Clinical Cancer Research

Lymph Node Metastases in Colon Cancer Are Polyclonal

Peter J. Ulintz[1], Joel K. Greenson[2], Rong Wu[2], Eric R. Fearon[2,3,4], and Karin M. Hardiman[5]

Abstract

Purpose: Recent studies have highlighted the existence of subclones in tumors. Lymph nodes are generally the first location of metastasis for most solid epithelial tumors, including colorectal cancer. We sought to understand the genetic origin of lymph node metastasis in colorectal cancer by evaluating the relationship between colorectal cancer subclones present in primary tumors and lymph nodes.

Experimental Design: A total of 33 samples from seven colorectal cancers, including two or three spatially disparate regions from each primary tumor and one to four matched lymph nodes for each tumor, underwent next-generation whole-exome DNA sequencing, Affymetrix OncoScan SNP arrays, and targeted deep confirmatory sequencing. We performed mapping between SNPs and copy number events from the primary tumor and matched lymph node samples, allowing us to profile heterogeneity and the mutational origin of lymph node metastases. The computational method PyClone was used to define subclones within each tumor. The method Clonality Inference in Tumors Using Phylogeny (CITUP) was subsequently used to infer phylogenetic relationships among subclones.

Results: We found that there was substantial heterogeneity in mutations and copy number changes among all samples from any given patient. For each patient, the primary tumor regions and matched lymph node metastases were each polyclonal, and the clonal populations differed from one lymph node to another. In some patients, the cancer cell populations in a given lymph node originated from multiple distinct regions of a tumor.

Conclusions: Our data support a model of lymph node metastatic spread in colorectal cancer whereby metastases originate from multiple waves of seeding from the primary tumor over time. *Clin Cancer Res; 24(9); 2214–24. ©2017 AACR.*

See related commentary by Gerlinger, p. 2032

Introduction

Death from colorectal cancer typically occurs due to distant metastasis (1). Lymph node (LN) metastases are thought to occur before distant metastasis. We and others have previously shown that colorectal cancer primary tumors are composed of multiple distinct genetic subclones (2, 3). Prior models have proposed that individual cells or groups of cells escape from the primary tumor and enter into the lymphatics or blood vessels to seed local LNs and potentially remote sites (4). Rare, individual metastatic cells escaping from the tumor would produce clonal metastatic lesions. In contrast, polyclonal metastatic lesions could arise from multiple different subclones from a primary cancer simultaneously seeding a metastatic site or by recurrent waves of seeding by different individual cancer cells and/or by polyclonal metastatic clusters of cells.

Prior limited studies of colorectal cancer LN metastasis have shown discordance between KRAS mutation status in primary tumors and associated LN metastasis in about 30% of patients, suggesting heterogeneity among metastatic colorectal cancer cells in individual patients (5). However, comparisons of mutational status between primary colorectal cancer and distant metastases have been mixed, with some showing concordance (6, 7) and others discordant mutation status (8). Recent studies of breast cancer lung metastasis in mice have shown that the lung metastases were polyclonal (9), but in human high-grade serous ovarian cancer, omental metastases are predominantly made of individual clones (10). Thus, the clonality of metastasis may be tumor-type or site specific or neither, depending when in time the tumor and the metastases are sampled.

In the work presented here, we studied multiple primary tumor regions and cancer-containing LNs from 7 colorectal cancer patients, with the goal of understanding whether a given LN metastasis was mono- or polyclonal in its composition. In addition, we sought to determine whether a given LN metastasis harbors cells from a single area of the cancer or whether the LN harbors cells from multiple spatially discrete areas of the primary tumor. Moreover, we wanted to assess the relationships among the cancer cell populations present in different metastatic LNs from individual patients and determine whether the metastatic subclones present in an LN originated early or late in tumor evolution. Our overall strategy for addressing these interrelated questions was to start by identifying individual somatic single-nucleotide changes that could be detected by sequencing, or copy number events that could be measured using SNP arrays, in specific regions of patient's

[1]Bioinformatics Core, University of Michigan, Ann Arbor, Michigan. [2]Department of Pathology, University of Michigan, Ann Arbor, Michigan. [3]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan. [4]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan. [5]Department of Surgery, University of Michigan, Ann Arbor, Michigan.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Karin M. Hardiman, University of Michigan, 1500 East Medical Center Drive, Ann Arbor, MI 48109. Phone: 734-647-9710; Fax: 734-647-9710; E-mail: kmha@med.umich.edu

doi: 10.1158/1078-0432.CCR-17-1425

Translational Relevance

The likelihood of cure is substantially decreased in patients with colorectal cancer with metastatic lesions. Regional lymph node metastasis is seen in about 35% of patients with colorectal cancer, but genetic factors leading to colorectal metastasis to lymph nodes are not well understood. Intratumor genetic heterogeneity and subclonal variants are present in primary cancers, but their relationship to metastases is largely uncertain. We utilized multiple in-depth studies in seven primary colorectal cancers and their matched lymph node metastases to define the relationships among primary cancers and metastases. We found that the relationships between individual primary tumors and associated metastatic lesions are complex; the metastases are polyclonal and can originate from more than one region of the primary tumor. The data support a model of lymph node metastases that likely reflects multiple waves of tumor seeding of a given lymph node over time. Our findings on the molecular evolution of metastasis in patients with colorectal cancer have implications for the origins and behavior of metastatic cancer cell populations and treatment strategies.

primary colorectal cancer specimen. Then, the somatic mutations and copy number changes found in the primary tumor regions were sought in the metastatic LN lesions of a given patient. We then analyzed the primary tumor and LN data in-depth to infer subclonal variant populations and their phylogenetic ordering.

We found that LN metastases are polyclonal, differ from one LN to another, and can originate from multiple geographic regions of the primary tumor. Additionally, we found that single LNs can harbor subclones from different geographic regions in the primary tumor. Our findings are consistent with a model of metastasis where multiple waves of metastatic cells escape the primary tumor over time and seed a given LN during tumor progression.

Materials and Methods

Sample collection and DNA extraction

For seven colon cancers, DNA was isolated from banked surgical resection materials present in the Michigan Medicine Tissue Procurement Core. A qualified gastrointestinal pathologist (J.K. Greenson) chose two or three spatially distinct areas of adenocarcinoma in each primary tumor specimen via inspection of hematoxylin and eosin (H&E)–stained glass slides. In addition, multiple matched formalin-fixed paraffin-embedded (FFPE) LNs containing tumor metastasis were identified. Each tumor region intended for analysis was at least 1 cm in distance from the other area(s). Normal colon tissue was also collected from the surgical specimens. The relevant areas of each frozen or FFPE tumor block were identified and manually dissected from the blocks. Clinical data on each patient were abstracted from the medical records. The Michigan Medicine Institutional Review Board approved this study. DNA was extracted using the AllPrep mini kit (QIAGEN) according to the manufacturer's instructions.

Whole-exome sequencing

An overview of the main workflow for this experiment is provided in Supplementary Fig. S1.

Genomic DNA samples were fragmented to a target size of 300 bp using a Covaris S2 fragmentation system. The samples were end-repaired and A-tailed, and custom adapters were ligated using the NEBNext DNA Library Prep Kit according to the manufacturer's recommended protocols. The custom adapters included 6-bp barcodes designed using BARCRAWL software (11) and synthesized by Integrated DNA Technologies. After ligation, the samples were size selected to 300 bp on a 2% agarose gel, and 1-mm gel slices were retained. Samples were isolated from the gel using the QIAGEN QIAquick gel extraction system. Either 10 μL or 15 μL of each ligation product was enriched using the Phusion master mix kit and custom polymerase chain reaction (PCR) primers for a total of 14 cycles of PCR amplification. The PCR products were purified using AMPure XP beads.

Library QC was performed using the Agilent Bioanalyzer and qPCR. Each pool was captured using the Nimblegen SeqCap EZ Exome v3.0 Kit. The captured pools were combined and sequenced on the lllumina HiSeq 2500 platform with paired-end 100-bp reads using v3 reagents. Paired reads were adapter trimmed and mapped to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA mem v0.7.15; ref. 12). Duplicates were removed using Picard v1.140, and local realignment and base quality recalibration performed using GATK v3.2-2. Tumor/normal pairs of the analysis-ready alignment files were analyzed with three somatic variant callers—MuTect v1.1.4, VarScan somatic v2.3.7, and Strelka v1.0.14—using default parameters, with the exception of VarScan for which the minimum variant allele frequency was reduced to 0.05. VarScan variants were filtered using the fpfilter tool. The resulting variant calls were compiled using a custom tool called Jacquard (GitHub repository: https://github.com/umich-brcf-bioinf/Jacquard). High-confidence somatic variants identified by any one of the three callers were retained as targets for deep-sequencing and validation. Variants were annotated using VarSeq (Golden Helix, v.1.4.1), classifying them by region (e.g., exon, intron, 5′-UTR) and amino acid impact (e.g., synonymous, nonsynonymous).

Deep targeted sequencing

To validate mutations identified from exome sequencing and to facilitate subclone profiling, we designed custom capture panels targeting somatic loci detected in the primary tumors using the Agilent SureSelect XT platform (13). The total target size across all capture panels was 1.772 Mb. Samples were sequenced on two lanes of an Illumina HiSeq 4000 using paired-end 150-bp reads.

Deep sequencing data were processed via a workflow identical to exome sequencing described above except that VarScan alternate allele frequency was permitted down to 0.01. Variants were retained from the deep sequencing data if at least two callers reported the mutation as somatic and if the variant was located in an exonic or UTR region of a RefSeq v105 gene. In addition to variant validation, these data were used as input into the PyClone and Clonality Inference in Tumors Using Phylogeny (CITUP) tools for assessing subclonality and phylogeny. Driver genes were annotated according to the recent publication by Giannakis and colleagues assessing recurrent mutations in 619 colorectal cancers (14).

OncoScan SNP arrays and copy number variation analysis

Thermo Fisher (Affymetrix) OncoScan v3 arrays were run on all samples. The assay detects copy number change by generating data at 50- to100-kb resolution across a set of 891 cancer genes and 300 to 400 kb across the rest of the genome. Raw array florescence intensity data generated on the Affymetrix scanners in the form of CEL files were loaded into the OncoScan Console software v.1.1.0 (Thermo Fisher) and processed using the standard Affymetrix reference control files based on sample type (normal or FFPE).

The copy number data were processed using the Nexus Copy Number software v7.5 (BioDiscovery) using their SNP-FASST2 algorithm for analysis and segmentation, generating a median $\log_2$ ratio (L2R) and a median B-allele frequency (BAF) for each genomic segment. The significance threshold for segmentation was set at 1.0E−5, also requiring a minimum of three probes per segment and a maximum probe spacing of 1000 kbp between adjacent probes before breaking a segment. Segments were classified as having gains or losses if the L2R exceeded or fell lower than thresholds manually set for each tumor depending on its inferred purity and the quality of the SNP array (see Supplementary Table S1 and the following section, "Purity estimation"). The reported median BAF is the median BAF of the markers identified as heterozygous; if the number of heterozygous markers in the segment was below 10 or the percentage of homozygous markers was above 85%, no value was reported. The BAF values are used to determine whether a segment is in a loss of heterozygosity (LOH) or an allelic imbalance state.

By default, probe sets were automatically centered to the median for all samples by the Nexus software. For individual samples where the median probe set value was not diploid, specified regions of balanced heterozygosity were manually identified by visual inspection of L2R and BAF plots and defined as diploid regions, permitting the Nexus software to recenter the entire probe set.

Purity estimation

The Affymetrix OncoScan Console software produces an estimate of tumor purity and ploidy using a custom version of the ASCAT algorithm (14) called TuScan. We used TuScan results as well as the ASCAT algorithm directly to infer purity and ploidy of the tumors, and we utilized the sunrise plots that are generated by the algorithm to assist in inferring or confirming the purity estimate for the tumor samples. In the case of multiple conflicting solutions, we chose solutions that either most closely corresponded with other samples from the same tumor or that favored lower overall ploidy (e.g., diploid rather than tetraploid). Purity estimates for each tumor are utilized as input into the PyClone algorithm.

PyClone analysis

PyClone datasets were assembled for each patient, including variant and reference read depth, an alternate read depth, as well as major and minor copy number values for the region in which the variant resides. The reference and alternate read depths were extracted from the alignment (bam) files for each deep sequencing sample using a custom software tool called Zither (https://github.com/umich-brcf-bioinf/Zither). Copy number data processed via Nexus as described above were compiled for each variant in a semimanual manner using the TAPS tool in the Patchwork software library (15). Copy number analysis results were exported from the Nexus software as text and imported into TAPS, and chromosomal plots were generated representing the log ratio versus allelic imbalance of every copy number segment (Supplementary Fig. S2). Chromosomal plots were manually inspected to assign copy number states to clustered regions on the plots. TAPS then uses this information to assign a copy number state and allelic ratio to each copy number segment. The segmental results from TAPS were mapped to detected somatic variants based on each variant's genomic coordinate using a custom Python script.

PyClone is a hierarchical Bayesian model that infers the cellular prevalence of each variant (the proportion of tumor cells in a sample that contain the variant), clustering variants based on covariance of those prevalence estimates across multiple samples of the same patient (refs. 10, 16–18; Supplementary File 1). The PyClone v0.13 Markov Chain Monte Carlo model was run for 10k iterations for each patient, discarding the first 1,000 as burn-in. Prior to downstream analysis, variants for which PyClone produced overly broad posterior cellular prevalence distributions —defined specifically as variants in a patient data set with cellular prevalence standard deviation estimates greater than 0.2 in at least 25% of the samples—were removed. PyClone clusters were retained for subsequent analysis and reporting if they contained four or more variants, or fewer variants if the cluster contained the variant of a driver gene or if the cluster's cellular prevalence measurement was the highest of all clusters for more than one sample.

Phylogeny analysis with CITUP

The CITUP tool (ref. 19; v0.1.0 of the Bitbucket version, https://bitbucket.org/dranew/citup/) was run for the assembled dataset of filtered cellular prevalence estimates for each variant in each patient generated by PyClone. CITUP enumerates all possible phylogenetic trees up to a given number of nodes, assigning variants to nodes in the tree and solving a quadratic inference problem that minimizes error in the assignment of variants to nodes in the tree. The QIP-based method of the tool was used and PyClone cluster assignments provided for each variant using 1,000 restarts and selecting the tree solution with the minimum Bayesian information criterion (BIC) score. The max number of nodes was set to eight. Higher max nodes counts were attempted for tumors for which PyClone predicted more subclones than eight but were computationally prohibitive.

Data sharing

The called mutation data can be found at the European Variation Archive website at https://www.ebi.ac.uk/ena/data/view/PRJEB23791. The OncoScan array data are available from the NCBI's Gene Expression Omnibus (20) at http://www.ncbi.nlm.nih.gov/geo/accession number GSE107225.

Results

Tumor characteristics

All of the tumors analyzed were stage III colon cancer. In total, 43 samples were collected from the seven colon cancers (five to eight samples per patient). Of these, seven normal, 15 primary tumor, and 18 lymph samples yielded sufficient amounts of DNA for complete analysis. Table 1 shows tumor and patient characteristics, including patient age, tumor location, size, and number of LNs containing tumor.

Table 1. Tumor sample characteristics

ID #	Age	Sex	Stage	Tumor location	Tumor samples, *N*	LN positive, *N*	LN studied, *N*
CP08	62	Male	T3N1M0	Transverse colon	2	2	1
CP11	38	Male	T3N2M0	Cecum	2[a]	8	3
CP14	66	Male	T2N2M0	Cecum	3	9	3
CP15	74	Male	T3N1M0	Cecum	2	3	2
CP17	72	Male	T3N2M0	Sigmoid	2	9	2
CP18	56	Female	T4N2M1	Sigmoid	2	8	4
CP19	56	Male	T3N1M0	Sigmoid	2	4	3

NOTE: Seven stage III colon cancer samples were analyzed, including a matching normal and two to three primary tumor samples, and one to four LN samples per patient. Of these, 15 primary tumor and 18 LN samples yielded sufficient material for analysis.
[a]SNP array analysis failed for one CP11 primary tumor sample.

Sequencing

Whole-exome sequencing was performed on tumor tissue samples and matched adjacent normal tissue samples, yielding an average read depth of 48×. Somatic variants were detected using three variant callers, selecting all variants identified by at least one caller. The number of somatic variants initially identified per patient across all tumor subsections was between 171 and 591 in 6 patients, with one hypermutated patient (CP11) yielding 4,718 somatic variants.

Custom targeted sequencing panels were designed around all somatic mutations in each patient to generate more accurate alternate allele measurements and to validate identified mutations. An exception was CP11 for which only loss-of-function (LOF), missense, synonymous, and splice-region loci were included (1,639 variants). LN metastatic samples were also assessed via these deep sequencing panels. An overall average read depth of 514× was achieved across target regions in all deep sequenced samples, with an average depth of 736× across detected variant locations. The overall result is a corpus of variants detected by sequencing, organized by patient, and containing 2,530 somatic variants: 196 LOF mutations, 1,454 missense alterations, and 880 silent mutations. Counts of variants by patient are shown in Supplementary Table S2.

Expected LOF *APC* mutations were detected in tumors CP14, CP15, CP17, CP18, and CP19 (Supplementary Table S2). CP11 is the only tumor of the seven examples without a detected APC variation, although it is a hypermutated sample with damaging mutations detected in *CTNNB1*, *KRAS*, and as well as several other driver genes (Supplementary Table S2). Predicted damaging *KRAS* mismatch mutations were also detected in CP14, CP15, and CP17, and LOF *TP53* mutations were present in CP08, CP14, and CP19. CP17 also had a predicted damaging FBXW7 mutation; CP15 had a SMAD4 mutation; and CP18 had a PIK3CA mutation. Mutation effect predictions were obtained through dbNSFP annotation available via our annotation software (VarSeq, Golden Helix; ref. 21), which aggregates results of six effect prediction tools: SIFT, Polyphen2 HumVar, MutationTaster, Mutation Assessor, FATHMM, and FATHMM-MKL Coding (18).

An annotated listing of the variants detected in this experiment is provided in Supplementary File 2.

Copy number alterations

Copy number variation data were analyzed using TuScan/ASCAT to determine purity and ploidy for each specimen (Supplementary Table S1). Tumor purity was quite diverse in this cohort, ranging from 20% to 95%. This diversity appears to be a general characteristic of colon cancer samples in our experience and complicates analysis. Purity was higher in the tumors than in the LNs (average purity of 0.65 vs. 0.50, respectively, two-sided *t* test 0.0532), and there was a higher diversity in purities of the LN samples versus the tumors in each patient: Standard deviation between the tumor samples averaged 0.06, whereas the LNs averaged 0.24, *t* test 0.0063.

The tumors were largely diploid with the exception of CP18 and CP15LN3, which have an overall triploid genome. These tumors showed substantial overall copy number changes across the genome, with an average of 36% of the genome modified by a copy number event (Supplementary Table S1; Supplementary File 3). The exception is CP11, which is a hypermutated tumor with characteristically little copy number change. A significant portion of the genome of these tumors was in allelic imbalance, with an average of 23.8% of the genomes designated as LOH regions. Over 50% of the samples exhibited copy number gains of Chr7, the q-arm of Chr8, and the q-arm of Chr20, and over 33% of the samples showed deletion events in the q-arm of Chr18. APC is in an LOH state in all CP08, CP15, and PC17 samples as well (Supplementary File 3). Primary tumor samples from CP19 show *APC* in LOH as well as in LN1 and LN2, but not in LN4. In addition, APC is in an amplified region in all CP18 samples. All CP19 samples contain a significant focal amplification encapsulating the EGFR gene (Supplementary File 3, Chr7: 55,086,725–55,238,738, p-arm near centromere). Supplementary File 3 shows log R and BAF profiles for all samples.

Primary tumors and LN metastases are genetically heterogeneous

All tumors had multiple examples of copy number heterogeneity between samples. Intratumor heterogeneity was observable between the primary tumor samples as well as between the primary samples and the metastases, and even between the metastases themselves. Copy number events in the primary tumors were often found in one or more of the LNs; however, a number of copy number variants (CNV) in the LNs were not found in the primary tumors and either arose after metastasis occurred or originated from an unsampled portion of the tumor.

Figure 1 shows an example of heterogeneity of copy number variation within a primary tumor and its matched LNs. In this example, LN2 and LN7 more closely resemble T1, and differences in allelic balance can be found amongst these two lymph samples themselves. This type of assessment was carried out for all regions of all tumors to assess the geographic origin of metastases and is discussed below.

Assessment of the deep targeted sequencing across samples also revealed heterogeneity across tumor samples and matched LNs in

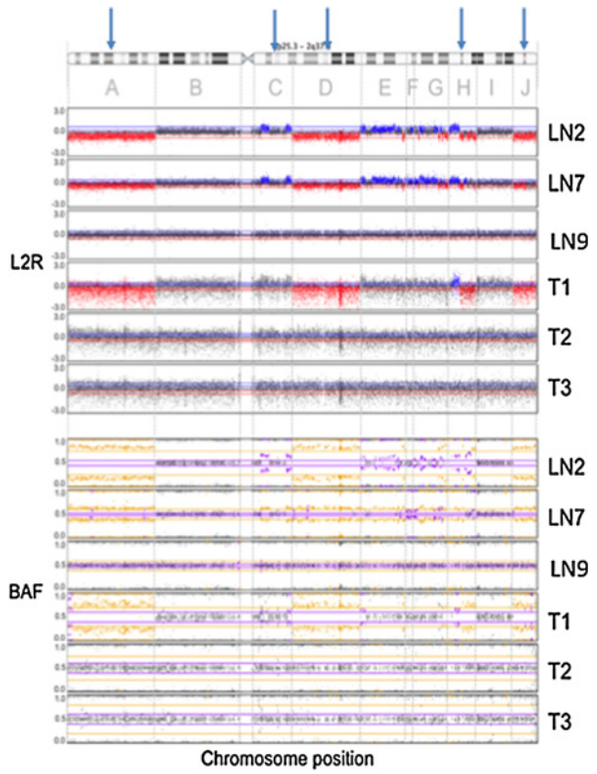

Figure 1.

Colorectal primary tumors and LN metastases are genetically heterogeneous. **A,** SNP array profile of patient CP14 chromosome 2. Top, L2R; bottom, BAF plots. Probes falling in copy gain and copy loss segments are colored blue and red, respectively; probes corresponding to allelic imbalance and LOH segments are colored purple and gold. The LN profiles LN2, LN7, and LN8 are of higher quality (higher signal-to-noise) than the primary tumor samples T1 to T3. Overall, the copy number variation and allelic imbalance patterns of LN2 and LN7 match those of T1, whereas LN9 matches T2 and T3. Regions A, D, and J indicate a deletion event resulting in LOH in T1, LN2, and LN7. Regions B and I are copy neutral with no allelic imbalance. Regions C and E contain copy gain events in LN2 and LN7, which are possibly present in T1 but uncalled by the segmentation algorithm. The copy gain segments in region C show corresponding allelic imbalance in T1, LN2, LN7, and, possibly, LN9 as well; differences in the BAF plots between LN2 and LN3 in this region (purple vs. gold) may be due to thresholding differences and may not reflect biological difference. Regions E and G contain copy gain and loss events in LN2 and LN7, with corresponding allelic imbalance in the BAF plots that are unique events in the lymphs; with the exception of several small subregions, these events are not present in T1 or the other tumor samples. Region F indicates a region of allelic imbalance in LN7, but not LN2, with no corresponding change in the primary tumors. Arrows at the top of the figure indicate regions used in the main copy number analysis (Table 3).

each patient. Using the presence or absence of variants between samples as a rough measure of heterogeneity indicated that between 1.2% and 89.1% of mutations detected in a tumor are shared across all samples in individual patients (Supplementary Table S2). Tumor CP14 appeared to have the highest degree of heterogeneity in detected variants between all samples, with only four mutations shared across all seven samples. Many of the variants in CP14 (160) were unique to individual samples. In tumor CP18, 66 out of 181 total variants were unique to a single sample, all in primary tumor samples, and 83 were shared by all. CP15 had 42 out of the total 297 private to individual samples, all but one in the primary tumor samples, and 231 shared by all. All of the unique variants of CP08 were in T1. Tumors CP08 and CP11 showed a relatively low amount of variant heterogeneity between samples, with 89% and 78% of the variants shared by all, respectively. Tumor CP19 had only 37 of 95 variants shared amongst all samples, but only seven variants private to individual samples; 27 variants were shared between T1, T2, LN1 and LN2. LN4 of CP19 (LN_C in Supplementary Table S3) was a low-purity sample with relatively fewer variants detected, as was LN9 of CP14 (LN_D); the private versus shared variant patterns in these tumors may be skewed as a result. As seen in Table 2 and described below, unique variants detected in the LN metastases can be mapped uniquely to specific regions of the primary tumor.

Primary tumors and LN metastases are polyclonal

Deep sequencing data and copy number data from SNP arrays were used to profile the subclonality of each tumor using PyClone. As discussed above, PyClone uses copy number data and tumor purity to estimate the cellular prevalence of each variant based on its alternate allele frequency, grouping variants into clusters based on their cellular prevalence profiles across samples (Supplementary File 1). Cellular prevalence plots along with inferred phylogenies for all tumors are shown in Supplementary Fig. S2. The cellular prevalence plots provide a profile of the subclones inferred in each sample and indicate the changing subclonal composition across samples. The ordering of the samples in the cellular prevalence plots is not temporal; that is, each profile is an assessment of the composition of a sample at a single time point.

All tumor samples showed evidence of subclonal structure, with the number of mutation clusters detected by PyClone varying between six and 15. However, inspection of the similarity matrices and prevalence plots indicates that clusters with very similar profiles may possibly be merged and considered a single subclone: for example, clusters 6 and 10 of CP08, clusters 14 and 15 of CP11, and clusters 8 and 9 of CP15 (Fig. 2; Supplementary File 1). The number of distinct clusters is not of primary importance for this analysis; of greater significance is the presence of distinctly definable subclones and their mapping between lymph and primary tumors.

Heterogeneity between primary tumor samples is evident not only in differing prevalence measurements of clusters across samples but also by mutations present in one but no other primary tumor samples, as discussed next. LNs also often have substantially different clonal profiles from one another, and between themselves and the primary tumor samples (Fig. 2). These subclonal profiles clearly indicate that LN metastases are not monoclonal.

LN metastases originate from single or multiple geographic regions of the primary tumor

We examined variant data to identify mutations uniquely present in individual primary tumor samples with the goal of using these variants as markers of the region of origin of LN metastasis in the matched primary tumor. For each tumor, variants that are private to only one of the primary tumor samples were mapped to the LNs in which they are detected. Counts of these private variant "marker" mappings are shown for each of the LN samples in Table 1. Multiple marker variants were manually validated by visualizing the raw pileup data for the region of the variant in a genome browser. The variant data indicate that for tumors CP08, CP18, and CP19, all private marker variants unique to a single area of the primary tumor found in the LN metastases

Table 2. Mapping of private sequencing variants

Tumor	Primary	LN	Count
CP08	T1	LN1	0
		LN2	0
		LN3	0
	T2	LN1	2
		LN2	7
		LN3	7
CP14	T1	LN2	21
		LN6	17
		LN7	19
		LN9	4
	T2	LN2	6
		LN6	18
		LN7	5
		LN9	8
	T3	LN2	8
		LN6	8
		LN7	10
		LN9	12
CP15	T1	LN2	2
		LN3	1
	T2	LN2	11
		LN3	2
CP17	T1	LN1	1
		LN5	1
		LN7	2
	T2	LN1	2
		LN5	2
		LN7	3
CP18	T1	LN5	0
		LN6	5
		LN7	4
		LN8	0
	T2	LN5	0
		LN6	0
		LN7	0
		LN8	0
CP19	T1	LN1	1
		LN2	6
		LN4	3
	T2	LN1	0
		LN2	0
		LN4	0

NOTE: Presence of variants that are unique to a single region in a primary tumor was assessed in each LN specimen from that tumor. In all tumors, individual LNs contained variants that were unique to single tumor regions. In tumors 14, 15, and 17, LNs contained unique mutations from multiple different regions, whereas the metastatic variants detected in the other four tumors are mapped to a single primary tumor region. Tumor CP11 is not listed because only a single primary tumor sample was available.

originate from a single geographic region of the primary tumor (Table 1). This is consistent with a group of cells carrying these unique mutations moving from a single region of the tumor to the LNs, or single cells from that same region moving over a series of events. However, in tumors CP14, CP15, and CP17, each LN harbors marker mutations originating from multiple regions of the primary tumor, indicating that multiple tumor regions contributed to each LN metastasis (Table 1). This is consistent with seeding of the LNs with the unique mutations from a single region of the primary tumor at one time and from a different region of the primary tumor during another event.

Next, we performed the same LN mapping using CNVs. Figure 1, introduced earlier, is an example of the method that was applied across the genome of each sample. The goal of the profiling was to identify copy number events private to individual regions of the primary tumors that could be used to indicate a possible clonal origin for each metastasis. The overall profile of LN2 and LN7 in Fig. 1 appears to clearly match primary tumor sample 1 (T1) and not the other samples from the primary tumor. However, there are unique regions of copy number and zygosity change present in these LN samples that are not reflected in any of the sampled regions of the primary tumor, as well as a small region of heterogeneity between the lymph LN2 and LN7 samples themselves (region F). LN9 exhibits similarity to T2 and T3 in this chromosome in that there are no gross copy number variation events displayed in these samples. Care must be taken in these analyses to distinguish between simple purity differences and actual copy number changes. LN9 has a lower fraction of aberrant cells; however, copy number variation events would easily be detectable if they were present in this region, as evidenced by a number of other copy number variation events clearly detected in this sample in other chromosomes (Supplementary File 3).

As with the sequencing variants, copy number events detected in the LN metastases which could be traced to a single primary tumor region were identified. A summary of these distinguishing events for each tumor is provided in Table 3, which also lists events that are unique to the LNs and not present in the primary tumor samples. The table summarizes distinguishing events at the chromosomal level. As can be seen in Table 3 and in Supplementary File 3, the copy number data for CP08 suggest a strong correlation between the LN and the T2 primary sample. All distinguishing copy number events appear to originate in T2 (Table 2) in agreement with the variant profile from Table 1. Tumor CP11 is a hypermutated sample with very little copy number change. The copy number data for CP14 suggest a specific match to T1 in many regions in apparent contrast to the variant profile that revealed distinguishing variants originating from all three tumor samples, which may reflect limitations of the data. The LN samples of CP15 show a similar pattern matching both T1 and T2 across the genome, but there are several specific regions in which the LN samples match T2 only and a region in LN2 (q-arm of Chr7) that maps uniquely to T1. A significant fraction of the CP15_LN2 genome has a unique copy number profile as well, and Chr14 signals an event present in T1 but not T2, which is missing in both LN samples. The copy number data for CP17 do not suggest a specific primary tumor sample of origin for the lymph metastases. There are a couple of regions in LN1 that are unique and not present in the primary tumor samples, however, and there are several regions in which an event is evident in T1 that are not reflected in the LNs as well as regions of heterogeneity between the LNs. In CP18, there are entire chromosome regions with copy number states that match T2, but not T1, in contrast to the variant data. However, there are also smaller regions of LNs LN7 and LN8 that indicate a T1 origin. The bulk of the genome of this tumor exhibits matching copy number events for both primary tumor samples, and significant regions of the LN genomes contain unique events not seen in the primary tumor samples. CP19 shows copy number similarity to both T1 and T2 across much of the genome in LN1 and LN2; LN4 is very impure and difficult to assess. All three LNs display a 150-kb deletion event in Chr20 that is present in T2, but not T1, and there is a region of amplification in Chr5 visible in LN1 and LN2 that matches T1 only. Regions of unique change are visible in these LNs as well.

Lastly, the unique subclones identified by PyClone can be utilized within the cellular prevalence plots (Fig. 2) to assess origin. For example, in CP8, subclone 2 (orange) is unique to T2 and is found in LN3, indicating the region of origin for the metastasis is T2. Another example is CP15, where subclone 19 (light green) found only in T2 is in LN2, but not LN3, indicating the origin for LN2 is T2.

These results suggest that LN metastasis can originate from one or more primary tumor regions. With some exceptions noted above, the origins of LN metastases are in general agreement across all methods tested, including variants, CNVs, and PyClone subclonality profiling. The results support a model of polyclonal metastasis (Fig. 3), the provenance of which is diverse, spanning one or more distinct tumor regions.

LN metastasis contains early and late subclones

Phylogenies were derived using CITUP and illustrate the sequence of the origins of each subclone in each patient (Fig. 2, right). Phylogenies reveal examples of LN metastasis that are early events such as subclone 13 in CP14, which is found in all samples as well as late events such as subclone 4 in CP14, which is a late subclone and is found in only one tumor sample (Fig. 2A). Subclones 7, 8, and 12 of CP17, as another example, are present in all four samples, whereas subclones 1 and 3 are present in primary samples T1 and T2, respectively. Subclones that occur late in the phylogenies are generally found in only a single tumor sample and at most a single LN; examples include subclone 15 in CP15, which is found only in primary tumor sample T1, and subclone 13 in CP19, which is found in primary tumor sample T1 at a low cellular prevalence and then found only in LN2. The most common driver genes are generally found in subclones in all samples from a tumor and are early in the phylogeny. For example, subclone 12, 21 (brown) in CP08 contains a *TP53* mutation and is in all samples and the green subclone in CP15, which contains *APC*, *KRAS*, and *TP53* mutations and is in all samples (Fig. 2). Less common drivers are more often found in subclones found only in a single region of the primary tumor, for example, the

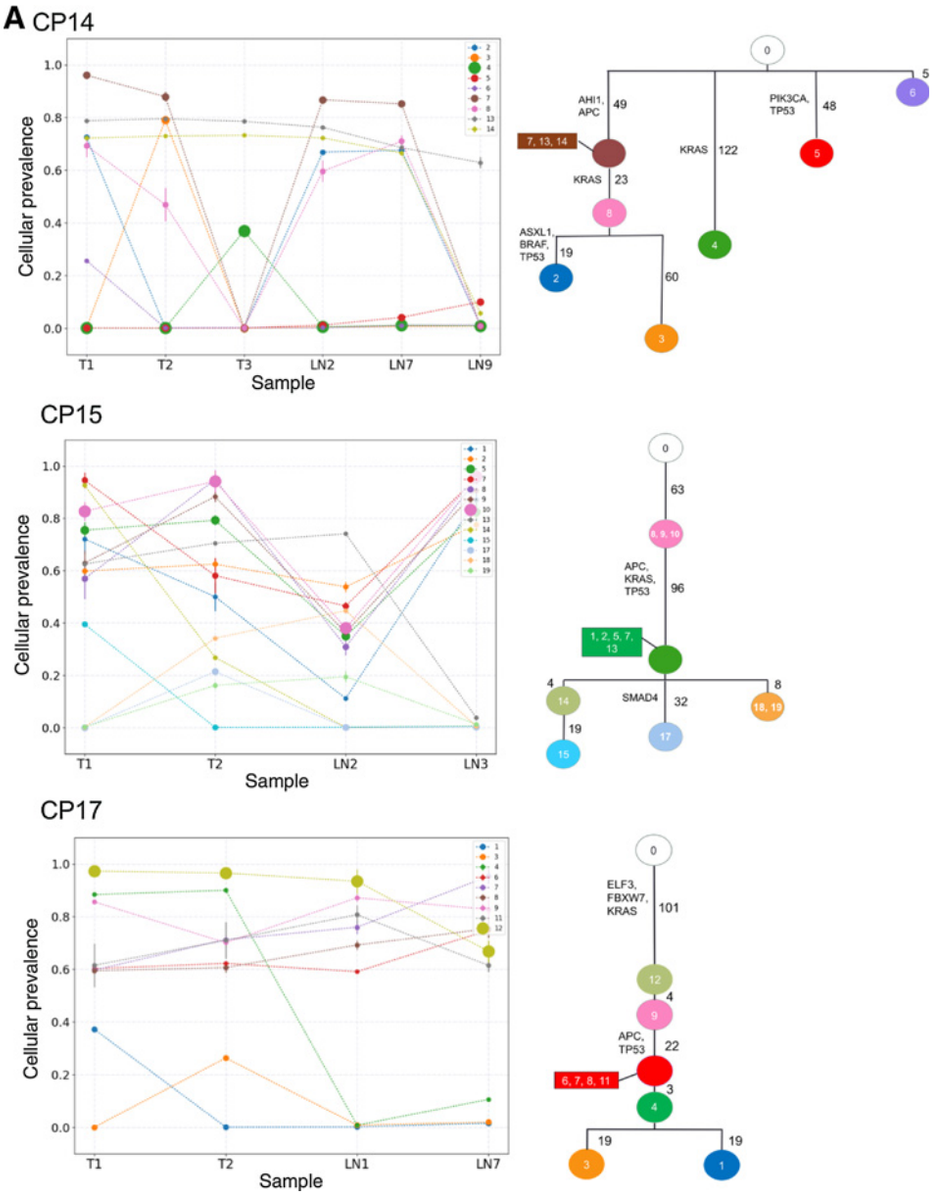

Figure 2.
Subclonal composition and phylogenies of tumors with LN metastases originating from multiple geographic regions of the primary tumor. Composition profiles generated by PyClone (left) plot the mean cellular prevalence of the variants in each cluster in each sample. The order of the samples along the *x*-axis is not temporal. The size of the marker for each mean cellular prevalence measurement is proportional to the number of variants in the cluster. Vertical lines at each point represent one standard deviation. The best scoring phylogenetic tree inferred by CITUP is shown for each tumor (right). Each tree node is labeled with the PyClone cluster ID(s) to which it corresponds. Vertical line lengths in the phylogenies correspond to the number of mutations occurring in the node below. Mutations occurring in known driver genes are indicated (17). Tumors for which metastases originate from multiple geographic regions of the primary tumor are shown in **A**, and those for which metastases originate from a single geographic region of the primary tumor are shown in
(*Continued on the following page.*)

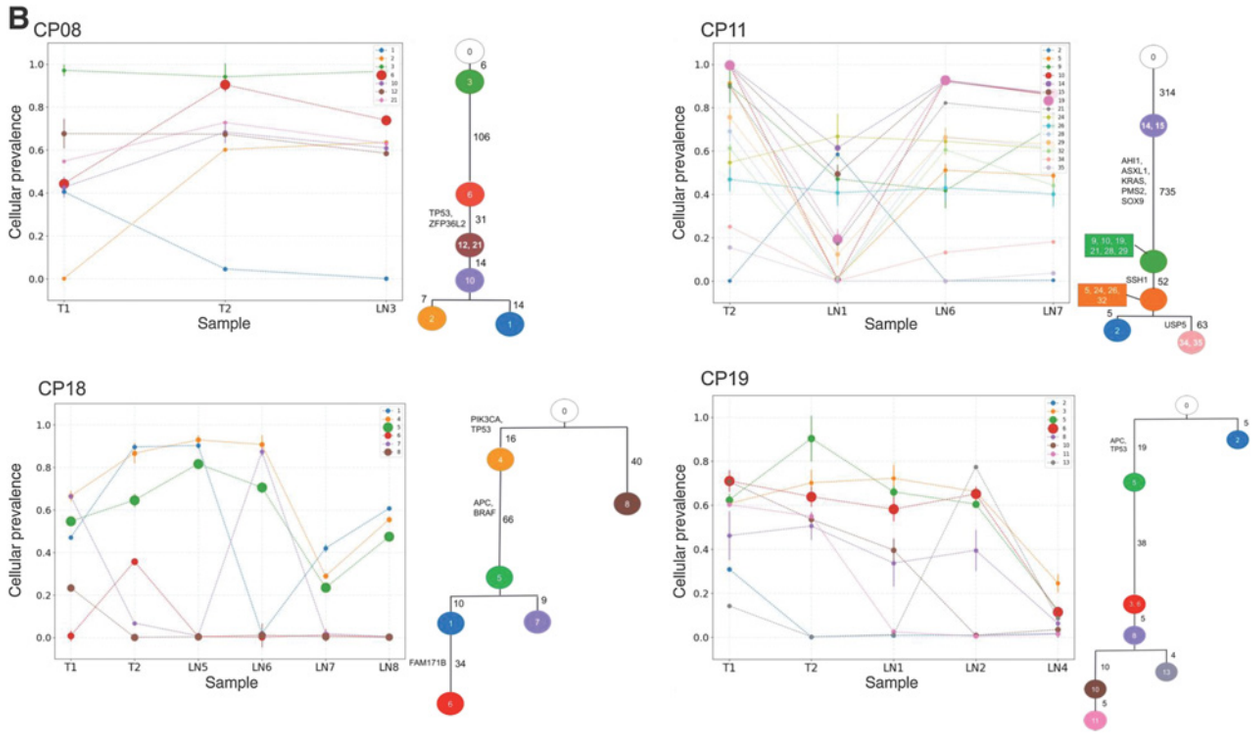

Figure 2.
(*Continued.*) **B**. Tumor CP08 has only one lymph sample, whereas CP11 has only one primary tumor sample.

subclone containing the *SMAD4* mutation in CP15 is only in T2. The drivers listed in Fig. 2 are from the study by Giannakis and colleagues that identified multiple recurrent drivers across 161 tumors (14). We find some examples of apparent subclones that do not contain these known drivers and yet are metastatic. When the composition of these subclones is assessed, putative drivers are identified. For example, the pink cluster (cluster 10) in CP15 contains two LOF mutations and 56 missense mutations, several of which could be candidate drivers. Another example of a tumor with subclones that metastasize but do not contain a known driver is subclone 6 (red) in CP8. This subclone does contain an LOF mutation (*SLC46A3*) and 57 missense mutations. In CP11, there is also a subclone (14, purple), which is metastatic but does not have a labeled driver in Fig. 2; however, it does have 16 LOF mutations and many missense mutations that could be drivers.

Multiple metastatic events seed LNs over time

We present a model of LN metastasis explaining the relationship between tumor subclones within the primary tumor and subclones found in the LNs (Fig. 3). As can be seen from the data above, each site in the primary tumor is polyclonal, and each LN is polyclonal. We have shown evidence supporting polyclonal LNs by assessing unique mutations in the primary tumor and noting that in three tumors, we find mutations unique to more than one geographic region of the tumor in the individual LNs. This, in addition to the PyClone data defining multiple tumor subclones in each tumor and LN, supports a model of polyclonal LN metastasis arising from multiple individual metastatic events from the primary tumor over time. Additional support for this model is found in the phylogenies showing that subclones that were early, mid, and late in tumor evolution are all found in the LNs, again, suggesting multiple clones leaving the primary tumor at different times.

Discussion

We have explored the genetics of primary colorectal cancer and matched LN metastases of seven tumors to uncover temporal and spatial patterns of metastatic spread. Using current informatics tools for inferring subclonality and phylogeny from bulk tumor samples, as well as comparative studies of mutations and CNVs in multiple regions of primary tumors and their matched metastasis, we have found that each LN contains multiple subclones derived from the primary tumor. Additionally, we showed that there is substantial heterogeneity of clones present in different LNs from any individual patient and within a single primary tumor. Together, these results support a model of metastasis with shedding of tumor, either as single cells or groups of cells, over time, as shown in Fig. 3. This model is similar to the one by Cheung and colleagues regarding seeding of distant metastasis, which proposes that polyclonal metastases are generated through polyclonal clusters (4). Our model differs in that it describes LN metastases and we hypothesize that seeding occurs in multiple waves over time because we find multiple tumors with evidence of seeding from multiple different geographic locations within the primary tumor. It is possible that these groups of seeding cells could be composed of single or multiple subclones. We do not find evidence of a superior "metastastic" subclone that is alone responsible for metastasis in any LN tested.

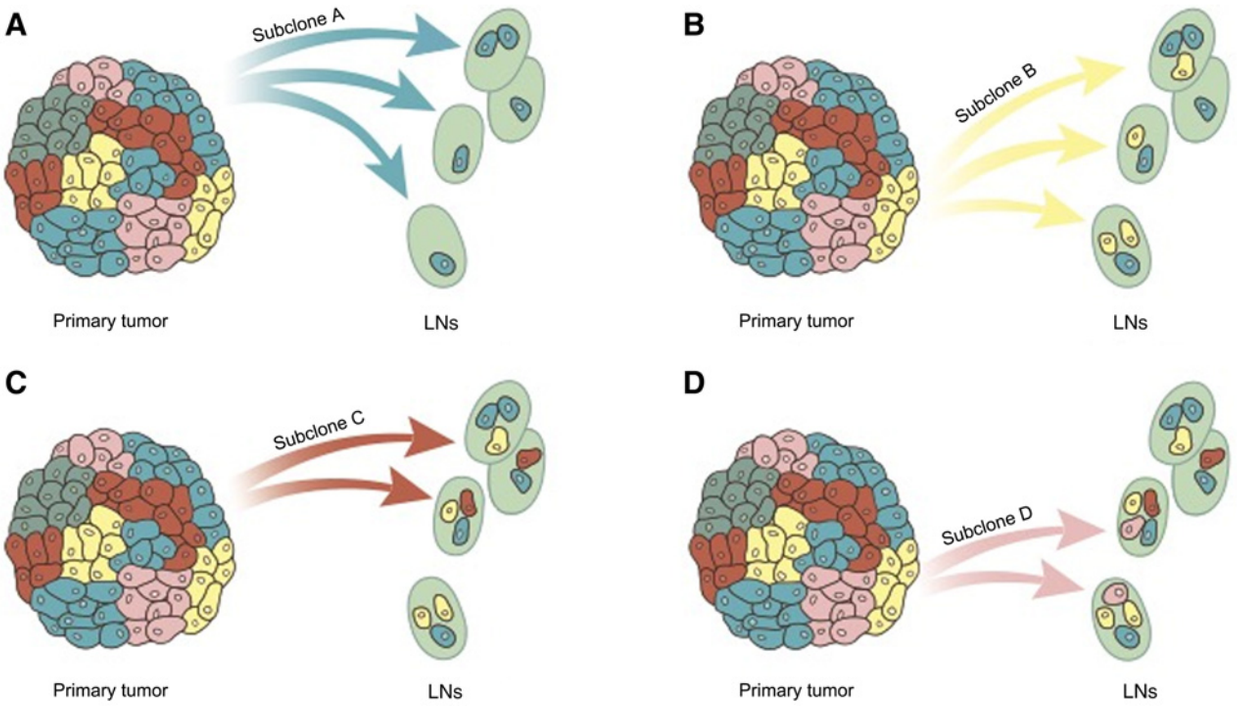

Figure 3.
Model of LN metastases with multiple waves of seeding of the LNs by subclones from the primary tumor over time. In **A–D,** subclones A through D move from the primary tumor to one or more LNs. **A,** A single clone populating multiple nodes. **B–D,** Subsequent subclones each successively populating subsets of LNs.

This model supported by our data may bear profound clinical consequences. Although LN metastases remain strong prognostic factors in colorectal cancer, it is clear from our work that a particular positive node may be quite different in its composition from another positive LN from a given patient. Thus, this work suggests at least two important future lines of investigation: first to couple clonal heterogeneity to the ability to understand the phenotypic consequences of individual clones and second to repeat this analysis on larger cohorts with known survival outcomes to better understand how tumor and LN heterogeneity may be used to prioritize actionable clones.

Very few studies have been performed assessing molecular genetics and mutational heterogeneity in LN metastases and none have assessed regional origin of metastasis. McPherson and colleagues studied clonal relationships in multiple samples from seven patients with metastatic high-grade serous ovarian cancer and found that the majority of metastatic sites in each patient were clonally pure, and polyclonal metastases were found in only two patients (10). Comparison between our study and the McPherson study raises the possibility that the mechanisms of metastases may be tumor (colorectal cancer vs. ovarian cancer) or site (LN vs. omentum) specific.

The data presented here offer support for polyclonal LN metastasis due to multiple waves of seeding of the LNs from the primary tumor over time. It is important to understand the relationship between tumor subclones and thus intratumor heterogeneity and metastasis in order to understand response to therapy. None of the patients whose tumors were studied here had previously received chemotherapy, revealing that colorectal cancer metastases are highly complex even prior to treatment. How LN metastases relate to distant metastases is yet to be determined.

Multiple recent studies, including this one, have called into question the details of the adenoma to carcinoma pathway: namely, clonal sweeps after APC, KRAS, and then TP53 mutations as proposed in 1990 (22). For example, CP14 does not follow any simple model of clonal evolution (Fig. 2A). In a recent study, Thirlwell and colleagues looked at multiple adenomas and sporadic cancers and assessed the mutation status of certain drivers in individual crypts in the neoplasms; they found that the lesions were polyclonal, and that in constructed phylogenies, there was not a clear obligate order of genetic events (23). Additionally, a recent study by Gausachs and colleagues performed deep nanofluidic PCR in multiple tumors and found that different crypts from the same tumor contained different APC and KRAS mutations, and even found crypts that did not contain APC mutations, calling into question a clonal origin of colorectal cancer (24). Query in cBioPortal of the 212 The Cancer Genome Atlas (TCGA) tumors with sequencing and copy number variation data for mutations and copy number alterations in *APC, KRAS*, and *TP53* reveals that although 91% of colorectal cancers have an alteration in at least one of these genes, only 16.5% harbor alterations in all three (25). We see subclones that do not include previously validated drivers, perhaps providing evidence for previously unrecognized drivers, or evidence for other driver events (e.g., copy number, epigenetic modification, posttranslational protein modification) that are not accounted for in a more direct manner in bulk sequencing or SNP array profiling. Explanations for this also include inaccuracies in the algorithm. As noted in the results,

Table 3. Copy number events in LNs mapping to unique primary tumors

	CP08		CP11						CP14						CP15				CP17				CP18								CP19					
	LN3		LN1		LN6		LN7		LN2		LN7		LN9		LN2		LN3		LN1		LN7		LN5		LN6		LN7		LN8		LN1		LN2		LN4	
	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF	LogR	BAF
Chr1		T2											U	U	U	U										T1							U	U		
Chr2	T2	T2		U					T1	T1	T1	T1				U																				
Chr3	T2	T2								T1			U	U		U																				
Chr4	T2	T2																					T2		T2				T2			T2		T1		
Chr5	T2	T2							T1	T1	T1	T1		U		U												U								
Chr6		T2											U	U											U			U					U			
Chr7									T1	T1	T1	T1			T1										U											
Chr8									T1	T1	T1	T1				T2		T2							U			U					U			
Chr9		T2												U		U																	U		U	
Chr10	T2	T2																							U		U						U			
Chr11									T1	T1				U		U								T2				T2			U					
Chr12	T2	T2							T1	T1	T1	T1							U	U			U		U		T1		T1							
Chr13	T2	T2																						U							U					
Chr14																			U	U							U		U							
Chr15		T2														U							U		U	U	U	U								
Chr16		T2														U								U		U			U							
Chr17		T2							T1	T1	T1	T1								U			U		U	U	U	U								
Chr18	T2	T2							T1	T1	T1	T1	T1	T1																							
Chr19		T2							T1	T1	T1	T1		T1	U	U											U		U								
Chr20									T1	T1	T1	T1				T2		T2																			
Chr21																								U					U		U					
Chr22									T1	T1	T1	T1																									

NOTE: For each chromosome, the table lists whether there exists a copy number event detected in a lymph sample that is private to an individual primary tumor location. If so, the primary tumor ID is indicated (T1 or T2 for each patient; there were no unique distinguishing CNVs detectable in CP14 T3). If the lymph sample contains a unique copy number event not present in any primary tumor, a "U" is indicated. See Supplementary File 3 for full copy number profiles for each tumor.

we do identify potential drivers in these subclones. One drawback of PyClone is that the algorithm does not include copy number variation events in defining subclones that could be driver events.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: E.R. Fearon, K.M. Hardiman
Development of methodology: P.J. Ulintz, R. Wu, K.M. Hardiman
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.K. Greenson, R. Wu, K.M. Hardiman
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): P.J. Ulintz, J.K. Greenson, K.M. Hardiman
Writing, review, and/or revision of the manuscript: P.J. Ulintz, J.K. Greenson, R. Wu, E.R. Fearon, K.M. Hardiman
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): P.J. Ulintz, K.M. Hardiman
Study supervision: K.M. Hardiman

Acknowledgments

K.M. Hardiman is funded by 5P50CA130810, American Surgical Association Foundation Fellowship; 5P30A046592, John S. and Suzanne C. Munn Cancer Research Fund, and K08CA190645 from the National Cancer Institute. E.R. Fearon is funded by R01CA082223 and P30CA046592. The authors would like to acknowledge Jun Li and Sofia Merajver for thoughtful criticism as well as Chris Gates, Bob Lyons, Jeanne Geskes, Melissa Coon, Chris Krebs, Peter Graf, Sam Dougaparsad, Patricial Beals, and Zhiwei Che. The authors would also like to thank the authors of the informatics tools used in this study—PyClone, CITUP, and TAPS—for their work.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 17, 2017; revised September 22, 2017; accepted November 28, 2017; published first December 4, 2017.

References

1. Edwards BK, Noone AM, Mariotto AB, Simard EP, Boscoe FP, Henley SJ, et al. Annual Report to the Nation on the status of cancer, 1975–2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. Cancer 2014; 120:1290–314.
2. Hardiman KM, Ulintz PJ, Kuick RD, Hovelson DH, Gates CM, Bhasi A, et al. Intra-tumor genetic heterogeneity in rectal cancer. Lab Invest 2016;96:4–15.
3. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nat Genet 2015;47: 209–16.
4. Cheung KJ, Ewald AJ. A collective route to metastasis: Seeding by tumor cell clusters. Science 2016;352:167–9.
5. Baldus SE, Schaefer KL, Engers R, Hartleb D, Stoecklein NH, Gabbert HE. Prevalence and heterogeneity of KRAS, BRAF, and PIK3CA mutations in primary colorectal adenocarcinomas and their corresponding metastases. Clin Cancer Res 2010;16:790–9.
6. Brannon AR, Vakiani E, Sylvester BE, Scott SN, McDermott G, Shah RH, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. Genome Biol 2014;15:454.
7. Kim TM, Jung SH, An CH, Lee SH, Baek IP, Kim MS, et al. Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. Clin Cancer Res 2015;21: 4461–72.
8. Sebagh M, Allard MA, Bosselut N, Dao M, Vibert E, Lewin M, et al. Evidence of intermetastatic heterogeneity for pathological response and genetic mutations within colorectal liver metastases following preoperative chemotherapy. Oncotarget 2016;7:21591–600.
9. Cheung KJ, Padmanaban V, Silvestri V, Schipper K, Cohen JD, Fairchild AN, et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. Proc Natl Acad Sci U S A 2016;113:E854–63.
10. McPherson A, Roth A, Laks E, Masud T, Bashashati A, Zhang AW, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nat Genet 2016;48:758–67.
11. Frank DN. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. BMC Bioinformatics 2009;10:362.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
13. Hagemann IS, Devarakonda S, Lockwood CM, Spencer DH, Guebert K, Bredemeyer AJ, et al. Clinical next-generation sequencing in patients with non-small cell lung cancer. Cancer 2015;121: 631–9.
14. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. Cell Rep 2016;17:1206.
15. Mayrhofer M, DiLorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol 2013;14:R24.
16. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods 2014; 11:396–8.
17. Lamy P, Nordentoft I, Birkenkamp-Demtroder K, Thomsen MB, Villesen P, Vang S, et al. Paired exome analysis reveals clonal evolution and potential therapeutic targets in urothelial carcinoma. Cancer Res 2016; 76:5894–906.
18. Findlay JM, Castro-Giner F, Makino S, Rayner E, Kartsonaki C, Cross W, et al. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. Nat Commun 2016;7: 11111.
19. Malikic S, McPherson AW, Donmez N, Sahinalp CS. Clonality inference in multiple tumor samples using phylogeny. Bioinformatics 2015;31: 1349–56.
20. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
21. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat 2016;37:235–41.
22. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759–67.
23. Thirlwell C, Will OC, Domingo E, Graham TA, McDonald SA, Oukrif D, et al. Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. Gastroenterology 2010;138:1441–54, 1454e1–7.
24. Gausachs M, Borras E, Chang K, Gonzalez S, Azuara D, Delgado Amador A, et al. Mutational heterogeneity in APC and KRAS arises at the crypt level and leads to polyclonality in early colorectal tumorigenesis. Clin Cancer Res 2017;23:5936–47.
25. Colorectal Adenocarcinoma (TCGA, Nature 2012): Tumors with Sequencing and CNA Data (212 Samples/3 Genes) [dataset]. July 9, 2012 [cited 2017 Nov 7]. Available from: http://www.cbioportal.org/index.do?session_id=5a01d0ae498e5df2e297d7ad&show_samples=false&.