (CGH) (Klein et al., 2002; 1999), a technique with a resolution below 20 megabases (Pinkel et al., 1998). Many improved technologies for single cell analysis of copy number have since become available, e.g. (Navin et al., 2010), and even single cell exome sequencing is now feasible (Xu et al., 2012). These methodologies await application to the study of DTC genomes. Regardless of analysis technique, one potential caveat when using DTCs to estimate the time point of metastasis is the uncertainty whether these cells truly are metastatic precursors, or rather represent indolent and thus clinically irrelevant remnants of early evolutionary stages of the primary tumor. The most instructive data therefore comes from the direct comparison of primary tumors and macroscopic metastases.

Comparative genomics in solid tumors

Few genome-wide analyses of solid primary tumors and their metastases have been conducted thus far, but they are crucial in providing empirical feedback to our models of metastatic progression. The primary end goal of existing studies typically has been the discovery of mutations that are causative of metastasis. Questions of lineage, i.e. when and by what route dissemination occurs and what the phylogenetic relationships between multiple tumor cell populations are, often are a secondary concern. It will be discussed later why experimental techniques designed for finding metastasis-causing mutations are not necessarily appropriate for inferring lineage relationships. Nevertheless, the available genome-wide portraits of solid primary tumors and matched

metastases represent the most relevant available data for distinguishing between different models of progression. Due to their importance, they will be reviewed in detail below.

Close evolutionary ties among primary tumor and metastases

A study dating to the earlier years of next generation sequencing used an “index lesion” approach to compare metastases to their respective primary tumors in 10 colorectal cancer patients (Jones et al., 2008). Exonic mutations discovered in the index lesion, which was a metastasis in all cases, were evaluated in the primary tumor in a site-specific manner, i.e. without generating a full exome sequence. 97% were present in both neoplasms. These highly convergent results were interpreted in support of linear progression. The authors even created a mathematical model to “translate” the mutation data into chronological time and estimated that while the development of the primary tumor took approximately 25 years, metastasis occurred only 3 years before diagnosis.

In contrast, an analogously designed study in pancreatic carcinoma (Yachida et al., 2010) found substantial genetic divergence. On average, 36% (range 17-52%) of mutations present in the metastatic index lesion could not be detected in the primary tumor or other metastases. At first glance this result suggested that metastasis in pancreatic cancer occurs relatively early. However, the authors went on to sequence DNA from multiple distinct regions of the primary tumor and through this spatially stratified approach were able to identify areas that had the same mutational profile as the index lesion. These “metastatic precursor areas” could be found in both primary tumors analyzed in this fashion. The conclusion was that metastatic subclones evolve within the primary cancer and give rise to metastases late in tumor progression.

It is worth noting here that results from these two studies might be regarded with some caution because the index lesion sequencing approach systematically neglects the evolutionary trajectory of the primary cancer. All mutations that arise after departure of the metastatic clone remain obscure if unbiased variant discovery does not occur in both tumors. Due to this limitation, the data observed in these analyses could also be explained with a scenario in which the metastatic clone disseminates early, enters a period of dormancy during which the mutational profile of the primary tumor at the time of departure is “frozen in time”, and finally undergoes a rapid clonal expansion at the new site. Since all mutations that the primary tumor acquired while the metastasis was dormant would remain invisible, late dissemination could erroneously be concluded.

In a study of metastatic prostate cancer, all lesions were investigated equally comprehensively (Liu et al., 2009). Using comparative genomic hybridization and single nucleotide polymorphism (SNP) arrays, the authors compared between two and eight synchronous metastases in 24 autopsy cases. They found that samples from the same patient typically shared a substantial number of copy number alterations, but also discovered subclonal changes. Unfortunately, no quantitative summary of clonal vs. subclonal aberrations was given, and no phylogenetic reconstruction was attempted: the magnitude of genomic discordance between metastases therefore remained somewhat unclear. A further limitation was that the primary tumor was available for comparison in 5 subjects only. In those cases, the authors reported “no signficant difference” between the primary tumor and its metastases, a finding that would appear to support the linear progression model in the setting of metastatic prostate cancer.

Notably, no specific genetic adaptation of metastases to ectopic microenvironments in different organs – a central prediction of parallel progression – was observed.

That tumor cell populations can thrive in dramatically different microenvironments without undergoing significant genetic adaptation was further shown in a metastatic triple-negative breast carcinoma (Ding et al., 2010). Ding et al. sequenced the whole genome of the primary breast tumor (post neoadjuvant chemotherapy), a pre-treatment biopsy that was propagated as a xenograft, and a matched cerebellar metastasis. 48 out of 50 discovered somatic mutations were present in all three tumors, indicating no significant genetic divergence. Interestingly, the allele frequencies of these mutations were broadly distributed in the primary tumor (ranging from <10% to 89%), while a heterogeneity reduction took place in the metastasis and the xenograft, with more than 50% of mutations showing enrichment in both these samples. This similar enrichment pattern showed that competing for survival in an ectopic microenvironment (regardless of whether this environment is the cerebellum or a mouse organism) can select for the same set of tumor-propagating cells. Notably, chemotherapeutic intervention did not seem to affect this shared clonal composition much. In both samples, the narrowing of the mutant allele frequency distribution was not accompanied by outright loss of any mutations, raising the possibility that the cerebellar metastasis – like the xenograft – was seeded by more than one cell from the primary tumor (polyclonal metastasis).

In contrast, a single cell sequencing study of a triple-negative breast carcinoma and its liver metastasis concluded that metastasis was monoclonal. Navin et al. compared genome-wide copy number alterations of 100 individual cells derived from the two lesions (Navin et al., 2011). Cells from the metastasis and the primary tumor were

very similar to each other, but separated into two distinct branches in an unsupervised neighbor-joining analysis. This indicated that one cell from the dominant clonal population of the primary tumor had founded the metastasis and that since then, no further mingling had taken place. Again these results were considered to be indicative of linear progression and late dissemination.

Genetic divergence among primary tumor and metastases

Relatively few genome-wide studies have found substantial genetic divergence between primary tumor and metastases. In one prominent example, an index lesion sequencing approach was used to compare the mutational spectrum of a pleural effusion metastasis and its primary tumor, a lobular breast carcinoma that was resected 9 years earlier (Shah et al., 2009). In contrast to the findings in pancreatic and colorectal cancer, only 11 of the 32 mutations that were discovered in the metastasis were also present in the primary tumor, indicating independent somatic evolution of the metastatic clone. Whether this result can be regarded as evidence of parallel progression is very debatable, given the long latency of 9 years. If the primary tumor had been sequenced as well, more rigorous conclusions could be drawn. For example, early dissemination would be a possibility if a large number of mutations that were present in the primary tumor could not be found in the metastasis. However, the more limited approach taken in this study is understandable given that sequencing costs at the time were significantly higher than today.

A more recent, very influential case study in renal cell carcinoma also found extensive differences between primary tumor and metastases (Gerlinger et al., 2012). In a thoughtful experimental design, Gerlinger et al. sequenced the exomes of nine

spatially distinct portions of a primary renal clear cell tumor, two synchronous metastases, and two pre-treatment biopsies. The data allowed for several important observations: 1) The evolutionary branches of primary and metastatic clones had diverged early on. Since the split, they had evolved at comparable rates, as shown by an almost identical number of metastasis-specific (n=28) and primary-specific (n=31) mutations. 2) A discrete region in the primary tumor harbored a precursor of the metastatic clone that contained some, but not all of the metastasis-specific mutations. 3) Pre-treatment biopsies of the primary tumor and the metastasis clustered with their post-treatment counterparts, suggesting that treatment with everolimus had not significantly affected clonal compositions. Taken together, the results in this renal cell carcinoma favor parallel progression of primary tumor and metastases.

Smaller-scale comparative genetics

While high-resolution genome- or exome-scale comparisons of primary tumors and metastases are still rare, hundreds such studies have been conducted using more limited marker panels or low-resolution metaphase CGH. Patient numbers in these studies are typically higher than in the genome-wide analyses summarized above, and cases supporting linear and parallel progression are usually found in varying proportions in the same study. Thought-provoking examples are: deep sequencing of a “cancer mini-genome” in primary colorectal cancers and matched metachronous liver metastases demonstrating that the number of concordant mutations vastly differs among patients (Vermaat et al., 2012); a CGH analysis of primary breast carcinomas and matched metachronous metastases demonstrating close clonal relationships in 69%, and almost completely unrelated genomic changes in 31% of patients

(Kuukasjarvi et al., 1997); finally, reports of varying frequencies of discordant mutations in therapeutically or prognostically important genes in lung adenocarcinoma (Schmid et al., 2009), melanoma (Colombino et al., 2012), colorectal (Baldus et al., 2010), and breast cancer (Niikura et al., 2013). Many more examples can be found in an excellent review by Stoecklein and Klein (Stoecklein and Klein, 2009).

Evidence in humans: lessons learned

The eclectic results presented above illustrate that we have yet to arrive at a definitive and coherent picture of metastasis in humans. With respect to the rival hypotheses of linear and parallel progression, the jury is still out on which model will prevail in the long run. It is likely that we will need both frameworks to describe metastasis in different cancer types and clinical scenarios.

Going forward, future genomics studies that aim to infer timelines of metastasis from sequence data should consider the following points. First, it will be germane to provide a detailed clinical context (patient age, anatomic location and extent of disease, treatment history etc.) for all analyzed samples. Current studies often omit this critical information and thus render meaningful comparisons with prior data impossible. How advanced a cancer is and how aggressively it developed likely is significantly associated with the shape of its genetic landscape. For example, as far as can be inferred from the provided clinical information, most findings of genetic concordance between primary tumor and metastases were obtained in patients who underwent extensive treatment and rapidly succumbed to aggressive disease. These samples are often easier to obtain than metachronous metastases, but probably do not accurately represent what happens in patients with more indolent cancers.

Second, a well-defined theoretical framework for the interpretation and comparison of results from different studies should be established. One example of potential misinformation arising in the absence of such a framework is widespread neglect of what could be termed the “founder effect” (illustrated in Figure 1.2). In current practice, the number of alterations that differ between two cancerous lesions is reported, and subjectively judged to be “large” or “small”, sometimes implicitly and often explicitly in relation to the number of overall detected changes. This may lead to misleading results as the number of mutations common to all cells within a tumor may vary widely depending on the mitotic history of the tumor founder cell. The mutational burden of any normal cell continually increases as it divides over a patient’s lifetime. Current estimates are that at least 50% (Tomasetti et al., 2013) or more (Welch et al., 2012) of the mutations found in a cancer represent the “fossil record” of the cell division history of the tumor founder cell. Depending on how frequently the founder cell divided, alterations that accumulated during carcinogenesis may represent different fractions of the total, even if tumors evolved exactly equally otherwise. These effects should be taken into consideration when inferring timelines of metastasis from genetic distances. Gaining a greater understanding of mutation prevalence in normal cells located in different human tissues will be crucial in this regard.

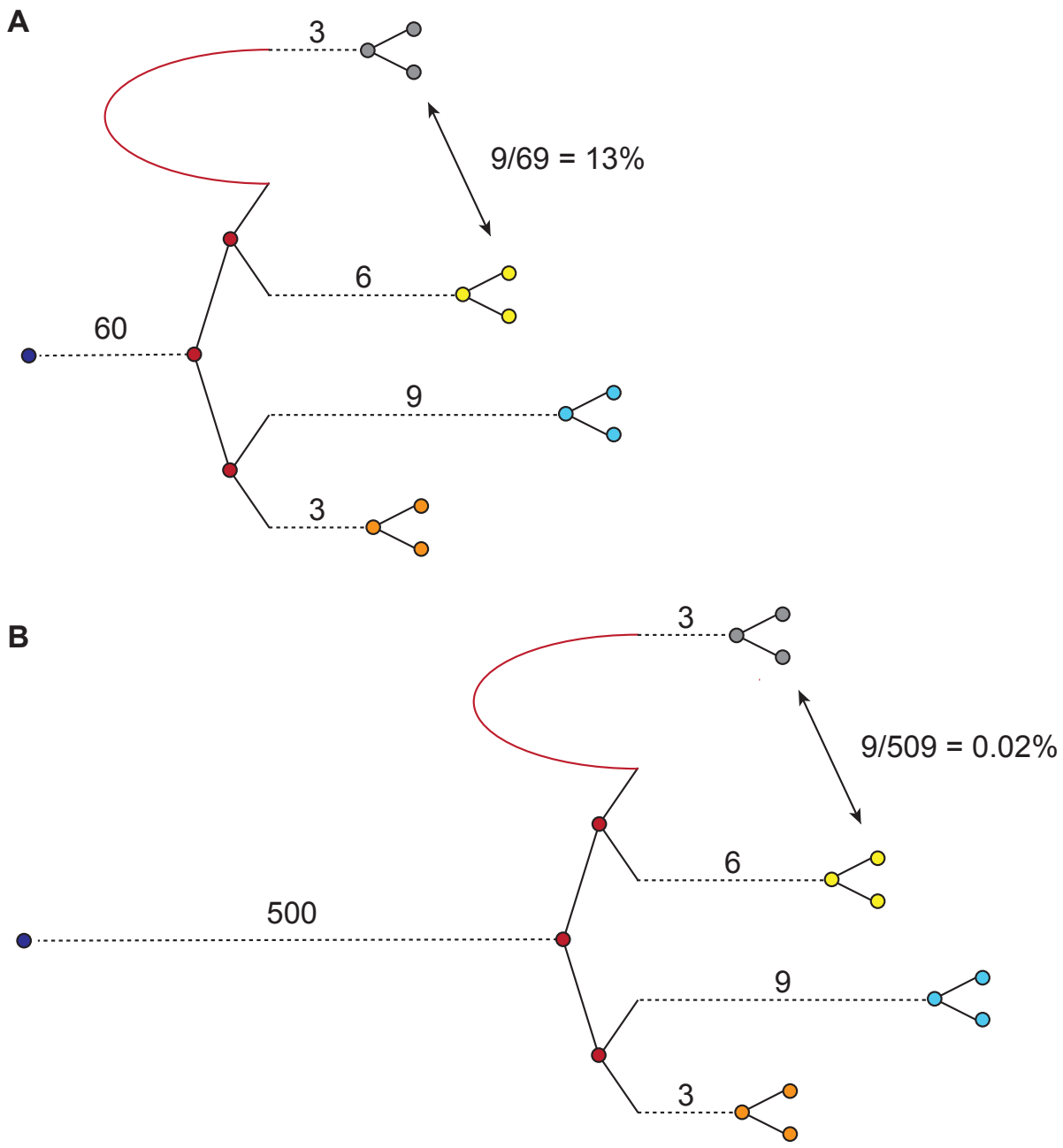

Figure 1.2: Founder Effect. ***(A)*** *A tumor mass arising from a mitotically young founder cell will have fewer clonal mutations than* ***(B)*** *a tumor arising from a mitotically advanced cell. The numbers of subclonal "progressor mutations" acquired during tumor development may appear large or small in comparison with the number of founder mutations.*

It will furthermore be important to determine whether tumor self-seeding plays a role in human disease because a substantial dynamic exchange of cells between synchronous lesions would make genetic reconstruction of tumor history very challenging. Tumor self-seeding seems to be a plausible explanation for some phenomena observed in genome-wide comparisons. For example, in cases in which a small localized patch in the primary tumor corresponds to a distant lesion (Yachida et al., 2010), but is distinct from the dominant clone in the primary, retrograde metastasis may be a more parsimonious explanation then late metastasis of small subclone. We know that gene expression profiles of bulk tumor tissue predict the risk of metastasis (van 't Veer et al., 2002). This finding is difficult to reconcile with metastatic properties confined to a very small portion of cells in the primary tumor. The tumor-self seeding model predicts that returning “seeds” will be more likely to inhabit the surface regions of the primary tumor (Comen et al., 2011). Detailed geographical analysis could elucidate whether this is the case in human tumors. If self-seeding does indeed occur at a significant level, we cannot hope to discover genetic variants that are required for early adaptation to specific microenvironments – an important prediction of parallel progression – in late stage disease.

1.4 Studying the metastatic lineage in humans: experimental approaches

The preceding parts have laid forward the motivations for studying the evolution of metastases and further refining our disease models. Here, a brief overview of experimental approaches toward this goal and their respective advantages and disadvantages will be presented. At the outset, a few conceptual considerations might be helpful. The primary aim of most comparative genetics studies, including those

reviewed above, is to identify alterations with functional significance. The hope is that these “drivers” or causal variants will help explain *why* metastases arise, and thereby inspire new treatment strategies. Resolving the conflict between the competing hypotheses of linear and parallel progression, on the other hand, will require determining which cells metastasize, by what routes, at what point in time: uncovering *what* happens *when*. These are fundamentally questions of tumor lineage or tumor phylogeny. Reflections on lineage are often appended to analyses designed to discover driver mutations, but in fact the questions of *why* vs. *what/when* are preferably addressed with different strategies. The first part of this section will discuss approaches that have been used in tumor lineage tracing, but were not specifically designed for this purpose (“Reappropriated methods”). An overview of methods that were exclusively developed to achieve accurate phylogenetic reconstruction in cancer (“Proper phylogenetic markers”) will follow.

Reappropriated methods

Histopathology

Inspection of a cancer’s histopathology is a critical step in determining prognosis and treatment. It is also the oldest and still most widely used method for lineage determination. Even in modern, molecular biology empowered clinical practice, a pathologist uses morphological examination to decide whether a malignancy is a metastasis or a new primary process, such as in multifocal lung or breast cancer. Treatment courses and prognostication can vary widely based on the outcome. The advantage of this “lineage tracing by eye” is mainly its convenience. However,

morphological comparison may not always reliably determine common descent. An interesting historical example is that Rudolf Virchow, the “father of cellular pathology”, firmly believed for a majority of his scientific career that primary tumors and metastases arise independently. He suspected that the primary cancer infuses the blood with “toxins” that trigger the formation of secondary growths at distant sites, but that no cellular exchange takes place (Weiss, 2000). The modern pathologist appears to be inversely biased: in a morphological evaluation of lung squamous cell carcinoma following head and neck squamous cell carcinoma, 86% of cases were diagnosed as metastases, while a molecular assay based on loss of heterozygosity (LOH) indicated that in fact only 43% were related lesions, and the rest represented independent transformations (Geurts et al., 2005). Also, some tumors can undergo profound histological changes in response to treatment (Sequist et al., 2011). Therefore, if alternatives are available, tumor morphology is not a preferred method for lineage tracing.

Chromosomal alterations

A rich literature documents the use of chromosomal alterations to study clonal relationships in metastatic cancer. Genome-wide approaches include metaphase CGH, its high-resolution variant arrayCGH and SNP microarrays. The most comprehensive view of chromosomal aberrations, including balanced translocations and inversions, can be inferred from deep paired-end sequencing, but almost no such data exists for primary tumors and matched metastases. Commonly used locus-specific techniques are LOH analysis of polymorphic DNA sequences and fluorescence in situ hybridization (FISH). Chromosomal alterations can be detected relatively easily and occur frequently

in many cancer types. However, it is debatable whether they represent good lineage markers because many if not most large-scale rearrangements are likely to have strong selective effects (Davoli et al., 2013). Convergent evolution, the independent occurrence of similar alteration patterns in two phylogenetically unrelated cells, cannot be excluded unless breakpoints are mapped very finely. For example, amplifications of chromosome 7 and deletions of chromosome 10 are present in more than 80% of primary glioblastomas (Beroukhim et al., 2007), even though they are obviously not related by descent. Some chromosomal alterations that are typically regarded as rare stochastic events can be induced by endogenous stimuli, such as the sharp increase of gene fusions between TMPRSS2 and ERG upon dihydrotestosterone exposure in prostate cancer cells (Mani et al., 2009). Finally, our incomplete understanding of how the number of chromosomal alterations relates to mitotic distance poses a problem. It is thought that the total burden of somatic mutations in a cell's genome is correlated with the number of divisions it underwent since fertilization (Welch et al., 2012). Hence, genetic divergence between two lesions as measured by single nucleotide variants (SNVs) can arguably be related to the number of mitoses that occurred since their separation. No such correlation is known for copy number variants (CNVs). For example, it was repeatedly reported that neurons, which are not a mitotically active tissue, are particularly enriched for CNVs (McConnell et al., 2013; Yurov et al., 2007). It seems that developmental and/or environmental factors influence CNVs in ways we do not yet fully appreciate. Chromosomal alterations should therefore be used with caution for inference of tumor phylogenetics.

Genome- or exome-wide somatic mutations

The problem of selective forces potentially causing artifacts in lineage reconstruction does not only arise with chromosomal alterations, but also with exome sequences. The exome is the 1% of the genome that is under most intensive evolutionary pressure and therefore arguably one of the least suitable targets for lineage analysis. Analogous emergence (or disappearance) of somatic variants could easily be misinterpreted as homology. The magnitude of this issue probably depends on the number of divergent mutations. If several dozens of variants are found that differ between primary tumor and metastases, such as in renal cell carcinoma (Gerlinger et al., 2012), many of them are likely to be neutral and reflect lineage relationships correctly. If limited divergent mutations are found – e.g. in a hypothetical scenario in which only a couple of mutations (perhaps even in cancer-related genes) are shared by multiple metastases, but not by the primary tumor – convergent evolution becomes a realistic concern. Ideally, the whole genome would be sequenced in all samples of interest: neutral intergenic regions could be used for lineage analysis, and functionally relevant information could be gleaned from SNVs, CNVs, and other structural variants. The (probably temporary) disadvantage is that whole genome sequencing of multiple tumor specimens in large numbers of patients still puts significant strain on financial and data analysis resources.

Proper phylogenetic markers

A molecular phylogenetic marker suitable for somatic lineage tracing should have several properties: First, it should be selectively neutral. Second, it should mutate at a high rate. Third, acquisition of mutations should occur during cell division so that total mutational burden measured in any cell is proportional to the number of cell divisions it

underwent since the zygote. The following section will present some phylogenetic markers that are compatible with some or all of these demands and thus appropriate for lineage analysis in cancer.

Epigenetic markers

Epigenetic modifications have a long history of being used as lineage markers. X inactivation, the random silencing of one X chromosome in females during early embryonic development, has been particularly useful in the study of tumor lineages. In normal tissues, both X chromosomes are inactivated in similar proportions. In 1965, Linder and Gartler discovered that in leiomyomas, all cells show inactivation of the same X chromosome, providing the first proof of monoclonality in tumors (Linder and Gartler, 1965). Since then, X inactivation has been used extensively to test clonality both within a tumor mass (Going et al., 2001) and between different lesions (Katona et al., 2007; Kuukasjarvi et al., 1997). X inactivation fulfills the first criteria of a good lineage marker because it is a random and presumably neutral event; in most tissues analyzed in bulk, the ratio between silenced alleles is indeed about 1:1, arguing against strong selective effects (Novelli et al., 2003). A further advantage is that silencing is stably heritable and unchanging: therefore, if two cell populations do not share the same X chromosome inactivation pattern, it can safely be concluded that they did not intermix since embryogenesis. However, the static nature of the marker is also its greatest limitation, because it cannot provide any information on evolutionary events that occurred after transformation of a tumor founder cell. Moreover, while a discordant pattern of X chromosome inactivation in two lesions is strong evidence of independent lineages, a concordant pattern can arise with a probability of 50% even if cell

populations are unrelated, fundamentally limiting the resolution of the assay. It therefore does not comply with the two latter characteristics of a good somatic lineage marker: high mutation rate and correlation of alterations with cell division history.

Methylation analysis of CpG dinucleotides fulfills these criteria. A majority of CpG loci are unmethylated in early development and acquire heritable cytosine methylation marks with successive rounds of cell division at a rate that is several orders of magnitude higher than the somatic nucleotide substitution rate (Shibata and Tavaré, 2006). Neutrality can be assumed when CpG loci in promoters of genes that are not expressed in the tissue of interest (e.g. heart-specific loci like CSX in colonic tissue) are examined. At least theoretically, CpG methylation represents an ideal somatic "molecular clock" (Shibata et al., 1996) and has been used extensively to study stem cell (Nicolas et al., 2007; Yatabe et al., 2001) and tumor (Siegmund et al., 2011; 2009; Woo et al., 2009) lineages in humans. However, an important concern is that cytosine methylation is a reversible mark and could potentially be affected by genome-wide methylation changes that occur during tumorigenesis (Feinberg and Vogelstein, 1983). A permanent change in DNA sequence would therefore be preferable to methylation for purposes of lineage tracing.

Microsatellites

Microsatellite sequences arguably come as close as possible to being optimal somatic lineage markers. Also called short tandem or simple sequence repeats, they are consecutive repetitions of one to four base pair units that are vastly overrepresented in the genome. Most are non-coding and show high levels of polymorphism in the population (Ellegren, 2004). Mutations typically occur in the form of insertion or deletion

of one or multiple units through slippage of DNA polymerase (Strand et al., 1993) and are thus tightly coupled to cell division. Mutation rates vary depending on the size and length of the repeat, but are generally much higher than the average genome-wide mutation rate, which is estimated to be approximately 10^{-9} per base per division in humans (Lynch, 2010). Across unique sequence, this number can vary to a limited degree (approximately five-fold according to recent estimates (Lawrence et al., 2013)). By contrast, the mutation rate of a typical $(CA)_{17}$ dinucleotide repeat in human cells is 100-times higher, on the order of 10^{-7} (Boyer et al., 2002).

Microsatellites first entered the spotlight in cancer genetics when frequent somatic length polymorphisms in these sequences were found in familial colorectal cancers (Aaltonen et al., 1993; Ionov et al., 1993) in patients with Lynch syndrome, also known as hereditary non-polyposis coli (HNPCC). The phenomenon, coined “microsatellite instability” (MSI), could also be observed in 10-15% of sporadic colorectal cancers and was associated with an improved prognosis (Samowitz et al., 2001). It was later discovered that MSI was associated with germline (HNPCC) or somatic (sporadic cases) mutations in DNA mismatch repair genes MLH1, MSH2, MSH6 or PMS2 (Bonadona et al., 2011; Fishel et al., 1993).

Microsatellites were subsequently used as “molecular clocks” of tumor evolution in MSI+ human cancers. Shibata and colleagues showed that dinucleotide repeat length distributions vary across tumor regions in HNPCC patients and suggested that heterogeneity is related to mitotic age, with older regions displaying more diversity (Shibata et al., 1996). Interestingly, they found similar mitotic ages in adenomas and carcinomas (Tsao et al., 2000). Randomly occurring replication slippage mutations were

used to reconstruct the phylogenetic relationships between single cells in MMR-deficient *Mlh1+/-* mice, both in normal tissues (Reizel et al., 2011; 2012; Wasserstrom et al., 2008) and a spontaneously arising lymphoma (Frumkin et al., 2008). While these studies attested to the power of somatic microsatellite alterations for phylogenetic inference, mutation rates of most simple repeats in MMR-proficient cells are too low to make this approach generalizable to normal tissues and microsatellite stable tumors.

In 2006, Salipante and Horwitz introduced a novel methodology for somatic lineage tracing that relied on a class of particularly mutable guanine mononucleotide repeats (Salipante and Horwitz, 2006). Polyguanine (poly-G) tracts are abundant in the human genome (Table 1.1) and can reach mutation rates of 10^{-6} per base per division in human cells (Boyer et al., 2002).; they mutate approximately 100 times faster than dinucleotide repeats and 1000 times faster than unique sequence. By genotyping merely 28 poly-G loci, the correct phylogenetic relationships between cells evolving *in vitro* for 66 divisions could be determined (Salipante and Horwitz, 2006). The technique was subsequently used to study various aspects of murine development in MMR-proficient animals (Salipante et al., 2010; 2008; Salk and Horwitz, 2010; Zhou et al., 2013). Importantly, poly-G analysis was also shown to be applicable in the human setting when pre-neoplastic clonal expansions marked by poly-G mutations were identified as a prodrome of cancer development in ulcerative colitis patients (Salk et al., 2009).

Table 1.1: Overview of poly-G tracts in the human genome (Hg19). *Based on custom sequence analysis.*

Total # of poly-G tracts (>10 bp)	9106
In introns	4781
Overlapping exons	296
Overlapping CpG islands	416
Intergenic, no CpG islands	4141

1.5 Research aims

The research presented in this dissertation shows how somatic mutations in hypermutable poly-G tracts can be used to reconstruct phylogenetic relationships in human cancer. The work was motivated by the many pressing questions that still surround the development of metastatic disease. While numerous large-scale, multi-institutional efforts to characterize causal variants in the cancer genome are underway, we still need to improve our understanding of the fundamental steps of metastatic progression in humans. In particular, it remains unclear when and by what route cancer cells spread throughout the organism. At the inception of this project, I extensively searched the literature for a methodology that would enable construction of “pedigrees” of tumor cell populations in human samples. Among the many available technologies, genotyping of poly-G tracts appeared to be the optimal solution: it allowed for quantitative as opposed to merely qualitative analysis; it was (at least theoretically) applicable to archival formalin fixed and paraffin embedded tissue specimens; it was scalable and cost-effective and thus well-suited for analysis of large numbers of

samples. In the following chapter, I will describe how I adapted poly-G tract profiling for the study of metastasis in human colorectal cancer.

References

Aaltonen, L.A., Peltomaki, P., Leach, F.S., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Jarvinen, H., Powell, S.M., Jen, J., and Hamilton, S.R. (1993). Clues to the pathogenesis of familial colorectal cancer. Science *260*, 812–816.

Aguirre-Ghiso, J.A., Bragado, P., and Sosa, M.S. (2013). Metastasis Awakening: Targeting dormant cancer. Nat. Med. *19*, 276–277.

Alberts, S.R., Sargent, D.J., Nair, S., Mahoney, M.R., Mooney, M., Thibodeau, S.N., Smyrk, T.C., Sinicrope, F.A., Chan, E., Gill, S., et al. (2012). Effect of Oxaliplatin, Fluorouracil, and Leucovorin With or Without Cetuximab on Survival Among Patients With Resected Stage III Colon CancerA Randomized Trial. Jama *307*, 1383–1393.

Baldus, S.E., Schaefer, K.L., Engers, R., Hartleb, D., Stoecklein, N.H., and Gabbert, H.E. (2010). Prevalence and Heterogeneity of KRAS, BRAF, and PIK3CA Mutations in Primary Colorectal Adenocarcinomas and Their Corresponding Metastases. Clinical Cancer Research *16*, 790–799.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences *104*, 20007–20012.

Bonadona, V., Bonaïti, B., Olschwang, S., Grandjouan, S., Huiart, L., Longy, M., Guimbaud, R., Buecher, B., Bignon, Y.-J., Caron, O., et al. (2011). Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. Jama *305*, 2304–2310.

Boyer, J.C., Yamada, N.A., Roques, C.N., Hatch, S.B., Riess, K., and Farber, R.A. (2002). Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum. Mol. Genet. *11*, 707–713.

Bross, I.D., Viadana, E., and Pickren, J. (1975). Do generalized metastases occur directly from the primary? J Chronic Dis *28*, 149–159.

Cairns, J. (1975). Mutation selection and the natural history of cancer. Nature *255*, 197–200.

Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., et al. (2011). Improved Survival with

Vemurafenib in Melanoma with BRAF V600E Mutation. N. Engl. J. Med. *364*, 2507–2516.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat Genet *45*, 1127–1133.

Colombino, M., Capone, M., Lissia, A., Cossu, A., Rubino, C., De Giorgi, V., Massi, D., Fonsatti, E., Staibano, S., Nappi, O., et al. (2012). BRAF/NRAS Mutation Frequencies Among Primary Tumors and Metastases in Patients With Melanoma. Journal of Clinical ….

Comen, E., Norton, L., and Massagué, J. (2011). Clinical implications of cancer self-seeding. Nature Publishing Group.

Cristofanilli, M., Budd, G.T., Ellis, M.J., Stopeck, A., Matera, J., Miller, M.C., Reuben, J.M., Doyle, G.V., Allard, W.J., Terstappen, L.W.M.M., et al. (2004). Circulating Tumor Cells, Disease Progression, and Survival in Metastatic Breast Cancer. N. Engl. J. Med. *351*, 781–791.

Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., Bets, D., Mueser, M., Harstrick, A., Verslype, C., et al. (2004). Cetuximab Monotherapy and Cetuximab plus Irinotecan in Irinotecan-Refractory Metastatic Colorectal Cancer. N. Engl. J. Med. *351*, 337–345.

Danila, D.C., Heller, G., Gignac, G.A., Gonzalez-Espinoza, R., Anand, A., Tanaka, E., Lilja, H., Schwartz, L., Larson, S., Fleisher, M., et al. (2007). Circulating Tumor Cell Number and Prognosis in Progressive Castration-Resistant Prostate Cancer. Clinical Cancer Research *13*, 7053–7058.

Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962.

Demicheli, R., Abbattista, A., Miceli, R., Valagussa, P., and Bonadonna, G. (1996). Time distribution of the recurrence risk for breast cancer patients undergoing mastectomy: further support about the concept of tumor dormancy. Breast Cancer Research and Treatment *41*, 177–185.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999–1005.

Disibio, G., and French, S.W. (2008). Metastatic patterns of cancers: results from a large autopsy study. Arch. Pathol. Lab. Med. *132*, 931–939.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.

Feinberg, A.P., and Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature *301*, 89–92.

Fishel, R., Lescoe, M.K., Rao, M.R., Copeland, N.G., Jenkins, N.A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell *75*, 1027–1038.

Fisher, B. (1980). Laboratory and clinical research in breast cancer--a personal adventure: the David A. Karnofsky memorial lecture. Cancer Research *40*, 3863–3874.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G., and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. Cancer Research *68*, 5924–5931.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Geurts, T.W., Nederlof, P.M., van den Brekel, M.W., van't Veer, L.J., de Jong, D., Hart, A.A., van Zandwijk, N., Klomp, H., Balm, A.J., and van Velthuysen, M.L. (2005). Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: metastasis or second primary? Clin. Cancer Res. *11*, 6608–6614.

Gianni, L., Dafni, U., Gelber, R.D., and Azambuja, E. (2011). Treatment with trastuzumab for 1 year after adjuvant chemotherapy in patients with HER2-positive early breast cancer: a 4-year follow-up of a randomised controlled trial. Lancet Oncol.

Giuliano, A.E., Hunt, K.K., Ballman, K.V., Beitsch, P.D., Whitworth, P.W., Blumencranz, P.W., Leitch, A.M., Saha, S., McCall, L.M., and Morrow, M. (2011). Axillary Dissection vs No Axillary Dissection in Women With Invasive Breast Cancer and Sentinel Node MetastasisA Randomized Clinical Trial. Jama *305*, 569–575.

Going, J.J., Abd El-Monem, H.M., and Craft, J.A. (2001). Clonal origins of human breast cancer. J. Pathol. *194*, 406–412.

Goss, G.D., O'Callaghan, C., Lorimer, I., Tsao, M.S., Masters, G.A., Jett, J., Edelman, M.J., Lilenbaum, R., Choy, H., Khuri, F., et al. (2013). Gefitinib Versus Placebo in Completely Resected Non-Small-Cell Lung Cancer: Results of the NCIC CTG BR19 Study. Journal of Clinical Oncology *31*, 3320–3326.

Holohan, C., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2013). Cancer drug resistance: an evolving paradigm. Nat. Rev. Cancer *13*, 714–726.

Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature *363*, 558–561.

Jones, S., Chen, W.D., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M.A., Siegel, C., Velculescu, V.E., et al. (2008). Comparative lesion sequencing provides insights into tumor evolution. Proceedings of the National Academy of Sciences *105*, 4283–4288.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Katona, T.M., Jones, T.D., Wang, M., Eble, J.N., Billings, S.D., and Cheng, L. (2007). Genetically heterogeneous and clonally unrelated metastases may arise in patients with cutaneous melanoma. Am J Surg Pathol *31*, 1029–1037.

Kim, M.Y., Oskarsson, T., Acharyya, S., Nguyen, D.X., Zhang, X.H., Norton, L., and Massague, J. (2009). Tumor self-seeding by circulating cancer cells. Cell *139*, 1315–1326.

Klein, C.A. (2009). Parallel progression of primary tumours and metastases. Nat. Rev. Cancer *9*, 302–312.

Klein, C.A. (2011). Framework models of tumor dormancy from patient-derived observations. Curr. Opin. Genet. Dev. *21*, 42–49.

Klein, C.A., Blankenstein, T.J., Schmidt-Kittler, O., Petronio, M., Polzer, B., Stoecklein, N.H., and Riethmuller, G. (2002). Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. Lancet *360*, 683–689.

Klein, C.A., Schmidt-Kittler, O., Schardt, J.A., Pantel, K., Speicher, M.R., and Riethmuller, G. (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proceedings of the National Academy of Sciences of the United States of America *96*, 4494–4499.

Klein, C.A. (2013). Selection and adaptation during metastatic cancer progression. Nature *501*, 365–372.

Krebs, M.G., Sloane, R., Priest, L., Lancashire, L., Hou, J.M., Greystoke, A., Ward, T.H., Ferraldeschi, R., Hughes, A., Clack, G., et al. (2011). Evaluation and Prognostic Significance of Circulating Tumor Cells in Patients With Non-Small-Cell Lung Cancer. Journal of Clinical Oncology *29*, 1556–1563.

Kuukasjarvi, T., Karhu, R., Tanner, M., Kahkonen, M., Schaffer, A., Nupponen, N., Pennanen, S., Kallioniemi, A., Kallioniemi, O.P., and Isola, J. (1997). Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. Cancer Research *57*, 1597–1604.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714–726.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Linder, D., and Gartler, S.M. (1965). Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. Science *150*, 67–69.

Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C.M., Eisenberger, M.A., Carducci, M.A., et al. (2009). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat. Med. *15*, 559–565.

Lynch, M. (2010). Evolution of the mutation rate. Trends Genet. *26*, 345–352.

Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., Gemma, A., Harada, M., Yoshizawa, H., Kinoshita, I., et al. (2010). Gefitinib or Chemotherapy for Non–Small-Cell Lung Cancer with Mutated EGFR. N. Engl. J. Med. *362*, 2380–2388.

Mani, R.S., Tomlins, S.A., Callahan, K., Ghosh, A., Nyati, M.K., Varambally, S., Palanisamy, N., and Chinnaiyan, A.M. (2009). Induced chromosomal proximity and gene fusions in prostate cancer. Science *326*, 1230.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? Nat. Rev. Cancer *12*, 323–334.

McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., et al. (2013). Mosaic copy number variation in human neurons. Science *342*, 632–637.

Mehlen, P., and Puisieux, A. (2006). Metastasis: a question of life or death. Nat. Rev. Cancer *6*, 449–458.

Meng, S., Tripathy, D., Frenkel, E.P., Shete, S., Naftalis, E.Z., Huth, J.F., Beitsch, P.D., Leitch, M., Hoover, S., Euhus, D., et al. (2004). Circulating tumor cells in patients with breast cancer dormancy. Clin. Cancer Res. *10*, 8152–8162.

Nagrath, S., Sequist, L.V., Maheswaran, S., Bell, D.W., Irimia, D., Ulkus, L., Smith, M.R., Kwak, E.L., Digumarthy, S., Muzikansky, A., et al. (2007). Isolation of rare circulating tumour cells in cancer patients by microchip technology. Nature *450*, 1235–1239.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., et al. (2010). Inferring tumor progression from genomic heterogeneity. Genome Res. *20*, 68–80.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Nicolas, P., Kim, K.M., Shibata, D., and Tavare, S. (2007). The stem cell population of the human colon crypt: analysis via methylation patterns. PLoS Computational Biology *3*, e28.

Niikura, N., Odisio, B.C., Tokuda, Y., Symmans, F.W., Hortobagyi, G.N., and Ueno, N.T. (2013). Latest biopsy approach for suspected metastases in patients with breast cancer. Nat. Rev. Clin. Oncol. advance online publication 22 October 2013; doi:10.1038/nrclinonc.2013.182

Novelli, M., Cossu, A., Oukrif, D., Quaglia, A., Lakhani, S., Poulsom, R., Sasieni, P., Carta, P., Contini, M., Pasca, A., et al. (2003). X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. Proceedings of the National Academy of Sciences of the United States of America *100*, 3311–3314.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet *20*, 207–211.

Polzer, B., and Klein, C.A. (2013). Metastasis Awakening: The challenges of targeting minimal residual cancer. Nat. Med. *19*, 274–275.

Reichardt, P., Blay, J.Y., Boukovinas, I., Brodowicz, T., Broto, J.M., Casali, P.G., Decatris, M., Eriksson, M., Gelderblom, H., Kosmidis, P., et al. (2012). Adjuvant therapy in primary GIST: state-of-the-art. Annals of Oncology *23*, 2776–2781.

Reizel, Y., Chapal-Ilani, N., Adar, R., Itzkovitz, S., Elbaz, J., Maruvka, Y.E., Segev, E., Shlush, L.I., Dekel, N., and Shapiro, E. (2011). Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. PLoS Genet. *7*, e1002192.

Reizel, Y., Itzkovitz, S., Adar, R., Elbaz, J., Jinich, A., Chapal-Ilani, N., Maruvka, Y.E., Nevo, N., Marx, Z., Horovitz, I., et al. (2012). Cell lineage analysis of the Mammalian female germline. PLoS Genet. *8*, e1002477.

Riethdorf, S., Wikman, H., and Pantel, K. (2008). Review: Biological relevance of disseminated tumor cells in cancer patients. International Journal of Cancer. Journal International Du Cancer *123*, 1991–2006.

Salipante, S.J., and Horwitz, M.S. (2006). Phylogenetic fate mapping. Proceedings of the National Academy of Sciences of the United States of America *103*, 5448–5453.

Salipante, S.J., Kas, A., McMonagle, E., and Horwitz, M.S. (2010). Phylogenetic analysis of developmental and postnatal mouse cell lineages. Evol Dev *12*, 84–94.

Salipante, S.J., Thompson, J.M., and Horwitz, M.S. (2008). Phylogenetic fate mapping: theoretical and experimental studies applied to the development of mouse fibroblasts. Genetics *178*, 967–977.

Salk, J.J., Salipante, S.J., Risques, R.A., Crispin, D.A., Li, L., Bronner, M.P., Brentnall, T.A., Rabinovitch, P.S., Horwitz, M.S., and Loeb, L.A. (2009). Clonal expansions in ulcerative colitis identify patients with neoplasia. Proceedings of the National Academy of Sciences *106*, 20871–20876.

Salk, J.J., and Horwitz, M.S. (2010). Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. Semin. Cancer Biol. *20*, 294–303.

Samowitz, W.S., Curtin, K., Ma, K.-N., Schaffer, D., Coleman, L.W., Leppert, M., and Slattery, M.L. (2001). Microsatellite Instability in Sporadic Colon Cancer Is Associated with an Improved Prognosis at the Population Level. Cancer Epidemiology ….

Schardt, J.A., Meyer, M., Hartmann, C.H., Schubert, F., Schmidt-Kittler, O., Fuhrmann, C., Polzer, B., Petronio, M., Eils, R., and Klein, C.A. (2005). Genomic analysis of single cytokeratin-positive cells from bone marrow reveals early mutational events in breast cancer. Cancer Cell *8*, 227–239.

Schmid, K., Oehl, N., Wrba, F., Pirker, R., Pirker, C., and Filipits, M. (2009). EGFR/KRAS/BRAF Mutations in Primary Lung Adenocarcinomas and Corresponding Locoregional Lymph Node Metastases. Clinical Cancer ….

Schmidt-Kittler, O., Ragg, T., Daskalakis, A., Granzow, M., Ahr, A., Blankenstein, T.J., Kaufmann, M., Diebold, J., Arnholdt, H., Muller, P., et al. (2003). From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. Proceedings of the National Academy of Sciences of the United States of America *100*, 7737–7742.

Sequist, L.V., Waltman, B.A., Dias-Santagata, D., Digumarthy, S., Turke, A.B., Fidias, P., Bergethon, K., Shaw, A.T., Gettinger, S., Cosper, A.K., et al. (2011). Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. Sci Transl Med *3*, 75ra26.

Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809–813.

Shaw, A.T., Yeap, B.Y., Solomon, B.J., Riely, G.J., Gainor, J., Engelman, J.A., Shapiro, G.I., Costa, D.B., Ou, S.-H.I., Butaney, M., et al. (2011). Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. Lancet Oncol. *12*, 1004–1012.

Shibata, D., Navidi, W., Salovaara, R., Li, Z.H., and Aaltonen, L.A. (1996). Somatic microsatellite mutations as molecular tumor clocks. Nat. Med. *2*, 676–681.

Shibata, D., and Tavaré, S. (2006). Counting divisions in a human somatic cell tree: how, what and why? Cell Cycle *5*, 610–614.

Siegmund, K.D., Marjoram, P., Tavare, S., and Shibata, D. (2011). High DNA

methylation pattern intratumoral diversity implies weak selection in many human colorectal cancers. PloS One *6*, e21657.

Siegmund, K.D., Marjoram, P., Woo, Y.J., Tavare, S., and Shibata, D. (2009). Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proceedings of the National Academy of Sciences *106*, 4828–4833.

Stoecklein, N.H., and Klein, C.A. (2009). Genetic disparity between primary tumours, disseminated tumour cells, and manifest metastasis. International Journal of Cancer. Journal International Du Cancer.

Strand, M., Prolla, T.A., Liskay, R.M., and Petes, T.D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature *365*, 274–276.

Tomasetti, C., Vogelstein, B., and Parmigiani, G. (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proceedings of the National Academy of Sciences *110*, 1999–2004.

Tsai, M.S., Su, Y.H., Ho, M.C., and Chen, T.P. (2007). Clinicopathological Features and Prognosis in Resectable Synchronous and Metachronous Colorectal Liver Metastasis - Springer. Annals of Surgical Oncology.

Tsao, J.L., Yatabe, Y., Salovaara, R., Jarvinen, H.J., Mecklin, J.P., Aaltonen, L.A., Tavare, S., and Shibata, D. (2000). Genetic reconstruction of individual colorectal tumor histories. Proceedings of the National Academy of Sciences of the United States of America *97*, 1236–1241.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530–536.

Vermaat, J.S., Nijman, I.J., Koudijs, M.J., Gerritse, F.L., Scherer, S.J., Mokry, M., Roessingh, W.M., Lansu, N., de Bruijn, E., van Hillegersberg, R., et al. (2012). Primary colorectal cancers and their subsequent hepatic metastases are genetically different: implications for selection of patients for targeted treatment. Clin. Cancer Res. *18*, 688–699.

Wasserstrom, A., Adar, R., Shefer, G., Frumkin, D., Itzkovitz, S., Stern, T., Shur, I., Zangi, L., Kaplan, S., Harmelin, A., et al. (2008). Reconstruction of cell lineage trees in mice. PloS One *3*, e1939.

Weckermann, D., Polzer, B., Ragg, T., Blana, A., Schlimok, G., Arnholdt, H., Bertz, S., Harzmann, R., and Klein, C.A. (2009). Perioperative activation of disseminated tumor cells in bone marrow of patients with prostate cancer. J. Clin. Oncol. *27*, 1549–1556.

Weinberg, R.A. (2008). Mechanisms of malignant progression. Carcinogenesis *29*, 1092–1095.

Weiss, L. (2000). Concepts of Metastasis - Springer. Cancer and Metastasis Reviews *19*, 219–234.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

Woo, Y.J., Siegmund, K.D., Tavare, S., and Shibata, D. (2009). Older individuals appear to acquire mitotically older colorectal cancers. J. Pathol. *217*, 483–488.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886–895.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Yatabe, Y., Tavare, S., and Shibata, D. (2001). Investigating stem cells in human colon by using methylation patterns. Proceedings of the National Academy of Sciences of the United States of America *98*, 10839–10844.

Yurov, Y.B., Iourov, I.Y., Vorsanova, S.G., Liehr, T., Kolotii, A.D., Kutsev, S.I., Pellestor, F., Beresheva, A.K., Demidova, I.A., Kravets, V.S., et al. (2007). Aneuploidy and confined chromosomal mosaicism in the developing human brain. PloS One *2*, e558.

Zhou, W., Tan, Y., Anderson, D.J., Crist, E.M., Ruohola-Baker, H., Salipante, S.J., and Horwitz, M.S. (2013). Use of somatic mutations to quantify random contributions to mouse development. BMC Genomics *14*, 39.

Chapter 2: Hypermutable DNA chronicles the evolution of human colon cancer

Statement of contribution: This chapter corresponds to a manuscript currently under review by Kamila Naxerova, Elena Brachtel, Jesse Salk, Aaron Seese, Karen Power, Bardia Abbasi, Matija Snuderl, Sarah Chiang, Simon Kasif and Rakesh K. Jain. I conceived of the research, designed and performed all experiments and wrote the manuscript. EB oversaw the collection of human tissue samples and provided invaluable clinical insight. JS contributed crucial theoretical considerations. AS, KP and BA provided important technical assistance. MS and SC contributed human tissue samples. SK offered computational advice. RKJ oversaw and guided all research and gave indispensible advice along the way. All co-authors provided helpful comments during writing.

2.1 Abstract

Intra-tumor heterogeneity in metastatic cancer influences treatment outcomes but has not been assessed in representative patient populations. Here we report that a simple PCR-based assay interrogating somatic variation in hypermutable polyguanine repeats can be used to study clonal architecture in human cancer. We find that 91% of tumors in a cohort of 22 colon carcinoma patients contain polyguanine variants. Mutation load positively correlates with patient age and inversely correlates with histological grade. We generate polyguanine mutation profiles of spatially separated samples from primary carcinomas and matched metastases to build well-supported phylogenetic trees that illuminate individual patient's path of progression. Our results show varying degrees of intra-tumor heterogeneity in different patients and suggest that metastasis occurs late in colon cancer development.

2.2 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 intra-tumor 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 metastasis in colon cancer is a late event and indicate 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.

2.3 Introduction

Human cancers are composed of a continually evolving population of genetically and phenotypically divergent cells (Marusyk et al., 2012). 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 (Fidler, 2003; Greaves and Maley, 2012; Merlo et al., 2006). In recent years, characterization of intra-tumor heterogeneity has received increased attention as advanced sequencing technologies have enabled more detailed analysis of tumor cell populations (Anderson et al., 2011; Ding et al., 2010; Gerlinger et al., 2012; Yachida et al., 2010).

Depending on the context, the term “intra-tumor 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. 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 (Klein, 2008; 2009; Weinberg, 2007). This aggressive clone generates new metastases in a so-called ‘metastasis shower’ (Weinberg, 2008). Linear progression predicts metastases will be genetically similar to the primary tumor and to each other. The alternative ‘parallel progression’ model (Klein, 2009) 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 (Yachida et al., 2010) and triple negative breast cancer (Ding et al., 2010) 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 (Gerlinger et al., 2012). 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 intra-tumor heterogeneity and mitotic history can reveal the evolution of systemic disease. Many clinically relevant questions in this area remain unanswered. Are highly diversified primary tumors more likely to give rise to genetically heterogeneous metastases? How does genetic diversity affect response to therapy? Heterogeneity is thought to increase resistance to therapy (Shibata, 2012). Metastases appear to be more homogeneous than primary tumors (Liu et al., 2009) – how does this affect their therapy response?

Addressing these important questions about the evolution of metastatic cancer will require analyzing large numbers of 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, though, only large genome centers can regularly generate and process data sets of this magnitude. To study intra-tumor 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 since 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 non-coding regions are especially suited for this purpose. Due to replication slippage (Viguera et al., 2001), mutations are introduced frequently but presumably have no effect on fitness. In patients with DNA mismatch repair (MMR) defects and resulting microsatellite instability, variation in dinucleotide repeats has been used to study several aspects of tumor progression (Shibata et al., 1996; Tsao et al., 2000; 1998), but mutation rates in tumors with intact MMR are too low to make this approach widely applicable (Laiho et al., 2002).

Recent research identified a particularly mutable class of polyguanine (poly-G) mononucleotide repeats as a hotspot of somatic variation even in normal cells (Salipante and Horwitz, 2006). Analysis of poly-G repeats has successfully been used to study phylogenetic relationships between single cells in mouse development (Salipante et al., 2010; 2008; Zhou et al., 2013) and has been adapted for detecting preneoplastic clonal expansions in ulcerative colitis patients (Salk et al., 2009).

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, in accordance with previous

studies (Jones et al., 2008), suggests that metastasis in colon cancer is a late event and demonstrates how a simple and highly scalable assay can be used to generate reliable maps of clonal architecture in formalin fixed and paraffin embedded (FFPE) tumor samples.

2.4 Results

Polyguanine tracts encode tumor lineage

We began by screening a cohort of 22 human colon cancers for somatic mutations in a panel of 20 poly-G tracts. Insertions/deletions of one or more base pairs 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 (Ellegren, 2004; Weber and Wong, 1993). Guanine homopolymers are particularly prone to replication slippage errors and can have mutation frequencies as high as 10^{-4} per base per cell division (Boyer et al., 2002). Figure 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 (Ellegren, 2004), and generally increases with repeat length (Brinkmann et al., 1998).

Figure 2.1: Propagation of neutral poly-G mutations in 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 stutter distributions for marker Sal45 for normal tissue, a primary colon cancer, and a metastasis to the ovary.*

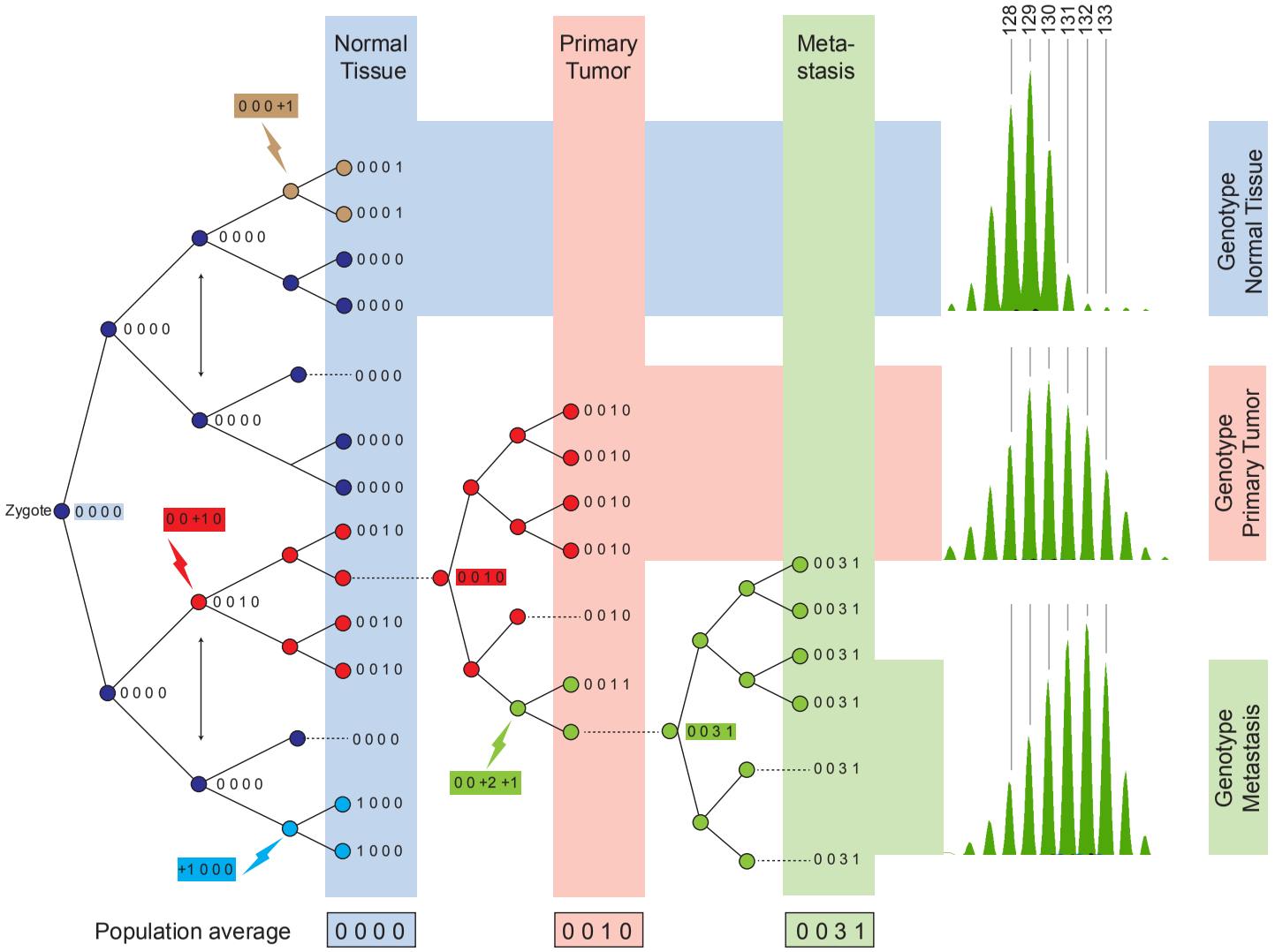

Figure 2.1 (Continued).

encodes its cell division history and its location in the organism's "cell lineage tree" (Frumkin et al., 2008; 2005; Wasserstrom et al., 2008a). 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 (Salipante et al., 2010) and cell culture (Salipante and Horwitz, 2006) using poly-G tracts, other microsatellites (Frumkin et al., 2008; Wasserstrom et al., 2008a), or random genomic regions (Carlson et al., 2012) 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 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 (Salipante et al., 2010). 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, i.e. the zygote or "germline" genotype in the case of normal tissue (Salk and Horwitz, 2010).

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 in turn rise above the “white noise” of genetic diversity 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 rightmost panel of Figure 2.1 shows examples of poly-G tract genotypes in normal (polyclonal) human tissue, a primary tumor, and its metastasis. Since poly-G tracts are inherently hypermutable, Taq polymerase slippage during PCR generates a fragment distribution instead of a single product. This fragment distribution can be precisely quantified, at single base pair resolution, by capillary electrophoresis following PCR with fluorescent primers. The distribution mode 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. We sought to determine whether mutations in poly-G sequences could be found in human colon cancer patients.

Polyguanine mutations are present in most colon cancers.

The 22 cases in our cohort were randomly selected from all patients who underwent colectomy and received a diagnosis of invasive carcinoma at Massachusetts General Hospital between 2010 and 2011. Complete patient information (pathological diagnosis, tumor size, histological grade, stage, anatomic location of the tumor, neoadjuvant therapy etc.) is presented in Supplementary Table S1. Since our ultimate goal was to study metastatic progression, we further sub-selected patients who had at least 3 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 formalin-fixed paraffin-embedded tissue cores and subjected to poly-G tract profiling. Mutated alleles were found in 91% of patients (Fig. 2.2A; complete genotype information in Supplementary Tables S2-S6). As expected, colon cancers with microsatellite instability (MSI) harbored the most alterations, and their mutation frequencies clearly separated them from microsatellite stable (MSS) tumors (Fig. 2.3A). Yet 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 mismatch repair. Loss of DNA mismatch repair 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 one or two base pairs 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. 2.2B).

We did not know *a priori* whether these new alleles were generated during tumor development or were already present in the tumor founder cell. In the latter case, the new variants would represent the mutational fingerprint a normal colonic stem cell acquired as it proliferated over a patient's lifetime. Colonic stem cells divide very frequently – every 30 hours by some estimates (Potten et al., 1992) – and would therefore be expected to accumulate large numbers of mutations over the years, with

Figure 2.2: Poly-G mutations in 22 human colon cancers. (A) *Mutation frequency plotted as mutations/number of interrogated loci for each patient. Clinical characteristics are listed in the table below, including microsatellite instability (red – unstable, green – stable), chemotherapy (red – neoadjuvant therapy, green – therapy naïve), extent of invasion (T4, red – through serosa; T3, orange – through muscularis propria into pericolorectal tissues; T2, yellow – through submucosa and extending into muscularis propria; T1, yellow – into submucosa; Rec, blue – recurrence, no primary tumor), lymph node status (N2b, red – metastasis in 7 or more regional lymph nodes; N2a, orange – 4 to 6 lymph nodes; N1b, yellow – 2 to 3 lymph nodes; N1a, yellow – 1 lymph node; N1c, yellow – tumor deposits in the subserosa, mesentery or non-peritonealized pericolic or perirectal tissues without regional lymph node metastasis) and distant metastasis (red – present, green – absent)* ***(B)*** *PCR stutter distribution for marker Sal52 in a microsatellite stable cancer (top) and a cancer with microsatellite instability (bottom). 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 microsatellite stable tumor.*

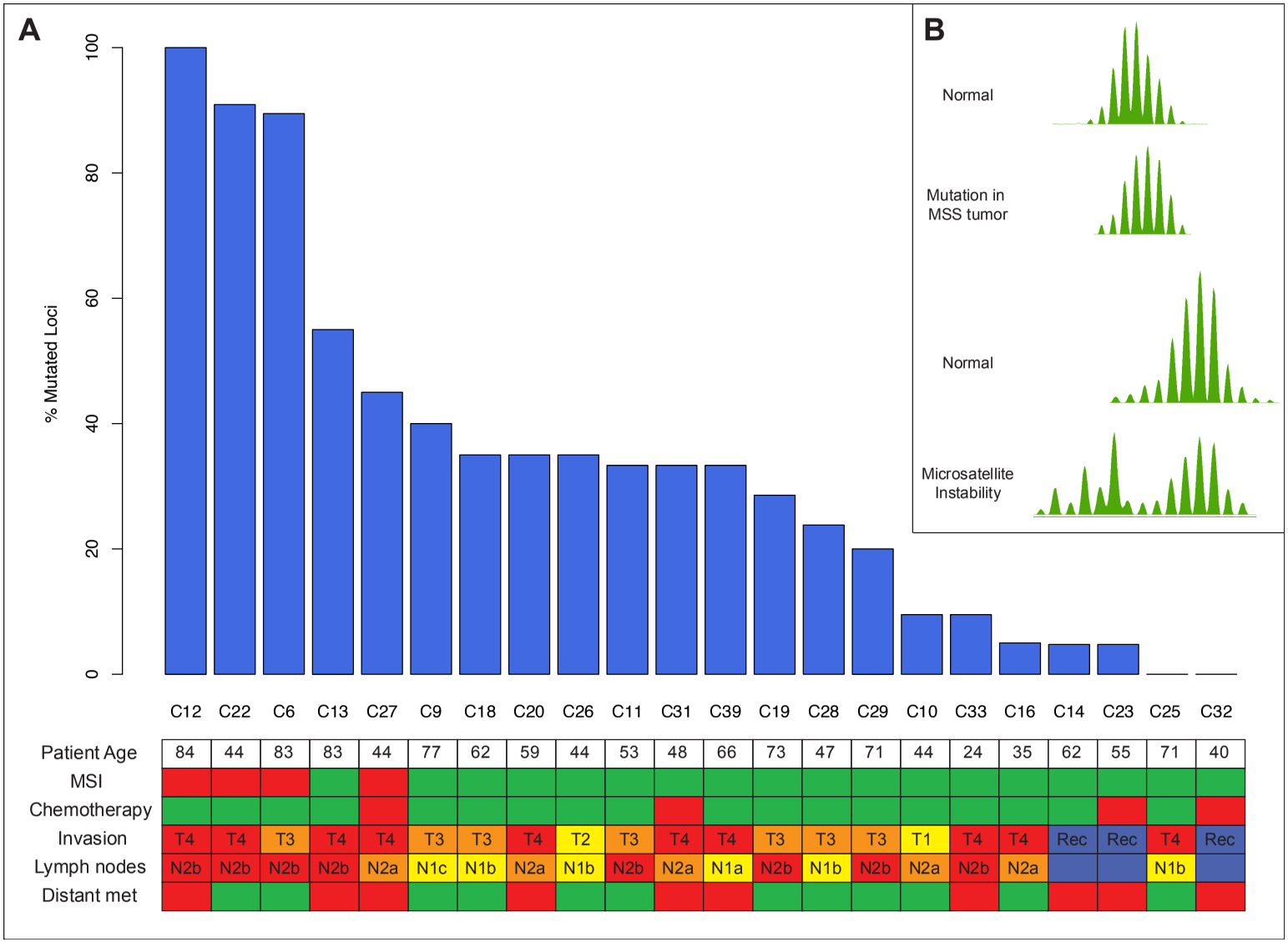

Figure 2.2 (Continued).

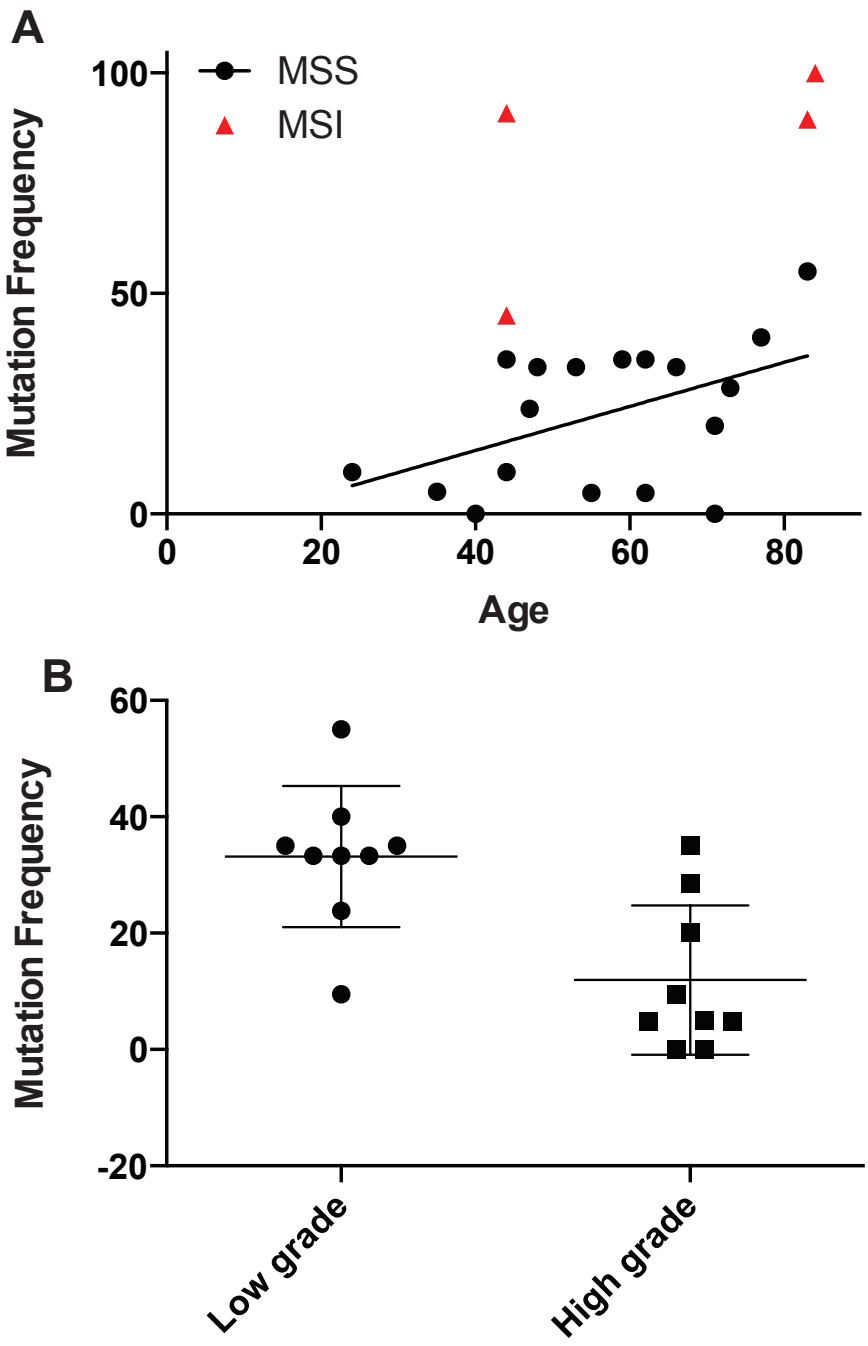

Figure 2.3: Association of mutation frequency with age and grade. (A) *Age is positively correlated with mutation load in microsatellite stable tumors, p=0.0416 (linear regression after exclusion of MSI cases, R^2=0.23, Pearson correlation coefficient=0.48)* ***(B)*** *Low grade tumors contain more mutations than high grade tumors, p=0.0041 (two tailed Mann-Whitney test)*

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 (Welch et al., 2012) and colorectal cancer (Laiho et al., 2002; Tomasetti et al., 2013). We tested this correlation in our data set after excluding MSI cases, since a distinct mutational mechanism is operational in these tumors. We found significant positive correlation between patient age and mutation frequency (Fig. 2.3A), suggesting that the poly-G tract mutation profile partially reflects the genotype of the tumor founder cell. Tumor size, lymph node status and presence of distant metastases were not significantly associated with mutation frequency, but since 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 tumor grade (Fig. 2.3B). Age and tumor grade did not correlate. Assuming the poly-G mutation profile is indeed shaped by the tumor founder cell's genotype, this observation suggests that poorly differentiated tumors derive from a cell with relatively short mitotic history, such as a rarely dividing tissue stem cell.

Polyguanine tract profiles generate a map of tumor evolution.

To determine whether poly-G mutations could be used to reconstruct lineage relationships between multiple tumor samples from the same patient, we selected four patients whose primary tumors had mutations at more than 30% of investigated loci. For each patient, we then collected between 8 and 15 spatially separated samples from different regions of the tumor (primary tumor mass, lymph node metastases, distant

metastases). We generated poly-G tract profiles for each sample using the same 20 markers used in our initial screen. To facilitate data analysis, we developed a semi-automated method for converting poly-G stutter distributions into genotypes (detailed in Experimental Procedures and Supplementary Figure S8). 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 as detailed below.

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-year-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. 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 identify lineage relationships in this patient by determining the two carcinomas'

Figure 2.4: Patient C39 with two synchronous adenocarcinomas of the colon. *CT – Cecal tumor, ST – Sigmoid tumor, LN – Lymph node metastasis, LM – Liver metastasis, N – Normal. 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. The phylogenetic tree was constructed by neighbor joining. Confidence values for each interior branch were calculated from 1000 bootstrap replicates. Branches with confidence values below 70% were collapsed into polytomies.*

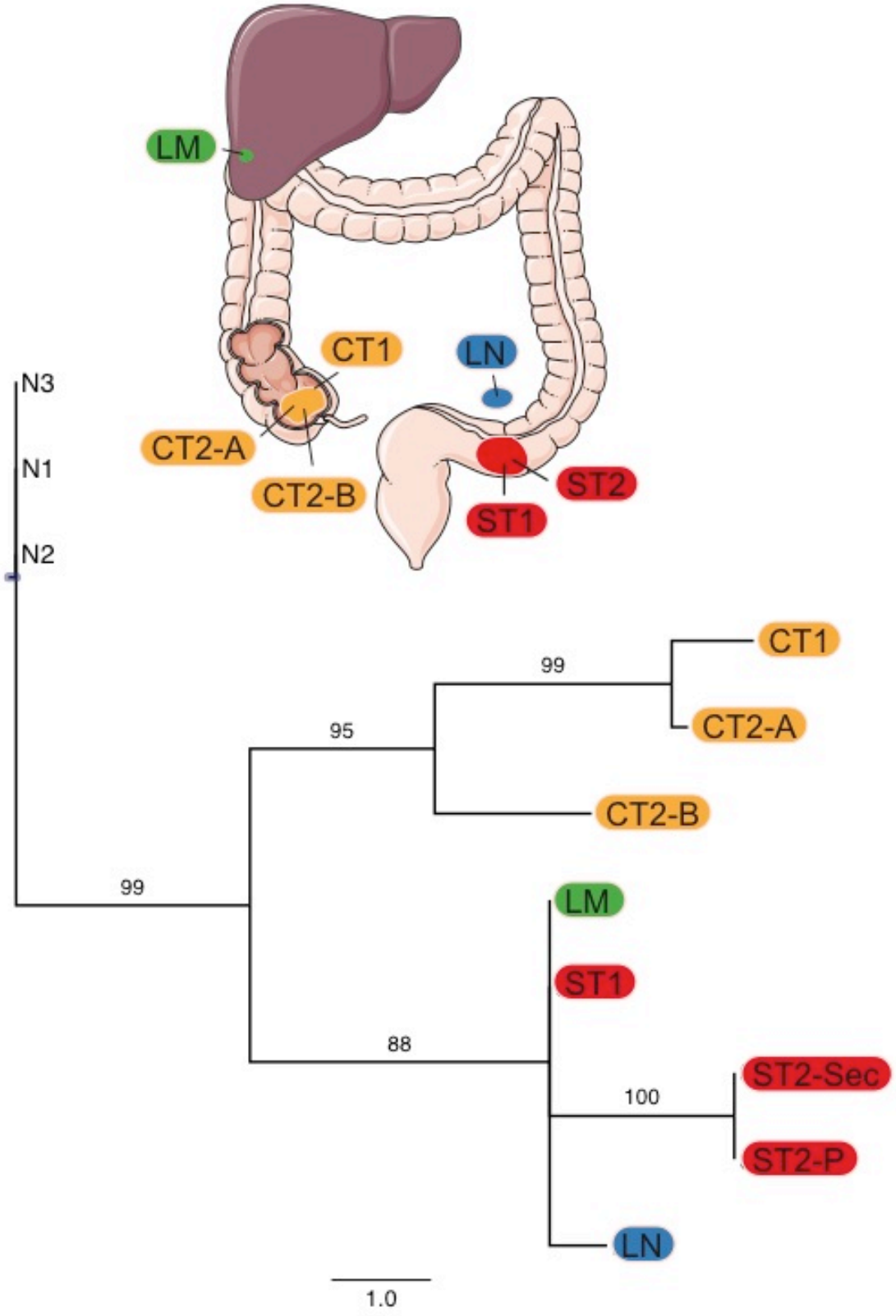

Figure 2.4 (Continued).

independent origins and linking the lymph node metastasis to its tumor of origin in the sigmoid colon. We found 7 variants in the most mutated parts of the cecal tumor, and 7 in the sigmoid lesion. That the tumors had the same number of mutations suggested similar mitotic ages, yet the mutations were largely mutually exclusive. (Full genotype data and a heatmap of all mutations can be found in Supplementary Table S3 and Supplementary Figure 1, respectively. Since both tumors had similar numbers of mutations, only the sigmoid tumor is depicted in the overview in Figure 2.2) The phylogenetic tree constructed from these data located the two tumors in two independent evolutionary branches with high confidence values based on 1000 bootstrap replicates (Figure 2.4). The lymph node metastasis was correctly assigned to the sigmoid tumor's branch. One year after the initial surgery, and after six cycles of adjuvant chemotherapy with FOLFOX (folinic acid, fluorouracil, oxaliplatin), 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.

Extensively diversified primary tumor gives rise to homogeneous metastasis.

Patient C13 was an 83-year-old female with a 7.0 cm invasive colonic adenocarcinoma and metastases to the left and right ovaries (Figure 2.5A). All lesions were removed in one surgery, and the patient did not receive any prior chemotherapy. The tumor was moderately differentiated (Figure 2.5B), microsatellite stable, and involved the ileum,

Figure 2.5: Patient C13 with invasive adenocarcinoma of the colon and metastasis to the ovaries. ***(A)*** *Approximate anatomical localization of all analyzed samples. PT – Primary tumor, RO – Right ovary Metastasis, LO – Left ovary Metastasis, N – Normal. 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)*** *Histological images of tumor regions PT4, PT5, PT6 and LO3. Scale bar 20 µm.* ***(C)*** *Neighbor-joining tree with bootstrap values, branches with bootstrap values below 70% collapsed into polytomies.* ***(D)*** *Complete mutation heatmap. Grey squares signify allele distributions that are indistinguishable from normal control. Colored squares indicate a shift in allele distribution, i.e. a poly-G mutation. If multiple different distributions exist per marker, they are indicated with additional colors. White squares indicate missing data due to amplification failure.*

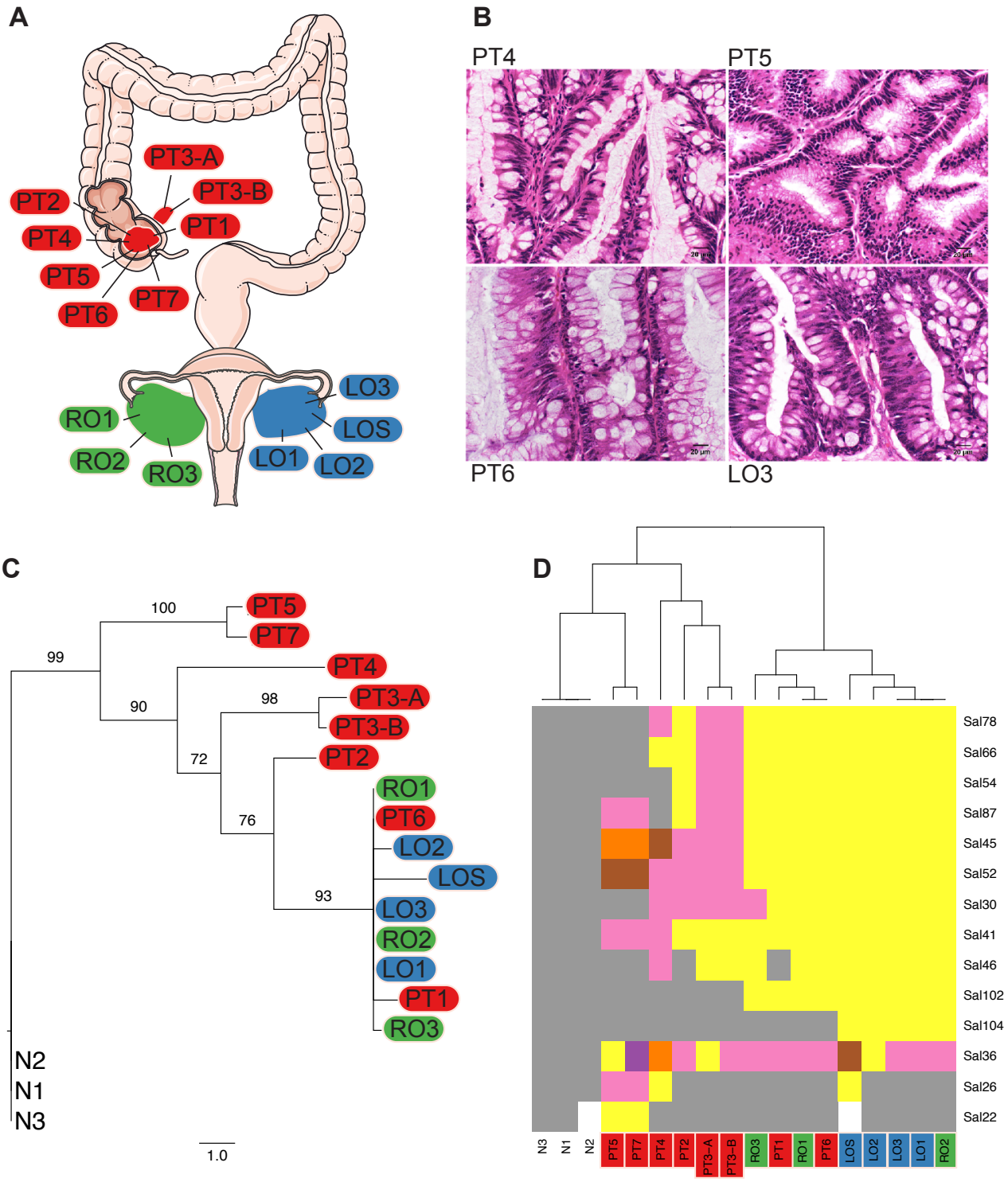

Figure 2.5 (Continued).

ileocecal valve, and cecum. We generated poly-G tract profiles for 3 normal tissue samples, 8 primary tumor samples, 3 right-ovary metastasis samples, and 4 left-ovary metastasis samples. (A detailed description of specimens based on the surgical pathology report and the full genotype data are provided in Supplementary Table S4.) 14 loci were mutated in at least one sample, and each sample contained at least 7 distinct mutations. (Figure 2.5C depicts patient C13's phylogenetic tree, and a complete heatmap of all mutations is shown in Figure 2.5D.) 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. 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 2 samples that were identical to each other, yet distinct from the cecal part of the tumor, and located on a separate branch of the tree. 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" (Wasserstrom et al., 2008b), i.e. 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 spatial heterogeneity within 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. This is consistent with a model of late metastasis, in which a genetically advanced clone (residing in PT1 or PT6) gives rise to all metastases in quick progression or in which one ovary metastasis gives rise to the other.

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-year-old female who received neoadjuvant FOLFOX chemotherapy and underwent surgery for a 3.2 cm microsatellite stable adenocarcinoma located at the hepatic flexure and a large 13 cm metastasis to the right ovary (Figure 2.6).

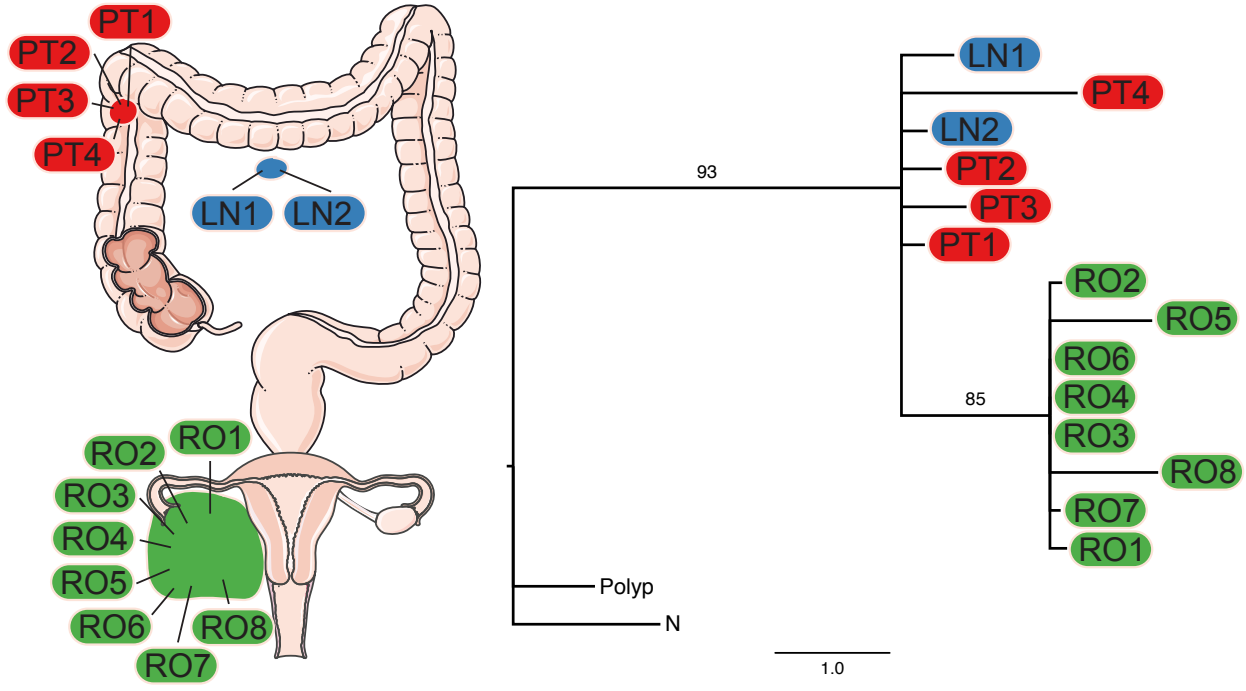

Figure 2.6 : Patient C31 with adenocarcinoma of the colon, lymph node and ovarian metastases. *PT – Primary tumor, RO – Right ovary Metastasis, LN – Lymph node metastasis N – Normal. Tumor sizes are drawn to scale. Neighbor-joining tree with bootstrap values, branches with bootstrap values below 70% collapsed into polytomies.*

The tumor had also metastasized to the mesenteric lymph nodes. We isolated four primary tumor samples, 8 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. (The full genotype data can be found in Supplementary Table S5; the mutation heatmap in Supplementary Figure S4). 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 (Figure 2.6). The metastasis had an approximately 40-fold larger volume, 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 substantial size). Given the relatively large number of mutations distinguishing the primary tumor and the ovarian metastasis, it is possible that they 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 that contained the metastatic clone. 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-year-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. 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 was microsatellite unstable, and the mutation rate was high with somatic alterations observed in 45% of interrogated loci. (Full genotype data provided in Supplementary Table S6; mutation heatmap in Supplementary Figure S6.) In contrast to patients C13 and C31, whose samples revealed substantial variation, all specimens from patient C27 had essentially the same poly-G tract profile, and the topology of the resulting phylogenetic tree was flat (Figure 2.7). 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 (Marusyk et al., 2012). 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.

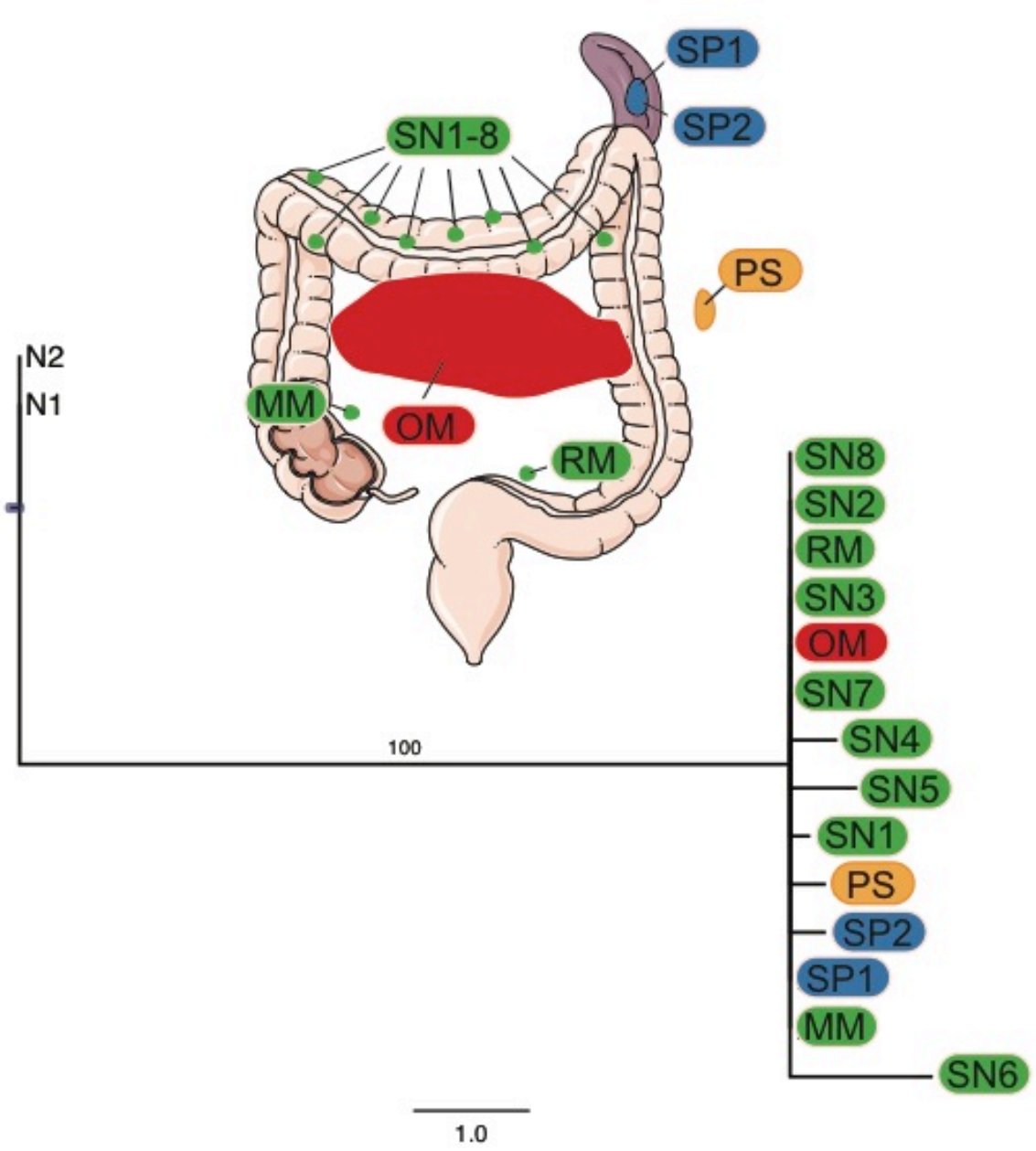

Figure 2.7: Patient C27 with mucinous adenocarcinoma of the colon. *SP – Spleen metastasis, SN – Serosal nodule, OM – Omentum, MM – Mesenteric margin, RM – Retroperitoneal margin, PS – Peritoneal side wall metastasis, N - Normal. Neighbor-joining tree with bootstrap values, branches with bootstrap values below 70% collapsed into polytomies.*

Polyguanine 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 polyguanine 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 (Supplementary Tables S7-S10). Our dataset is not comprehensive enough to determine average tumor mutation frequency in cancers other than colon, though 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 (Supplementary Table S9). Analysis of a breast cancer (patient B1, Figure 2.8A) 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, while tumor focus TF4 contained a distinct mutational profile and had segregated early on in its evolution. By contrast, patient O1's (Figure 2.8B) 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 one year later he underwent excision of a 6.5 cm lung metastasis. Poly-G tract profiling revealed

Figure 2.8: Patient B1 with multifocal breast cancer and patient O1 with malignant peripheral nerve sheath tumor. ***(A)*** *Patient B1. TF – Tumor focus, LN – Lymph node metastasis, N – Normal. 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.*

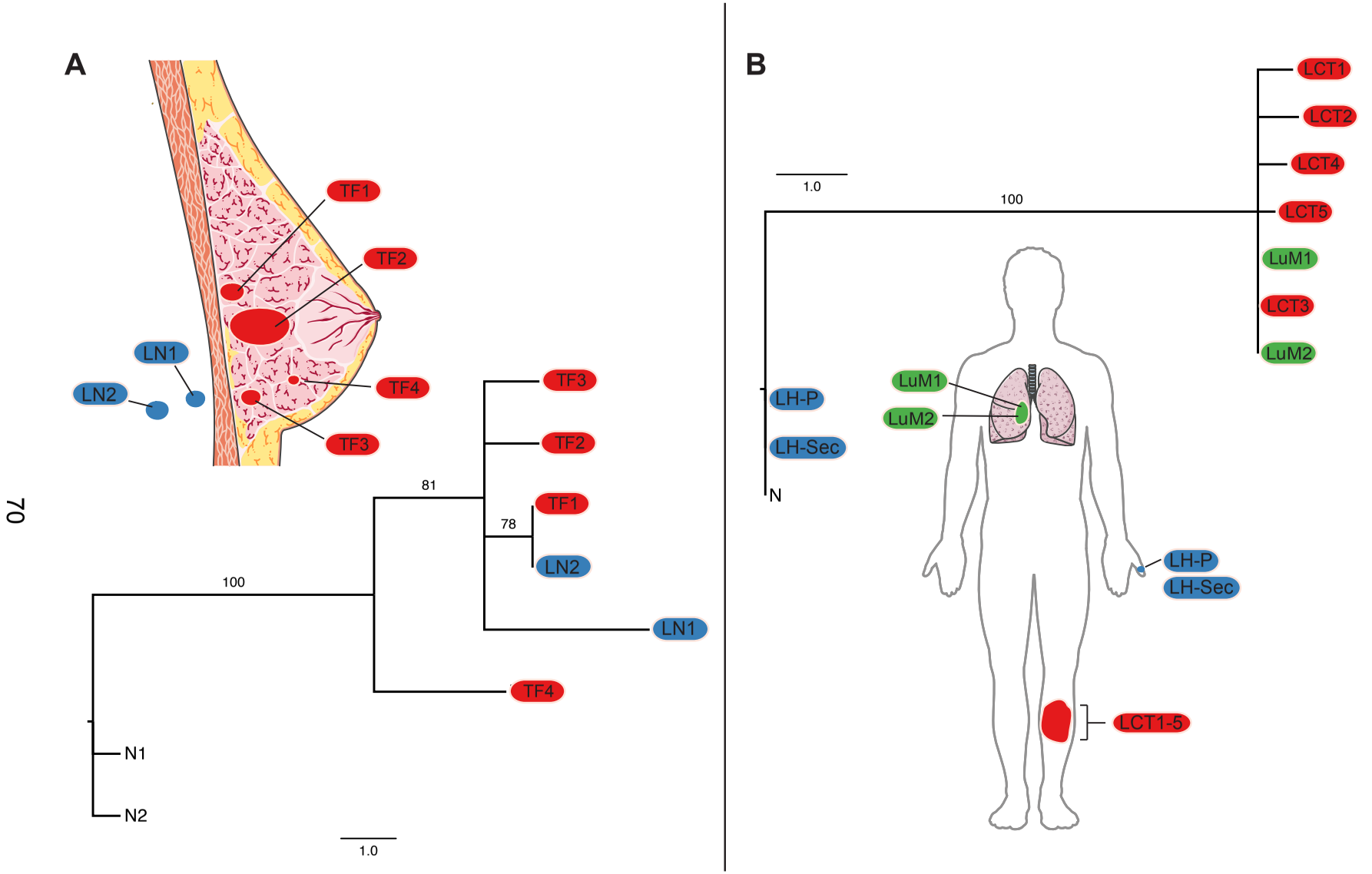

Figure 2.8 (Continued).

identical mutations in all 8 calf tumor and two lung metastasis samples, which suggests that the calf tumor was the source of the lung metastasis, while the tumor in his left hand showed no alterations and likely represented an independent transformation.

2.5 Discussion

We have shown that somatic mutations in non-coding 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 occur 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 (Jones et al., 2008) and pancreatic cancer (Yachida et al., 2010). Patient C13's cancer supports the late metastasis paradigm. Patient C31 could potentially represent a case of parallel progression because relatively few mutations are shared between the distant metastasis and the primary tumor, but with the caveat that sampling of the primary tumor might have missed the metastatic clone. 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, while a distant ovarian metastasis had a distinct mutational profile and contained many private alterations. Two

plausible explanations exist for this result. 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), while others shared the same genotype across all primary and metastatic tumor samples (O1), in one case despite an elevated mutation rate caused by microsatellite instability (C27). We did not find any obvious connection between administration of chemotherapy prior to surgery and intra-tumor heterogeneity. For example, both patients C31 (diversified) and C27 (homogeneous) received neoadjuvant chemotherapy with FOLFOX. While the “flat” clonal expansions (Siegmund et al., 2009) clearly represent younger entities than the diversified cancers (Shibata et al., 1996), 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 (Maley et al., 2006), which suggests that heterogeneity promotes malignancy, but the situation may be different in established cancers and/or differ by cancer type. In breast cancer, intra-tumor heterogeneity, as defined by cell surface marker expression, correlates with histopathological stage (Park et al., 2010), 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% (Salk et al., 2009). Subclonal diversity below this threshold should therefore be evaluated with complementary techniques such as fluorescence in situ hybridization (Snuderl et al., 2011) or deep sequencing (Ding et al., 2010).

Our data further show a positive correlation between age at diagnosis and mutation frequency. This result accords with growing evidence that a large proportion of mutations (more than 50% by some estimates (Tomasetti et al., 2013)) 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 (Welch et al., 2012). Since 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 (Lawrence et al., 2013).

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 determined by the mitotic history of the tumor founder cell, this observation suggests that less differentiated tumors derive from a rarely dividing cell. In the colon, two distinct progenitor populations have been identified, one displaying the characteristics of an actively dividing tissue stem cell that continuously replenishes the epithelial compartment, the other showing signs of quiescence (Li and Clevers, 2010). It may be that poorly differentiated tumors arise from the latter population. Future studies will also determine if somatic mutation load can predict colon cancer patient survival.

In summary, we have shown that a highly scalable 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 unique biological insights into the origin and progression of metastatic colon cancer. 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 in order to improve clinical decision processes. Since detecting mutated alleles in poly-G tracts does not require sequencing, patient privacy would be protected. In addition to its diagnostic potential, poly-G tract profiling can be used in its own right to study tumor evolution or to efficiently screen samples for deeper analysis

by next-generation sequencing. Successfully developing and applying targeted cancer therapies will depend on accurately understanding clonal architecture in human tumors. A neoplasm's mitotic history, as captured by neutral lineage markers, will provide an important backdrop on which to project the distribution of numerous therapeutically relevant mutations.

2.6 Experimental Procedures

Patient selection and tissue collection

This study was approved by the IRB of Massachusetts General Hospital, Boston, MA. 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. 18 patients were treatment naïve, three had received neoadjuvant chemotherapy, and one patient received neoadjuvant chemotherapy and radiation. Detailed patient information is provided in Supplementary Table 1. For each patient, we reviewed all available histology slides and FFPE tissue blocks and selected areas of homogeneous tumor for sampling. Tumors with 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 (Shibata, 1994). DNA was extracted with phenol-chloroform and precipitated with ethanol. We estimate that the average tissue sample had a volume of 3 mm^3 and contained $9x10^6$ cells.

Genotyping

A panel of primers flanking 35 poly-G tracts in the human genome was previously published (Salk et al., 2009). We used a randomly selected subset of primers from this panel (20 loci were sufficient to generate reliable phylogenies in our patients). Marker identification numbers are provided alongside full genotype data in all Supplementary Tables. Forward primers incorporated a fluorescent dye (HEX or 6-FAM) on their 5' end. Reverse primers contained a 5' GTTTCTT "pigtail" sequence (Brownstein et al., 1996). Since 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 1µM forward and reverse primers, 200µM of each dNTP, 2.5 units Taq Polymerase, 1x PCR buffer, and 1x 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. Microsatellite instability was tested using the Bethesda Markers as described in (Loukola et al., 2001). We did not distinguish between MSI-low and microsatellite stable 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 base pairs 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 (described below). 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. Supplementary Figure 5 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 (http://www.R-project.org). For each patient and marker, we calculated pairwise correlation coefficients among all stutter distributions and used these as input 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: 1) Normal samples had to cluster separately from mutated tumor samples, 2) replicates had to cluster within the same clade (allowing for some variation due to PCR failure), and 3) 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. Since we did not want to make assumptions about the likelihood of a particular allele distribution occurring, we treated mutations as unordered characters. This data set 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 (Frumkin et al., 2005) and employed neighbor-joining (Saitou and Nei, 1987) in R to infer the phylogenetic relationships between samples. We used bootstrapping with 1000 replicates to test the reliability of the resulting trees (Felsenstein, 1985) 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 Supplementary Figures 1-4. We used the software MrBayes (Huelsenbeck and Ronquist, 2001) with the same model parameters that were previously used for the analysis of poly-G tract mutation profiles (Salipante and Horwitz, 2006).

Other statistical analyses

Statistical analysis was performed in Prism (Graphpad). 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, 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).

References

Anderson, K., Lutz, C., van Delft, F.W., Bateman, C.M., Guo, Y., Colman, S.M., Kempski, H., Moorman, A.V., Titley, I., Swansbury, J., et al. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature *469*, 356–361.

Boyer, J.C., Yamada, N.A., Roques, C.N., Hatch, S.B., Riess, K., and Farber, R.A. (2002). Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum. Mol. Genet. *11*, 707–713.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. *62*, 1408–1415.

Brownstein, M.J., Carpten, J.D., and Smith, J.R. (1996). Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. Biotechniques *20*, 1004–1006, 1008–1010.

Carlson, C.A., Kas, A., Kirkwood, R., Hays, L.E., Preston, B.D., Salipante, S.J., and Horwitz, M.S. (2012). Decoding cell lineage from acquired mutations using arbitrary deep sequencing. Nat. Methods *9*, 78–80.

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999–1005.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution *39*, 783–791.

Fidler, I.J. (2003). Timeline: The pathogenesis of cancer metastasis: the “seed and soil” hypothesis revisited. Nat. Rev. Cancer *3*, 453–458.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G., and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. Cancer Research *68*, 5924–5931.

Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U., and Shapiro, E. (2005). Genomic variability within an organism exposes its cell lineage tree. PLoS Computational Biology *1*, e50.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754–755.

Jones, S., Chen, W.D., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M.A., Siegel, C., Velculescu, V.E., et al. (2008). Comparative lesion sequencing provides insights into tumor evolution. Proceedings of the National Academy of Sciences *105*, 4283–4288.

Klein, C.A. (2008). Cancer. The metastasis cascade. Science *321*, 1785–1787.

Klein, C.A. (2009). Parallel progression of primary tumours and metastases. Nat. Rev. Cancer *9*, 302–312.

Laiho, P., Launonen, V., Lahermo, P., Esteller, M., Guo, M., Herman, J.G., Mecklin, J.P., Jarvinen, H., Sistonen, P., Kim, K.M., et al. (2002). Low-level microsatellite instability in most colorectal carcinomas. Cancer Research *62*, 1166–1170.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Li, L., and Clevers, H. (2010). Coexistence of quiescent and active adult stem cells in mammals. Science *327*, 542–545.

Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C.M., Eisenberger, M.A., Carducci, M.A., et al. (2009). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat. Med. *15*, 559–565.

Loukola, A., Eklin, K., Laiho, P., Salovaara, R., Kristo, P., Jarvinen, H., Mecklin, J.P., Launonen, V., and Aaltonen, L.A. (2001). Microsatellite marker analysis in screening for hereditary nonpolyposis colorectal cancer (HNPCC). Cancer Research *61*, 4545–4549.

Maley, C.C., Galipeau, P.C., Finley, J.C., Wongsurawat, V.J., Li, X., Sanchez, C.A., Paulson, T.G., Blount, P.L., Risques, R.A., Rabinovitch, P.S., et al. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet *38*, 468–473.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? Nat. Rev. Cancer *12*, 323–334.

Merlo, L.M.F., Pepper, J.W., Reid, B.J., and Maley, C.C. (2006). Cancer as an evolutionary and ecological process. Nat. Rev. Cancer *6*, 924–935.

Park, S.Y., Lee, H.E., Li, H., Shipitsin, M., Gelman, R., and Polyak, K. (2010). Heterogeneity for Stem Cell-Related Markers According to Tumor Subtype and Histologic Stage in Breast Cancer. Clinical Cancer Research *16*, 876–887.

Potten, C.S., Kellett, M., Roberts, S.A., Rew, D.A., and Wilson, G.D. (1992). Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. Gut *33*, 71–78.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Salipante, S.J., and Horwitz, M.S. (2006). Phylogenetic fate mapping. Proceedings of the National Academy of Sciences of the United States of America *103*, 5448–5453.

Salipante, S.J., Kas, A., McMonagle, E., and Horwitz, M.S. (2010). Phylogenetic analysis of developmental and postnatal mouse cell lineages. Evol Dev *12*, 84–94.

Salipante, S.J., Thompson, J.M., and Horwitz, M.S. (2008). Phylogenetic fate mapping: theoretical and experimental studies applied to the development of mouse fibroblasts. Genetics *178*, 967–977.

Salk, J.J., Salipante, S.J., Risques, R.A., Crispin, D.A., Li, L., Bronner, M.P., Brentnall, T.A., Rabinovitch, P.S., Horwitz, M.S., and Loeb, L.A. (2009). Clonal expansions in ulcerative colitis identify patients with neoplasia. Proceedings of the National Academy of Sciences *106*, 20871–20876.

Salk, J.J., and Horwitz, M.S. (2010). Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. Semin. Cancer Biol. *20*, 294–303.

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*, 561–563.

Shibata, D., Navidi, W., Salovaara, R., Li, Z.H., and Aaltonen, L.A. (1996). Somatic microsatellite mutations as molecular tumor clocks. Nat. Med. *2*, 676–681.

Shibata, D. (2012). Cancer. Heterogeneity and tumor history. Science *336*, 304–305.

Siegmund, K.D., Marjoram, P., Tavare, S., and Shibata, D. (2009). Many colorectal cancers are “flat” clonal expansions. Cell Cycle *8*, 2187–2193.

Snuderl, M., Fazlollahi, L., Le, L.P., Nitta, M., Zhelyazkova, B.H., Davidson, C.J., Akhavanfard, S., Cahill, D.P., Aldape, K.D., Betensky, R.A., et al. (2011). Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. Cancer Cell *20*, 810–817.

Tomasetti, C., Vogelstein, B., and Parmigiani, G. (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proceedings of the National Academy of Sciences *110*, 1999–2004.

Tsao, J.L., Yatabe, Y., Salovaara, R., Jarvinen, H.J., Mecklin, J.P., Aaltonen, L.A., Tavare, S., and Shibata, D. (2000). Genetic reconstruction of individual colorectal tumor histories. Proceedings of the National Academy of Sciences of the United States of America *97*, 1236–1241.

Tsao, J.L., Zhang, J., Salovaara, R., Li, Z.H., Jarvinen, H.J., Mecklin, J.P., Aaltonen, L.A., and Shibata, D. (1998). Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture. The American Journal of Pathology *153*, 1189–1200.

Viguera, E., Canceill, D., and Ehrlich, S.D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. Embo J *20*, 2587–2595.

Wasserstrom, A., Adar, R., Shefer, G., Frumkin, D., Itzkovitz, S., Stern, T., Shur, I., Zangi, L., Kaplan, S., Harmelin, A., et al. (2008a). Reconstruction of cell lineage trees in mice. PloS One *3*, e1939.

Wasserstrom, A., Frumkin, D., Adar, R., Itzkovitz, S., Stern, T., Kaplan, S., Shefer, G., Shur, I., Zangi, L., Reizel, Y., et al. (2008b). Estimating cell depth from somatic mutations. PLoS Computational Biology *4*, e1000058.

Weber, J.L., and Wong, C. (1993). Mutation of human short tandem repeats. Hum. Mol. Genet. *2*, 1123–1128.

Weinberg, R.A. (2007). The Biology of Cancer (Garland Science).

Weinberg, R.A. (2008). Mechanisms of malignant progression. Carcinogenesis *29*, 1092–1095.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Zhou, W., Tan, Y., Anderson, D.J., Crist, E.M., Ruohola-Baker, H., Salipante, S.J., and Horwitz, M.S. (2013). Use of somatic mutations to quantify random contributions to mouse development. BMC Genomics *14*, 39.

Chapter 3: Discussion and future directions

3.1 Summary

The first chapter of this dissertation lays out a number of important questions regarding metastatic progression in humans. The development of effective treatment strategies for both localized and advanced disease requires precise knowledge of when metastatic precursors disseminate, by which routes they travel, and what their genetic relationship to the primary tumor is. It is likely that different cancer types metastasize along divergent paths and timelines. Reconstructing the “tree of life of metastatic cancer” in representative patient populations will help us formulate more precise disease models, which will in turn guide the design of improved treatment schemes.

In the second chapter, I show how a simple PCR assay can be used to determine the lineage relationships between primary tumors and metastases in human colorectal cancer. The proposed methodology relies on small insertions and deletions in hypermutable polyguanine (poly-G) tracts that are introduced into the genome at high frequency in a replication-dependent manner. Poly-G variants that are distinct from the germline are present in 91% of colorectal cancers. A positive correlation of the tumor mutational load with age at diagnosis indicates that at least a portion of them are acquired during proliferation of normal colonic stem cells during the patient’s lifetime. Poorly differentiated tumors have fewer mutations than well-differentiated tumors,

possibly pointing to a shorter mitotic history of the tumor founder cell in high grade cancers.

More poly-G mutations are acquired during clonal evolution. These variants can be used for lineage tracing. In four patients, I use poly-G fingerprints of various parts of the primary tumor and its metastases to construct phylogenetic trees that reflect evolutionary relationships among tumor cell populations. The phylogenies show that metastasis occurs relatively late in colon cancer progression. The level of regional genetic heterogeneity varies substantially among patients, implying that some tumors are created by rapid clonal expansions, while others undergo spatially discrete speciation events. Finally, I demonstrate the presence of poly-G variants in a number of other human cancer types. Poly-G tract profiling therefore has applicability beyond colorectal cancer.

3.2 Methodological perspective

Analysis of poly-G tracts for phylogenetic reconstruction of human cancer histories has several advantages over existing methods. First, the technique is highly scalable, making it suitable for analysis of large patient cohorts. Dozens of samples can be processed in parallel at low cost, using relatively common equipment: a PCR machine and a capillary sequencer. The sample preparation procedure – tissue digestion followed by DNA extraction with phenol-chloroform and precipitation – can be done in bulk. Pooling of PCR products labeled with three or more different fluorophores enables work-efficient fragment analysis by capillary electrophoresis. Multiplexing primers during amplification can further optimize the workflow, but this possibility has not been implement in our studies so far, partly because extensive reaction optimization is

required. Establishing a reliable multiplex protocol is one of our future goals.

Another advantage is that DNA from formalin fixed paraffin embedded (FFPE) tissues is suitable for poly-G profiling. FFPE-derived DNA is fragmented and contains an abundance of artifactual single-nucleotide changes (Do et al., 2013). Obtaining good quality sequence data from FFPE tissue requires specialized approaches (Wagle et al., 2012). Since the insertions and deletions that are characteristic of replication slippage in microsatellites are not typical artifacts of fixation, poly-G tract profiling is a reliable methodology for the analysis of FFPE-derived DNA. The problem of nucleic acid fragmentation, on the other hand, remains. Generation of reproducible poly-G tract genotypes therefore requires high template concentration (~ 90 ng) in each reaction.

Finally, poly-G tract analysis does not produce any identifiable information that could be used to breach patient privacy. Most institutional guidelines prohibit whole genome or even exome sequencing of archival tissues without explicit consent. Hence, the large numbers of valuable cancer samples stored in pathology departments cannot be used for this purpose because patients would have to be contacted retrospectively, which is time-consuming for the investigator and potentially psychologically demanding for the patient. Poly-G tract profiling can be performed under most discarded tissue protocols, thereby opening up tissue resources for analysis of intra-tumor heterogeneity that would remain untapped otherwise.

However, poly-G tract profiling also suffers from a number of significant limitations. First and foremost, fragment analysis is an analog technology with very low resolution. Alleles that are present at a frequency below 40% in a sample cannot be detected (Salk et al., 2009). Biologically relevant variation within a sample (local

intermingling of different clones) remains invisible below this threshold, but we know that many subclonal mutations are common in tumors and may play an important role in therapy resistance and relapse (Ding et al., 2010; Landau et al., 2013; Nik-Zainal et al., 2012). Genetic analyses of di- and trinucleotide microsatellites have circumnavigated this limitation by performing digital PCR on extensively diluted samples, essentially amplifying single alleles (Tsao et al., 1998). For poly-G tract profiling, we have not been able to establish a similar approach to date. The mutation rates in poly-G tracts are exceedingly high, and the noise introduced during PCR appears to preclude reliable amplification of single molecules.

An additional problem of fragment analysis is that in some cases mutations cannot be distinguished from loss of heterozygosity (LOH) (illustrated in Figure 3.1). Poly-G tract mutations most often occur in the form of 1 base pair (bp) insertions or deletions. If a clearly heterozygous allele (e.g. displaying two peaks at 120 and 130 bp) in normal tissue is reduced to a single peak at 120 bp in a tumor, it is reasonable to assume LOH. However, if an allele is heterozygous with two peaks at 120 and 121 base pairs, the probability of a mutation vs. LOH is unknown. For the purposes of all analyses in this dissertation, a change was scored as a poly-G tract mutation unless it represented an unambiguous case of LOH. Fortunately many heterozygous alleles are easily distinguishable from each other, so it is unlikely that this limitation generated significant biases in the data.

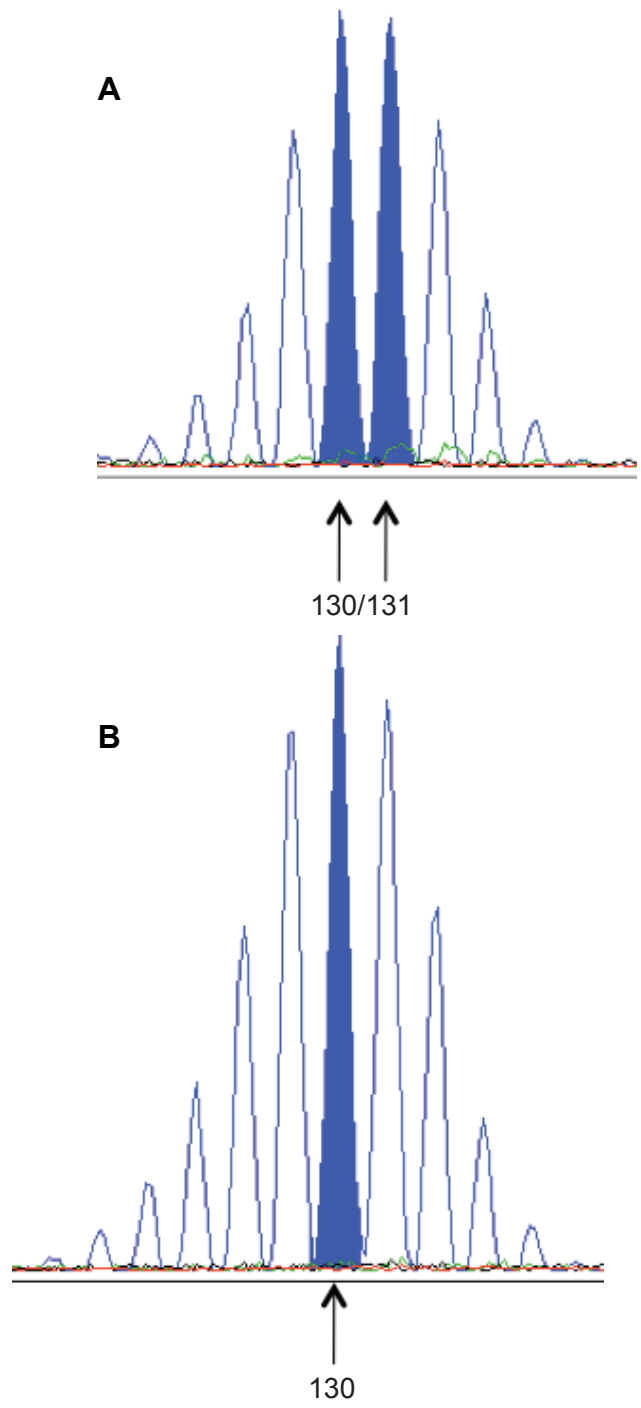

Figure 3.1: Ambiguity between one base pair deletion and LOH. (A) *Normal tissue control and* ***(B)*** *tumor genotype for marker Sal104. The distribution shift could be caused a one base pair contraction of the poly-G tract, or represent loss of the upper allele.*

Finally, while poly-G tracts, due to their neutrality and high mutability, are of great interest for somatic lineage tracing, they do not lend themselves to parallelized measurement by current sequencing technologies. Ideally, all ~4000 intergenic poly-G tracts (see Table 1.3) would be captured and sequenced simultaneously in a sample of interest to yield a high-resolution fingerprint of somatic lineage. However, the typical PCR stutter that accompanies amplification of homopolymers prevents the generation of truly clonal clusters that are the basis of high quality sequence reads. It would be worth exploring whether parallelized interrogation of less mutable microsatellites (e.g. dinucleotide repeats) could be used for this purpose instead. Alternatively, development of single molecule sequencing technologies that can reliably interrogate homopolymers (perhaps Nanopore sequencing (Hayden, 2012)) might enable genome-wide evaluation of somatic poly-G variants in the future.

3.3 Biological perspective: insights and ongoing follow-up studies

In our study of poly-G mutation patterns in colorectal cancer, we made multiple intriguing observations that serve as sources of novel hypotheses.

A large majority (91%) of patients in our cohort had at least one poly-G mutation, but the percentage of mutant genotypes (excluding cases with microsatellite instability) varied from 4.76% to 55%. To define the factors driving this large variation is one of our ongoing efforts. The total mutational burden of a sample consists of two components (summarized in Figure 3.2): mutations that were already present in the tumor founder cell at the time of transformation (fully clonal “founder mutations”, inherited by all cells)

Figure 3.2: Founder and progressor mutations. *Mutations found in a tumor can be subdivided into those that were acquired during embryogenesis and normal tissue homeostasis (founder mutations) and those that accumulated during tumor evolution (progressor mutations). Under the simplifying assumption of similar mutation rates across tumors, the total number of variants is a function of the mitotic age of the founder cell (top panel) and the number of divisions that occurred before the tumor reached a given mass (bottom panel). Cells with different genetic alterations are indicated in different colors. The larger the number of divisions that occurred during tumor progression, the larger the potential for genetic heterogeneity is.*

Figure 3.2 (Continued).

and mutations that emerged during tumor evolution ("progressor mutations" (Yachida et al., 2010): these can be subclonal, or clonal if completely sweeping expansions took place during tumor development (Greaves and Maley, 2012)). The number of divisions the tumor founder cell underwent before it became neoplastic mainly determines the former category. As discussed in chapter 2, a large portion of alterations in a tumor are probably founder mutations (Tomasetti et al., 2013; Welch et al., 2012). Age and patient-specific epithelial proliferation characteristics – inflammation and the associated cell turnover, for example – likely play a role in determining how large the contribution of founder mutations to the total burden is. Subclonal progressor mutations are relevant for lineage tracing purposes, and their number is intimately associated with the growth dynamics of a tumor. If a cancer grows aggressively, i.e. proliferation levels far exceed apoptosis levels, relatively few cell divisions are required to generate a mass of a certain size and intra-tumor diversification will be limited (Figure 3.2, lower right panel). If proliferation and apoptosis are more balanced, on the other hand, more cell divisions will be required for the development of a detectable tumor, and more pronounced regional speciation might take place (Figure 3.2, lower left panel).

At least two exceptions to this logic are worth noting here. First, "regional speciation", which has been documented in multiple cancer types (Gerlinger et al., 2012; Yachida et al., 2010), is contingent on limited mobility of emergent clones. (This possibility is not depicted in Figure 3.2.) If cells within the tumor travel and intermingle extensively, new mutations will be diluted and thereby obfuscated. A lack of detectable progressor mutations can therefore be caused by rapid clonal expansion or increased cell mobility. Second, fully clonal progressor mutations cannot be distinguished from

founder mutations.

Founder and progressor mutations are products of two fundamentally different biological processes: normal cell renewal during tissue homeostasis vs. tumor growth. Despite the caveats noted above, it will be of high interest to separate these two mutation types as cleanly as possible, determine whether their proportions differ among tumors and assess whether their relative characteristics are correlated with clinical variables.

We are currently exploring the provenance of both founder and progressor mutations in more detail. In collaboration with the Department of Pathology at Massachusetts General Hospital, we have begun analyzing a cohort of 75 colon carcinomas that were resected at stage T3N0M0 (tumor invades through muscularis propria into pericolorectal tissues, no lymph node or distant metastasis). Detailed survival information is available for all patients. For our analysis, we are relying on the following simplified assumption: founder mutations are those that are present in all tumor regions, progressor mutations are only found in some locations. We plan to sample each tumor in multiple distinct places, generate poly-G tract genotypes, and calculate two distinct measures: 1) the “founder mutation score” which corresponds to the number of alterations that are found in all samples and therefore are clonal at our level of resolution and 2) a “heterogeneity score” corresponding to subclonal mutations that reflects the degree of regional diversification that has taken place.

The working hypotheses related to these measures of interest are: 1) Low founder mutation frequency characterizes aggressive cancers and is correlated with shorter survival. 2) Homogeneous tumors with low heterogeneity scores are more lethal

than heterogeneous tumors and associated with shorter survival. They arise from the rapid expansion an inherently aggressive clone and do not have time to diversify.

Both these hypotheses are driven by our observation that a low overall mutation frequency (consisting of founder mutations + progressor mutations) is correlated with high histological grade and therefore indicative of a more malignant phenotype (Figure 2.3B). However, in our single sample analysis we could not distinguish between founder and progressor mutations. Consequently, it is possible that the inverse correlation of grade and mutation frequency is caused *either* by founder *or* by progressor mutations *or* by both. Given the unambiguous association of overall mutations frequency with patient age, we favor founder mutations as the driving factor (and this interpretation is mainly embraced in chapter 2), but a more thorough analysis of multiple tumor regions in the manner proposed above will shed more definitive light on this question.

One caveat is that our approach is likely to overestimate the contribution of founder mutations. Some mutations will be categorized as clonal even though finer sampling might reveal that they are not present everywhere. Also, our technique cannot precisely estimate mutant allele frequencies; some mutations will be classified as clonal even though they are not present in all cancer cells, but only in a large fraction. Ideally, the experiment would be accompanied by deep sequencing of individual samples to capture intra-sample heterogeneity and quantify mutant allele frequencies more precisely, but this is likely impossible due to the patient privacy protections described above.

Independently, to further characterize founder mutations in the colon, we would like to assess how many poly-G variants are normally present in colonic stem cells in

subjects of varying ages. Understanding the “baseline mutational load” of normal cells will give us a more precise idea of the average percentage contribution of founder mutations to the total mutational burden of a colon cancer. To this end, we are collaborating with a group in the Netherlands that specializes in culturing organoids derived from single human colonic crypts (Jung et al., 2011). These organoids are close progeny of a single colonic stem cell. We will genotype organoids from young children and elderly subjects and determine i) how their mutational burden compares to the number of alterations observed in colon cancer and ii) whether the positive correlation between mutational load and age is reproduced in normal tissues.

3.4 Future directions

A plethora of clinically relevant questions related to the work described here await resolution. Many could be addressed effectively with a combination of poly-G tract profiling and deep sequencing.

For example, a systematic analysis of how the degree of intra-tumor heterogeneity changes at different progression stages would be very instructive. Very few studies have been conducted in this area. High levels of heterogeneity are connected with increased malignant potential in the early stages of tumorigenesis. For example, clonal diversity in Barrett’s esophagus, a premalignant condition, is linked with a higher likelihood of progression to cancer (Maley et al., 2006). The final step of systemic disease advancement, metastasis, on the other hand, appears to go hand in hand with a steep drop in heterogeneity. This makes intuitive sense, as dissemination represents a natural evolutionary bottleneck, and it is evident on multiple levels of observation: in the narrowing of the mutant allele frequency distribution in metastasis,

observed in breast carcinoma (Ding et al., 2012), apparent monoclonality of multiple lesions at the time of death in prostate cancer patients (even though localized prostate cancer often is multifocal and highly heterogeneous) (Liu et al., 2009), increased homogeneity of genomic rearrangements in DTCs derived from patients with overt metastasis vs. those with MRD (Klein et al., 2002; Weckermann et al., 2009), and our results presented in Chapter 2, in particular, the homogeneity of metastases in comparison with the primary tumor in patient C13. Parallel comparisons of cancer at different stages of advancement, conducted with standardized techniques, will further elucidate whether dynamic changes in diversity accompany progression.

Also, it is unclear how the drop in heterogeneity suspected to occur in metastatic disease relates to treatment response. Darryl Shibata, a pioneering scientist in the field of intra-tumor heterogeneity, wrote in a recent perspective: “Targeted therapies should be directed at public driver mutations present in all cancer cells, but the probability of drug-resistant variants increases with cell division. Therefore, a more heterogeneous tumor with many private mutations is more likely to fail chemotherapy” (Shibata, 2012). Metastases (at least those arise synchronously or almost synchronously) are younger clonal expansions than primary tumors, as such they should be more responsive to therapy. This appears to be true for some cancer types. Clinical studies have shown that primary lung carcinomas, for example, are much less likely to respond to chemotherapy than corresponding metastases (11.8% vs. 32.8% response rate). For breast cancer, however, this pattern is reversed (40% vs. 19.8% response rate for primary tumors vs. metastases) (Slack and Bross, 1975). It is known that microenvironmental cues at different sites contribute to differential therapy response

(Kodack et al., 2012), but more rigorous investigation of how intra-tumor heterogeneity relates to these treatment outcomes will also be important in the future.

Another interesting and unanswered question is how spatial diversification – the presence of distinct clones in different regions of a tumor – relates to strictly local heterogeneity (intermingling of clones), and how these forms of diversity associate with clinical behavior. More invasive and motile cells might generate more “dispersed” forms of genetic heterogeneity. It has furthermore been suggested that locally co-existing clones might be able to develop commensal relationships (Merlo et al., 2006) and cooperate to achieve greater fitness. One speculation could therefore be that high levels of local heterogeneity are a hallmark of aggressive disease, while regional clonal expansions that stay delineated from each other might signify more indolent phenotypes. Currently, no data exist to corroborate or refute this hypothesis.

The answers to these and many other crucial questions in the field of intra-tumor heterogeneity are now within our reach. This is in large part due to the development of revolutionary next generation sequencing technologies and other important technical advances. Poly-G tract profiling is a small, but potentially useful addition to the growing repertoire of techniques that can collectively help us shine a bright light on the cells that are responsible for the deaths of far too many cancer patients.

References

Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature *464*, 999–1005.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Do, H., Wong, S.Q., Li, J., and Dobrovic, A. (2013). Reducing Sequence Artifacts in Amplicon-Based Massively Parallel Sequencing of Formalin-Fixed Paraffin-Embedded DNA by Enzymatic Depletion of Uracil-Containing Templates. Clinical Chemistry *59*, 1376–1383.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Hayden, E.C. (2012). Nanopore genome sequencer makes its debut. Nature News.

Jung, P., Sato, T., Merlos-Suárez, A., Barriga, F.M., Iglesias, M., Rossell, D., Auer, H., Gallardo, M., Blasco, M.A., Sancho, E., et al. (2011). Isolation and in vitro expansion of human colonic stem cells. Nat. Med. *17*, 1225–1227.

Klein, C.A., Blankenstein, T.J., Schmidt-Kittler, O., Petronio, M., Polzer, B., Stoecklein, N.H., and Riethmuller, G. (2002). Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. Lancet *360*, 683–689.

Kodack, D.P., Chung, E., Yamashita, H., Incio, J., Duyverman, A.M.M.J., Song, Y., Farrar, C.T., Huang, Y., Ager, E., Kamoun, W., et al. (2012). Combined targeting of HER2 and VEGFR2 for effective treatment of HER2-amplified breast cancer brain metastases. Proceedings of the National Academy of Sciences *109*, E3119–E3127.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714–726.

Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C.M., Eisenberger, M.A., Carducci, M.A., et al. (2009). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat. Med. *15*, 559–565.

Maley, C.C., Galipeau, P.C., Finley, J.C., Wongsurawat, V.J., Li, X., Sanchez, C.A., Paulson, T.G., Blount, P.L., Risques, R.A., Rabinovitch, P.S., et al. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet *38*,

468–473.

Merlo, L.M.F., Pepper, J.W., Reid, B.J., and Maley, C.C. (2006). Cancer as an evolutionary and ecological process. Nat. Rev. Cancer *6*, 924–935.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The Life History of 21 Breast Cancers. Cell 1–14.

Salk, J.J., Salipante, S.J., Risques, R.A., Crispin, D.A., Li, L., Bronner, M.P., Brentnall, T.A., Rabinovitch, P.S., Horwitz, M.S., and Loeb, L.A. (2009). Clonal expansions in ulcerative colitis identify patients with neoplasia. Proceedings of the National Academy of Sciences *106*, 20871–20876.

Shibata, D. (2012). Cancer. Heterogeneity and tumor history. Science *336*, 304–305.

Slack, N.H., and Bross, I.D. (1975). The influence of site of metastasis on tumour growth and response to chemotherapy. British Journal of Cancer *32*, 78–86.

Tomasetti, C., Vogelstein, B., and Parmigiani, G. (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proceedings of the National Academy of Sciences *110*, 1999–2004.

Tsao, J.L., Zhang, J., Salovaara, R., Li, Z.H., Jarvinen, H.J., Mecklin, J.P., Aaltonen, L.A., and Shibata, D. (1998). Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture. The American Journal of Pathology *153*, 1189–1200.

Wagle, N., Berger, M.F., Davis, M.J., Blumenstiel, B., DeFelice, M., Pochanard, P., Ducar, M., Van Hummelen, P., MacConaill, L.E., Hahn, W.C., et al. (2012). High-Throughput Detection of Actionable Genomic Alterations in Clinical Tumor Samples by Targeted, Massively Parallel Sequencing. Cancer Discovery.

Weckermann, D., Polzer, B., Ragg, T., Blana, A., Schlimok, G., Arnholdt, H., Bertz, S., Harzmann, R., and Klein, C.A. (2009). Perioperative activation of disseminated tumor cells in bone marrow of patients with prostate cancer. J. Clin. Oncol. *27*, 1549–1556.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Appendix A – Supplementary tables and figures

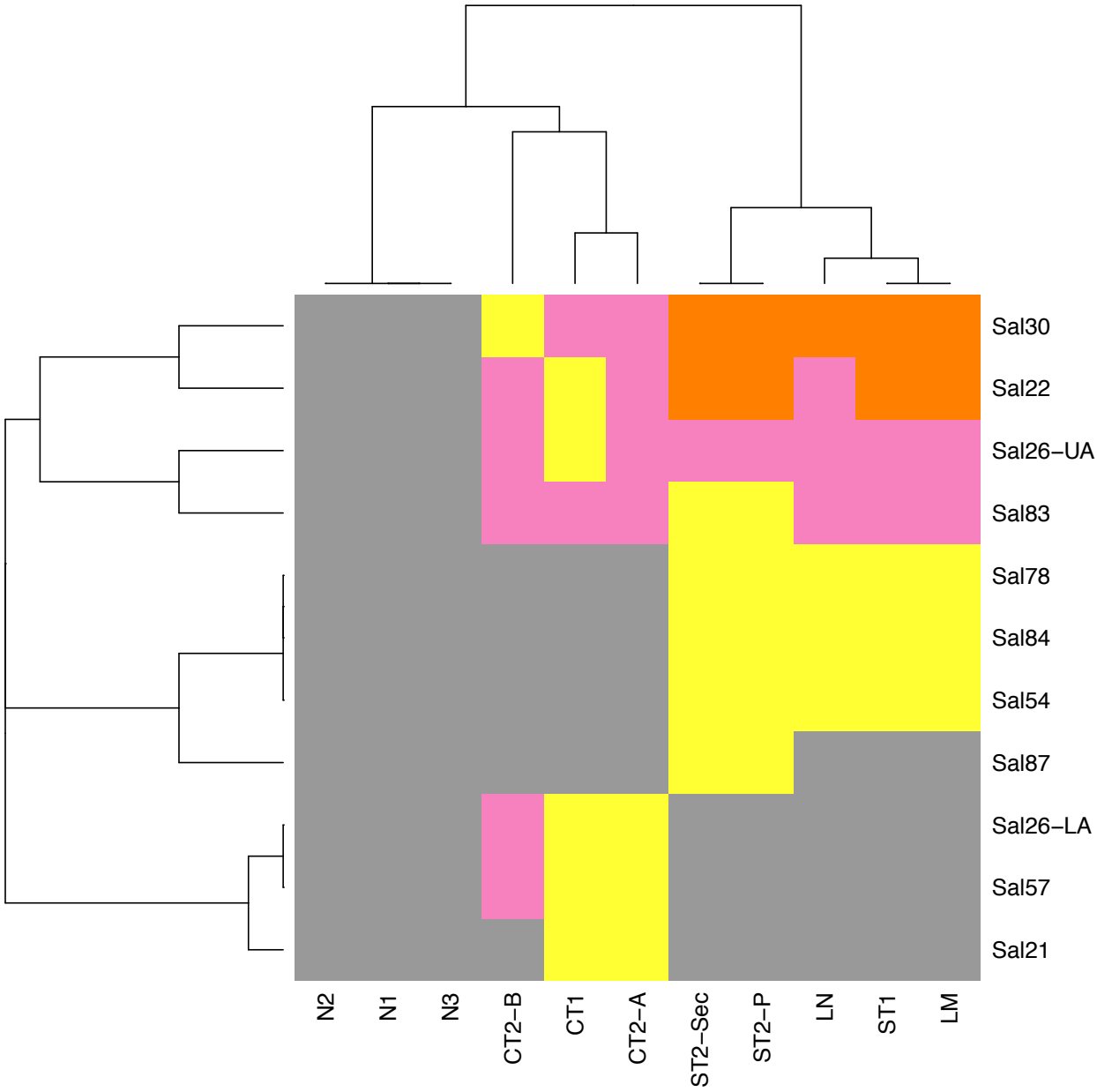

Figure S1: Mutation heatmap patient C39. *Grey squares signify allele distributions that are indistinguishable from normal control. Colored squares indicate a shift in allele distribution, i.e. a poly-G mutation. If multiple different allele distributions exist per marker, they are indicated with additional colors.*

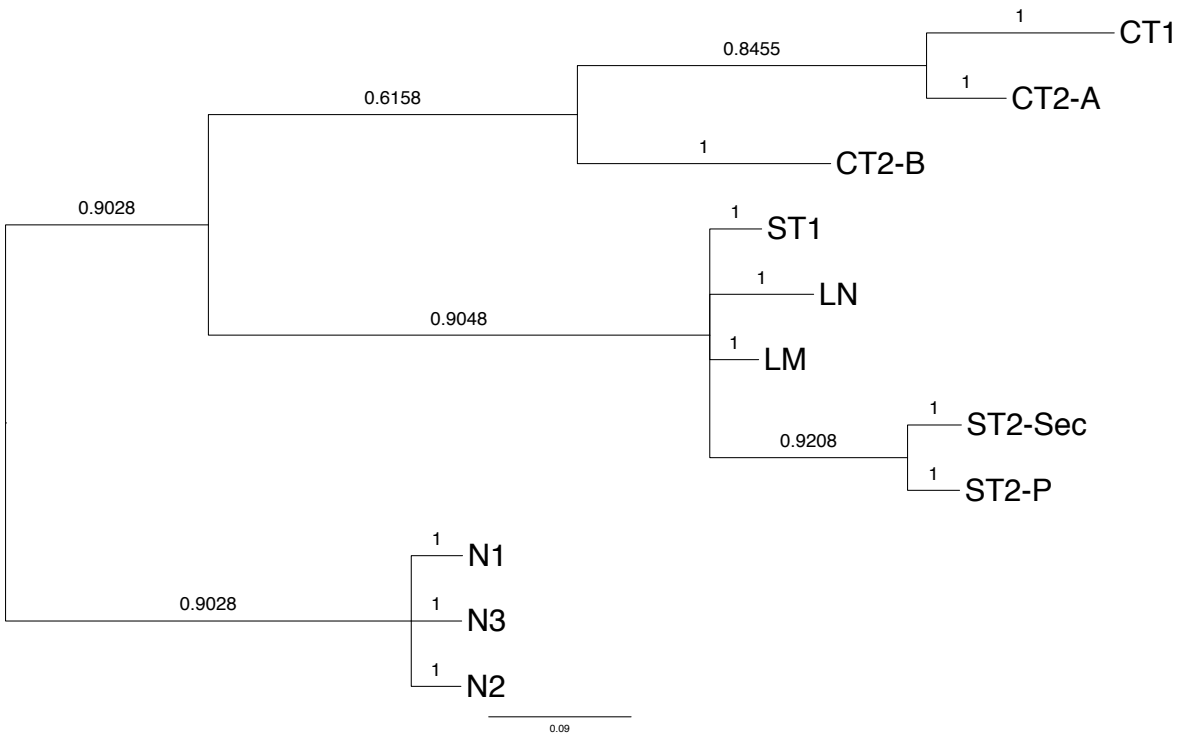

Figure S2: Bayesian phylogeny patient C39. *Posterior probability values for each clade are given as branch labels.*

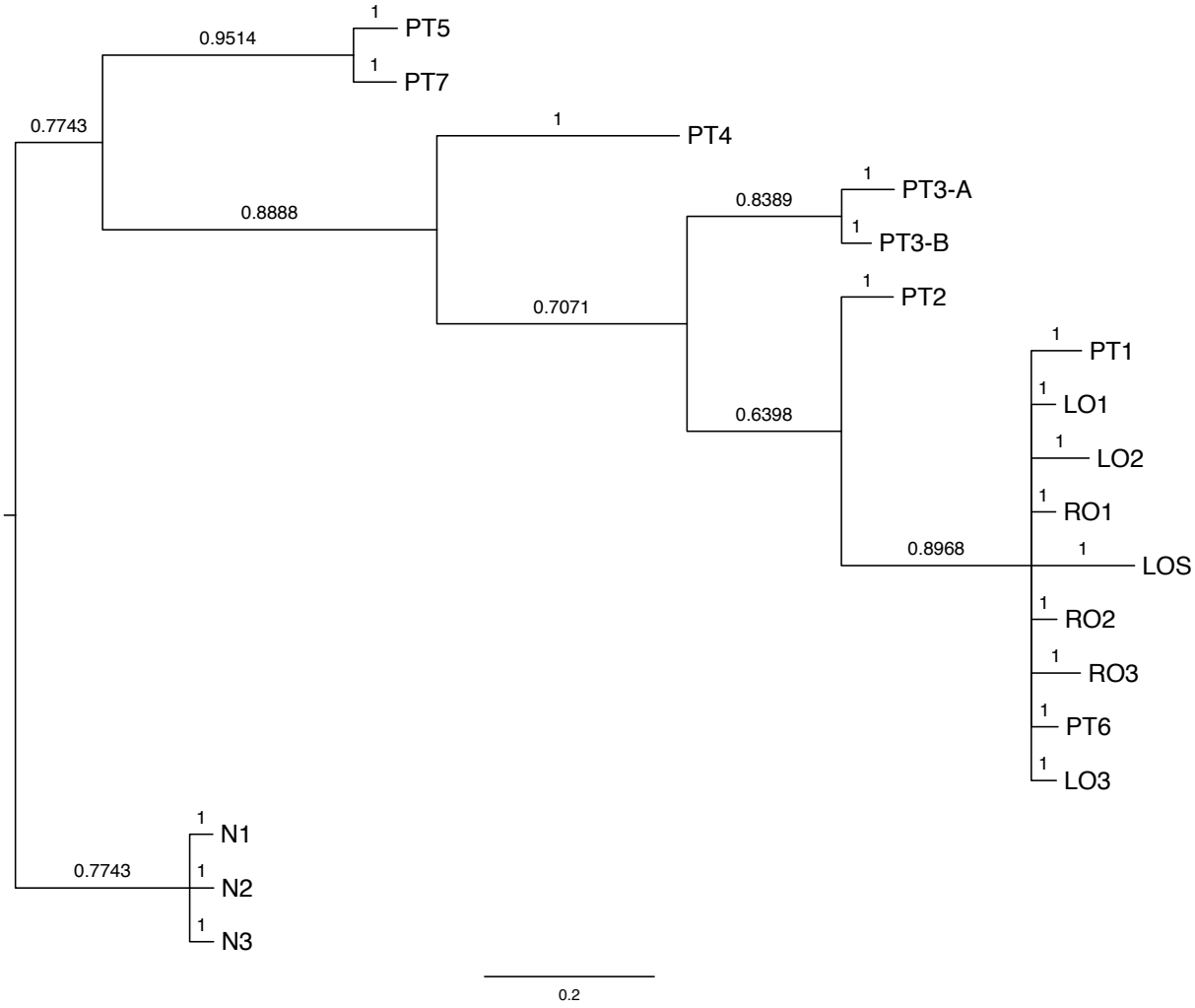

Figure S3: Bayesian phylogeny patient C13. *Posterior probability values for each clade are given as branch labels.*

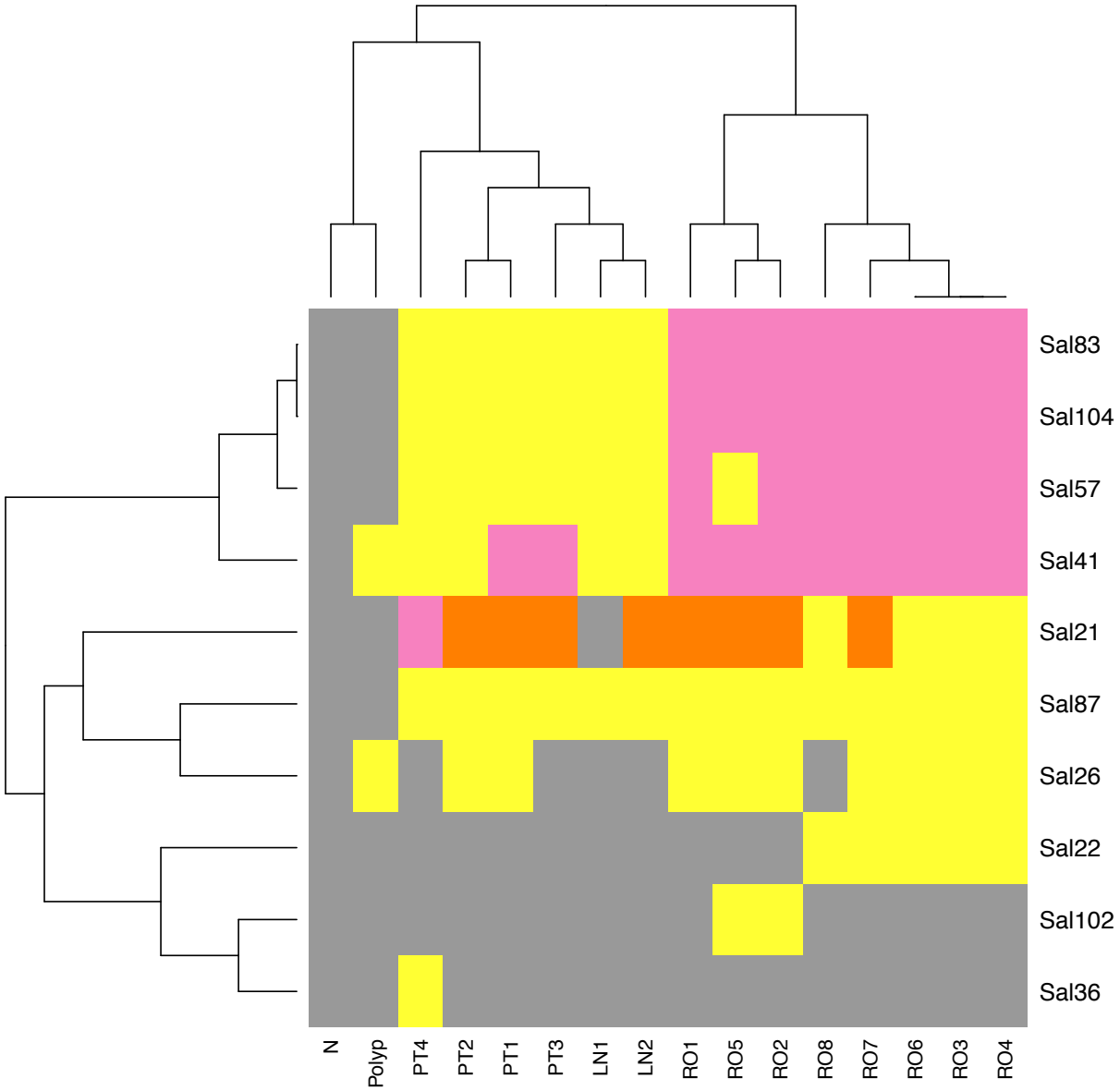

Figure S4: Mutation heatmap patient C31. *Grey squares signify allele distributions that are indistinguishable from normal control. Colored squares indicate a shift in allele distribution, i.e. a poly-G mutation. If multiple different allele distributions exist per marker, they are indicated with additional colors.*

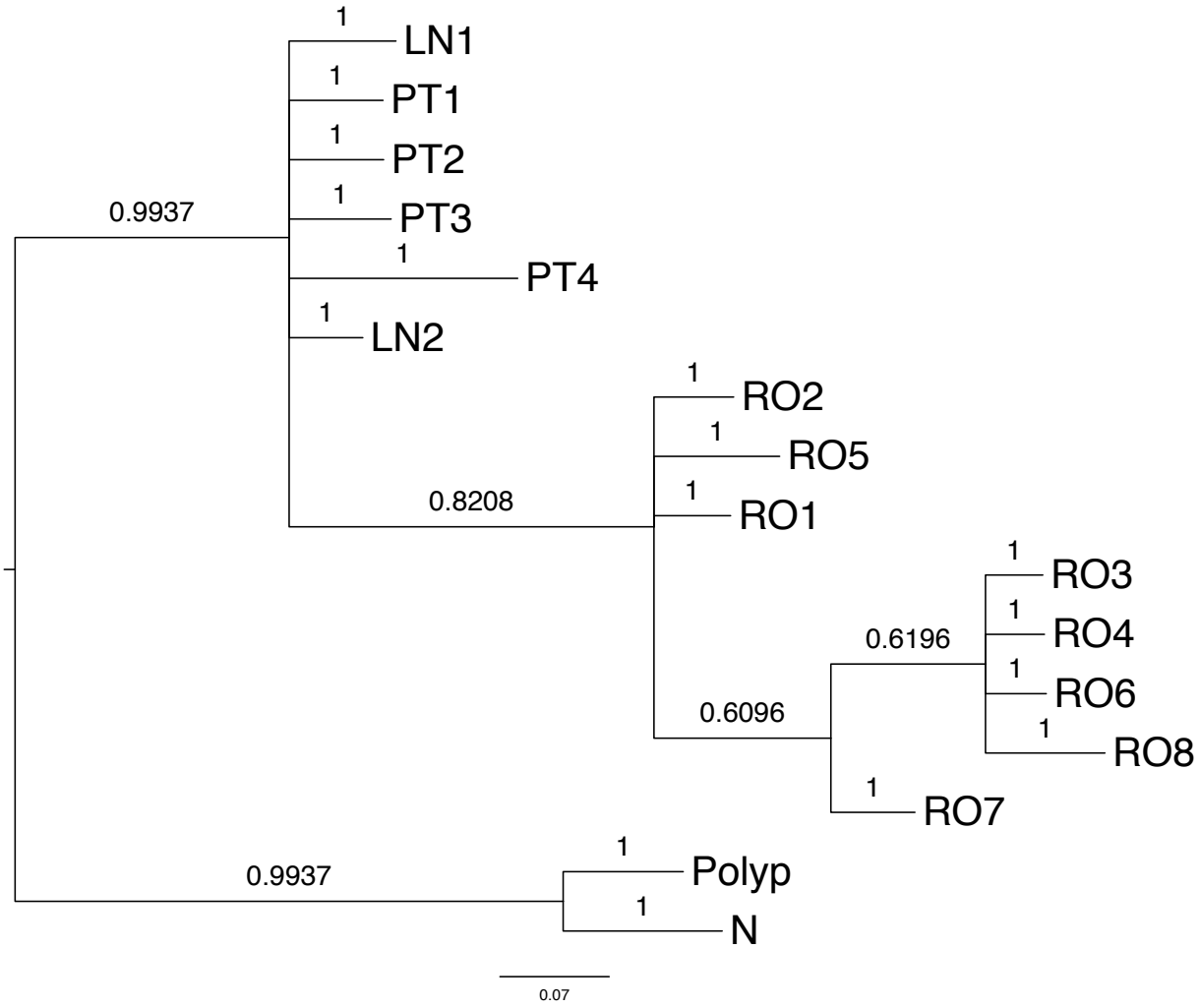

Figure S5: Bayesian phylogeny patient C31. *Posterior probability values for each clade are given as branch labels.*

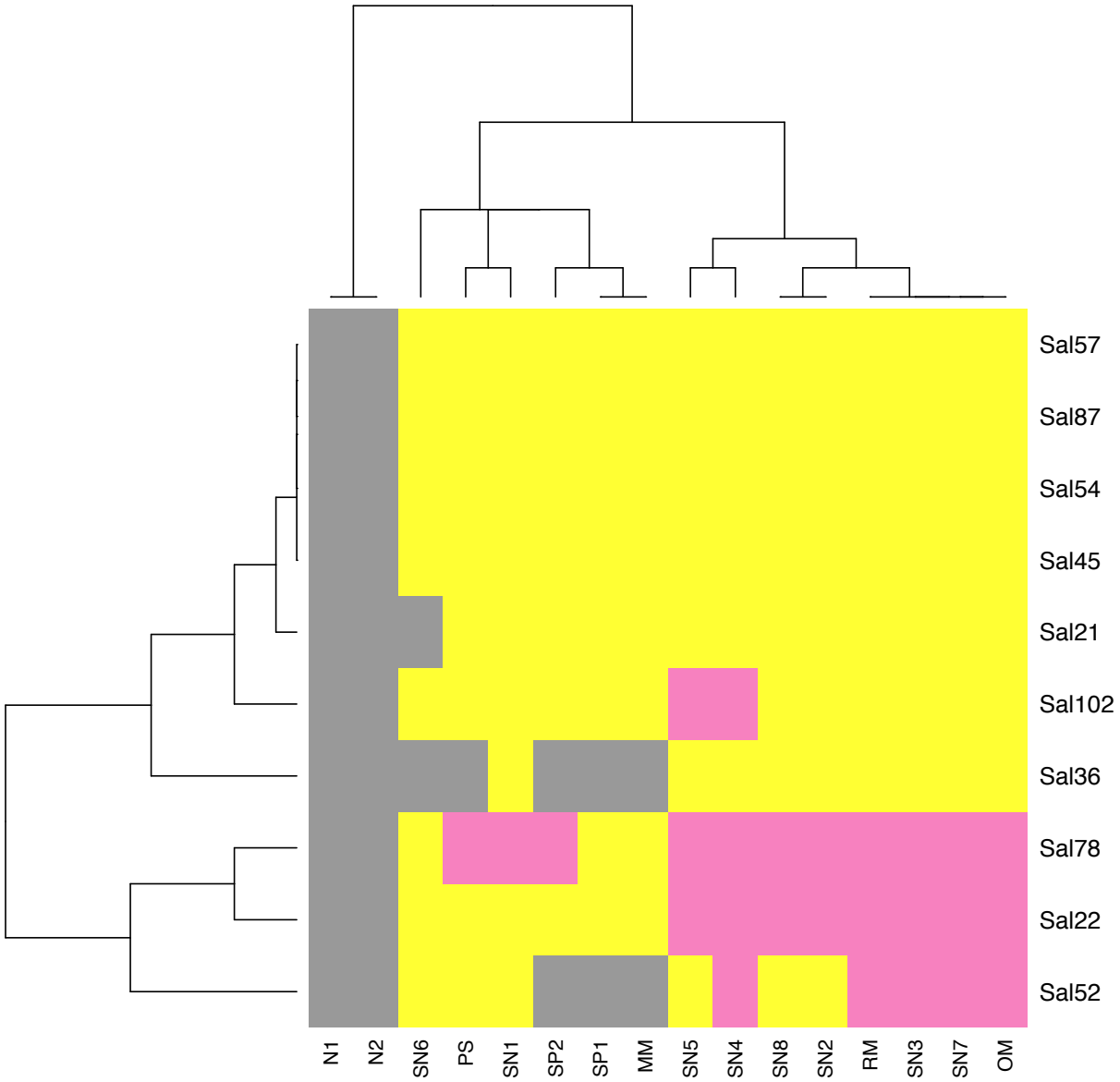

Figure S6: Mutation heatmap patient C27. *Grey squares signify allele distributions that are indistinguishable from normal control. Colored squares indicate a shift in allele distribution, i.e. a poly-G mutation. If multiple different allele distributions exist per marker, they are indicated with additional colors.*

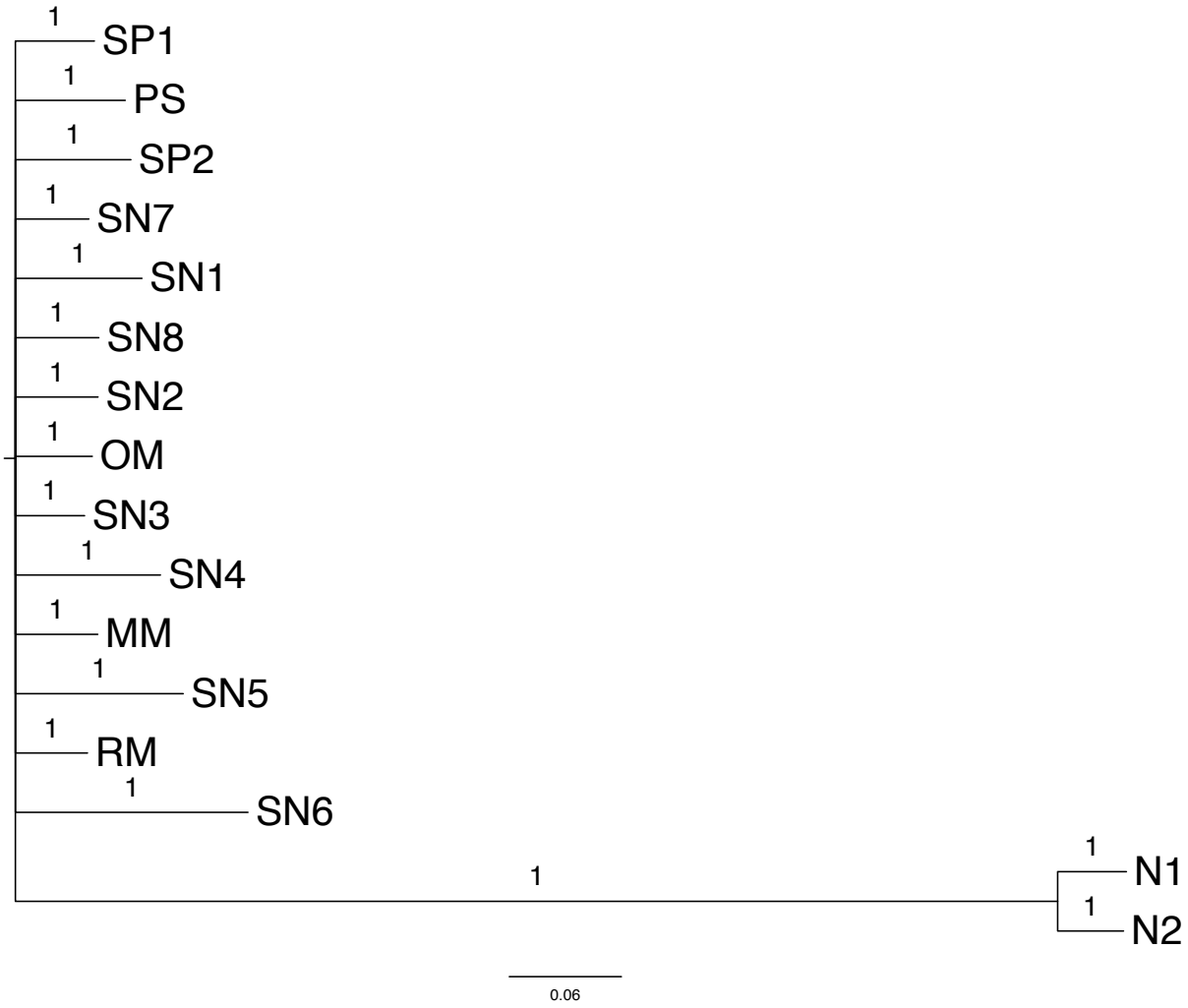

Figure S7: Bayesian phylogeny patient C27. *Posterior probability values for each clade are given as branch labels.*

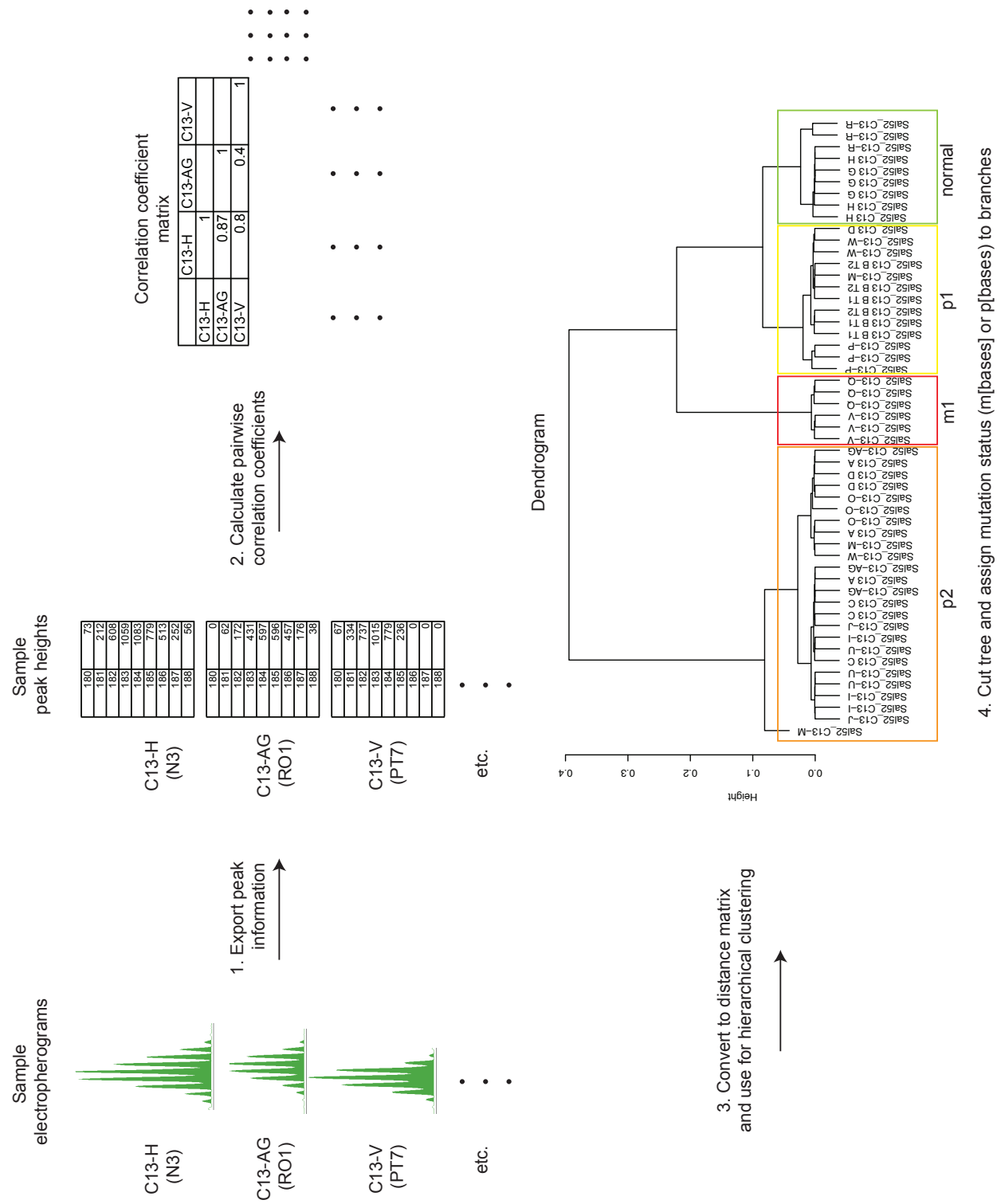

Figure S8: Analysis algorithm flowchart. *Detailed description in Chapter 2, Experimental Procedures*

Table S1: Clinical characteristics of 22 colon cancer patients. *FOLFOX - Chemotherapy regimen of folinic acid, fluorouracil, oxaliplatin. FOLFIRI - Chemotherapy regimen of folinic acid, fluorouracil,irinotecan. IHC -Immunohistochemistry. Bat25, D17S250, Bat26, D5S346, D2S123 - Bethesda Panel microsatellite markers.*

Table S1 (Continued).

	ID	Pathological Diagnosis	Size (cm)	Histological Grade	Sample from re-currence or meta-stasis	Anatomic location	T	N	M	Age	Neo-adjuvant therapy	MSI classi-fication by PCR	MSI classi-fication by IHC	Mutation Frequency (% mutated poly-G loci)
1	C6	INVASIVE ADENOCARCINOMA WITH MUCINOUS FEATURES	4.3	High	No	SIGMOID	3	N2b	0	83	None	MSI-HI	unknown	89.47
2	C9	INVASIVE ADENOCARCINOMA OF THE TRANSVERSE COLON	7	Low	No	TRANS-VERSE	3	N1c	0	77	None	MSI-LO	unknown	40
3	C10	INVASIVE ADENOCARCINOMA	1	Low	No	SIGMOID	1	N2a	0	44	None	MSS	preserved	9.52
4	C11	INVASIVE ADENOCARCINOMA	5	Low	No	SIGMOID	3	N2b	0	53	None	MSS	unknown	33.33
5	C12	ADENOCARCINOMA OF THE COLON WITH MUCINOUS AND SIGNET RING CELL FEATURES	9	High	No	HEPATIC FLEXURE	4	N2b	1	84	None	MSI-HI	unknown	100
6	C13	INVASIVE ADENOCARCINOMA, WITH MUCINOUS AND FOCAL SIGNET RING CELL FEATURES	7	Low	No	CECUM/ ILEUM	4	N2b	1	83	None	MSS	unknown	55
7	C14	INVASIVE ADENOCARCINOMA OF THE SIGMOID COLON, MUCINOUS AND SIGNET RING CELL TYPE	5.3	High	Yes	SIGMOID	4	NA	1	62	None	MSS	unknown	4.76
8	C16	SIGNET RING CELL AND MUCINOUS CARCINOMA OF COLON	3.9	High	No	TRANSVERSE	4	N2a	0	35	None	MSS	preserved	5
9	C18	COLORECTAL ADENOCARCINOMA	5.3	Low	No	SIGMOID	3	N1b	0	62	None	MSS	preserved	35
10	C19	ADENOCARCINOMA OF COLON	3.1	High	No	RECTO-SIGMOID	3	N2b	0	73	None	MSS	preserved	28.57
11	C20	COLONIC ADENOCARCINOMA	4.8	High	No	CECUM	4	N2a	1	59	None	MSS	preserved	35
12	C22	MUCINOUS ADENOCARCINOMA	8	Low	No	CECUM/ ILEUM	4	N2b	0	44	None	MSI-HI	loss of MLH1 and PMS2	90.91
13	C23	METASTATIC MUCIN PRODUCING ADENOCARCINOMA	NA	High	Yes	DIFFUSE	NA	NA	1	55	FOLFOX & FOLFIRI	MSS	preserved	4.76
14	C25	COLONIC ADENOCARCINOMA	9	High	No	SIGMOID	4	N1b	0	71	None	MSS	preserved	0
15	C26	COLONIC ADENOCARCINOMA	2	Low	No	RECTO-SIGMOID	2	N1b	0	44	None	MSS	preserved	35
16	C27	MUCINOUS ADENOCARCINOMA	34.5	Low	No	DESCENDING	4	N2a	1	44	FOLFOX & RADIATION	MSI-HI	loss of MLH1 and PMS2	45
17	C28	COLONIC ADENOCARCINOMA	4.5	Low	No	SIGMOID	3	N1b	0	47	None	MSS	preserved	23.81
18	C29	ADENOCARCINOMA OF THE RECTUM	4.8	High	No	RECTUM	3	N2b	0	71	None	MSS	preserved	20
19	C31	COLONIC ADENOCARCINOMA	3.2	Low	No	HEPATIC FLEXURE	4	N2a	1	48	FOLFOX	MSS	preserved	33.33
20	C32	RECURRENT RECTAL ADENOCARCINOMA	NA	High	Yes	DIFFUSE	NA	NA	1	40	FOLFOX & RADIATION	MSS	unknown	0
21	C33	SIGNET RING CELL CARCINOMA OF COLON	9	High	No	SIGMOID	4	N2b	1	24	None	MSS	preserved	9.52
22	C39	COLONIC ADENOCARCINOMA	6	Low	No	SIGMOID	4	N1a	1	66	None	MSS	unknown	33.33

Table S2: Genotyping information for all patients (separated into four parts). *Four cases that were analyzed in more detail are excluded and instead presented in tables S3-S6. The 20 interrogated poly-G loci are listed in the rows of the table (Sal104-Sal21). For each patient, the length of the PCR product in tumor and normal control tissue is shown. Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. If a locus was clearly heterozygous in a patient (6 or more base pairs length difference), we marked the locus in blue, scored the two alleles independently and increased the "total # of loci scored" count by one if both alleles could be evaluated in both tumor and normal (i.e. if one allele was not lost in the tumor, such as Sal30 in C18). Unambiguous loss of heterozygosity is highlighted with a blue background and not counted as a poly-G mutation.*

Table S2 (Continued).

	C6-Normal	C6-Tumor	C9-Normal	C9-Tumor	C10-Normal	C10-Tumor	C11-Normal	C11-Tumor	C12-Normal	C12-Tumor
	Uninvolved colonic mucosa	Mass to include central ulceration	Representative sections of appendix	Mass to mesenteric fat	Distal donut	Exophytic lesion	distal margin	center of sigmoid mass	Proximal resection margin	Full-thickness colonic tumor
1 Sal104	**126**	**m1**	**127/128**	**127/128**	**126**	**126**	**126/129**	**126/129**	**125/126**	**m3/125/126**
2 Sal102	**155**	**m2**	**156/159**	**m1/159**	**157/158**	**157/158**	**156**	**156**	**157/158**	**m3**
3 Sal87	**NA**	**194/200**	**199/200**	**m1**	**203**	**203**	**196/200**	**196/m0.5**	**199/200**	**m4**
4 Sal84	**199**	**m1**	**199/202**	**199/202**	*200*	**200**	**199**	**199**	**200**	**m2**
5 Sal83	**205**	**m0.5**	**208**	**208**	*210/211*	**210/211**	**205**	*205*	**208**	**m4**
6 Sal78	**182**	**m1/m3**	**182**	**182**	**182**	**182**	**182**	**182**	**181**	**m2/181**
7 Sal74	**190**	**m2**	**187/188**	**m0.5**	*190*	**190**	**189**	**m2**	**189**	**m2**
8 Sal66	**177**	**m3/177**	**177**	**m1**	*177*	**m1**	**177**	**p0.5**	**176**	**m4**
9 Sal57	**201**	**m2**	**201**	**201**	**200**	**200**	**201**	**201**	**200**	**m0.5**
10 Sal54	**167**	**167**	**167/172**	**167/m1**	*166*	**166**	**170**	**m1**	**167**	**m4**
11 Sal52	**180**	**m2**	**180/181**	**180/181**	**180/181**	**180/181**	**181/182**	**181/182**	**182/185**	**m5**
12 Sal47	**179**	**179**	**179**	**179**	**179**	**179**	**179**	**179**	**179**	**m3**
13 Sal46	162/163	**m2**	*162/163*	*162/163*	162/163	162/163	162/163	162/163	163	**m4**
14 Sal45	*127*	*m2*	**126/127**	*126/127*	**128**	**128**	**127/128**	**127/128**	**127/128**	**m3/m5/127/128**
15 Sal41	**181**	**m1/m13**	**176**	**176**	*175*	**m2**	**175**	**m1**	*176*	**m9**
16 Sal36	**125**	**m6**	**121/125**	**m1/del**	**121**	**121**	**125**	**m1/m4**	**118/125**	**m3/del**
17 Sal30	**117**	**m6**	**NA**	**NA**	**114/116**	**114/116**	**110/115**	**m1/del**	**116**	**m5**
18 Sal26	**157**	**m3**	**159**	**m1**	**172**	**172**	*158/173*	**158/173**	**158/173**	**m1/m5**
19 Sal22	**190**	**m2**	**172/195**	**m1/195**	**179/190**	**179/190**	**179/190**	**179/del**	**190**	**m4**
20 Sal21	**150**	**m3**	**NA**	**NA**	**150/153**	**150/153**	**150/154**	**del/154**	**150**	**m3**
# of loci scored		19		20		21		21		21
Mutation freq		89.4736842		40		9.52380952		33.3333333		100

Table S2 (Continued).

	C14-Normal	C14-Tumor	C16-Normal	C16-Tumor	C18-Normal	C18-Tumor	C19-Normal	C19-Tumor
	anasto-mosis ring	central lesion			resection margins	mass	donuts	mass
1 Sal104	**126**	**126**	**126/127**	**126/127**	**130/131**	**130/131**	**126**	**126**
2 Sal102	**155**	**155**	**156**	**156**	**155**	**155**	**157/158**	**157/158**
3 Sal87	**200**	**200**	**200**	**200**	**200/203**	**m1**	**199/200**	**199/200**
4 Sal84	**199**	**199**	**200**	**200**	**199**	**199**	**200**	**200**
5 Sal83	**208**	**208**	**211**	**211**	**206LH**	**p0.5**	**208**	**208**
6 Sal78	**182**	**182**	**182**	**182**	**182**	**182**	**182**	**182**
7 Sal74	**190**	**190**	**190**	**190**	**190**	**190**	**189**	**189**
8 Sal66	**178**	**178**	NA	NA	**177**	**m2**	**177**	**p0.5**
9 Sal57	**201**	**201**	**202**	**202**	**202LH**	**p1**	*200*	**m0.5**
10 Sal54	**167**	**p1**	**168/171**	**168/171**	**167**	**m1**	**169/170**	**m1/m3**
11 Sal52	**181/185**	*181/185*	**184**	**184**	**183/185**	**m2**	**181/184**	**181/184**
12 Sal47	**179**	**179**	179	*179*	**179**	**179**	*179/183*	*179/183*
13 Sal46	**163**	**163**	**163**	**163**	**162**	**162**	**163**	**163**
14 Sal45	**127**	**127**	*126/127*	*126/127*	**128/134**	**m1/del**	**125/126**	**p0.5**
15 Sal41	**177**	**177**	*173*	*173*	**181**	**181**	**177/181**	**177/181**
16 Sal36	**113/127**	**113/127**	**122**	**122**	**121**	**121**	**125**	**m4/p1**
17 Sal30	**114/116**	*114/116*	**116**	**116**	**110/117**	**110/del**	**110**	**110**
18 Sal26	**157**	**157**	**157/164**	**157/164**	**158**	**158**	**157**	**p2**
19 Sal22	**180**	**180**	**179**	**179**	**179**	**179**	**179/190**	*179/190*
20 Sal21	**154**	**154**	**154**	**p0.5**	**154**	**154**	**150/154**	**150/154**
# of loci scored		21		20		20		21
Mutation freq		4.76190476		5		35		28.5714286

Table S2 (Continued).

	C20-Normal	C20-Tumor	C22-Normal	C22-Tumor	C23-Normal	C23-Tumor	C25-Normal	C25-Tumor	C26-Normal	C26-Tumor
	Colonic mucosa	Tumor	small bowel	mass	Uninvolved ascending colon	Cecum and pericecal mass	colon distal margin	mass	rectosigmoid margins	mass
1 Sal104	**126/127**	**126/127**	**126**	**m4/126**	**126/131**	**126/131**	**127/130**	**127/130**	**127/130**	**m1**
2 Sal102	**157/158**	**m2**	**158**	**m1**	*155*	*155*	*156*	*156*	*155*	*155*
3 Sal87	**199/203**	**199/203**	**199**	**m5/199**	**201**	**201**	**200**	**200**	**199**	**m0.25**
4 Sal84	**199**	**199**	**200**	**m1**	**200**	**200**	**199**	**199**	**199**	**199**
5 Sal83	**205/206**	**m1**	**206/212**	**m3/212**	**205**	**205**	**207**	**207**	**207/208**	*m0.5*
6 Sal78	**182**	*m0.5*	182	**m3/182**	**182LH**	**m1**	*182*	*182*	*182*	*182*
7 Sal74	**190**	**190**	*190*	**m3/190**	**189**	**189**	**189/190**	**189/190**	**190LH**	**m1**
8 Sal66	**177RH**	**177RH**	*177*	**m3/177**	**177**	**177**	**176/177**	**176/177**	**177**	**177**
9 Sal57	*201*	*m0.5*	**202**	**m3/202**	*201*	**201**	**200**	**200**	**201**	**201**
10 Sal54	**167/168**	**167/168**	*167*	**m0.5**	**167**	**167**	**167/172**	**167/172**	**167**	**167**
11 Sal52	**181LH**	**p0.5**	**183**	**m4/183**	**181/185**	**181/185**	**181**	**181**	**181/185**	**181/185**
12 Sal47	**179**	*179*	179	**m1**	**178/179**	**178/179**	**177**	**177**	**177**	**177**
13 Sal46	**163**	**163**	**163**	**m4/163**	**162/163**	**162/163**	**163**	**163**	**162/163**	**162/163**
14 Sal45	**127RH**	**m1**	**126/127**	**m4/126/127**	**127/128**	**127/128**	**128**	**128**	*127/128*	**m1**
15 Sal41	**181**	**181**	**177**	**m12/177**	**181**	**181**	**173**	**173**	**176**	**176**
16 Sal36	**125/126**	**125/126**	**121**	**m3/121**	121	*121*	*126*	*126*	*121*	*121*
17 Sal30	*110*	**110**	**110/117**	**m3/del**	*110*	**110**	*114/118*	*114/118*	**116**	*m2*
18 Sal26	**159**	**159**	**158/172**	**m6/158/m6**	**158/176**	**158/176**	**159/166**	**159/166**	**159**	**m0.5**
19 Sal22	**180**	**p0.25**	*179/190*	**179/190**	*179*	*179*	*179*	*179*	*180*	*180*
20 Sal21	**154**	**154**	*151*	*m4*	*154/155*	*154/155*	*155*	*155*	154/155	154/155
# of loci scored		20		22		21		21		20
Mutation freq		35		90.9090909		4.76190476		*0*		35

Table S2 (Continued).

	C28-Normal	C28-Tumor	C29-Normal	C29-Tumor	C32-Normal	C32-Tumor	C33-Normal	C33-Tumor
	donut margins	mass	rectosigmoid, proximal margin	Tumor to closest posterior margin			Proximal margin, en face.	Representative section of primary mass
1 Sal104	**127**	**127**	**126/127**	**126/127**	**126/127**	**126/127**	**125/126**	*125/126*
2 Sal102	*157/158*	*157/158*	**156**	**156**	**155**	**155**	*155*	**155**
3 Sal87	**196/200**	**196/200**	*201*	**m0.5**	**200**	**200**	**199**	**m0.25**
4 Sal84	**200**	**200**	**200**	**p0.5**	**199**	**199**	**200**	**200**
5 Sal83	**200/206**	**200/206**	**208/209**	**208/209**	**209**	**209**	**208**	**208**
6 Sal78	**182**	*182*	**182**	**m1**	**182**	**182**	**182**	**182**
7 Sal74	**190**	**m1**	**189**	**189**	**189**	**189**	**189**	**189**
8 Sal66	**177**	**177**	**176/177**	**m2**	**177**	**177**	**176**	**176**
9 Sal57	**201**	**201**	**201**	**201**	**201**	**201**	**200**	**200**
10 Sal54	**168/170**	**168/m1**	**168**	**168**	**168**	**168**	**169/171**	**169/171**
11 Sal52	**183**	**183**	**180/181**	**180/181**	**181/187**	**181/187**	**181**	**p02.5**
12 Sal47	**178/179**	**178/179**	179	179	179	179	NA	177
13 Sal46	**162/163**	**162/163**	**163**	**163**	**163**	**163**	**163**	**163**
14 Sal45	**129**	**129**	NA	NA	*127*	*127*	*128/129*	*128/129*
15 Sal41	**178/179**	**p1**	*181*	*181*	*177/181*	*177/181*	*174/177*	*174/177*
16 Sal36	*124*	*m3*	**120**	**120**	**126/127**	**126/127**	**121**	**121**
17 Sal30	*110*	**110**	**110/116**	**110/116**	**117**	**117**	**117**	**117**
18 Sal26	**158**	**m1**	**158**	**158**	**158**	**158**	**158/164**	**158/164**
19 Sal22	*180/190*	*180/190*	**190**	**190**	**180/190**	**180/190**	**179/190**	**179/190**
20 Sal21	*152*	*152*	**150**	**150**	**150**	**150**	**152/155**	**152/del**
# of loci scored		21		20		21		21
Mutation freq		23.8095238		20		0		9.52380952

Table S3: Genotyping information for all samples collected from patient C39. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. If a locus was clearly heterozygous (6 or more base pairs length difference), we marked the locus in blue and scored the two alleles independently. Unambiguous loss of heterozygosity is highlighted with a blue background and not counted as a poly-G mutation. The sample description from the surgical pathology report is given along with the name of the sample. Samples appear in the same order as in the corresponding heatmap.*

Table S3 (Continued).

	Sample description from surgical pathology report	Proximal resection margin, en face	Distal resection margin, en face	Unremarkable liver parenchyma	Lesion #1 (arising in adenoma) to deepest invasion	Lesion #1 (arising in adenoma) to deepest invasion	Lesion #1 (arising in adenoma) to deepest invasion	Lesion #2 to deepest invasion (pericolic fat), bisected	Lesion #2 to deepest invasion (pericolic fat), bisected	IMA lymph node	Lesion #2 to deepest invasion (pericolic fat), bisected	Segment 6 liver metastasis
		N2	**N1**	**N3**	**CT2-B**	**CT1**	**CT2-A**	**ST2-Sec**	**ST2-Punch**	**LN**	**ST1**	**LM**
1	Sal104	125	125	125	125	125	125	125	125	125	125	125
2	Sal102	158RH	158RH	158RH	158RH	158RH	158RH	158RH	158RH	158RH	158RH	158RH
3	Sal87	199/202	199/202	199/202	199/202	199/202	199/202	199/del	199/del	199/202	199/202	199/202
4	Sal84	202RH	202RH	202RH	202RH	202RH	202RH	m0.5	m0.5	m0.5	m0.5	m0.5
5	Sal83	210/211	210/211	210/211	m2	m2	m2	m1	m1	m2	m2	m2
6	Sal78	181	181	181	181	181	181	m0.5	m0.5	m0.5	m0.5	m0.5
7	Sal74	188/189	188/189	188/189	188/189	188/189	188/189	188/189	188/189	188/189	188/189	188/189
8	Sal66	175/177	175/177	175/177	175/177	175/177	175/177	175/177	175/177	175/177	175/177	175/177
9	Sal57	202	202	202	m0.5	m1	m1	202	202	202	202	202
10	Sal54	168	168	168	168	168	168	m0.5	m0.5	m0.5	m0.5	m0.5
11	Sal52	181/185	181/185	181/185	181/185	181/185	181/185	181/185	181/185	181/185	181/185	181/185
12	Sal46	165	165	165	165	165	165	165	165	165	165	165
13	Sal45	126	126	126	126	126	126	126	126	126	126	126
14	Sal41	174	174	174	174	174	174	174	174	174	174	174
15	Sal36	122	122	122	122	122	122	122	122	122	122	122
16	Sal30	110/119	110/119	110/119	110/m2	110/m1	110/m1	m1/del	m1/del	m1/del	m1/del	m1/del
17	Sal26-UA	169	169	169	m1	m1.5	m1	m1	m1	m1	m1	m1
18	Sal26-LA	158	158	158	m0.5	m1	m1	158	158	158	158	158
19	Sal22	190LH	190LH	190LH	m0.5	m1	m0.5	p0.5	p0.5	m0.25	p0.5	p0.5
20	Sal21	152	152	152	152	m1	m1	152	152	152	152	152

Table S4: Genotyping information for all samples collected from patient C13. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. The sample description from the surgical pathology report is given along with the name of the sample. Samples appear in the same order as in the corresponding heatmap.*

Table S4 (Continued).

	Sample description from surgical pathology report	Representative full-thickness sections of posterior endomyometrium	Representative full-thickness sections of posterior endomyometrium	Proximal margin	Adenocarcinoma, deeply invasive	Mass to retroperitoneal margin	Adenocarcinoma, deeply invasive	Mass to retroperitoneal margin, Ileocecal valve	Small intestine to cecum	Small intestine to cecum	Right tube and ovary, mass	Mass to retroperitoneal margin	Right tube and ovary, mass	Mass to retroperitoneal margin, Mass to uninvolved mucosa	Granular surface of left ovary	Left tube and ovary, mass	Left tube and ovary, mass	Left tube and ovary, mass	Right tube and ovary, mass
		N1	**N3**	**N2**	**PT5**	**PT7**	**PT4**	**PT2**	**PT3-A**	**PT3-B**	**RO3**	**PT1**	**RO1**	**PT6**	**LO-S**	**LO2**	**LO3**	**LO1**	**RO2**
1	Sal104	127	127	127	127	127	127	127	127	127	127	127	127	127	m1	m1	m1	m1	m1
2	Sal102	156LH	156LH	156LH	156LH	156LH	156LH	156LH	156LH	156LH	p1	p1	p1	p1	p1	p1	p1	p1	p1
3	Sal87	201	201	201	m0.5	m0.5	201	m1	m0.5	m0.5	m1	m1	m1	m1	m1	m1	m1	m1	m1
4	Sal84	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
5	Sal83	207	207	207	207	207	207	207	207	207	207	207	207	207	207	207	207	207	207
6	Sal78	182	182	182	182	182	m0.5	m1	m0.5	m0.5	m1	m1	m1	m1	m1	m1	m1	m1	m1
7	Sal74	189	189	189	189	189	189	189	189	189	189	189	189	189	189	189	189	189	189
8	Sal66	176/177	176/177	176/177	176/177	176/177	m2	m2	m1	m1	m2	m2	m2	m2	m2	m2	m2	m2	m2
9	Sal57	201	201	201	201	201	201	201	201	201	201	201	201	201	201	201	201	201	201
10	Sal54	167RH	167RH	167RH	167RH	167RH	167RH	m1	m0.5	m0.5	m1	m1	m1	m1	m1	m1	m1	m1	m1
11	Sal52	183/184	183/184	183/184	m1	m1	p1	p1	p1	p1	p2	p2	p2	p2	p2	p2	p2	p2	p2
12	Sal47	179	179	179	179	179	179	179	179	179	179	179	179	179	179	179	179	179	179
13	Sal46	163LH	163LH	163LH	163LH	163LH	m1	163LH	m0.5	m0.5	m0.5	163LH	m0.5	m0.5	m0.5	m0.5	m0.5	m0.5	m0.5
14	Sal45	129	129	129	129/p4	129/p4	m1	p1	p1	p1	p3	p3	p3	p3	p3	p3	p3	p3	p3
15	Sal41	176/177	176/177	176/177	m2.5	m2.5	m2.5	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2
16	Sal36	125/126	125/126	125/126	m2.5	m0.5	m4	m3	m2.5	m3	m3	m3	m3	m3	m3.5	m2.5	m3	m3	m3
17	Sal30	116	116	116	116	116	m0.5	m0.5	m0.5	m0.5	m0.5	m1	m1	m1	m1	m1	m1	m1	m1
18	Sal26	158	158	158	m2	m2	m1	158	158	158	158	158	158	158	m1	158	158	158	158
19	Sal22	180	180	NA	p0.25	p0.25	180	180	180	180	180	180	180	180	NA	180	180	180	180
20	Sal21	154	154	NA	154	154	154	154	154	154	154	154	154	154	NA	154	154	154	154

Table S5: Genotyping information for all samples collected from patient C31. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. If a locus was clearly heterozygous (6 or more base pairs length difference), we marked the locus in blue and scored the two alleles independently. The sample description from the surgical pathology report is given along with the name of the sample. Samples appear in the same order as in the corresponding heatmap.*

Table S5 (Continued).

	Sample description from surgical pathology report	Appendix	Polyp	Colon mass	Colon mass	Colon mass	Colon mass	Matted lymph node mass abutting mesenteric margin	Matted lymph node mass abutting mesenteric margin	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass	Right ovary and fallopian tube, mass
		N	**Polyp**	**PT4**	**PT2**	**PT1**	**PT3**	**LN1**	**LN2**	**RO1**	**RO5**	**RO2**	**RO8**	**RO7**	**RO6**	**RO3**	**RO4**
1	Sal104	127RH	127RH	m1	m1	m1	m1	m1	m1	m2	m2	m2	m2	m2	m2	m2	m2
2	Sal102	156LH	156LH	156LH	156LH	156LH	156LH	156LH	156LH	156LH	m1	m1	156LH	156LH	156LH	156LH	156LH
3	Sal87	200RH	200RH	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1
4	Sal84	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
5	Sal83	209/210	209/210	m1	m1	m1	m1	m1	m1	m2	m2	m2	m2	m2	m2	m2	m2
6	Sal78	182	182	182	182	182	182	182	182	182	182	182	182	182	182	182	182
7	Sal74	189	189	189	189	189	189	189	189	189	189	189	189	189	189	189	189
8	Sal66	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178	177/178
9	Sal57	201/202	201/202	m1	m1	m1	m1	m1	m1	m1.5	m1	m1.5	m1.5	m1.5	m1.5	m1.5	m1.5
10	Sal54	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167	166/167
11	Sal52	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180	180
12	Sal47	179	179	179	179	179	179	179	179	179	179	NA	179	179	179	179	179
13	Sal46	163	163	163	163	163	163	163	163	163	163	163	163	163	163	163	163
14	Sal41	175	m1	m1	m1	m1.5	m1.5	m1	m1	m1.5	m1.5	m1.5	m1.5	m1.5	m1.5	m1.5	m1.5
15	Sal36	113/121	113/121	113/m1	113/121	113/121	113/121	113/121	113/121	113/121	113/121	113/121	113/121	113/121	113/121	113/121	113/121
16	Sal30	117	117	117	117	117	117	117	117	117	117	117	117	117	117	117	117
17	Sal26	158	m1	158	m1	m1	158	158	158	m1	m1	m1	158	m1	m1	m1	m1
18	Sal22-UA	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190
19	Sal22-LA	179	179	179	179	179	179	179	179	179	179	179	p0.5	p0.5	p0.5	p0.5	p0.5
20	Sal21	152/155	152/155	m1.5/155	152/m1	152/m1	152/m1	152/155	152/m1	152/m1	152/m1	152/m1	m1/155	152/m1	m1/155	m1/155	m1/155

Table S6: Genotyping information for all samples collected from patient C27. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. If a locus was clearly heterozygous (6 or more base pairs length difference), we scored the two alleles independently. UA - Upper allele, LA – Lower allele. The sample description from the surgical pathology report is given along with the name of the sample. Samples appear in the same order as in the corresponding heatmap.*

Table S6 (Continued).

	Sample description from surgical pathology report	Uninvolved spleen	Uninvolved colon	Serosal nodules	Peritoneal sidewall, larger fragment	Serosal nodules	Splenic mass	Splenic mass	Tumor at mesenteric margin (terminal ileum)	Serosal nodules	Serosal nodules	Serosal nodules	Serosal nodules	Tumor at retroperitoneal margin (rectum)	Serosal nodules	Serosal nodules	Omental mass
		N1	**N2**	**SN6**	**PS**	**SN1**	**SP2**	**SP1**	**MM**	**SN5**	**SN4**	**SN8**	**SN2**	**RM**	**SN3**	**SN7**	**OM**
1	Sal104	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129	126/129
2	Sal102	156LH	156LH	m2	m2	m2	m2	m2	m2	m2.5	m2.5	m2	m2	m2	m2	m2	m2
3	Sal87	199/200	199/200	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1
4	Sal84	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
5	Sal83	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212	208/212
6	Sal78	182	182	m0.5	m1	m1	m1	m0.5	m0.5	m1	m1	m1	m1	m1	m1	m1	m1
7	Sal74	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190	190
8	Sal66	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176	176
9	Sal57	201LH	201LH	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1
10	Sal54	171/172	171/172	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1	m1
11	Sal52	181LH	181LH	p0.5	p0.5	p0.5	181LH	181LH	181LH	p0.5	p1	p0.5	p0.5	p1	p1	p1	p1
12	Sal47	179	179	179	179	179	179	179	179	179	179	179	179	179	179	179	179
13	Sal46	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH	163LH
14	Sal45	127	127	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2	m2
15	Sal41	177	177	177	177	177	177	177	177	177	177	177	177	177	177	177	177
16	Sal36	127	127	127	127	m.05	127	127	127	m.05	m.05	m.05	m.05	m.05	m.05	m.05	m.05
17	Sal30-UA	116	117	117	117	117	117	117	117	117	117	117	117	117	117	117	117
18	Sal30-LA	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110
19	Sal26-UA	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172	172
20	Sal26-LA	159	159	159	159	159	159	159	159	159	159	159	159	159	159	159	159
21	Sal22	179	179	p0.5	p0.5	p0.5	p0.5	p0.5	p0.5	p1	p1	p1	p1	p1	p1	p1	p1
22	Sal21	154	154	154	p1	p1	p1	p1	p1	p1	p1	p1	p1	p1	p1	p1	p1

Table S7: Genotyping information for all samples collected from patient O1. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. Unambiguous loss of heterozygosity is highlighted with a blue background and not counted as a poly-G mutation. The sample description from the surgical pathology report is given along with the name of the sample.*

Table S7 (Continued).

Sample description from surgical pathology report	Left calf, Nodule to soft t. mass to margin	Left calf, Nodule to superior margin	Left calf, Nodule to deep margin	Left calf, Nodule to posterior margin	Left calf, Nodule to infiltrative mass	Lung metastasis, further section of tumor to margin	Lung metastasis, additional sections of tumor to pleura	Left hand nodule	Left hand nodule	Normal lung
	LCT1	**LCT2**	**LCT3**	**LCT4**	**LCT5**	**LuM1**	**LuM2**	**LH-P**	**LH-S**	**N**
Sal104	m1	m1	m1	m1	m1	m1	m1	126	126	126
Sal102	155	155	155	155	155	155	155	155	155	155
Sal87	199/200	199/200	199/200	199/200	199/200	199/200	199/200	199/200	199/200	199/200
Sal84	p1	p1	p1	p1	p1	p1	p1	200	200	200
Sal78	182/pardel	pardel/185	182/pardel	pardel/185	pardel/185	182/pardel	182/185	182/185	182/185	182/185
Sal74	190	p0.5	190	190	p0.5	190	190	190	190	190
Sal66	p1	p1	p1	p1	p1	p1	p1	176/177	176/177	176/177
Sal54	m1	m1	m1	m1	m1	m1	m1	168	168	168
Sal52	m2	m2	m2	m2	m2	m2	m2	179	179	179
Sal47	179	179	179	179	179	179	179	179	179	179
Sal22	179/del	179/del	179/del	179/del	179/del	179/del	179/del	179/190	179/190	179/190
Sal21	del/154	del/154	del/154	del/154	del/154	del/154	del/154	151/154	151/154	151/154

Table S8: Genotyping information for all samples collected from patient B1. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. Unambiguous loss of heterozygosity is highlighted with a blue background and not counted as a poly-G mutation. The sample description from the surgical pathology report is given along with the name of the sample.*

Table S8 (Continued).

Sample description from surgical pathology report	Normal breast stroma	Normal lympho-cytes	Mass #4	1 lymph node	Mass #3	Mass #2	Mass #1	1 lymph node
	N1	**N2**	**TF4**	**LN1**	**TF3**	**TF2**	**TF1**	**LN2**
Sal104	130	p0.5	m0.5	m1	m0.5	m0.5	m0.5	m0.5
Sal102	156	156	p0.5	p0.5	156	156	156	156
Sal87	200	NA	200	200	200	200	200	200
Sal84	200	200	m0.25	m0.5	m1	m0.75	m1	m1
Sal83	206/211	206/211	pardel/m2	p1/del	206/m2	206/m2	pardel/m2	pardel/m2
Sal78	182	182	182	182	182	182	182	182
Sal74	189	189	189	189	189	189	189	189
Sal66	177	177	177	177	177	177	177	177
Sal57	201	201	201	m0.5	m0.5	m0.5	m0.5	m0.5
Sal54	166/170	166/170	166/170	166/pardel	166/pardel	166/pardel	166/pardel	166/pardel
Sal52	183	183	183	183	183	183	183	183
Sal46	146	146	146	146	146	146	146	146
Sal45	129	129	m0.5	m2	m1	m1	m1	m1
Sal41	177	177	m4/177	m4/177	m4/177	m4/177	m4/177	m4/177
Sal36	122	122	122	122	122	122	122	122
Sal30	117	117	117	117	117	117	117	117
Sal26	159	159	m1	m1	m1	m1	m1	m1
Sal22	179	179	m0.5	179	179	179	179	179
Sal21	154	154	p0.5	p0.5	p0.5	p0.5	p0.5	p0.5

Table S9: Genotyping information for all samples collected from a renal cell carcinoma. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. Unambiguous loss of heterozygosity is highlighted with a blue/purple background and not counted as a poly-G mutation. The sample description from the surgical pathology report is given along with the name of the sample*

Sample description from surgical pathology report	Ureter and vascular margin	Representative uninvolved parenchyma	Representative section of tumor to soft tissue margin	Mass to inked margin	Mass to inferior calyx	Mass to inferior calyx
	N1	N2	PT1	PT2	PT3	PT4
Sal104	129/130	129/130	m1	129/130	m1	129/130
Sal102	156	156	NA	156	156	156
Sal87	199/200	199/200	NA	199/200	199/200	199/200
Sal84	199/201	199/201	del/201	pardel/201	del/201	pardel/201
Sal78	182	182	182	182	182	182
Sal74	190	190	190	190	190	190
Sal66	177	177	p1	p0.5	p1	p0.5
Sal54	167/170	167/170	167/del	167/170	167/del	167/170
Sal52	181	181	181	181	181	181
Sal47	179	179	179	179	179	179
Sal22	180	180	p0.5	180	p0.5	180
Sal21	154	154	NA	154	m0.25	154

Table S10: Genotyping information for various other tumor types. *Poly-G loci are listed in the rows of the table (Sal104-Sal21). Deletions are recorded as m[number of deleted bases] and marked in red, insertions as p[number of inserted bases] and marked in orange. Unambiguous loss of heterozygosity is highlighted with a blue/purple background and not counted as a poly-G mutation.*

	Normal	Glioblastoma	Normal	Cholangio-carcinoma	Normal	Esophageal carcinoma
Sal104	126	126	128	m1	127	127
Sal102	158	158	156	p1	156	p0.25
Sal87	198/199	m1	199/202	199/del	199	NA
Sal84	199	199	201	201	200	200
Sal78	182	182	182	182	182	182
Sal74	189/190	189/190	189/190	189/190	189	189
Sal66	177	177	177	m0.25	176/177	176/177
Sal54	167/171	167/del	171	m0.5	169	169
Sal52	180/181	180/181	180/181	180/181	185	185
Sal47	179	179	179	m0.5	177	177
Sal22	190	190	179/189	179/189	179	m0.25
Sal21	152	152	150	150	154/155	m1

	Normal	Mesothelioma	Normal	Pancreatic islet cell tumor	Normal	Lung carcinoid tumor
Sal104	126/129	126/129	126	m0.5	127	127
Sal102	155	155	158	158	156	156
Sal87	199	199	201	201	199	p1/p3
Sal84	201	201	199	199	200	200
Sal78	183/184	183/184	182	182	182	182
Sal74	189	189	189	189	189	p0.25
Sal66	177	177	177	177	177	177
Sal54	167/168	167/168	170/171	m0.5	170/171	m0.5
Sal52	180/181	180/181	181LH	m0.5	182/186	182/186
Sal47	179	179	179	179	179	179
Sal22	178	178	190	190	179	179
Sal21	152/154	152/154	149	149	154	m2

Appendix B – Protocols and primer sequences

Protocol for DNA extraction & precipitation from FFPE tissue blocks

1. Punch specimen with a 1.5 mm or 2 mm biopsy punch
2. (Wash punch in 100% ethanol and wipe to re-use, optional)
3. (Cut core in smaller pieces with scalpel, optional)
4. Add 1 ml xylene and incubate at 56°C for up to 30 min
5. Centrifuge and discard xylene
6. (Repeat, optional)
7. Wash the pellet twice with 100% ethanol
8. Air dry the pellet
9. Add 784 µl Shibata buffer (100 mM TrisHCl, pH 8; 4 mM EDTA, pH 8) to each sample
10. Add 16 µl Proteinase K to each sample
11. Incubate overnight (or longer) at 56°C until all tissue is dissolved, adding Proteinase K as needed
12. Let sample cool to room temperature (RT), add equal volume Phenol : Chloroform : IAA (25:25:1, pH 8, RT)
13. Vortex for 1 minute
14. Centrifuge at 14000 rpm for at least 5 minutes at RT.
15. Retrieve aqueous phase, note precise volume
16. Add 1/10 volume 3M Sodium Acetate
17. Add 2.5 volumes 100% ethanol
18. Add 1-2 µl glycogen
19. Precipitate overnight at -80°C

20. Centrifuge at 4°C for 45 min
21. Wash DNA pellet with 1 ml ice-cold 70% ethanol. Pipet up and down to make sure pellet is properly immersed.
22. Incubate at -80°C for one hour
23. Centrifuge at 14000 rpm for 1 minute at 4°C.
24. Repeat step 9 and 10
25. Air dry pellet
26. Resuspend in at least 50 µl nuclease-free water

Primer sequences for amplification of poly-G loci

Locus ID	Chr	Forward primer 5'->3'	Reverse primer 5' -> 3'	Exp. size	Comments
Sal102	2	ttggtattctattatagcagcctgaac	GTTTCTTcattacacatacttattaccaccagga	153	intergenic
Sal103	2	gggcagtattaaaaactatagaataccc	GTTTCTTtacactcttgtgcattttccttttc	168	intergenic
Sal104	1	agttacgacaatcaaaaatgtctctg	GTTTCTTgagatgcctagaccactgattctc	129	MTR intron
Sal105	1	ttaccttaacattcagtcttcctcttg	GTTTCTTtagatatgccacttttgtcatctacag	167	ENAH & PARP1 intron
Sal107	1	ctctcatgacctagctaaaaatgattc	GTTTCTTgcccagacttttatttcttattttgtc	114	PLD5 intron
Sal2	X	tcatcaggttactaggcaatattagg	GTTTCTTctctgtccctgaccaggtctac	108	MAGED4B intron
Sal21	17	tacccaggtgtaagatcttgaaaag	GTTTCTTaggaaacctctactcatgctgaaag	143	GPATCH8 intron
Sal22	17	cctatattcccagctacagctacac	GTTTCTTgggtatatagtgatagtggtttgtttc	188	GPATCH8 intron
Sal26	16	gactgacactgttgtaataccaagg	GTTTCTTggtttcaaacattacaagatcaagg	154	intergenic
Sal27	16	ctgatgagggacaggaatctcac	GTTTCTTatgacccagggacaggtacag	101	intergenic
Sal30	15	ggagtattgctaggagggagttttc	GTTTCTTcgctatatgggtagtcactatctgg	112	intergenic
Sal36	14	gggcattcaggaccactagg	GTTTCTTgttcagagcgtctcttggtttc	125	abPARTs intron
Sal41	13	tcttttgactctaagtcccttagcc	GTTTCTTgtttatagtccgctttttgtaaagg	173	intergenic
Sal45	12	aaggtctgagataagctccagaatc	GTTTCTTaccttagagttcggtgtcatgaag	125	ANO6 intron
Sal46	12	ccgttattaaaaagtctcacgtttg	GTTTCTTatatctaacctctcctcaggtttcc	156	SPATS2 intron
Sal47	12	tttggttaaggccctaaatttgaac	GTTTCTTtttctgcatttttatagtgctttcc	175	intergenic
Sal52	11	cagctaatttttctgtttttagtacagg	GTTTCTTtgcagtcaagaacccatcac	175	TCIRG1 intron
Sal54	10	ctaagtgttaaggacacagactgaagg	GTTTCTTgagaccttacaggaacagaagaata	161	CXCL12 intron
Sal57	10	ccgaatcttaaattgaaaacacaaag	GTTTCTTtttttagtagaaatggggtttcacc	199	intergenic
Sal58	10	gtaagtaaatcaatgaatgtggttgtg	GTTTCTTataaatttttattggattttcgtttgg	113	intergenic
Sal64	9	gtaatcaccatcaatttggcaatttac	GTTTCTTgactaagggaggagaatcactagaa	162	RNF38 intron
Sal66	8	acatgtacattcagttcactgttaagc	GTTTCTTtagctttgttctagttttgtgtgtgtg	171	intergenic
Sal74	7	taacaagggaatgtaaaggaacttatg	GTTTCTTttatttagtccagatttaatgacaaagg	183	intergenic
Sal75	7	catgagttcaattgtttttatttttagc	GTTTCTTcatttctgagataagggttcaaatg	181	intergenic
Sal78	7	caaagagtgaaacagactatcgacttc	GTTTCTTaacctttagatttacagaaaaattgag	175	intergenic
Sal81	6	gtgaactgtgtttctgtcactacactc	GTTTCTTtacaaaaatcatggtttagttttctcc	158	RANBP9 intron
Sal83	6	cagtgtctcattcatctttgtcattc	GTTTCTTcaaaaactacaaaatgtcttaatgga	200	JARID2 intron
Sal84	6	aggtgtctgagaataaagaagatgaag	GTTTCTTatggattcctggtgagatgttg	195	intergenic
Sal87	4	tacatgaaattctcaatgattacaacg	GTTTCTTaagatctattccatcccattgactc	197	ARHGAP10 intron
Sal88	4	aattttcagtctctgagtgtgatcc	GTTTCTTcatttgcgagcaatttctctttag	187	WWC2 intron

These sequences have been previously published in:

Salk, J. J., Salipante, S. J., Risques, R. A., Crispin, D. A., Li, L., Bronner, M. P., et al. (2009). Clonal expansions in ulcerative colitis identify patients with neoplasia. *Proceedings of the National Academy of Sciences*, *106*(49), 20871–20876. doi:10.1073/pnas.0909428106