Single cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer

Marco L. Leung, Alexander Davis, Ruli Gao, et al.

Genome Res. published online May 25, 2017
Access the most recent version at doi:10.1101/gr.209973.116

Single Cell DNA Sequencing Reveals a Late-Dissemination Model in Metastatic Colorectal Cancer

Marco L. Leung[1,2,*], Alexander Davis[1,2,*], Ruli Gao[1], Anna Casasent[1,2], Yong Wang[1], Emi Sei[1], Eduardo Sanchez[3], Dipen Maru[3], Scott Kopetz[4] and Nicholas E. Navin[1,2,5]

[1]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[2]Graduate School in Biological Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[3]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[3]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[5]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX 77030.

* Authors contributed equally to the work

Corresponding author:

Nicholas E. Navin, Ph.D.

nnavin@mdanderson.org

Abstract

Metastasis is a complex biological process that has been difficult to delineate in human colorectal (CRC) cancer patients. A major obstacle in understanding metastatic lineages is the extensive intratumor heterogeneity at the primary and metastatic tumor sites. To address this problem, we developed a highly-multiplexed single cell DNA sequencing approach to trace the metastatic lineages of two CRC patients with matched liver metastases. Single cell copy number or mutational profiling was performed, in addition to bulk exome and targeted deep-sequencing. In the first patient we observed monoclonal seeding, in which a single clone evolved a large number of mutations prior to migrating to the liver to establish the metastatic tumor. In the second patient we observed polyclonal seeding, in which two independent clones seeded the metastatic liver tumor after having diverged at different time points from the primary tumor lineage. The single cell data also revealed an unexpected independent tumor lineage that did not metastasize, and early progenitor clones with the 'first hit' mutation in *APC* that subsequently gave rise to both the primary and metastatic tumors. Collectively, these data reveal a late-dissemination model of metastasis in two CRC patients, and provide an unprecedented view of metastasis at single cell genomic resolution.

Introduction

Metastasis is the primary cause of death in most human cancer patients (Mehlen and Puisieux 2006). Colorectal cancer (CRC) patients with primary tumors detected during colonoscopy often have good survival rates, but patients with late-stage (IV) disease have poor 5-year survival rates of only 11% (American Cancer Society 2015). Large-scale cancer genome sequencing efforts have identified genes that are frequently mutated in primary CRC tumors, including *APC*, *KRAS*, *NRAS*, and *TP53* (2012). In addition to these common mutations, many low-frequency

mutations have also been identified, suggesting extensive inter-patient heterogeneity *(2012).* Further work has begun to investigate the mutational concordance of matched primary and metastatic tumors in CRC patients by next-generation sequencing. In a study that profiled microsatellite-stable (MSS) CRC patients, a large number of mutations were reported as being concordant between the primary and metastatic tumors, in addition to a small number of metastasis-specific mutations (Brannon et al. 2014, Tan et al. 2015).

The metastatic cascade is a complex biological process in which tumor cells escape the primary organ site, intravasate the circulation, and disseminate to distant organs (Valastyan and Weinberg 2011). Several competing models of metastasis have been proposed: (1) late-dissemination, (2) early dissemination, and (3) self-seeding (Supplemental Fig. S1). The *late dissemination* model is a unidirectional model, in which tumor cells evolve for an extended period of time at the primary tumor site, before acquiring specific mutations that enable the clones to disseminate. The *early dissemination* model posits that tumor cells disseminate at the earliest stages of primary tumor growth, and that primary and metastatic tumors evolve in parallel (Klein 2009). An alternative model is *self-seeding*, which posits that tumor cells disseminate from the primary tumor, establish distant metastatic tumor sites, and then travel bi-directionally back to the primary tumor to promote its growth (Norton and Massague 2006).

Single cell DNA sequencing methods have emerged as powerful new tools for resolving intratumor heterogeneity and tracing clonal lineages during tumorigenesis (Navin 2015, Wang and Navin 2015). Our group reported the development of the first single cell DNA sequencing method (Single-Nucleus-Sequencing) and used this method to delineate aneuploidy evolution in breast tumors (Navin et al. 2011). Subsequent work from our group, and others, has led to the development of high-coverage single cell sequencing methods to detect genome-wide mutations at base-pair resolution (Xu et al. 2012, Zong et al. 2012, Wang and Navin 2015, Wang et al. 2014, Leung et al. 2016, Leung et al. 2015, Gawad, Koh and Quake 2016). Computational methods can be used to infer phylogenetic trees from single-cell sequencing

data (Davis and Navin 2016, Jahn, Kuipers and Beerenwinkel 2016, Ross and Markowetz 2016). However, a major challenge is that current single cell DNA sequencing methods are low-throughput and expensive. To address this challenge we developed a high-throughput single cell DNA sequencing method that utilizes library barcoding and a 1000 cancer gene panel to study clonal evolution during metastasis in two CRC patients.

Results

Experimental Approach

We selected frozen primary colon cancer and matched liver samples from two CRC patients with metastatic disease (Fig. 1A). Both patients were classified as microsatellite-stable (MSS) with invasive adenocarcinomas and late-stage (IV) disease (Methods). Nuclear suspensions were prepared and stained with DAPI for flow-sorting by ploidy. Cellular fractions were isolated by gating diploid (D) or aneuploid (A) distributions. In patient CRC1, the cell count histogram revealed a diploid (2N) and aneuploid (2.6N) distribution in the primary tumor and a diploid (2N) and aneuploid (2.9N) distribution in the liver metastasis (Fig. 1B). In patient CRC2, we identified a diploid (2N) and aneuploid (3.3N) distribution in the primary tumor, and a diploid (2N) and aneuploid (3N) distribution in the liver metastasis (Fig. 1B). Millions of cells from the D and A peaks were gated and flow-sorted for exome and targeted cancer gene panel sequencing in CRC1 and CRC2. Single nuclei were isolated by FACS for single cell copy number profiling or single cell mutational profiling (Fig. 1C). Single cell libraries were barcoded and pooled together (48 cells) for copy number profiling using SNS (Navin et al. 2011), or barcoded (96 cells) for highly-multiplexed targeted sequencing using a 1000 cancer gene panel (T1000) that captures 12,500 exons and promoter regions (Leung et al. 2016). The exome capture platform and the T1000 cancer gene panel only overlap within the exonic regions. The resulting libraries were

used for sequencing on the Illumina platform (Methods) and somatic variants were detected (Supplemental Fig. S2).

Bulk Primary and Metastatic Mutations Are Concordant

To investigate mutational concordance between the primary and metastatic liver tumors, we performed deep-exome sequencing of millions of diploid or aneuploid cells that were flow-sorted cells. To distinguish germline from somatic mutations, we also sequenced matched normal tissue (Methods). The exome libraries were sequenced at high coverage depth (75.5x) and breadth (97.33%), where *breadth* is defined as the percentage of the targeted region with physical coverage of 1X or higher read depth (Supplemental Table S1). We detected 127 mutations in patient CRC1, of which 90 were nonsynonymous, and 80 were shared between the primary and metastatic tumors (Fig. 2A). Shared mutations included *APC* and *KRAS*, while metastasis-specific mutations included *BOD1, TRRAP, GSTCD,* and *SNX19* (Supplemental Table S2). In patient CRC2 we identified 131 mutations of which 107 were nonsynonymous and 68 were shared between the primary and metastasis (Fig. 2B). Common mutations in CRC2 included mutations in *APC*, *TP53*, *CDK4*, *TOX, NRAS and MYH11*, while metastasis-specific mutations included *FUS*, *SPEN*, *DAPK1* and *FBN* (Supplemental Table S2). Our data also identified a number of nonsynonymous mutations that changed VAF between the primary and metastatic sites (Figs. 2C and 2D). These mutations may reflect clonal selection during metastatic dissemination or may be due differences in copy number. To distinguish between these two possibilities, we applied PyClone to normalize the VAFs by copy number events and calculate clonal frequencies. The resulting data suggest that a number of SNVs changed in frequency during metastatic dissemination, possibly due to selection at the metastatic tumor site (Supplemental Fig. S3).

Metastasis-specific Mutations Were Acquired After Dissemination

The metastasis-specific mutations may have occurred in a rare subclone of the primary tumor prior to dissemination, or alternatively after dissemination to the liver. To address this question, we performed ultra-deep targeted sequencing on a subset of metastasis-specific mutations in the primary tumor. From these data, we investigated whether the metastasis-specific mutations existed at low frequencies in the primary tumor mass. Targeted amplicon sequencing was performed at 1,368,403X mean coverage depth for 3 metastasis-specific mutations in CRC1 and 18 metastasis-specific mutations in CRC2 (Fig. 2E). Bayesian hypothesis testing and DeepSNV were used independently to determine the significance of each mutation frequency, relative to the background noise in the matched normal samples (Methods) (Gerstung et al. 2012). This analysis identified no significant increases in the variant read counts in the primary tumors relative to the matched normal tissue sample, with the exception of *GREB1*, in which DeepSNV reported a significant p-value (Supplemental Table S3, Fig. 2E). In contrast, all of VAFs between the metastasis and primary tumor were found to be significant ($p < 0.05$). These data suggest that most of the metastasis-specific mutations evolved after disseminating to the metastatic liver site. However, we cannot exclude the possibility that some of these mutations exist at frequencies below our detection sensitivity (1e-3) in the primary tumor.

Mutational Substructure of the Primary and Metastatic Tumors

To resolve the clonal substructure of the primary and metastatic tumors, we applied highly-multiplexed single-cell DNA sequencing (Leung et al. 2016) to profile point mutations in 372 single cells using a 1000 cancer gene (T1000) panel. The single cell sequencing data resulted in a mean coverage depth of 137x and average coverage breadth of 0.92 (Supplemental Table S4). In parallel, we sequenced millions of flow-sorted aneuploid tumor and normal cells using the T1000 cancer gene panel. To ensure the quality of single cell analysis, we filtered single cell data with low coverage depth and annotated variants based on variant/reference genotype read ratios (Methods, Supplemental Fig. S2). In total, we analyzed 178 and 182 single cells for

CRC1 and CRC2, respectively. We first compared the single cell mutation data on the T1000 platform to the bulk exome data (V2 exome capture platform), and found 100% concordance for mutations in the exonic regions (Supplemental Table S5).

To broadly identify subpopulations of cells that shared common mutations, we performed multi-dimensional-scaling (MDS) analysis using the cells sequenced with the T1000 platform (Figs. 3A and 3B). In both patients, we identified 3 major clusters of cells that corresponded to normal cells (N), primary tumor cells (P) and metastatic tumor cells (M). The normal cell clusters included diploid cells from both the primary colon and liver metastasis. The P clusters consisted mainly of aneuploid cells from the primary tumor, while the M cluster consisted of aneuploid tumor cells from the liver. However, a few cells sorted from the diploid fractions clustered with the primary aneuploid tumor cells in both patients, suggesting that they may have been missorted during FACS.

To more carefully delineate the clonal architecture, we used 2-dimensional hierarchical clustering to identify groups of single cells with similar mutational profiles (Figs. 3C and 3D, Supplemental Fig. 4A). Consistent with the MDS analysis, hierarchical clustering identified three major clusters of tumor cells in CRC1: the normal diploid cells (N), the primary aneuploid cells (P) and the metastatic aneuploid cells (M). Additionally, we identified a small subcluster of diploid cells (E) that were not evident in the MDS analysis (Fig. 3C). The aneuploid tumor cells from the primary and metastatic sites shared 10 common mutations, including driver mutations in *APC*, *TP53* and *KRAS*. These data also identified five metastatic-specific mutations (*ZNF521*, *RBFOX1*, *TRRAP*, *GATA1*, *EYS*) and 1 primary-specific mutation (*TPM4*). Most of the diploid cells did not have mutations, suggesting that they are normal stromal cells. However, we did identify a rare subcluster (E) that consisted of three diploid cells with a single heterozygous nonsense mutation in *APC* (c.4012C>T).

In patient CRC2, the clustered heatmap identified 6 major subpopulations: normal diploid cells (N), primary aneuploid cells (P) and three metastatic subpopulations (MP1, M2, M3), in

addition to a minor independent subpopulation (I) (Fig. 3D, Supplemental Fig. 4B). In total, we identified 14 common mutations that were shared between the primary and metastatic tumors, including driver mutations in *NRAS*, *APC*, *TP53*, *FHIT* and *CDK4*. We also identified two primary specific mutations (*LINGO2*, *LRP1B*) and 14 metastasis-specific mutations (including *SPEN, PIK2CG, FUS* and *HELZ*). Multiple mutations were detected in *LINGO2*, which made us suspect that they might be technical artifacts from PCR or sequencing, however we found no evidence of this by analysis of strand bias, low coverage depth or coinciding with regions of poor mappability. In CRC2 the metastatic tumor was composed of three major subpopulations (MP1, M2 and M3). The MP1 subpopulation consisted of both primary tumor cells and metastatic tumor cells, while M2 and M3 were composed of only metastatic tumor cells.

Identification of a Rare *APC* Progenitor Subclone

Unexpectedly, the single cell mutational data in CRC1 identified a rare subpopulation of three tumor cells (PD16, PD41, PDD93) that had diploid copy number and contained a single heterozygous mutation in *APC* (c.4012C>T, p.Gln1338Ter). This mutation was present in all of the subsequent primary and metastatic tumor cells (Fig. 3C). The three early tumor cells did not show evidence of harboring any of the other point mutations (e.g. *KRAS, TP53, TCFL2*) that were present in the major primary and metastatic tumor cells (Fig. 4). Interestingly, the heterozygous *APC* mutation was found to be homozygous in the aneuploid tumor cells, likely due to a hemizygous copy number loss that occurred in the later stages of tumorigenesis. These data suggest that *APC* was likely the first ‘hit’ that initiated the colorectal tumor in this patient. The ancestral clones subsequently underwent genome-wide aneuploidy and expanded to form both the primary and metastatic tumors.

Copy Number Substructure of the Primary and Metastatic Tumors

To investigate the copy number substructure of the primary and metastatic tumors we performed SNS (Navin et al. 2011). In total, 32 single nuclei were analyzed from CRC1 and 42 single nuclei from CRC2. Single cell copy number profiles were calculated from read depth at 220kb resolution (Methods). To identify clusters of cells that shared similar profiles, we applied MDS which revealed three major clusters, representing normal diploid cells (N), primary tumor cells (P) and metastatic tumor cells (M) in both patients (Figs. 5A and 5B). In CRC1, the primary and metastatic clusters were discrete, suggesting only minor genomic variation. However in patient CRC2, the metastatic cluster showed considerable cell-to-cell variation (rho = 0.80, Mean Spearman Correlation) compared to the primary tumor cell cluster (Rho = 0.88, Mean Spearman Correlation), indicating a significant amount of intratumor heterogeneity in the metastasis.

We performed a more detailed analysis of the copy number substructure using 1-dimensional hierarchical clustering (Methods). In patient CRC1, the primary and metastatic cells shared highly similar profiles, including amplifications of several known oncogenes *(EGFR, MET, CDK6, CDX2, WNT2 CDK8, ZNF217*) and deletions of tumor suppressors (*CTNNB1, APC, TP53, SMAD4, TP53*) that have previously been reported in colon cancer (Xie et al. 2012) (Fig. 5C). However, the primary tumor cells in CRC1 also contained an additional amplification of chromosome 17q (*ERBB2*) and a 1.4 Mb homozygous deletion on chromosome 4q32.3 that were not present in the metastasis. Similarly, the metastatic tumor cells showed an additional 47 Mb amplification on the X Chromosome that included the androgen receptor.

In patient CRC2, we identified a single cluster of normal diploid cells (N), a single cluster of major clones in the primary tumor (P), and two major clones (M1, M2) in the liver metastasis. The primary and metastatic tumor cells shared a large number of common CNAs, including amplification of oncogenes including *CDX2*, *CDK8*, *JAK3* and *ZNF217* (Fig. 5D). The CNAs distinguishing the primary and metastatic tumor cells included an additional amplification of chromosome 9 (*JAK2, CDKN2A*) in the primary cells, and amplifications of chromosomes 3q, 8q

and 13p. While the primary tumor cells were highly clonal, the metastatic tumor cells clustered into two major subpopulations (M1, M2) that were distinguished by an amplification on 3q (*ETV5*, *PIK3CA, BCL6*) in M1 and amplification of chr8 in M2. To further investigate the genetic relationship between single cell copy number profiles, we constructed phylogenetic trees using FastME (Lefort, Desper and Gascuel 2015, Nilsen et al. 2012), which were highly consistent with the topologies of the hierarchical trees. (Supplemental Fig. S5).

Phylogenetic Analysis Reveals Late Dissemination and Polyclonal Seeding

To reconstruct clonal lineages during metastatic dissemination, we computed phylogenetic mutation trees using SCITE (Jahn et al. 2016). SCITE uses a Markov Chain Monte Carlo (MCMC) algorithm to construct optimal mutation trees and then reattaches single cells at the nodes (See methods). In patient CRC1, the mutation tree shows a linear series of mutations that occurred as the primary tumor mass evolved and seeded the metastatic tumor (Fig. 6A). The tumor initiated through a first 'hit' in *APC* and subsequently evolved mutations in the *KRAS* oncogene, *TP53* tumor suppressor and *CCNE1* oncogene, as well as 6 additional somatic mutations, and expanded to form the primary tumor mass. In the late stages of the primary tumor lineage, monoclonal seeding occurred, in which a single clone diverged and migrated to the liver, where it established the metastatic tumor. The point of metastatic divergence occurred after the acquisition of *POU2AF1* mutation in the primary tumor lineage.

In patient CRC2, we observed a more complex metastatic lineage in which late dissemination occurred as well as polyclonal seeding of two independent clones that established the metastatic liver tumor (Fig. 6B). The primary tumor initiated from the normal cells via mutations in *TP53*, *APC*, *NRAS* and *CDK4*. These early truncal mutations (and others eg. *TOX*, *MYH11*) lead to the expansion of the primary tumor mass. Data from the bulk exome sequencing also supports that these heterozygous mutations are truncal and occurred early in the lineage, with mutations frequencies of approximately 0.5. The first clone disseminated after

acquiring a mutation in *MN1* in the primary tumor and seeded the metastatic liver tumor where the tumor cells continued to evolve a number of metastasis-specific mutations (eg. *IL7R, PIK3CG*, *SPEN* and *F8*, *PTPRD*). During this time, the primary tumor cells continued to evolve in parallel with the first metastasis and acquired additional mutations in *CHN1, FHIT, ATP7B* and a second nonsense mutation in the *APC* tumor suppressor. The advanced primary tumor cells subsequently underwent a second seeding event after acquiring the *ATP7B* mutation. The second clone evolved in parallel to the first clone in the metastatic liver site, and acquired additional mutations in *NR4A3, FUS, PRKCB*, *HELZ,* and *TSHZ3* leading to further expansion of the liver tumor mass.

To more rigorously evaluate the accuracy of the SCITE tree and evidence for polyclonal seeding in CRC2, we performed a statistical analysis of the 4 'bridge mutations' in the primary tumor (*CHN1*, *FHIT*, *APC* and *ATP7B*) that occurred between the first and second metastatic seeding events (Supplemental Fig. S6). We performed a mixture-model Bayesian binomial test (Methods) of the reference and variant read counts to determine if the bridge mutations were present in the primary tumor and the second metastasis, but absent in the first metastasis as indicated by the SCITE tree. The resulting probability heatmap and read count data suggest that all four mutations were present in the primary tumor, and provided strong evidence that *FHIT* and *ATP7B* were present in 10/13 and 13/13 tumor cells in the second metastasis and absent in the first metastasis (detected in only 1/15 and 1/15 cells), supporting a two independent seeding events. However, this analysis also showed some uncertainty regarding the placement of the *APC* and *CHN1* in the SCITE tree lineages, which may have not occurred between the first and second metastatic seeding events. We also investigated whether the 4 bridge mutations may have been lost in the second metastasis due to chromosomal deletions or LOH. Our data show that the copy number states did not change in the first and second metastasis (APC, CN=3; ATP7B, CN=4, CHN1, CN=3, FHIT, CN=2), and that the B-allele frequencies did not support copy-neutral LOH for these mutations in metastasis 1

(CHN1=0.279, FHIT=0.245, APC=0.268, ATP7B=0.33), suggesting that their absence is unlikely to be explained by chromosomal loss.

To better understand the potential error rates based on the ordering of the SCITE tree we calculated genotype matrices (Supplemental Fig. S7, Methods). Our data suggest that the false negative error rate for CRC1 is 7.89%, while the false positive rate is 1.52%. In CRC2 these data suggested a false negative error rate of 12.56% and a false positive error rate of 1.74%. These error rates are low for single cell DNA sequencing data and suggest that the technical noise does not greatly confound the inference of the tree topologies.

Integrated Phylogenetic Trees

To better understand the timing of the CNA events relative to the mutational lineages, we integrated the two phylogenetic trees (Supplemental Fig. S8). In CRC1 these data suggest that the majority of CNA events were acquired early in the tumor lineage, after the *APC* mutations occurred. However to integrate the copy number and mutation data in CRC2, we first needed to determine which copy number subpopulations (M1 and M2) matched the two mutation subpopulations (first, second). To address this question, we performed a statistical analysis of the sequence read density data for a marker (chr3q) in the single cell mutation data that distinguished the CNA profiles. Our data showed that the coverage depth was significantly increased on chr 3q ($p = 0.006081$) in the tumor cells from the first metastasis relative to the second metastasis, suggesting that it corresponded to the M1 copy number subpopulation (Supplemental Fig. S9). After integrating the two trees, the inferred copy number and mutation tree in CRC2 suggested that at least two major genomic instability events occurred: one event occurred at the earliest stages of tumor evolution, while the other event occurred in the primary tumor, after both metastatic seeding events.

Evolution of an Independent Primary Tumor Lineage

In patient CRC2, the single cell mutation trees revealed an unexpected lineage that evolved independently and in parallel to the main tumor lineages (Fig. 6B). This rare subpopulation consisted of 9 diploid tumor cells that evolved mutations in *ALK, ATR, EPHB6, NR3C2* and *SPEN* and did not share any mutations with the major primary or metastatic aneuploid tumor cells (eg. *APC, NRAS* or *TP53*). These diploid tumor cells did not achieve prevalence in the primary tumor mass, nor did they metastasize to the liver. In summary, these data suggest that 8 tumor cells represent a completely independent lineage that can be traced back to a different initiating cell in the normal colon tissue, and evolved in parallel to the main tumor lineage.

Discussion

In this study we applied single cell DNA sequencing, exome sequencing and targeted deep-sequencing to study clonal evolution during metastatic dissemination in two colon cancer patients. In both patients, our data support a late-dissemination model of metastasis, in which the primary tumor cells evolved for an extended period of time and acquired many mutations (e.g. *KRAS, NRAS, APC* and *TP53*) and CNAs prior to disseminating to distant organ sites. The late-dissemination model is consistent with genomic data from pancreatic cancers (Yachida et al. 2010) and prostate cancers (Gundem et al. 2015) that report metastatic clones emerging in the later stages of primary tumor growth. In contrast to bulk sequencing methods, our single cell data was able to distinguish between the self-seeding (bi-directional migration) and early dissemination models of metastasis, for which we found no empirical evidence.

A major question in the field is whether metastatic tumors are seeded from a single clone (monoclonal seeding) or from multiple clones (polyclonal seeding) over the course of the disease. The data from CRC1 was consistent with monoclonal seeding, however, in CRC2, we observed polyclonal seeding of two independent clones that established the metastatic liver tumor. The first clone disseminated after acquiring many of the salient driver mutations (*APC, NRAS, TP53, CDK4)* in the middle of the primary tumor lineage, while the second clone evolved

337 additional mutations prior to disseminating to the liver. These data are consistent with a multi-
338 region sequencing study in which both monoclonal and polyclonal seeding were observed in
339 different prostate cancer patients during metastasis (Gundem et al. 2015).
340 Our single cell sequencing data revealed several unexpected findings. In CRC1 we
341 identified a rare subpopulation of diploid cells (3/112) that carried a heterozygous nonsense
342 mutation in *APC*, but showed no evidence of any other somatic mutations. This *APC* mutation
343 represents the 'first hit' that initiated tumorigenesis in the colon epithelium and subsequently
344 gave rise to the primary tumor and liver metastasis. Interestingly, these cells were diploid,
345 suggesting that they had not yet undergone the complex aneuploid rearrangements observed in
346 tumor cells. These data are consistent with the original model of colon cancer progression
347 proposed over two decades ago, which posited that *APC* was the first hit that initiated colon
348 cancer, prior to *KRAS* and *TP53* mutations (Fearon and Vogelstein 1990). What is surprising is
349 that these progenitor subclones remained in the advanced carcinoma at a relatively high
350 frequency (2.6%) and were not outcompeted by other tumor clones, suggesting that they had a
351 high fitness.
352 Another unexpected observation was an independent tumor lineage in CRC2. In the
353 primary tumor, we observed a small subpopulation of diploid tumor cells that harbored a
354 completely different set of mutations than the main tumor lineage. This independent
355 subpopulation did not achieve prevalence in the primary tumor mass and did not metastasize to
356 the liver. Phylogenetic analysis suggests that the tumor cells can be traced back to a different
357 initiating normal cell in the colon tissue. These data are in contrast to the vast majority of tumor
358 lineage studies published to date which frequently (98.4% in 312 patients), report a set of
359 truncal mutations that can be traced back to a single initiating normal cell (Gerlinger et al. 2012,
360 Yates and Campbell 2012, Newburger et al. 2013, Wang et al. 2014, Zhang et al. 2014,
361 McPherson et al. 2016). However our independent lineages data are consistent with a few
362 uncommon reports (~1.6% of 312 patients) on tumor lineages, including deep-sequencing data

of eyelid skin (Martincorena et al. 2015) and multi-region sequencing data from a single patient with lung cancer (de Bruin et al. 2014), a single patient with prostate cancer (Boutros et al. 2015) and 2 patients with multifocal prostate cancer (Cooper et al. 2015).

A late-dissemination model has several important clinical implications. This model is consistent with the clinical observation that treatment and surgical excision of local disease (even when the primary tumor is very advanced) can prevent the development of metastatic disease. Such intervention would not be possible in the context of an early dissemination model, in which tumor cells would have already disseminated to distant organ sites at the earliest stages of the local disease. Another important clinical implication is that late-dissemination implies that the primary and metastatic tumors share the majority of clinically relevant mutations. This is an important feature, since it means that a diagnostic biopsy of the primary tumor will be representative of the metastatic sites. Indeed, this was the case in both CRC patients in which the driver mutations (*APC, KRAS, TP53, NRAS, CDK4*) were found in both the primary and metastatic organ sites. These data are also consistent with previous NGS data that have reported a high concordance of primary and metastatic tumor mutations (Brannon et al. 2014, Tan et al. 2015).

While pioneering, our study also has several limitations. One notable limitation is that we analyzed only two CRC patients, and therefore our study represents a proof-of-concept that late-dissemination models of metastasis can occur in colon cancer, but should not be interpreted as a common model in all CRC patients yet. Second, our studies examined only a single metastatic site (in the liver) and therefore we did not investigate seeding events to other common organ sites, such as the lung, brain, bones or peritoneum. This will require samples collected from a warm autopsy program (Lindell, Erlen and Kaminski 2006).

In closing, this study provides an unprecedented view of metastasis in colon cancer patients at single cell genomic resolution. Our study provides a comprehensive framework for studying the complexities of metastatic lineages that can be extended to many human cancer

types. Such studies will soon become feasible as the cost and time for analyzing the genomes of thousands of single cells in parallel is realized through the development of new high-throughput technologies (Baslan et al. 2015, Leung et al. 2016, Zahn et al. 2017, Vitak et al. 2017). In the near future, the translation of these technologies into clinical practice will undoubtedly have a profound impact on reducing morbidity in cancer patients with metastatic disease.

Methods and Materials

Patient samples

Frozen tumor samples from two CRC patients (CRC1 and CRC2) were obtained from the MD Anderson Tumor bank. CRC1 is a 77-year-old CRC patient with invasive moderately to poorly differentiated adenocarcinoma with liver metastasis. CRC2 is a 64-year-old CRC patient with invasive moderately differentiated adenocarcinoma with liver and lung metastasis. Both patients had metastatic disease diagnosed synchronously with the primary tumor. Neither patient received chemotherapy until after resection of both the primary and metastatic tumors.

Single Cell Isolation

Nuclear suspensions were prepared from frozen tumors using an NST/DAPI buffer (800mL of NST (146mM NaCl, 10mM Tris base at pH 7.8, 1mM $CaCl_2$, 0.05% BSA, 0.2% Nonidet P-40 and 21mM $MgCl_2$)), 200mL of 106 mM $MgCl_2$ and 10mg DAPI. Sectioned tumors were cut and minced using surgical blades in a Petri dish in NST/DAPI buffer in the dark. Samples were filtered through a 36-µm plastic mesh to a 5-mL polystyrene tube. Nuclei were then sorted using FACS Aria II (BD Biosciences) and single nuclei were deposited into individual wells on a 96-well plate for whole-genome amplification.

Single Cell Genome Amplification

For copy number profiling, single cells were amplified using DOP-PCR following the SNS protocol as previously described (Navin et al. 2011, Baslan et al. 2012). For mutational profiling, single cell multiple-displacement-amplification (MDA) was performed using a 2:3 ratio of lysis buffer (200mM KOH, 50mM DTT):1xPBS solution. 3.5µL of solution was loaded into each well of a 96-well plate. After flow sorting, the plate was centrifuged at 130g for 1 minute at room temperature. 1.5µL of neutralization buffer (900mM Tris-HCl, 300mM KCl, 200mM HCl) was added into each well and centrifuged. MDA was performed using Φ29 polymerase (NEB, M0269L) with 1mM hexamers (with phosphorothioate modification at the last 2 bases) and 1mM dNTP (NEB, N0446S). Final reaction volume was 50µL per well. The MDA incubation was time-limited to 3 hours at 30°C and 65°C for 3 minutes. A detailed description of the protocol and buffers was published by Leung *et al*. (Leung et al. 2016)

Library Construction

Whole-genome amplified DNA was fragmented using the Covaris Sonicator to 250 bp and purified by Zymo DNA Clean & Concentrator Column Kit (Zymo, D4004) according to manufacturer's instructions. Barcoded next-generation sequencing libraries were constructed using the NEBNext end repair model (NEB, E6050L), dA-tailing module (NEB, E6053L) and quick ligation module (NEB, E6056L). Libraries were amplified via PCR using NEBNext HiFi2x PCRmix (NEB, M0541L). Targeted capture for single cells was performed using Nimblegen SeqCap EZ Choice Library, according to Leung *et al*. (Leung et al. 2016). Exome capture for CO5 population was performed on single cell sequencing libraries using the TruSeq Exome Enrichment Kit (Illumina, 15013230) following manufacturer's instructions. Exome capture for CO8 population was performed using Nimblegen SeqCap EZ Exome V2 kit (Roche, 05860482001). For exome or targeted-capture sequencing, samples were sequenced on a 100 pair-end flowcell on the Illumina HiSeq4000 system. For copy number profiling, barcoded

libraries were pooled using equimolar concentrations and sequenced at 76 single-read flowcell on Illumina HiSeq 2000 system.

Sequencing Data Alignment and Processing

The FASTQ file was demultiplexed for each single cell library using our custom software (deplexer.pl). Individual FASTQ files were aligned to the human genome reference assembly (HG19) using Bowtie 2 (Langmead and Salzberg 2012), and converted to BAM files using SAMtools (Li et al. 2009). BAM files were processed by Picard to remove PCR duplicates. Re-alignment was performed around indel regions using the Genome Analysis Toolkit (GATK) (McKenna et al. 2010). Sequencing reads with mapping quality lower than 40 were removed. To calculate coverage metrics, we used a custom Perl script (cal-coverage_metrics.pl), which uses BEDTools (Quinlan and Hall 2010) to calculate coverage depth and breadth. *Coverage breadth* is defined as the percent of the genome or targeted regions with at least 1X depth. The aforementioned scripts can be downloaded from the Leung *et al*. paper published in *Nature Protocols*.

Variant Detection and Filtering

GATK was used to detect variants and generate a multi-cell VCF file. GATK was also used to recalibrate variant quality scores. We ran GATK with default parameters for depth (maximum read coverage = 250x). Mutations were filtered out and removed from analysis by consensus filtering (mutation must occur in at least three cells) and clustered regions (multiple mutations are detected within a 10-bp window). Variant annotation was performed on the VCF4 file using ANNOVAR (Wang, Li and Hakonarson 2010). For matched normal bulk sample sequencing, each site with less than or equal to 100x was required to have at least 6x coverage, in which at least 3 reads were required to have variants. For sites with more than 100x coverage depth, at

least 3% of reads were required to have variants. For single cell samples, sites were required to have a minimum of 10x coverage. For 10-20x coverage, we required at least 10 variant reads. For 20-100x coverage depth, at least 30% of reads were required to have variants. For sites with 100-250x, at least 20% of reads were required to have variants. Sites excluded due to low (<10X) coverage were labeled as missing values (NA), whereas other non-variant sites were labeled as reference. Please see Supplemental Fig. S2 for a detailed flow chart of these steps.

Clustered Mutation Heatmaps

Single cell mutation heatmaps were constructed using 2-dimensional hierarchical clustering, using the heatmap.2 function from the 'gplots' package available on CRAN (Team 2013)(www.cran.r-project.org). The row and column distance was calculated by using dist(method = "Euclidean") function and clustering was performed using hclust(method = 'ward'). A passcode was assigned, with trinary values (0,1,2, representing homozygous reference, heterozygous, and homozygous variant, respectively) to generate a genotype matrix from the VCF file. The single cell genotype matrix was filtered to reduce technical errors. First, variants were removed if they appeared three times or less across all single cells. Second, a variant was retained only if it has coverage in at least 75% of cells in the heatmap. Variants were then filtered if they occurred four times or less across all cells. For the remaining variant sites, we recovered true mutations at low coverage regions, requiring a minimum of at least 3 variant reads to call the mutation. For sites with more than 100x, at least 3% of reads were required to be variant reads. Finally, false-positive errors in regions of poor mappability were annotated as errors (dark grey) in the final heatmap.

Single Cell Integer Copy Number Calculation

Single cell copy number profiles were calculated from sequence read depth as previously described using a 'variable bin' method (Navin et al. 2011, Baslan et al. 2012). The variable binning intervals reduce mappability bias and false deletion of CNA events when compared to scaffolds using fixed length-fixed bins. The median genomic length spanned by each bin is 220 kb. A blacklist of systematic aberrant bins was filtered to remove false-positive amplifications near the centromeric and telomeric regions. Absolute ratios were calculated as read counts per bin divided by the median read counts across the whole genomic bins, followed by Loess normalization to correct for GC bias (Baslan et al. 2012). For population segmentation, bincounts were divided by the mean bincount for each cell, and log2 was taken, to produce log-ratio values. For each patient, all log-ratio profiles were segmented by estimating shared changepoints using the R "copynumber" package, version 1.10.0 with regularization parameter $\gamma = 40$ (Nilsen et al. 2012). Copy number profiles were scaled to have mean equal to the ploidy of the originating tumor, as estimated by flow cytometry. Profiles that lacked CNAs were assumed to be tumor stroma, and were scaled to have ploidy 2. Scaled values were rounded to the nearest integer to yield integer copy numbers using custom R scripts (Supplemental Scripts).

Single Cell Copy Number Clustering

Pairwise Euclidean distances were calculated from the single cell copy number data matrix (log2(ratio+0.1)) and then used for hierarchical clustering using ward-linkage in R using the heatmap.3 function from the 'gplots' package available on CRAN (www.cran.r-project.org). (www.r-project.org).

Balanced Minimum Evolution Copy Number Tree

To normalize segment size and prevent large segments from contributing too much weight, the vector of segment means was used to construct an event matrix for phylogenetic inference.

Pairwise distances were calculated using Manhattan distance rather than Euclidean distance to avoid large contributions from measurement error in small segments. Phylogenetic inference was performed using the balanced minimum evolution algorithm (Lefort, Desper and Gascuel 2015), implemented in the R package “ape”, version 3.5 (Paradis, Claude and Strimmer 2004).

Multi-dimensional-Scaling Analysis

MDS plots were constructing in R using the single cell genotype binary matrix with columns as single cells and rows as mutations. Classical Multidimensional Scaling was performed with the following command: cmdscale(x, eig=TRUE, k=2).

Inference of Single Cell Mutation Trees

Mutational trees of single cells were calculated using SCITE and redrawn using Cytoscape (Shannon et al. 2003, Cline et al. 2007, Jahn et al. 2016). The binary genotype matrix of single cells and point mutations with missing values was used for tree inference. SCITE was run using a false positive rate of 10%, a prior for allelic dropout rate with mean 30% and standard deviation 10%, one repetition, a chain length of 500,000, a 10% chance of proposing a new allelic dropout rate in each MCMC step, and a seed of 225 for the random number generator. Cells were attached to the resulting mutation tree in their maximum likelihood positions, breaking ties by placing them closer to the root, using a modified version of SCITE’s output code and a custom R script (Supplemental Scripts). The resulting phylogenetic tree was plotted using Cytoscape.

Bayesian Probabilities for Deep-Sequencing Variants

The significance of differences in amplicon deep sequencing of the normal and primary was

determined using Bayesian hypothesis testing. Variant read counts were modeled using the beta-binomial distribution:

$$V_{N,i} \sim \text{Beta-Binomial}(\alpha = \tau p_{N,i}, \beta = \tau(1 - p_{N,i}), n = n_{N,i})$$

$$V_{P,i} \sim \text{Beta-Binomial}(\alpha = \tau p_{P,i}, \beta = \tau(1 - p_{P,i}), n = n_{P,i})$$

where i is the index of a mutation, $V_{N,i}$ and $V_{P,i}$ are the number of variant reads observed in the normal and primary respectively, $n_{N,i}$ and $n_{P,i}$ are the total number of reads sequenced, $p_{P,i}$ and $p_{N,i}$ are the unknown true variant read frequencies, and τ is an unknown shared overdispersion parameter. If neither the normal nor the primary have the variant, then $p_{P,i}$ and $p_{N,i}$ are expected to be equal and represent the false positive rate of the sequencing experiment.

For Bayesian hypothesis testing, the prior distribution used was:

$$I_i = \text{Bernoulli}\left(\frac{1}{2}\right)$$

$$p_{N,i} \sim \text{Uniform}(0,1)$$

$$\begin{aligned} &\text{if } I_i = 1: \quad p_{P,i} = && p_{N,i} \\ &\text{if } I_i = 0: \quad p_{P,i} \sim && \text{Uniform}(0,1) \end{aligned}$$

$$\tau \sim \text{Exponential}(\lambda = 0.01)$$

where I_i is the indicator function of $p_{P,i} = p_{N,i}$. τ has a vague prior appropriate without prior information about its likely values. The probability that there is no true difference is $P(I_i = 1)$. $P(I_i = 1)$ was calculated jointly for all mutations i with an MCMC algorithm. The MCMC was computed with rJAGS (https://cran.r-project.org/web/packages/rjags/index.html) (Supplemental Scripts). One chain was used, with 1000 adaption iterations, and a chain length of 1000. Significant difference was defined as $P(I_i = 1) \leq .05$. For comparing amplicon deep

sequencing of the normal to exome sequencing of the metastasis, the same method was used to determine significance, but with separate overdispersion parameters for the two samples to reflect the difference between the experiments.

DeepSNV for Deep-Sequencing Variants

Statistical significance of observed variants was calculated using deepSNV version 1.16.0, which detects variants assuming a beta-binomial model (Gerstung et al. 2012). To estimate the overdispersion parameter of the model, data from the targeted sites plus flanking regions of 20bp on either side were used. DeepSNV was used to calculate p values for the null hypothesis that the targeted variant was equally frequent in primary tumor and control using separate one-tailed likelihood ratio tests for each strand orientation, and combining the p-values using Fisher's method. The code applying DeepSNV included in Supplemental Scripts.

Posterior Probability for Bridge Mutations

"Bridge mutations" were defined as those mutations occurring between the two metastatic seeding events in CRC2, and estimated as the mutations between the two branchpoints in the mutation tree. Cells sequenced with the T1000 panel were used for the analysis. Cells sorted from the aneuploid peak were grouped into categories of "primary", "first metastasic seeding", and "second metastasic seeding" on the basis of their attachment positions in the mutation tree. Reference and variant read counts were retrieved for these mutations. For ease of visualization, cells in the matrix of read count were colored according to an estimate of posterior probability that a variant is present. Posterior probabilities were calculated using the following statistical mixture model:

$$V_{ij} \sim \begin{cases} \text{Beta-Binomial}\left(\alpha = \tau\phi_j, \beta = \tau\left(1-\phi_j\right), n = n_{ij}\right) \text{ if } I_{ij} = 0 \\ \text{Beta-Binomial}\left(\alpha = \tau\psi_j, \beta = \tau\left(1-\psi_j\right), n = n_{ij}\right) \text{ if } I_{ij} = 1 \end{cases}$$

where i is the index of a cell, j is the index of a mutation, V_{ij} is the number of variant reads in cell i at mutation j, τ is an overdispersion parameter, ϕ_j is the false positive rate (probability that a read carries mutation j given that it is from a reference site to), ψ_j is the true variant allele frequency of mutation j in individual cells (assumed to be the same for each cell carrying the mutation), n_{ij} the total number of reads sequenced at a site, and I_{ij} the indicator of cell i carrying mutation j.

The prior was as follows:

$$\phi_j \sim \text{Beta}(1,7)$$

$$\psi_j \sim \text{Beta}(4,4)$$

$$\tau \sim \text{Exponential}(\lambda = 0.01)$$

$$I_{ij} = 0 \text{ if cell } i \text{ is from the diploid FACS peak}$$

$$I_{ij} \sim \text{Bernoulli}\left(\frac{1}{2}\right) \text{ if cell } i \text{ is from the aneuploid FACS peak}$$

For each mutation j, $P(I_{ij} = 1|\text{data})$ was calculated jointly for all cells i with an MCMC algorithm. The MCMC was computed with rJAGS (https://cran.r-project.org/web/packages/rjags/index.html) (Supplemental Scripts). One chain was used, with a chain length of 10,000. These posterior probabilities were used to determine colors of sites in visualizing bridge mutation variant read counts in the heatmap.

Inference of Errors in Single Cell Genotypes

Theoretical single cell genotype matrices were constructed by considering a cell to have a mutation if the cell node is a descendant of the mutation node on the single-cell mutation tree. A

site was considered to be a false negative if it is marked as mutated in the theoretical genotype matrix but as not mutated in the observed genotype matrix inferred from the data. A site was considered to be a false positive if it is marked as not mutated in the theoretical genotype matrix but as mutated in the observed genotype matrix. The matrix of errors was plotted using ggplot2 (Wickham 2009). The R code for inferring errors is provided in the Supplemental Scripts.

Copy number and LOH analysis of T1000 cells in CRC2

Classification of cells in the first or second metastasis was defined by the SCITE lineage tree. To determine which metastatic subclone carried a 3q amplification, ratio values for each exon were calculated (read depth divided by average read depth of exons in cell), median ratio values within the 3q amplification region (defined as segments 30 and 31) were calculated for each cell, and the difference between cells in the first and second metastasis was tested using a Wilcoxon rank sum test. Calculation of read depths was performed using GNU Parallel (Tange 2011) and BEDTools (Quinlan and Hall 2010). Heterozygous SNPs were defined as those with at least 10 reads supporting both variant and reference in combined bulk exome from matched normal samples. For each bridge mutation, B-allele frequency in the first metastasis was calculated as the average of minimum(p,1-p) across first metastasis cells for each heterozygous SNP site on the same copy number segment.

Figure Legend

Figure 1 - Single Cell and Bulk Population Experimental Workflow

(A) The frozen primary tumors and liver metastasis from two CRC patients were dissociated into nuclear suspensions and stained with DAPI (B) Single nuclei and populations of cells were gated and flow-sorted by ploidy distribution (C) To detect mutations, single nuclei were amplified by MDA and libraries were captured using the T1000 cancer gene panel, while copy number

623 detection was performed on single nuclei using DOP-PCR. Millions of cells were isolated in
624 parallel for standard exome sequencing. Barcoded libraries were constructed and captured for
625 targeted cancer gene panels or exome panels. Libraries were pooled for next-generation
626 sequencing on the Illumina platform.
627
628 **Figure 2 - Concordance of mutations in bulk primary and metastatic tumors**
629 (A,B) Scaled Venn diagrams reflect the total number mutations (synonymous and
630 nonsynonymous) identified by exome sequencing of the bulk flow-sorted tumor cells from the
631 primary and metastatic tumors. (C,D) Dot plots showing the variant allele frequencies of the
632 nonsynonymous mutations in the primary and metastatic tumors. (E) Targeted deep amplicon
633 sequencing of the metastasis-specific mutations in the primary tumor and matched normal
634 tissue. Significance of the mutations based on the variant read counts was determined using
635 DeepSNV and a Bayesian hypothesis test (methods).
636
637 **Figure 3 - Single Cell Mutational Profiling of Matched Primary and Metastatic Tumors**
638 Targeted cancer gene panel (T1000) sequencing data of point mutations in 372 single cells from
639 the primary colon and liver metastatic tumors from patients CRC1 and CRC2. (A-B)
640 Multidimensional scaling analysis, in which each dot represents a single cell. Cells are colored
641 by the flow-sorting distribution from which they were isolated. (C-D) 2-dimensional clustered
642 heatmaps of the single cell mutation data (T1000), with clusters labeled by color above.
643 Nonsynonymous mutations are labeled in bold, while synonymous mutations are labeled in
644 regular text. Populations of flow-sorted aneuploid tumor cells that were sequenced on the
645 T1000 panel from the primary and metastatic tumors are shown on the right-hand side and
646 labeled as ‘pop’. Blue bars represent mutations, light grey bars represent reference alleles, dark
647 grey bars represent false positives and white bars represent sites with low or no coverage (NA).
648 Bold font on the gene names indicate exonic mutations that were captured by both exome and

T1000 platforms.

Figure 4 - Early APC Progenitor Cells Detected in CRC1

Raw sequencing reads and variant alleles are plotted for 3 diploid *APC* progenitor cells (PD16, PD41, PDD93) and one representative primary tumor cell (PA74) from the major tumor population at genomic regions where mutations were detected in *APC, RAS, TP53* and *TCF7L2*. Plots and read counts were generated using the Integrated Genome Viewer.

Figure 5 - Single Cell Copy Number Profiling of Primary and Metastatic Tumors

(A,B) MDS plots of single cell copy number profiles from patients CRC1 and CRC2. (C,D) Hierarchical 1-dimensional clustered heatmaps of single cell integer copy number profiles from patients CRC1 and CRC2. Heatmap colors correspond to the integer copy number values in the single cells. Clusters of cells with similar profiles are labeled in colored bars on the left-hand side and cancer genes are annotated on the x-axis.

Figure 6 - Mutational Lineage Trees of Single Cells During Metastasis

Mutational trees calculated from single cell mutation data using SCITE showing clonal lineages during tumor progression and metastasis. (A) Mutational lineage tree from patient CRC1 with a monoclonal seeding event (B) Mutational lineage tree from patient CRC2 with polyclonal seeding events and an independent tumor lineage. Grey circles represent single cells, while blue boxes represent mutations.

Data Access

The sequencing data from this study has been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP074289.

Acknowledgements

This work was supported by the MD Anderson Colon Cancer Moonshot project and the Eric & Liz Lefkofsky Family Foundation. The research was also supported by grants to N.N. from NCI (1R01CA169244-01) and the American Cancer Society (129098-RSG-16-092-01-TBG). N.E.N. is an Andrew Sabin Family Fellow. The study was supported by the MD Anderson Cancer Moonshot Knowledge Gap Award and the Center for Genetics & Genomics. M.L.L. is supported by a Research Training Award from the Cancer Prevention and Research Institute of Texas (CPRIT RP140106), and is also supported by the American Legion Auxiliary (ALA) and Hearst Foundations. A.D. is supported by the ALA and by the National Library of Medicine Training Program in Biomedical Informatics (4T15LM007093-25). This work was also supported by an RO1 grant to S.K. from NCI (RO1CA184843). This study was supported by the MD Anderson Sequencing Core Facility Grant (no. CA016672) and the Flow Cytometry Facility grant from NIH (no. CA016672). We thank Niko Beerenwinkle, Jack Kuipers and Jahn Katharina for their assistance with SCITE.

Author Contributions

M.L.L. performed experiments, analyzed data and wrote the manuscript. A.D., R.G., A.C., and Y.W. analyzed data. E.S. reviewed and wrote the manuscript. D.M. provided tumor tissues. S.K. provided tumor tissues and analyzed data. N.E.N. analyzed data and wrote the manuscript.

Disclosure Declaration

The authors declare no competing financial interests.

References

TCGA (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330-7.

American Cancer Society. 2015. What are the survival rates for colorectal cancer by stage?

Baslan, T., J. Kendall, L. Rodgers, H. Cox, M. Riggs, A. Stepansky, J. Troge, K. Ravi, D. Esposito, B. Lakshmi, M. Wigler, N. Navin & J. Hicks (2012) Genome-wide copy number analysis of single cells. *Nat Protoc,* 7**,** 1024-41.

Baslan, T., J. Kendall, B. Ward, H. Cox, A. Leotta, L. Rodgers, M. Riggs, S. D'Italia, G. Sun, M. Yong, K. Miskimen, H. Gilmore, M. Saborowski, N. Dimitrova, A. Krasnitz, L. Harris, M. Wigler & J. Hicks (2015) Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Res,* 25**,** 714-24.

Boutros, P. C., M. Fraser, N. J. Harding, R. de Borja, D. Trudel, E. Lalonde, A. Meng, P. H. Hennings-Yeomans, A. McPherson, V. Y. Sabelnykova, A. Zia, N. S. Fox, J. Livingstone, Y.-J. Shiah, J. Wang, T. a. Beck, C. L. Have, T. Chong, M. Sam, J. Johns, L. Timms, N. Buchner, A. Wong, J. D. Watson, T. T. Simmons, C. P'ng, G. Zafarana, F. Nguyen, X. Luo, K. C. Chu, S. D. Prokopec, J. Sykes, A. Dal Pra, A. Berlin, A. Brown, M. a. Chan-Seng-Yue, F. Yousif, R. E. Denroche, L. C. Chong, G. M. Chen, E. Jung, C. Fung, M. H. W. Starmans, H. Chen, S. K. Govind, J. Hawley, A. D'Costa, M. Pintilie, D. Waggott, F. Hach, P. Lambin, L. B. Muthuswamy, C. Cooper, R. Eeles, D. Neal, B. Tetu, C. Sahinalp, L. D. Stein, N. Fleshner, S. P. Shah, C. C. Collins, T. J. Hudson, J. D. McPherson, T. van der Kwast & R. G. Bristow (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nature Genetics,* 47**,** 1-14.

Brannon, A. R., E. Vakiani, B. E. Sylvester, S. N. Scott, G. McDermott, R. H. Shah, K. Kania, A. Viale, D. M. Oschwald, V. Vacic, A. K. Emde, A. Cercek, R. Yaeger, N. E. Kemeny, L. B. Saltz, J. Shia, M. I. D'Angelica, M. R. Weiser, D. B. Solit & M. F. Berger (2014) Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol,* 15**,** 454.

Cline, M. S., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker & G. D. Bader (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc,* 2**,** 2366-82.

Cooper, C. S., R. Eeles, D. C. Wedge, P. Van Loo, G. Gundem, L. B. Alexandrov, B. Kremeyer, A. Butler, A. G. Lynch, N. Camacho, C. E. Massie, J. Kay, H. J. Luxton, S. Edwards, Z. Kote-Jarai, N. Dennis, S. Merson, D. Leongamornlert, J. Zamora, C. Corbishley, S. Thomas, S. Nik-Zainal, M. Ramakrishna, S. O'Meara, L. Matthews, J. Clark, R. Hurst, R. Mithen, R. G. Bristow, P. C. Boutros, M. Fraser, S. Cooke, K. Raine, D. Jones, A. Menzies, L. Stebbings, J. Hinton, J. Teague, S. McLaren, L. Mudie, C. Hardy, E. Anderson, O. Joseph, V. Goody, B. Robinson, M. Maddison, S. Gamble, C. Greenman, D. Berney, S. Hazell, N. Livni, I. P. Group, C. Fisher, C. Ogden, P. Kumar, A. Thompson, C. Woodhouse, D. Nicol, E. Mayer, T. Dudderidge, N. C. Shah, V. Gnanapragasam, T. Voet, P. Campbell, A. Futreal, D. Easton, A. Y. Warren, C. S. Foster, M. R. Stratton, H. C. Whitaker, U. McDermott, D. S. Brewer & D. E. Neal (2015) Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet,* 47**,** 367-72.

Davis, A. & N. E. Navin (2016) Computing tumor trees from single cells. *Genome Biol,* 17**,** 113.

de Bruin, E. C., N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Gronroos, M. A. Muhammad, S. Horswell, M. Gerlinger, I. Varela, D. Jones, J. Marshall, T. Voet, P. Van Loo, D. M. Rassl, R. C. Rintoul, S. M. Janes, S.-M. Lee, M. Forster, T. Ahmad, D. Lawrence, M. Falzon, A. Capitanio, T. T. Harkins, C. C. Lee, W. Tom, E. Teefe, S.-C. Chen, S. Begum, A. Rabinowitz, B. Phillimore, B. Spencer-Dene, G. Stamp, Z. Szallasi, N. Matthews, A. Stewart, P. Campbell & C. Swanton (2014) Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science,* 346**,** 251-256.

Fearon, E. R. & B. Vogelstein (1990) A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759-67.

Gawad, C., W. Koh & S. R. Quake (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet,* 17**,** 175-88.

Gerlinger, M., A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal & C. Swanton (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* 366**,** 883-92.

Gerstung, M., C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch & N. Beerenwinkel (2012) Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun,* 3**,** 811.

Gundem, G., P. Van Loo, B. Kremeyer, L. B. Alexandrov, J. M. Tubio, E. Papaemmanuil, D. S. Brewer, H. M. Kallio, G. Hognas, M. Annala, K. Kivinummi, V. Goody, C. Latimer, S. O'Meara, K. J. Dawson, W. Isaacs, M. R. Emmert-Buck, M. Nykter, C. Foster, Z. Kote-Jarai, D. Easton, H. C. Whitaker, I. P. U. Group, D. E. Neal, C. S. Cooper, R. A. Eeles, T. Visakorpi, P. J. Campbell, U. McDermott, D. C. Wedge & G. S. Bova (2015) The evolutionary history of lethal metastatic prostate cancer. *Nature,* 520**,** 353-7.

Jahn, K., J. Kuipers & N. Beerenwinkel (2016) Tree inference for single-cell data. *Genome Biol,* 17**,** 86.

Klein, C. A. (2009) Parallel progression of primary tumours and metastases. *Nat Rev Cancer,* 9**,** 302-12.

Langmead, B. & S. L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.

Lefort, V., R. Desper & O. Gascuel (2015) FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol Biol Evol,* 32**,** 2798-800.

Leung, M. L., Y. Wang, C. Kim, R. Gao, J. Jiang, E. Sei & N. E. Navin (2016) Highly multiplexed targeted DNA sequencing from single nuclei. *Nat Protoc,* 11**,** 214-35.

783 Leung, M. L., Y. Wang, J. Waters & N. E. Navin (2015) SNES: single nucleus exome
784 sequencing. *Genome Biol,* 16**,** 55.

785 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R.
786 Durbin & S. Genome Project Data Processing (2009) The Sequence Alignment/Map
787 format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

788 Lindell, K. O., J. A. Erlen & N. Kaminski (2006) Lessons from our patients: development of a
789 warm autopsy program. *PLoS Med,* 3**,** e234.

790 Martincorena, I., A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A.
791 Fullam, L. B. Alexandrov, J. M. Tubio, L. Stebbings, A. Menzies, S. Widaa, M. R.
792 Stratton, P. H. Jones & P. J. Campbell (2015) Tumor evolution. High burden and
793 pervasive positive selection of somatic mutations in normal human skin. *Science,* 348**,**
794 880-6.

795 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D.
796 Altshuler, S. Gabriel, M. Daly & M. A. DePristo (2010) The Genome Analysis Toolkit: a
797 MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*
798 *Res,* 20**,** 1297-303.

799 McPherson, A., A. Roth, E. Laks, T. Masud, A. Bashashati, A. W. Zhang, G. Ha, J. Biele, D.
800 Yap, A. Wan, L. M. Prentice, J. Khattra, M. A. Smith, C. B. Nielsen, S. C. Mullaly, S.
801 Kalloger, A. Karnezis, K. Shumansky, C. Siu, J. Rosner, H. L. Chan, J. Ho, N. Melnyk, J.
802 Senz, W. Yang, R. Moore, A. J. Mungall, M. A. Marra, A. Bouchard-Cote, C. B. Gilks, D.
803 G. Huntsman, J. N. McAlpine, S. Aparicio & S. P. Shah (2016) Divergent modes of
804 clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat*
805 *Genet*.

806 Mehlen, P. & A. Puisieux (2006) Metastasis: a question of life or death. *Nat Rev Cancer,* 6**,** 449-
807 58.

808 Navin, N., J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D.
809 Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks & M. Wigler
810 (2011) Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.

811 Navin, N. E. (2015) The first five years of single-cell cancer genomics and beyond. *Genome*
812 *Res,* 25**,** 1499-507.

813 Newburger, D. E., D. Kashef-Haghighi, Z. Weng, R. Salari, R. T. Sweeney, A. L. Brunner, S. X.
814 Zhu, X. Guo, S. Varma, M. L. Troxell, R. B. West, S. Batzoglou & A. Sidow (2013)
815 Genome evolution during progression to breast cancer. *Genome Res,* 23**,** 1097-108.

816 Nilsen, G., K. Liestøl, P. Van Loo, H. K. Moen Vollan, M. B. Eide, O. M. Rueda, S. F. Chin, R.
817 Russell, L. O. Baumbusch, C. Caldas, A. L. Børresen-Dale & O. C. Lingjaerde (2012)
818 Copynumber: Efficient algorithms for single- and multi-track copy number segmentation.
819 *BMC Genomics,* 13**,** 591.

820 Norton, L. & J. Massague (2006) Is cancer a disease of self-seeding? *Nat Med,* 12**,** 875-8.

Paradis, E., J. Claude & K. Strimmer (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20**,** 289-90.

Quinlan, A. R. & I. M. Hall (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

Ross, E. M. & F. Markowetz (2016) OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol,* 17**,** 69.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski & T. Ideker (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res,* 13**,** 2498-504.

Tan, I. B., S. Malik, K. Ramnarayanan, J. R. McPherson, D. L. Ho, Y. Suzuki, S. B. Ng, S. Yan, K. H. Lim, D. Koh, C. M. Hoe, C. Y. Chan, R. Ten, B. K. Goh, A. Y. Chung, J. Tan, C. X. Chan, S. T. Tay, L. Alexander, N. Nagarajan, A. M. Hillmer, C. L. Tang, C. Chua, B. T. Teh, S. Rozen & P. Tan (2015) High-depth sequencing of over 750 genes supports linear progression of primary tumors and metastases in most patients with liver-limited metastatic colorectal cancer. *Genome Biol,* 16**,** 32.

Tange, O. (2011) GNU Parallel - The Command-Line Power Tool. *USENIX Magazine,* 36**,** 42-47.

Team, R. C. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Valastyan, S. & R. A. Weinberg (2011) Tumor metastasis: molecular insights and evolving paradigms. *Cell,* 147**,** 275-92.

Vitak, S. A., K. A. Torkenczy, J. L. Rosenkrantz, A. J. Fields, L. Christiansen, M. H. Wong, L. Carbone, F. J. Steemers & A. Adey (2017) Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods,* 14**,** 302-308.

Wang, K., M. Li & H. Hakonarson (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.

Wang, Y. & N. E. Navin (2015) Advances and Applications of Single-Cell Sequencing Technologies. *Mol Cell,* 58**,** 598-609.

Wang, Y., J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, A. Multani, H. Zhang, R. Zhao, F. Michor, F. Meric-Bernstam & N. E. Navin (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature,* 512**,** 155-160.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Xie, T., D. A. G, J. R. Lamb, E. Martin, K. Wang, S. Tejpar, M. Delorenzi, F. T. Bosman, A. D. Roth, P. Yan, S. Bougel, A. F. Di Narzo, V. Popovici, E. Budinska, M. Mao, S. L. Weinrich, P. A. Rejto & J. G. Hodgson (2012) A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PLoS One,* 7**,** e42001.

Xu, X., Y. Hou, X. Yin, L. Bao, A. Tang, L. Song, F. Li, S. Tsang, K. Wu, H. Wu, W. He, L. Zeng, M. Xing, R. Wu, H. Jiang, X. Liu, D. Cao, G. Guo, X. Hu, Y. Gui, Z. Li, W. Xie, X. Sun, M. Shi, Z. Cai, B. Wang, M. Zhong, J. Li, Z. Lu, N. Gu, X. Zhang, L. Goodman, L. Bolund, J. Wang, H. Yang, K. Kristiansen, M. Dean & Y. Li (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell,* 148**,** 886-95.

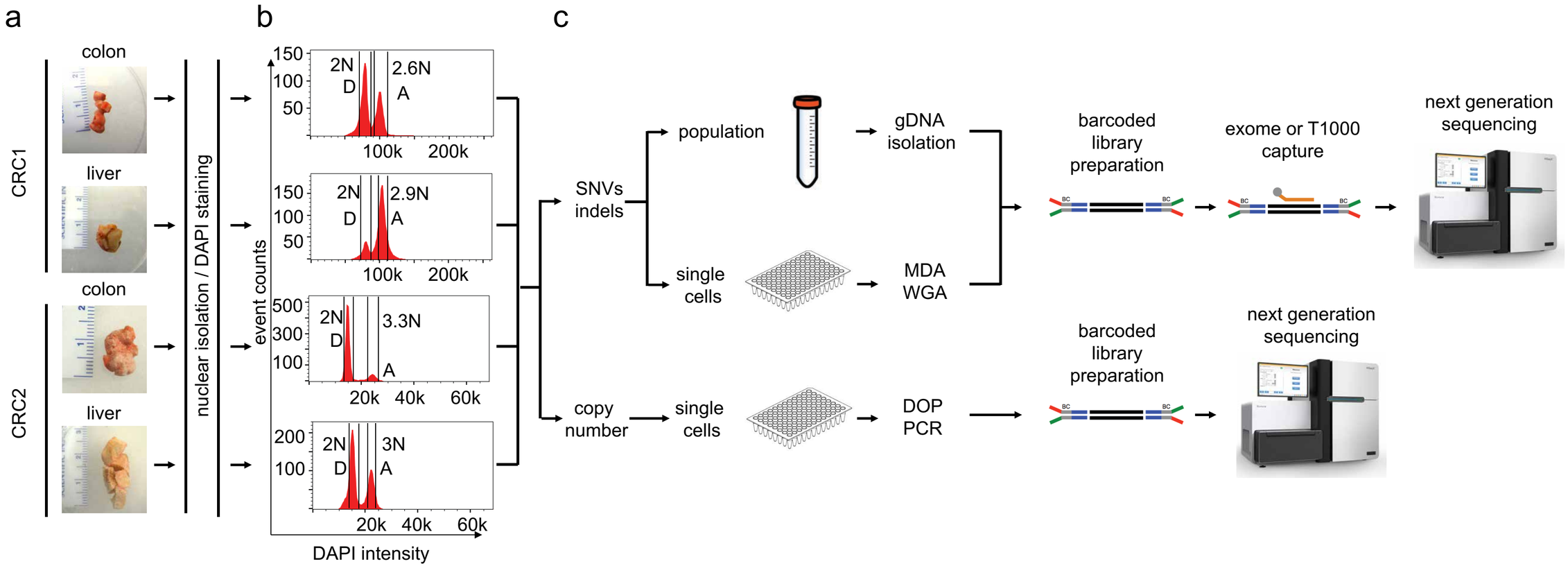
Yachida, S., S. Jones, I. Bozic, T. Antal, R. Leary, B. Fu, M. Kamiyama, R. H. Hruban, J. R. Eshleman, M. A. Nowak, V. E. Velculescu, K. W. Kinzler, B. Vogelstein & C. A. Iacobuzio-Donahue (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature,* 467**,** 1114-7.

Yates, L. R. & P. J. Campbell (2012) Evolution of the cancer genome. *Nat Rev Genet,* 13**,** 795-806.

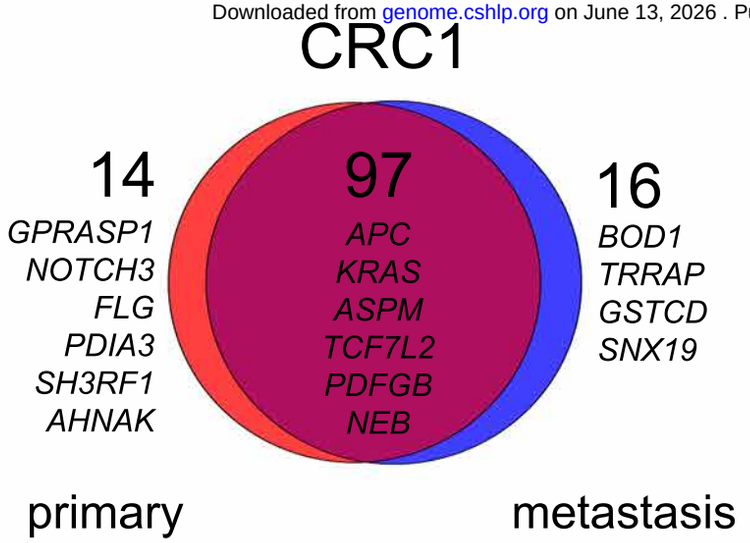
Zahn, H., A. Steif, E. Laks, P. Eirew, M. VanInsberghe, S. P. Shah, S. Aparicio & C. L. Hansen (2017) Scalable whole-genome single-cell library preparation without preamplification. *Nat Methods,* 14**,** 167-173.

Zhang, J., J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C. W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, Wistuba, II & P. A. Futreal (2014) Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science,* 346**,** 256-9.

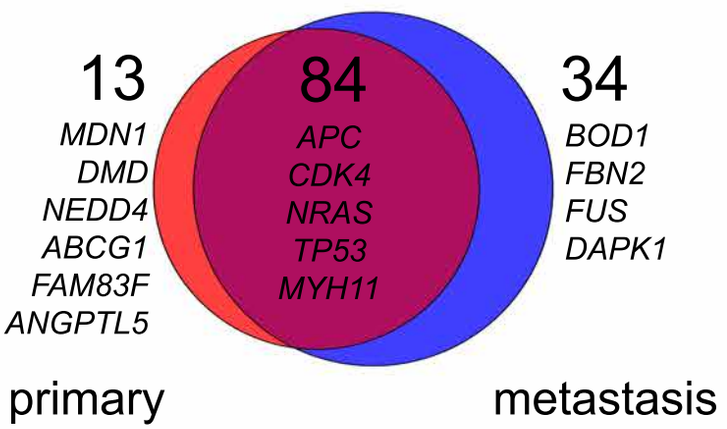
Zong, C., S. Lu, A. R. Chapman & X. S. Xie (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science,* 338**,** 1622-6.


a
b
c
CRC1
CRC2
colon
liver
colon
liver
nuclear isolation / DAPI staining
event counts
DAPI intensity
2N
D
2.6N
A
2.9N
3.3N
3N
100k
200k
20k
40k
60k
SNVs
indels
copy
number
population
single
cells
single
cells
gDNA
isolation
MDA
WGA
DOP
PCR
barcoded
library
preparation
barcoded
library
preparation
exome or T1000
capture
next generation
sequencing
next generation
sequencing

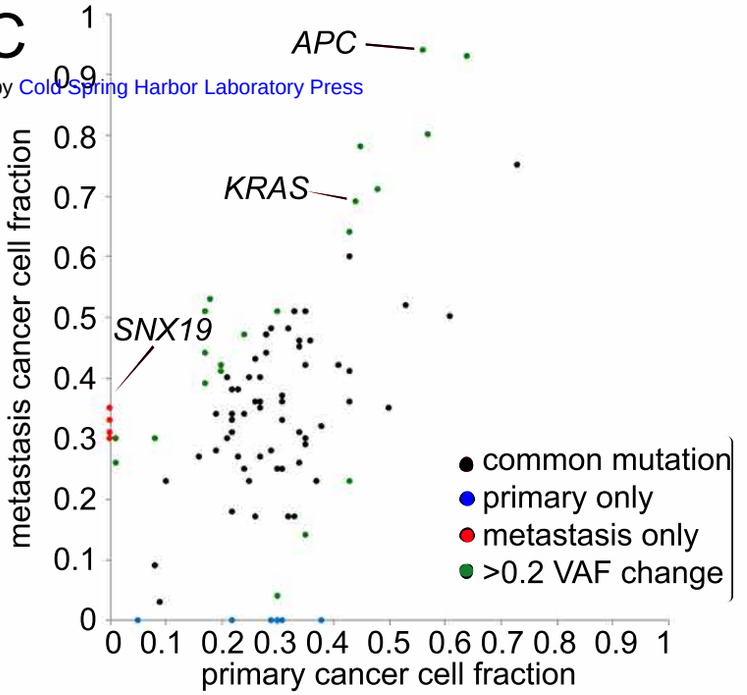
A



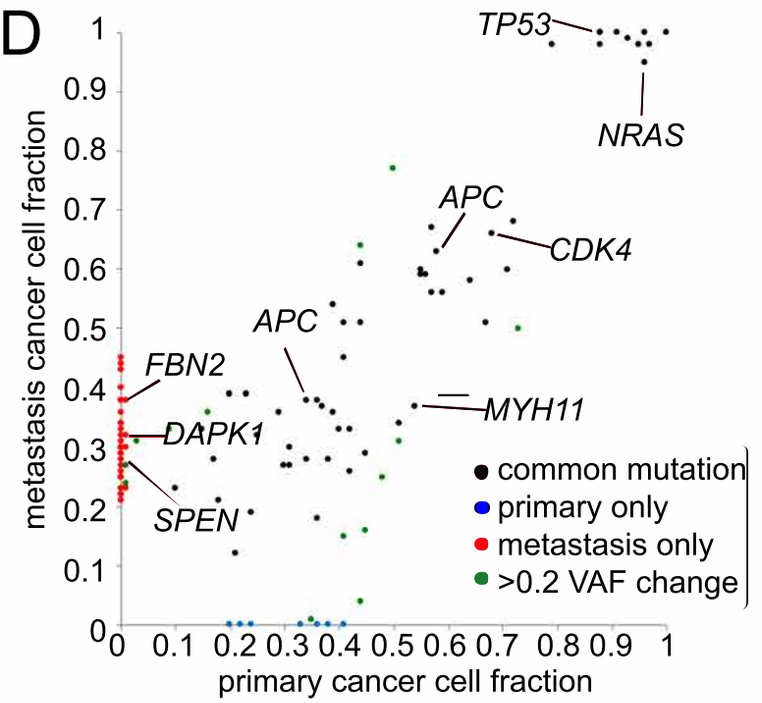
B



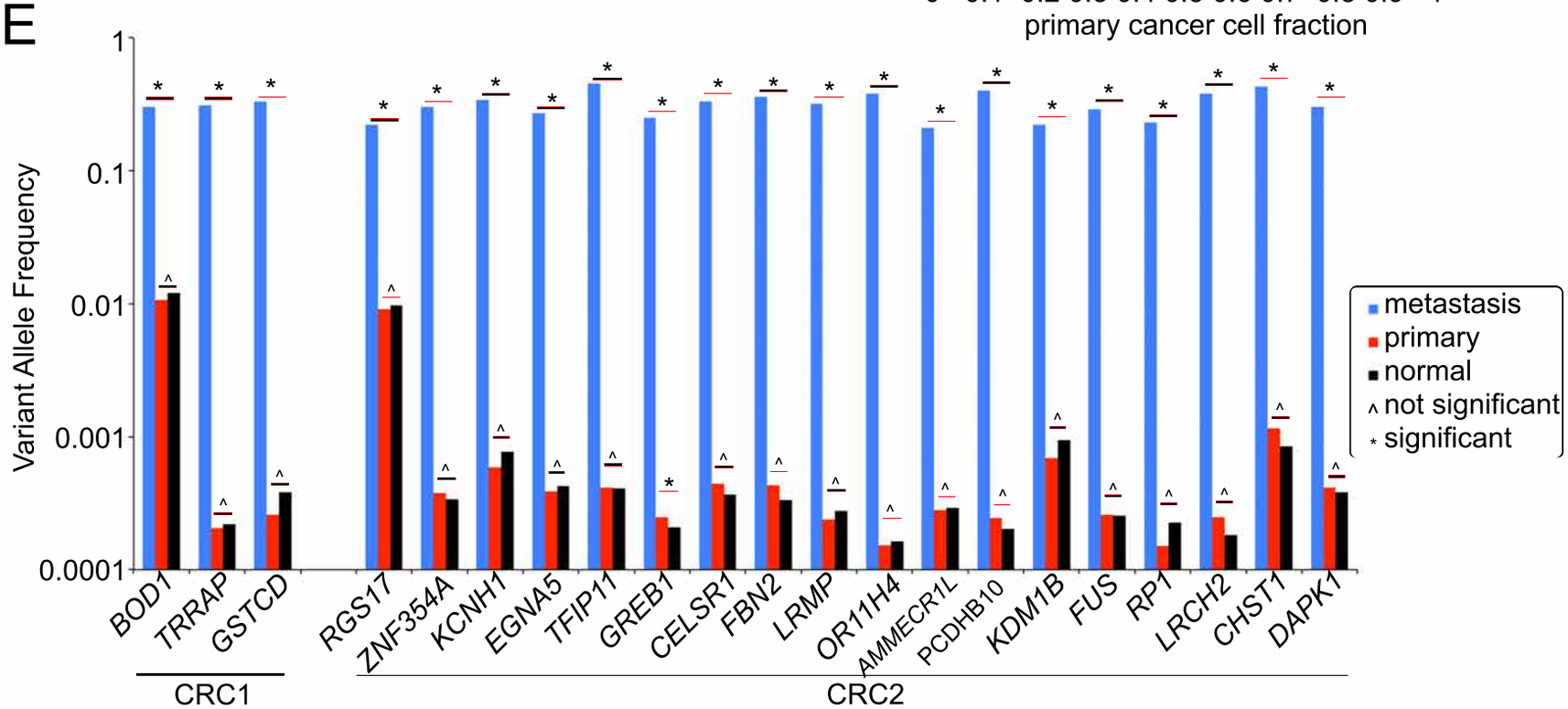
C

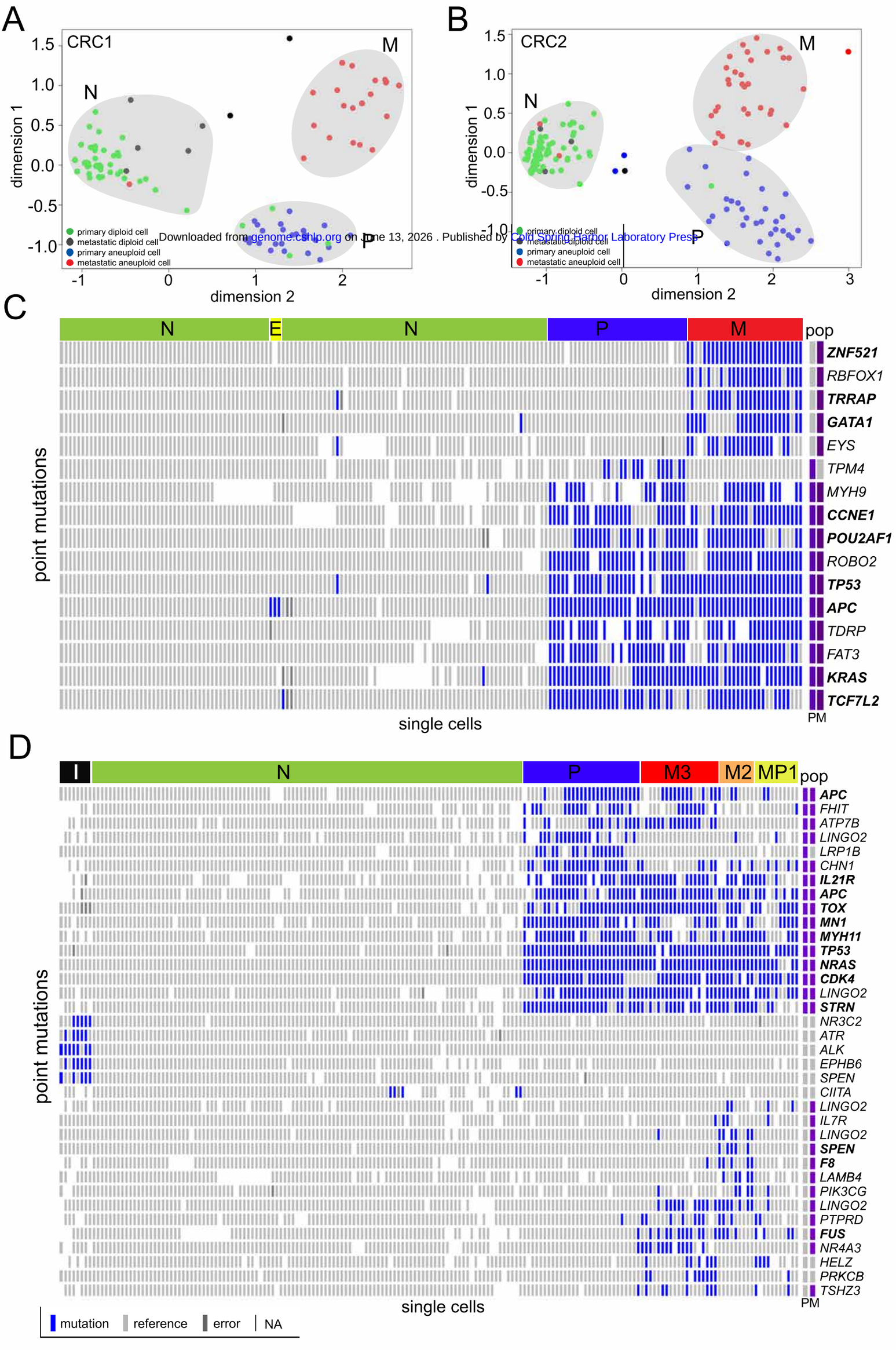


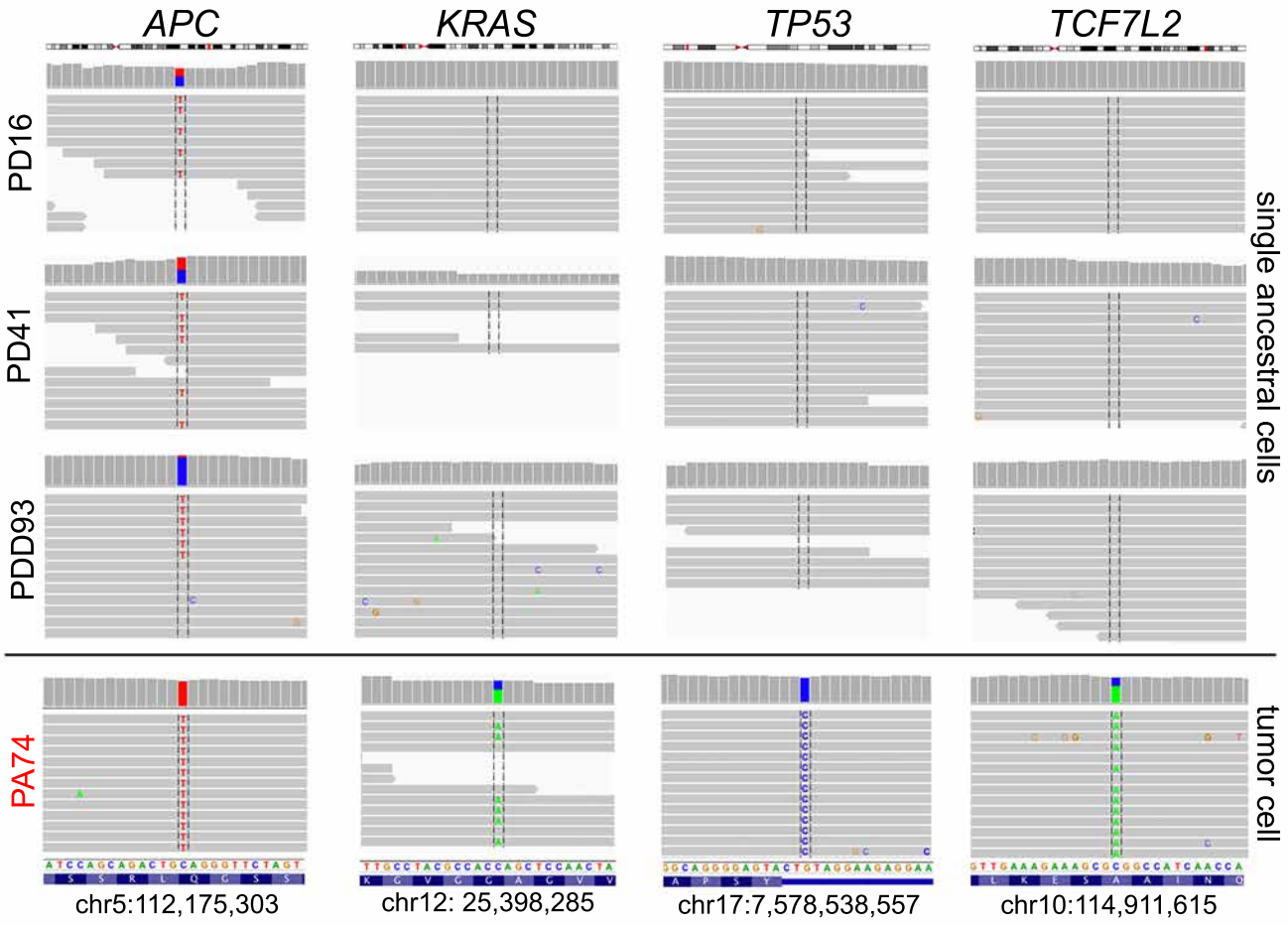
D

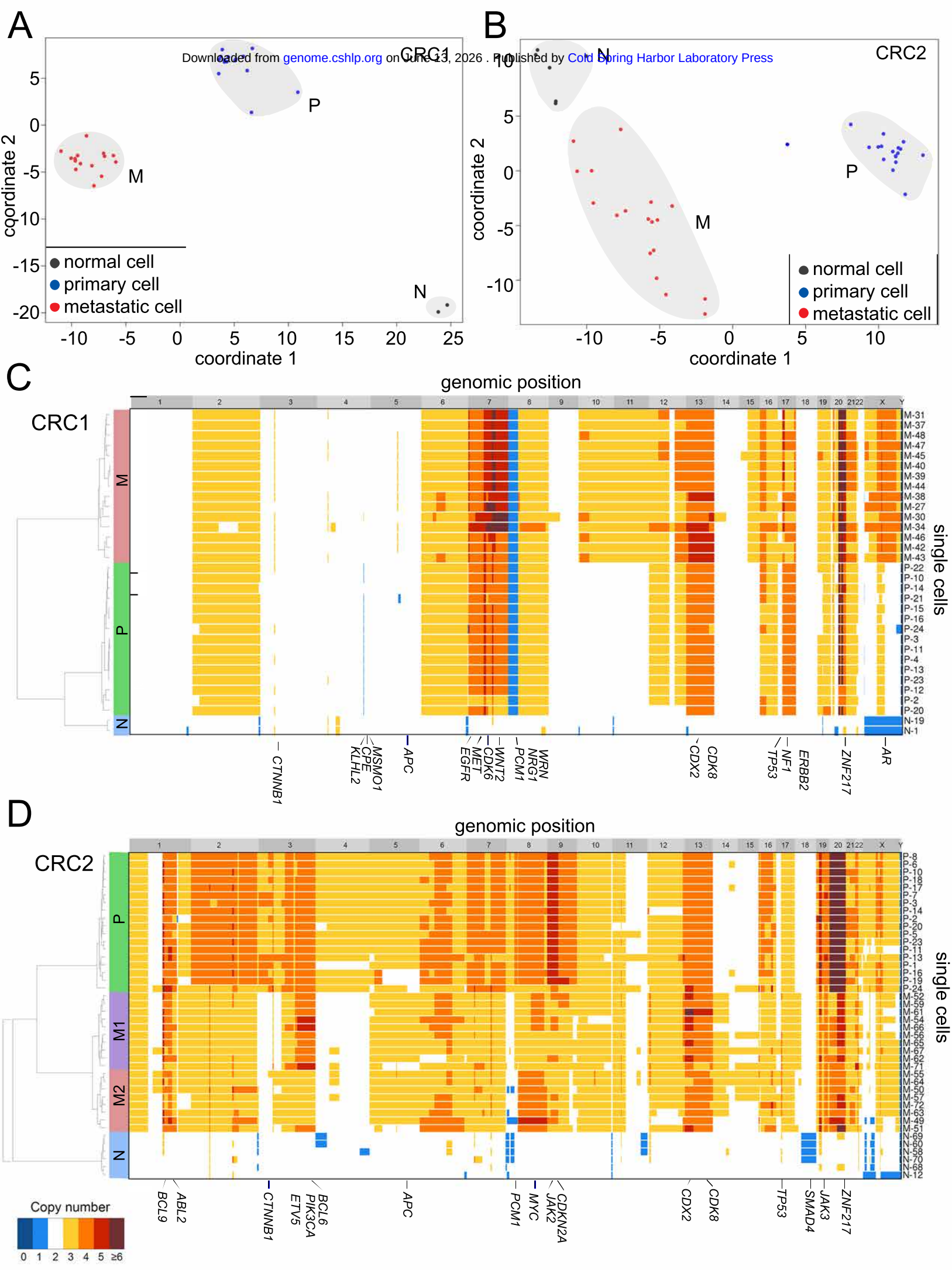


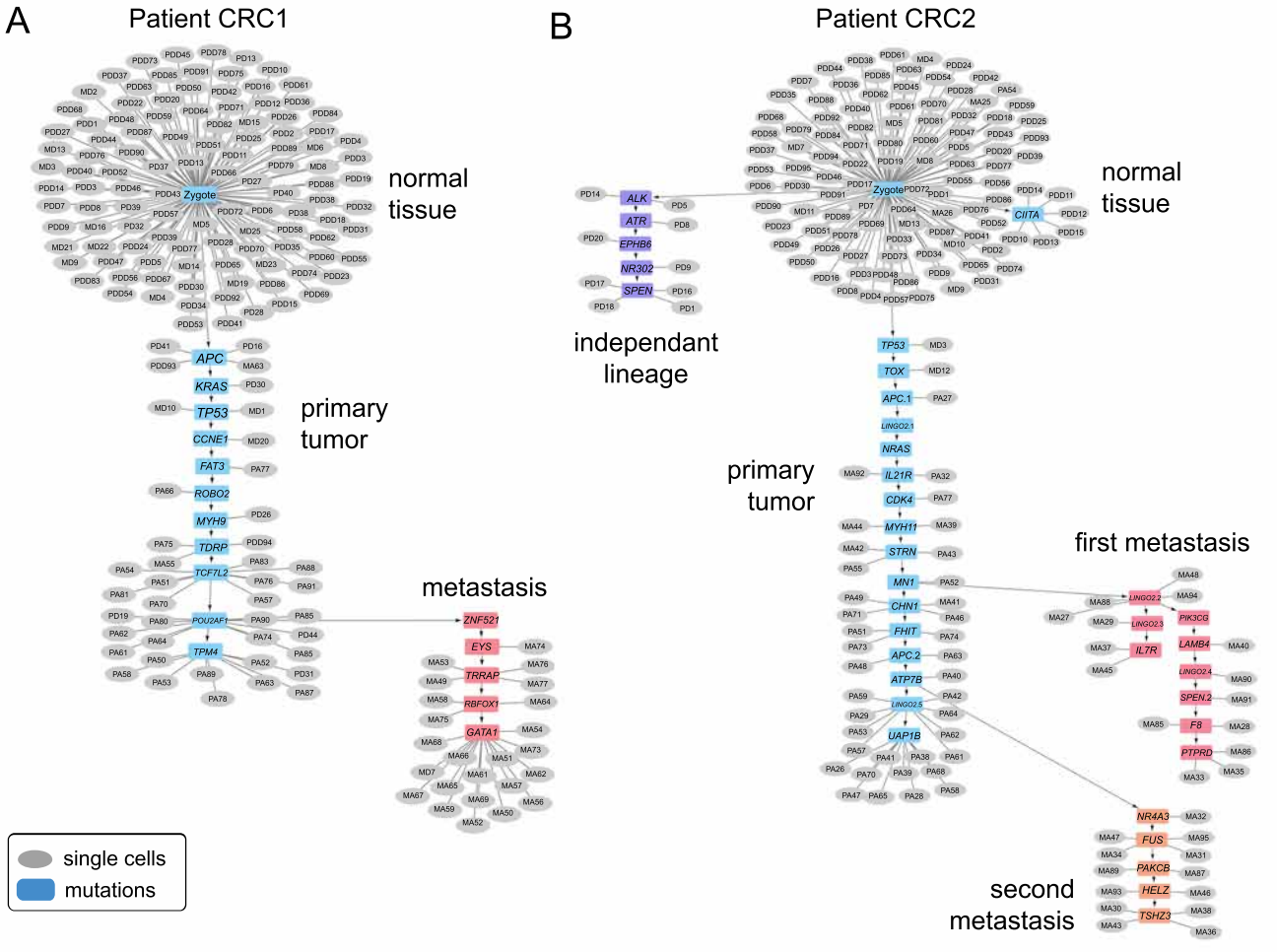
E






APC
KRAS
TP53
TCF7L2
PD16
PD41
PDD93
PA74
single ancestral cells
tumor cell
chr5:112,175,303
chr12: 25,398,285
chr17:7,578,538,557
chr10:114,911,615
mutation sites




A
Patient CRC1
normal tissue
primary tumor
metastasis
Zygote
APC
KRAS
TP53
CCNE1
FAT3
ROBO2
MYH9
TDRP
TCF7L2
POU2AF1
TPM4
ZNF521
EYS
TRRAP
RBFOX1
GATA1
B
Patient CRC2
normal tissue
independant lineage
primary tumor
first metastasis
second metastasis
ALK
ATR
EPHB6
NR302
SPEN
Zygote
CIITA
TP53
TOX
APC.1
LINGO2.1
NRAS
IL21R
CDK4
MYH11
STRN
MN1
CHN1
FHIT
APC.2
ATP7B
LINGO2.5
UAP1B
LINGO2.2
LINGO2.3
IL7R
LINGO2.4
SPEN.2
PIK3CG
LAMB4
F8
PTPRD
NR4A3
FUS
PAKCB
HELZ
TSHZ3
single cells
mutations