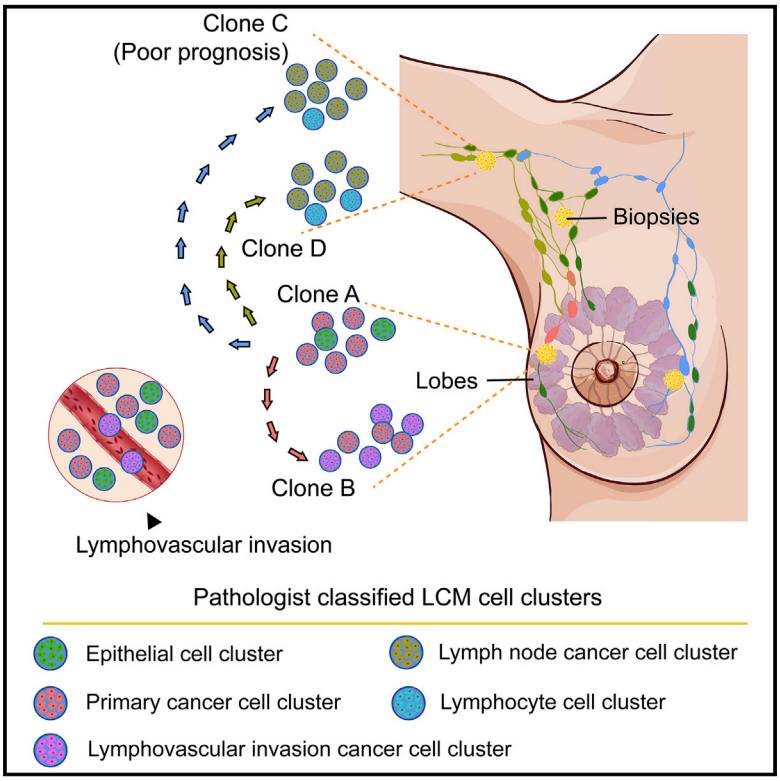
Cell Reports
Medicine

Article

Genome profiles of pathologist-defined cell clusters by multiregional LCM and G&T-seq in one triple-negative breast cancer patient

Graphical abstract



Highlights

- Pathologically diverse cell clusters share genomic and transcriptomic profiles
- Transcriptome-defined clones are more complex than genome-defined clones
- Three distinct pathways were inferred, each with disparate survival outcomes
- Lower expression of ribosomal proteins may be an indicator of poor prognosis

Authors

Zhongyi Zhu, Weiwei Wang, Feng Lin, ..., John R. Mackey, Bo Li, Gane Ka-Shu Wong

Correspondence

john.mackey@ahs.ca (J.R.M.), libo@genomcs.cn (B.L.), gane@ualberta.ca (G.K.-S.W.)

In brief

Zhu et al. reveal the complexity of tumor development, metastasis, and prognosis by simultaneous DNA and RNA sequencing of pathologist-defined cell clusters, excised through laser capture microdissection from multiregional frozen sections of a triple-negative breast cancer patient.

Zhu et al., 2021, Cell Reports Medicine 2, 100404
October 19, 2021
https://doi.org/10.1016/j.xcrm.2021.100404

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Genome profiles of pathologist-defined cell clusters by multiregional LCM and G&T-seq in one triple-negative breast cancer patient

Zhongyi Zhu,[1,8,9] Weiwei Wang,[2,3,9] Feng Lin,[1,9] Tracy Jordan,[2] Guibo Li,[1] Sveta Silverman,[4] Si Qiu,[1] Anil Abraham Joy,[5] Chao Chen,[1] Deanna L. Hockley,[5] Xi Zhang,[1] Qing Zhou,[1] Lynne M. Postovit,[6] Xiuqing Zhang,[1] Yong Hou,[1] John R. Mackey,[5,*] Bo Li,[1,*] and Gane Ka-Shu Wong[1,2,7,10,*]

[1]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China
[2]Department of Medicine, University of Alberta, Edmonton, AB T6G 2E1, Canada
[3]Geneis, Bldg A, 5 Guangshun North Street, Beijing 100102, China
[4]Department of Pathology and Laboratory Medicine, University of Alberta, Edmonton, AB T6G 2H7, Canada
[5]Division of Medical Oncology, Department of Oncology, University of Alberta, Cross Cancer Institute, Edmonton, AB T6G 1Z2, Canada
[6]Department of Oncology, University of Alberta, Edmonton, AB T6G 2H7, Canada
[7]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada
[8]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[9]These authors contributed equally
[10]Lead contact
*Correspondence: john.mackey@ahs.ca (J.R.M.), libo@genomcs.cn (B.L.), gane@ualberta.ca (G.K.-S.W.)
https://doi.org/10.1016/j.xcrm.2021.100404

SUMMARY

Pathological examination is the gold standard for cancer diagnosis, and breast tumor cells are often found in clusters. We report a case study on one triple-negative breast cancer (TNBC) patient, analyzing tumor development, metastasis, and prognosis with simultaneous DNA and RNA sequencing of pathologist-defined cell clusters from multiregional frozen sections. The cell clusters are isolated by laser capture microdissection (LCM) from primary tumor tissue, lymphatic vessels, and axillary lymph nodes. Data are reported for a total of 97 cell clusters. A combination of tumor cell-cluster clonality and phylogeny reveals 3 evolutionarily distinct pathways for this patient, each associated with a unique mRNA signature, and each correlated with disparate survival outcomes. Hub gene analysis indicates that extensive downregulation of ribosomal protein mRNA is a potential marker of poor prognosis in breast cancer.

INTRODUCTION

Pathological examination is the gold standard for cancer diagnosis. Assessments of frozen sections determine the degree of tumor differentiation, immune differentiation, radiotherapy sensitivity, chemotherapy sensitivity, and even gene mutation.[1] Tumorigenesis itself occurs by way of genetic aberrations that alter the function or expression of specific driver genes and tumor suppressors. For disease progression to occur, cancer cells are believed to acquire additional mutations that promote expansion, invasion of surrounding tissues, and, finally, metastasis.[2] Selection pressure from the local microenvironment or systemic drug treatment can aid this process by promoting the expansion of tumor clones with genetic or transcriptional abnormalities that are advantageous for disease progression.[2–4] Hence, at any time in the tumor life cycle, there exists the possibility of detecting multiple tumor clones in a single patient.[5] These clonal populations contribute to intratumor heterogeneity and are relevant for medicine as they have the potential to become drug resistant or metastatic.[6]

The detection of clonal populations has been greatly advanced by the advent of single-cell sequencing, which allows us to attribute unique DNA and RNA signatures to tumor cells based on their presence in specific microenvironments. In the case of breast cancer, single-cell sequencing has helped define gene expression signatures related to metastatic burden,[7] metastatic subtype,[8] and even spatial orientation within primary breast cancer tissue.[9,10] However, the extant methods for isolating single cells, as well as the subsequent DNA and RNA sequencing steps, vary widely, and few groups have analyzed both genomic aberrations and gene expression changes together in a particular cancer cell type. Furthermore, sequencing single cells may not be the ideal method for characterizing metastatic tumor populations because tumor cells that circulate as clusters exhibit higher metastatic potential and are associated with worse disease prognosis.[11,12] Circulating tumor cell clusters are often polyclonal, exhibiting a mix of primary tumor and epithelial-like characteristics that are believed to enhance their invasiveness, dissemination, and metastatic colonization.[11–13] Thus, analyzing invasive tumor cells as a collective multicellular unit, rather than on a single-cell level, may provide a more physiological representation of cancer cells with an aggressively metastatic phenotype.

We report here a case study of one triple-negative breast cancer (TNBC) patient, with simultaneous genomic and transcriptomic sequencing (G&T-seq)[14] of pathologist-defined cell clusters excised from multiregional frozen sections by laser-capture microdissection (LCM). This combination of methodologies allowed us to extract and characterize cell clusters, not only from the primary and lymph node (LN) tumors of this patient, but also from the lymphatic vessels, enabling the analysis of cells discernably *en route* from the primary tumor to the axillary LNs (i.e., lymphovascular invasion [LVI]). G&T-seq was performed on these cell clusters to assess the chromosome aberration patterns, genetic mutations, and gene expression profiles associated with LVI and LN metastasis. Not only did we identify genomic aberration patterns and RNA expression profiles associated with LVIs but, intriguingly, we also found evidence of three transcriptionally distinct pathways of metastatic spread in this one patient that, when mapped to a much larger dataset of patient outcomes, were varying predictors of survival outcome. We believe the combined utilization of these methods may help identify new targets for preventing metastatic disease and may become a valuable tool in the development of precision medicine for patients with genetically heterogeneous diseases such as TNBC.

RESULTS

To avoid misunderstanding, we must first clarify what we are trying to do versus what we are not trying to do, and the terminology that we have adopted. Our analyses are primarily focused on the progression of breast-derived cell clusters from the primary tumor, through the LVI, to the LNs. We are not trying to ascertain when a cell cluster should be declared a cancer or metastasis, which in any case is surprisingly difficult to do, since a large number of apparently normal cells in seemingly healthy individuals will often contain driver mutations.[15,16] Although the term "clone" has been used to define genetically identical cells asexually derived from a common ancestor, our analyses revealed that cell clusters with different pathologist-defined classifications, based on the morphology of their dominant cells, can share a common genomic or transcriptional profile. In that sense, they can legitimately be called quasi-clones, but for brevity, we simply used the word "clone." This also reveals the inherent limitations of morphology-based classifications, which are not perfectly correlated with genomic or transcriptional profiles. As the premise of this article is that we must study cancer progression on a collective level, using multicellular units excised by LCM, we believe that this is an appropriate, if unconventional, terminology.

Cell cluster isolation from primary breast tumor and axillary LNs using LCM

To isolate and characterize cancer cells with an inherent ability to invade the lymphovasculature and form distant metastatic lesions, we obtained fresh tissue samples from a 24-year-old woman with T3 N1 M1, ER(−), PR(−), HER-2(−), grade 3, invasive ductal breast carcinoma following palliative mastectomy and full axillary LN dissection. This patient had treatment refractory disease (two cycles of docetaxel and four cycles of doxorubicin/cyclophosphamide chemotherapy), which, at the time of sample collection, had metastasized to the patient's local regional axillary LNs and liver. To maximize the likelihood of detecting LVI, we collected 2 tissue samples from the tumor-stromal interface of the primary tumor, as well as 2 samples from her carcinoma-positive axillary LNs (3 of the 15 removed LNs were positive for metastatic disease). These tissue samples were flash-frozen, sectioned, H&E-stained, and screened by 2 anatomic pathologists for the presence of cancer cell clusters defined as histologically malignant groups of cells of at least 10 μm in diameter.[12,17] Laser capture microdissection, which permits the acquisition of select cells, while preserving anatomical structures such as lymphatic vessels, was then used to dissect cell clusters from select tissue sections. A total of 186 cell clusters were collected for processing, with 97 (∼52%) passing both DNA and RNA quality control (QC) checks for sequencing analysis. These clusters comprised 17 morphologically normal breast epithelial cell clusters, 17 primary tumor cancer cell clusters, and 20 clusters found within lymphatic vessels, which, due to the unidirectionality of lymphatic flow, represent cells categorically *en route* to the axillary LNs from the primary tumor. In addition, there were 36 cancer cell clusters and 7 lymphocyte cell clusters from the tumor-infiltrated axillary LN samples. G&T-seq was then performed as described[14] to simultaneously extract, amplify, and create DNA- and mRNA-seq libraries from each LCM isolated cell cluster. In brief, multiple-displacement amplification (MDA) was performed on the genomic DNA (gDNA) isolated from each cluster, followed by whole-genome (WGS) and whole-exome sequencing (WES) to detect copy-number variants (CNV) and single-nucleotide variants (SNV), respectively. WGS was performed at a depth of 1.12×, while WES was performed at a usage depth of 272× (raw depth >1,000×). Whole-transcriptome sequencing (WTS) was performed on mRNA isolated from each cluster at a sequencing depth of 4.5 G clean bases with a quality score of ≥30. An overview of the experiments and analyses is presented in Figure 1.

LVI-associated cell clusters share a CNV profile abundant in chromosome amplification and deletion events

We performed WGS on the gDNA isolated from each cluster to detect CNVs. Using a single-cell CNV calling method previously described[18] to avoid amplification bias, we detected chromosome amplification and deletion events in every cell cluster analyzed (Figure S1). This included clusters comprised of histologically normal cells positioned in a well-organized breast duct, indicative of the heterogeneous nature of the cell clusters isolated from this patient. Next, we performed hierarchical clustering using the Ward.D2 algorithm to generate a CNV heatmap and evaluate the clonal architecture of these cell clusters, using bootstrap to identify the stable clades. Heatmap analysis showed the existence of three distinct CNV profiles that we labeled CNV clone A, CNV clone B, and CNV clone C (Figure 2A). Notably, the cell clusters obtained from lymphatic vessels mapped together, along with several primary cancer cell clusters, as CNV clone B. In contrast, cell clusters associated with CNV clone

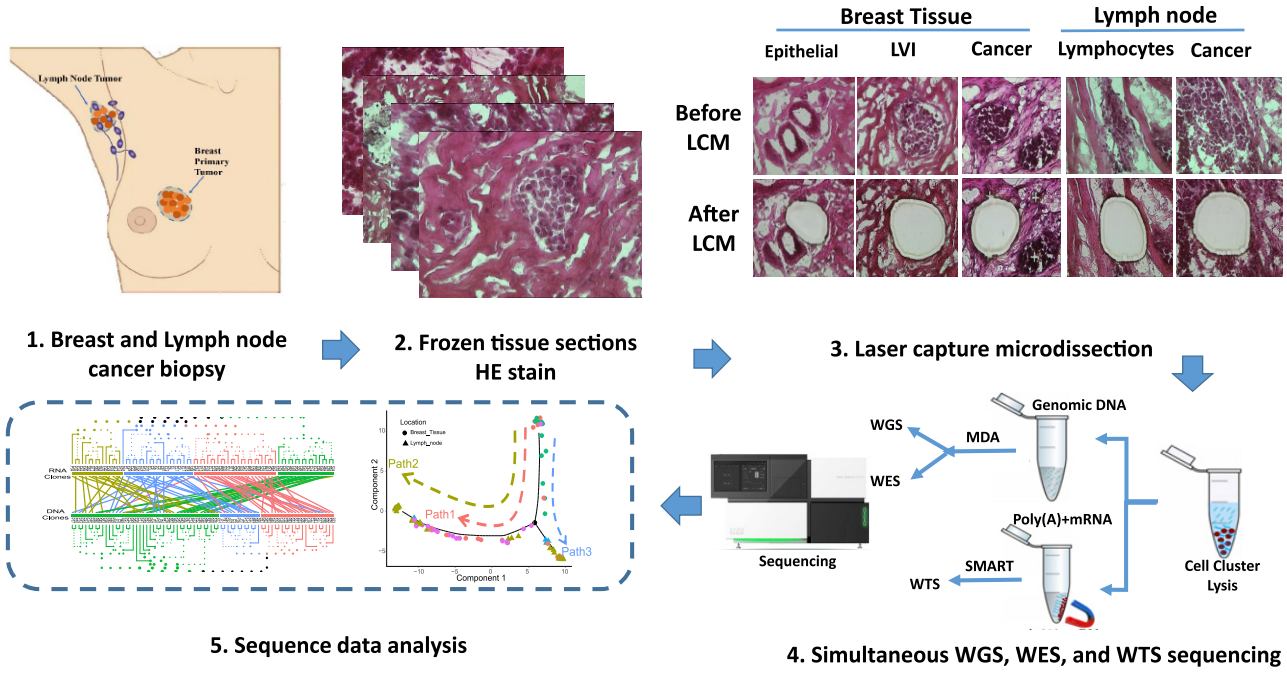


Figure 1. Schematic representation of LCM cell cluster isolation and G&T-seq from TNBC patient samples

Tissue samples were collected from the tumor-stromal interface of the primary tumor and carcinoma-positive axillary lymph nodes (LNs) of a single TNBC patient. Tissue samples were flash-frozen, sectioned, and H&E stained for pathological examination. LCM was used to isolate epithelial, LVI, lymphocyte, and cancer cell clusters measuring >10 μm in size from anatomically intact tissue sections. G&T-seq was then performed to extract and sequence gDNA and mRNA from each cell cluster. gDNA was amplified with MDA, and mRNA was amplified using a modified Smart-seq2 protocol. WGS and WES was performed on the amplified gDNA for CNV and SNV detection, respectively, while WTS was performed on amplified mRNA transcripts for transcriptome analysis.

C were isolated exclusively from the LNs of this patient. Because no cell type heterogeneity was associated with CNV clone C, this indicates that we either did not sample the closest ancestor of this LN clone, or that the signature of its ancestor changed significantly over the course of disease progression. Lastly, CNV clone A was found to be a combination of epithelial, lymphocyte, primary, and LN cancer cell clusters. These cell clusters exhibited relatively few CNVs (4.1%), but shared a common amplification of the X chromosome. Notably, because the cell clusters associated with the mixed-epithelial CNV clone included LN cancer cells, this suggests that relatively few CNV changes were required to promote LN metastasis. The LVI and LN CNV clones by contrast had a substantial number of CNV changes, characteristic of chromosome instability (CIN). The LVI clone contained CNVs that covered ~71.7% of the genome and included amplification of the X chromosome. The LN clone also had extensive CNV changes covering ~53.1% of the genome, but did not show a significant X chromosome amplification (Table S1.1).

Cell clusters associated with LVI have a high homozygous mutational burden

Next, we performed WES on the gDNA isolated from each cell cluster to detect SNVs. We again performed unsupervised clustering and constructed SNV heatmaps based on the homozygous and heterozygous SNVs detected in each cell cluster.[19] Similar to our CNV findings, analysis of the SNVs revealed three distinct SNV clones with shared gene mutation patterns (Figure 2B). SNV clone A was comprised of LCM cell clusters from every cell type tested, with the exception of lymphatic vessel-associated cell clusters (as in CNV clone A), and harbored SNVs occurring at a range of frequencies (6.7%–100%) from >750 genes (Table S1.2). SNV clone B contained cell clusters isolated from the primary tumor and lymphatic vessels, indicating the presence of an LVI clone. Finally, some of the LCM cell clusters from the LNs exhibited a similar gene mutation pattern, indicative of a SNV LN clone. Mutations were found in the LVI and LN SNV clones at a range of frequencies (LVI SNV clone 11.1%–100%, LN SNV clone 18.2%–100%), affecting >540 and 450 genes, respectively. Furthermore, cell clusters associated with the SNV LVI clone were found to have, on average, a higher homozygous mutation burden, suggesting that the degree of SNV homozygosity may be related to LVI invasiveness (Figure S2.1).

LVI cell clusters contain numerous oncogenic mutations, including *BRCA1* and *TP53*

Using our WGS and WES results, we next sought to identify potentially significant oncogenic events that may have occurred in this patient. We compiled a list of >89 known transformation-associated genes and assessed each cell cluster for the frequency of SNV missense, splice site, and frameshift mutations, as well as for CNV-associated chromosome deletion and amplification events. In total, 59 of the genes were found to harbor either SNV or CNV mutations (Figure 2C). As commonly

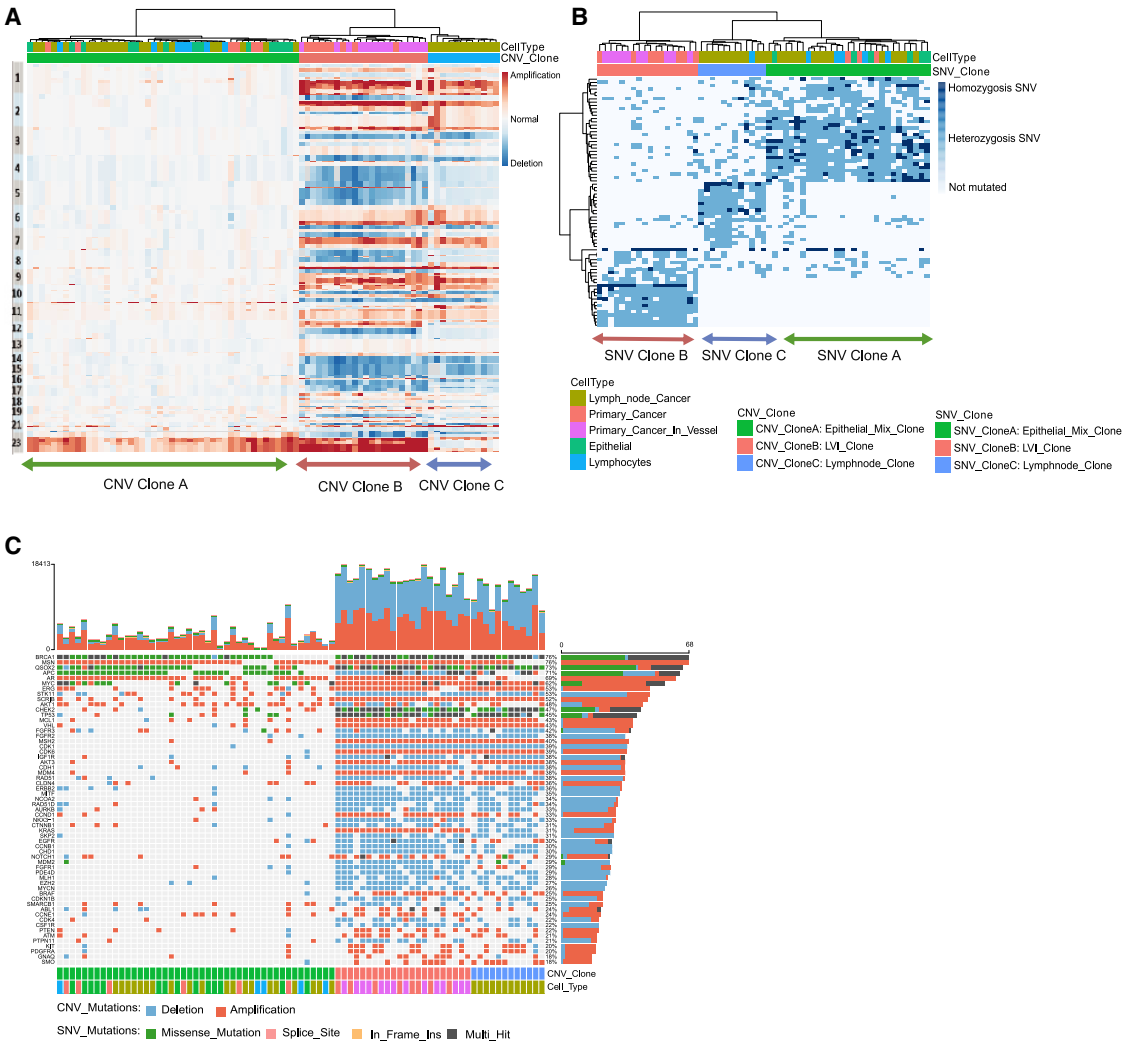


Figure 2. CNV and SNV sequencing reveal the presence of a distinct LVI cancer cell clone among a vast mutational landscape

(A) CNV heatmap generated by unsupervised hierarchal clustering of 80 cell clusters following WGS. Cell clusters are plotted along the x axis, and CNVs are plotted in genomic order along the y axis. Color bars (top) represent cell type (yellow, LN; red, primary tumor; purple, primary tumor-associated vessels; green, epithelial; blue, lymphocytes) and clone designation (green, mixed clone; red, intravasation clone; blue, LN clone). Chromosomal amplifications are depicted in red, while deletions are shown in blue.

(B) SNV heatmap generated by unsupervised hierarchal clustering of 59 cell clusters following WES. Cell clusters are plotted along the x axis, and SNVs are plotted according to clone prevalence along the y axis. The phylogeny on the left shows an unsupervised clustering to classify the most abundant SNVs in each clone. Cell type and clone designation are indicated as in (A). Homozygous SNVs are depicted as dark blue, heterozygous SNVs are light blue, and genes with no mutations are white. Only SNVs present in more than half of the cell clusters of each clone were shown. All SNVs can be found in Table S1.2 and are shown in Figure S2A.

(C) Oncoplot depicting chromosome alterations and/or genetic mutations detected in 59 common oncogenes. Each vertical column represents the CNV (amplification, red; deletion, blue), SNV (missense, green; splice site, pink; frameshift, yellow) or multihit (black) mutations detected in each oncogene from the gDNA of each cell cluster following WGS and WES sequencing. Stacked bars (top) show the accumulated alterations (both CNV and SNV) across all 23,588 reference genes in each cell cluster. Stacked bars (right) show the accumulated number of CNV and SNV mutations detected in each oncogene across all 80 cell clusters and the percentage of cell clusters harboring mutations in each oncogene. Individual cell clusters are arranged based on CNV clone association, and oncogenes are arranged by mutation/chromosome alteration frequency.

observed in young TNBC patients,[20–22] the DNA repair-associated gene *BRCA1* either contained a missense mutation, was deleted, or exhibited multiple transformative hits in ~76% of all cell clusters isolated, including those from the mixed-epithelial clone A. Other genes affected by such variabilities included the tumor suppressor *TP53*, which was mutated or exhibited multiple hits in the LVI and LN clones, and *APC*, which was mutated in ~71% of all of the cell clusters isolated. In accordance with an amplification of the X chromosome, two genes, *AR* and *MSN*, located on the X chromosome, were amplified in a high proportion of mixed-epithelial and LVI clones, but to a lesser degree in the LN (Table S1.3) clone. Furthermore, *MYC*, *ERG*,

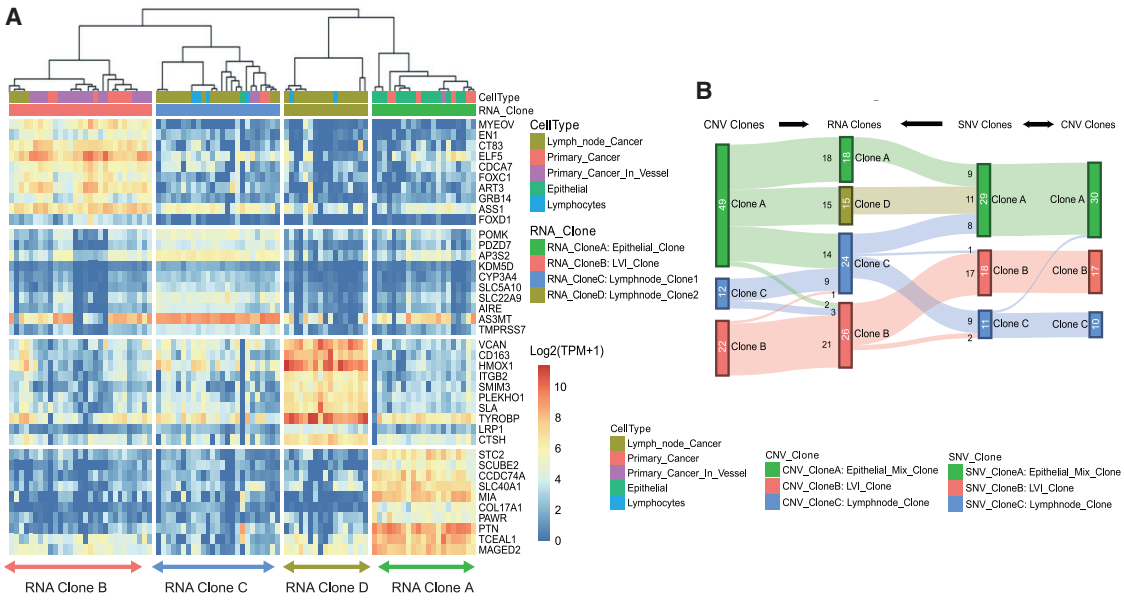


Figure 3. G&T-seq identifies 4 RNA-based clones in a single TNBC patient and permits direct comparison between DNA and RNA clones

(A) RNA heatmap depicting the unsupervised hierarchal clustering of 92 cell clusters based on their relative expression of select marker genes. Cell clusters are plotted along the x axis and marker genes are plotted along the y axis. Color bars (top) represent the cell type (yellow, LN; red, primary tumor; purple, primary tumor associated lymphatic vessels; green, epithelial cells; blue, lymphocytes) and RNA clone designation (green, epithelial clone; red, LVI clone; blue, LN subclone 1; yellow, LN subclone 2). Increases in relative gene expression are represented on a Log2 (TPM+1) scale where 10 (dark red) is equivalent to >1,000-fold increase and 0 (dark blue) indicates no relative increase in gene expression.

(B) Sankey diagram comparison of CNV clones, SNV clones, and RNA clones. For any pairwise comparison, we show only those cell clusters that were shared by both pairs. RNA had 92 cell clusters pass quality control (QC), CNV had 84 cell clusters pass QC, but only 83 were shared. SNV had 59 cell clusters pass QC, and of these, 57 were shared with either RNA or CNV. Clones are shown by the colors defined in (A).

MCL1, *VHL*, *MSH2*, and *CDK6* were amplified, while *STK11*, *FGFR2*, and *IGF1R* were deleted in nearly all of the LVI and LN clones. Overall, this demonstrates the complex CNV and SNV landscape in this patient, with genetic alterations affecting not only the CIN-associated LVI and LN clones but also the mixed-epithelial clone.

Two distinct LN RNA clones are identified by whole-transcriptome sequencing

In metastatic breast cancer, the tumor microenvironment varies considerably among the primary tumor, the tumor-associated lymphatic vessels, and the sentinel LNs. This influences the gene expression patterns. WTS was performed on the mRNA extracted from each cell cluster to detect the relative expression of 23,588 RefSeq-curated genes. RNA clustering proved more challenging than CNV and SNV clustering insofar as the previously deployed algorithms did not provide robust bootstrap values. We therefore used the single-cell consensus clustering (SC3) algorithm,[23] which generates high accuracy and robustness by combining multiple unsupervised clustering solutions using a consensus approach. SC3 computes a silhouette width and a stability index, which lets users evaluate the optimal number (k) of clones. Both methods concluded that four was the optimum for our dataset (Figure S3). The first clone identified was called an epithelial clone (RNA clone A) because cell clusters with this gene expression pattern were predominantly composed of histologically normal breast epithelial cells. Again, LCM cell clusters isolated from lymphatic vessels grouped together and were found to exhibit a unique RNA expression pattern, indicating the presence of a LVI clone (RNA clone B). Although the SC3 algorithm uses all the genes for clustering purposes, some genes are more statistically informative, and these are the marker genes depicted in Figure 3A. Marker genes highly expressed by cell clusters with an LVI clone phenotype included several related to tumor invasion and cell differentiation, such as *FOXC1/D1/Q1*, *NOTCH1*, *ART3*, *BIRC7*, *RAB40B*, *PTP4A3*, *CDK1*, *CLDN4/7*, *FGFR3*, *QSOX2*, *AURKB*, *SCRIB*, and *CCNB1* (Table S2.1). Unlike our CNV and SNV analysis, the transcriptional assessment detected the presence of two distinct LN RNA clones, each with a unique pattern of gene expression. RNA LN clone 2 (RNA clone D) exhibited a very high expression of a number of immune-related genes such as *CD163*, *HMOX1*, and *TYROBP*, as well as several genes involved in cell migration such as *VCAN* and *CTSH*. RNA LN clone 1 (RNA clone C) by comparison demonstrated a relatively muted level of gene expression, with *POMK*, *CYP3A4*, and several solute carriers being among the most highly expressed genes. Thus, while these LN clones are related in terms of physical location, their gene expression profiles are distinct, suggesting that they are genetically unrelated.

G&T-seq enables direct comparison between DNA and RNA clones

G&T-seq facilitates the analysis of DNA and RNA from the same sample, but it is still important to validate the clonal assignments

and confirm the relationships between the CNV, SNV, and RNA clones identified. To visualize these relationships, we used a Sankey diagram (Figure 3B). In Figure 3B, we present the flow from CNV clones → RNA clones in the left panel, the flow from SNV clones → RNA clones in the center panel, and compare CNV clones with SNV clones in the right panel. Almost all (56/57 ≈ 98%) of the CNV clones and SNV clones were perfected correlated. Therefore, in comparing to RNA clones, we need only focus on CNV clones and can ignore SNV clones. For RNA clone B, 21/26 ≈ 81% of the cell clusters matched CNV clone B, both of which came from the LVI. Interestingly, RNA clones A and D were both entirely matched to CNV clone A, demonstrating that the same CNV (and SNV) profile can match to completely different RNA profiles, perhaps as the result of the different microenvironments around the cell clusters (i.e., RNA clone A came from the primary tumor and RNA clone D came from the LN). Lastly, for RNA clone C, 14/24 ≈ 58% of the cell clusters matched CNV clone A, while 9/24 ≈ 38% matched CNV clone C and 1/24 ≈ 4% matched CNV clone B. The overall lesson is that while CNV (and SNV) profiles can be matched to RNA profiles, they are not by themselves sufficient to predict RNA profiles.

For individual cell clusters that matched consistently to a CNV, an SNV, and an RNA clone, we also generated Venn diagrams to depict the number of shared genes (Figure S2.2). Little mutual overlap was observed for genes associated with either the mixed-epithelial clone or LN clone 2. However, the number of overlapping genes, particularly those affected by both CNV and SNV changes, was higher in cell clusters associated with LN clone 1 and the LVI clone. This is consistent with these clones exhibiting a large number of genomic alterations, and the LVI clone being the most homogeneous in terms of cell-type composition. Furthermore, the number of overlapping genes between CNV and RNA clones was in general much higher compared to SNV and RNA clones. This suggests that while point mutations may have played a role in shaping the transcriptome, CNVs appeared to have a larger influence.

G&T-seq conveys the evolutionary history of tumor cell clones and pathways of metastatic spread

Having defined LVI, LN, and mixed/epithelial clones from G&T-seq data, we infer how they are evolutionarily related. To begin, we constructed a CNV maximum parsimony tree to trace descent from common ancestors, and measure evolutionary distances between each CNV clone.[9,19,24] LCM cell clusters were plotted against hamming distance and categorized based on cell type and location (LN versus breast tissue). Overall, the CNV phylogenetic tree was rooted by the mixed-epithelial clone with relatively few CNVs, and had evidence of both branched and gradual evolution (Figure 4A). Two early divergent subpopulations were identified branching from the mixed epithelial clone, representing the LN and LVI clones, both of which are highly CIN. Once diverged, these two clones experienced many copy-number changes. By comparison, the mixed-epithelial clone exhibited a relatively flat evolutionary profile with just a few minor diverging subpopulations. These likely represent random changes in genetic copy number that ultimately were not advantageous to the tumor. A second maximum parsimony tree constructed from the SNV data showed a similar result, in which the LN and LVI clones evolved separately and gradually from an ancestor of mixed-epithelial origin (Figure 4B). In contrast to the CNV phylogenetic tree, the evolutionary distances between the three SNV clones were much smaller, and the SNV LVI clone seems to have arisen from a more closely related SNV LN clone. Overall, this phylogenetic analysis suggests that each of the CNV and SNV clones are evolutionarily distinct, and provides evidence for the LVI and LN clones being seeded by a mixed/epithelial clone originally present in the primary tumor.

As chromosomal aberrations and genetic polymorphisms do not readily translate into predictable gene expression changes, we assessed the evolution of our clones at a transcriptional level. For this analysis, we used the Monocle 2.0 algorithm to organize the cell clusters based on their transcriptional similarity and to construct transcriptional trajectories representative of the gene expression pattern exhibited by individual cell clusters during differentiation.[25,26] Cell clusters were mapped against a pseudotime that served as a quantitative measure of developmental progress. Branch points in the trajectories appear when variations in gene expression were found to yield distinct differentiation events. Using the marker genes defined by the SC3 algorithm, we plotted the RNA expression trajectories from Monocle 2.0. This analysis revealed the presence of two transcriptionally distinct tumor cell differentiation fates, or trajectories, each originating from cell clusters with an RNA epithelial clone phenotype (Figure 4C). Notably, the RNA LVI clone and the RNA LN2 clone were found to differentiate along a similar trajectory, suggesting that they are related to each other developmentally. Because the Sankey diagram (Figure 3B) showed that the RNA LVI clone maps to the CNV LVI clone with extensive chromosome aberrations, and the RNA LN2 clone maps to the CNV mixed-epithelial clone with relatively few chromosome aberrations, it is not possible for the RNA LVI clone and the RNA LN2 clone to be genetically related. It is unlikely that the RNA LVI clone seeded either of the two RNA LN clones. We concluded that there were three metastatic paths in this one patient, despite there being only two RNA trajectories. Pseudotime gene expression profiles and Venn diagram analysis confirmed that these paths were largely unique, each with a distinct gene expression pattern (Figure S4; Table S2.2).

Combining these metastatic paths with oncogene assessments, we can begin to reconstruct a simplified history of disease progression in this patient, as illustrated in Figure 4D. The original primary tumor likely had a *BRCA1* driver mutation that was inherited by most successor cells. Furthermore, because the X chromosome is amplified in the RNA epithelial, LVI, and LN2 clones but not in the RNA LN1 clone, we conclude that the RNA LN1 clone must have evolved from an ancestral clone in the primary tumor that was either not sampled by us, or is no longer present. Duplication of the X chromosome occurred later and led to the RNA epithelial clone that is still present in the primary tumor, and metastasized to the patient's LNs as the RNA LN2 clone. Based on the CNV phylogenetic tree, the RNA epithelial clone also seeded the RNA LVI clone, whose high mutational burden and CIN are most likely related to the chemotherapeutic regimen that was administered to this patient before our sample acquisition. This may have also played a role

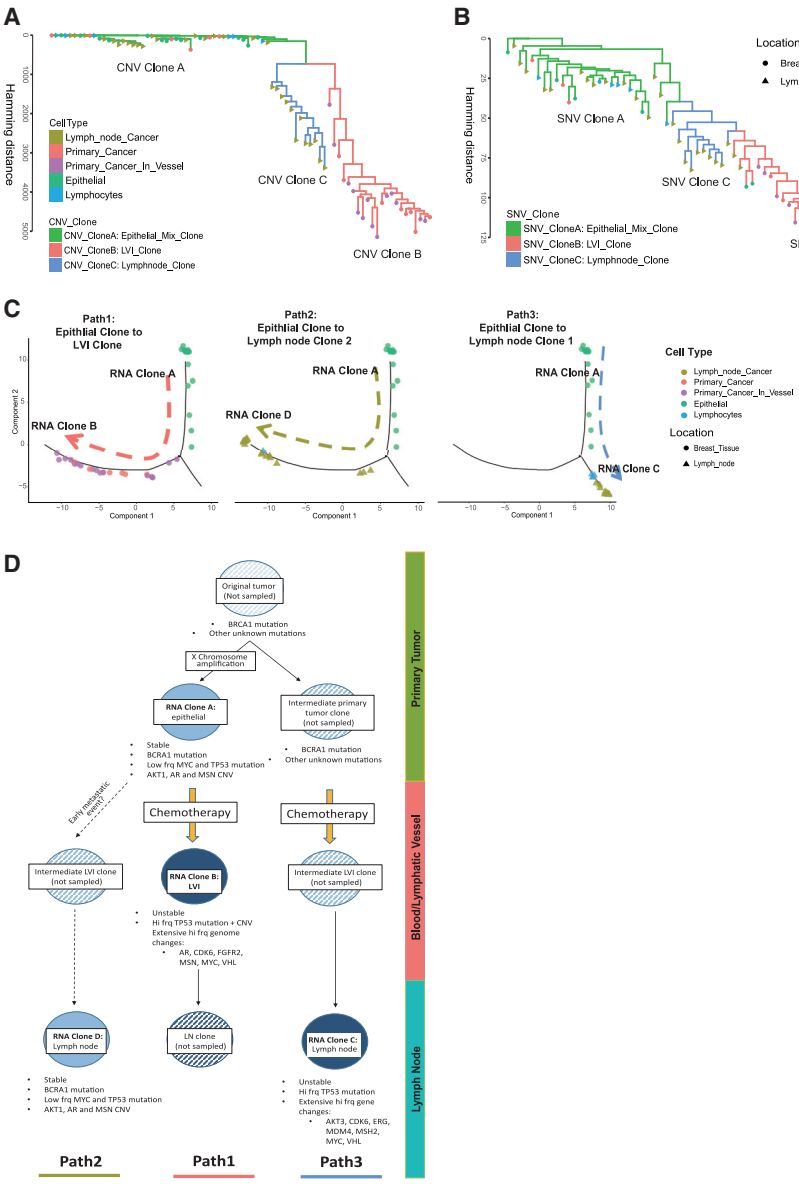


Figure 4. Evolutionary history of cancer cell clones reveals 3 distinct pathways of metastatic disease in one patient

(A and B) Maximum parsimony tree constructed based upon the CNV (A) and SNV (B) detection. Each marker represents an individual cell cluster isolated from either breast tissue (circle) or LN tissue (triangle). Markers are colored according to their cell type. Lines between cell cluster markers correspond to CNV/SNV clone type and represent the evolutionary distance between cell clusters.

(C) Metastatic transcriptional pathways of each RNA clone as defined by the Monocle 2.0 algorithm. Individual cell clusters are defined by cell type and tissue location as described in (A) and (B) and plotted against pseudotime.

(D) Combined reconstruction of the evolutionary history of the tumor cell-clusters isolated from our patient. Clones detected by G&T-seq are depicted by solid circles, while unsampled clones are represented with hash marks. The key genes potentially affected by oncogenic events are shown below each clone. Clone location is shown on the right (epithelial/primary tumor, green; lymphatic vessel, red; LN, blue) and metastatic pathway (Path1, Path2, Path3) is depicted on the bottom.

in the development of the genomically unstable RNA LN1 clone, derived from an unsampled primary tumor clone that lacked the X chromosome amplification and metastasized to the LNs. In summary, with limited tissue sampling, G&T-seq allowed us to reconstruct a complex model of tumor progression and metastasis in this patient.

Gene enrichment identifies distinct metastatic and immune-related biological processes for each of three paths

Having deduced the existence of three transcriptionally distinct metastatic paths in this patient, we sought to determine the key biological processes and genes associated with each of these paths. Following gene-set enrichment of significantly changing genes associated with each path by Metascape[27] (Table S3.1), we selected the 20 most representative biological processes across all 3 metastatic paths and generated a heatmap to compare their relative significance to each path (Figure 5A). Similar to other reports, all three paths were enriched in biological processes related to epithelium morphogenesis,[28] suggesting that epithelial changes provided a foundation for metastasis. Furthermore, the two metastatic paths exhibiting genome instability (path 1, RNA LVI clone; path 3, RNA LN1 clone) shared biological processes related to cell cycle and oxidative stress, indicative of rapid proliferation and a role for hypoxia. Metastatic paths involving the primary cancer cells (path 1, RNA LVI clone; path 2, RNA LN2 clone) were enriched in biological processes for blood vessel development, epithelial cell migration, and regulation of cell adhesion, all of which are related to tumor cell invasion and migration.[29–31] Each metastatic path was also associated with unique biological processes such as chromosome segregation and T-helper cell differentiation for path 1; extracellular matrix organization, Toll-like receptor signaling, and response to interferon-γ (IFN-γ) for path 2. Path 3 was enriched in biological processes associated with immune regulation—for example, antigen processing and presentation via major histocompatibility complex class I (MHC class I), cellular metabolism (e.g., oxidative phosphorylation), and translation termination. These results are detailed in Figures 5A and S5 and Table S3.1.

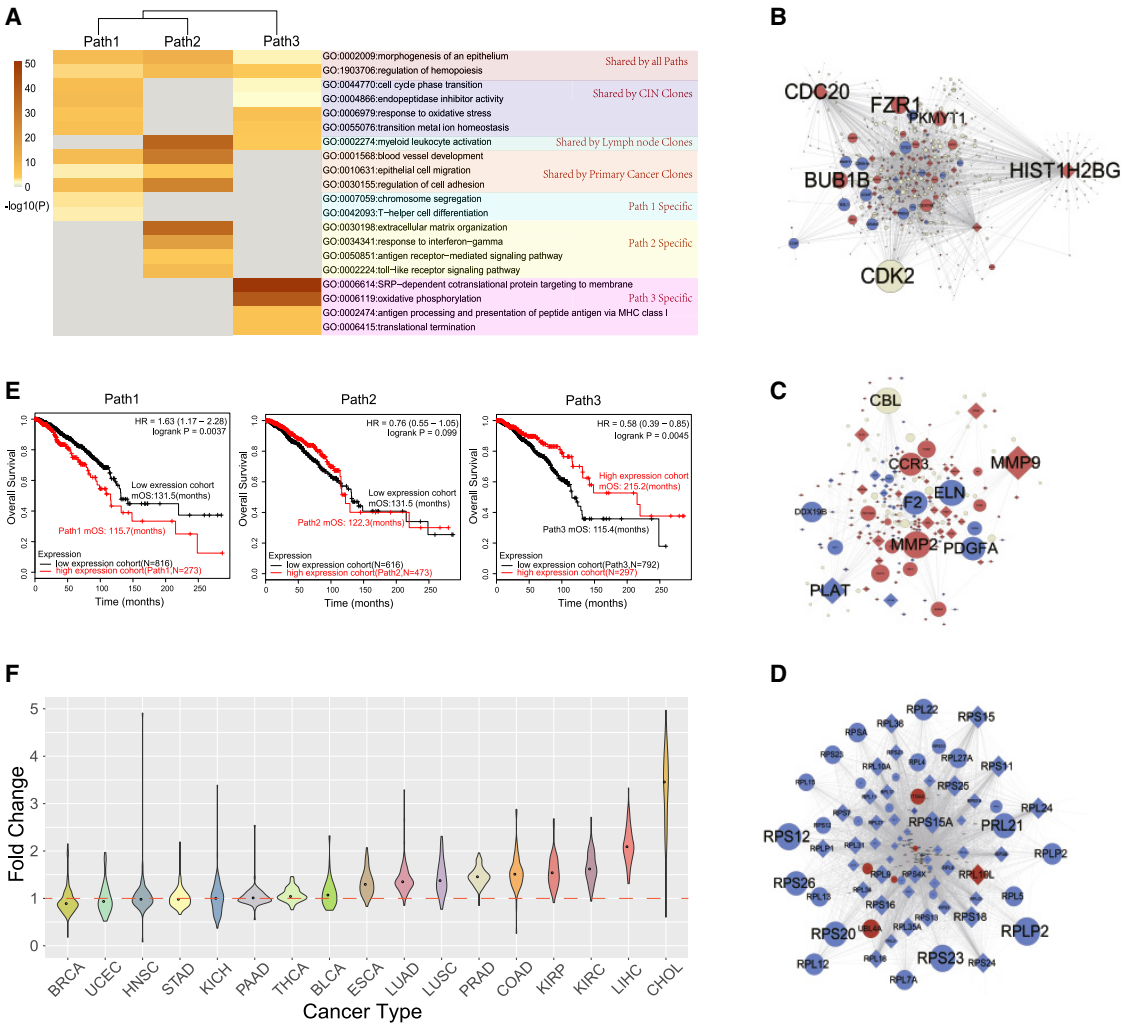


Figure 5. Biological process, hub gene, and overall survival (OS) analysis of each metastatic pathway reveals distinct gene signatures associated with variable patient outcomes

(A) Heatmap depicting the relative significance of the top 20 Metascape-defined biological process groupings to each metastatic pathway. Biological processes are arranged based upon being (1) shared by all paths, (2) shared by CIN clones, (3) shared by LN clones, (4) shared by primary cancer clones, or (5) path specific. Scale coloring represents the log p value for the indicated biological process grouping.

(B–D) PPI network map of hub genes associated with path 1 (B), path 2 (C), and path 3 (D). Nodes represent individual genes, and edges denote a known association between genes. Red nodes represent genes exhibiting an increase in expression, while blue nodes represent genes with decreased expression. Hub genes, which were among the top 10 hits from each metastatic path but did not demonstrate a significant change in expression, are depicted in gray. Node size denotes the relative importance of each gene to the network.

(E) Kaplan-Meier plots depicting OS outcomes of breast cancer patients with high (red) versus low (black) expression for the top 20 hub genes associated with each metastatic path. OS curves were generated using KM-plotter from a cohort of 1,089 breast cancer tumor samples and corresponding patient survival data. Hazard ratios and log-rank p values (0.0037, 0.099, 0.0045 for paths 1 to 3) are shown in the top right-hand corner of each Kaplan-Meier curve. mOS, median overall survival.

(F) Violin plots depicting the fold change in expression of 80 RPL and RPS genes across 17 cancer types in comparison to normal tissue. RPL and RPS expression data were derived from the TCGA database incorporating ENCORI. Tumor types are sorted from lowest to highest median fold change in ribosomal protein expression. The median fold change value of each cancer type is presented by a dot in the center of each violin plot. The horizontal orange dashed line marks a fold change value of 1.0. The violin shape corresponds to the density of data (more bulbous means more data). The tumor types assessed are as follows: BRCA, breast-invasive carcinoma; UCEC, uterine corpus endometrial carcinoma; HNSC, head and neck squamous cell carcinoma; STAD, stomach adenocarcinoma; KICH, kidney chromophobe; PAAD, pancreatic adenocarcinoma; THCA, thyroid carcinoma; BLCA, bladder urothelial carcinoma; ESCA, esophageal carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PRAD, prostate adenocarcinoma; COAD, colon adenocarcinoma; KIRP, kidney renal papillary cell carcinoma; KIRC, kidney renal clear cell carcinoma; LIHC, liver hepatocellular carcinoma; CHOL, cholangiocarcinoma.

We used the Cytoscape application CHAT[32] to identify hub genes (i.e., highly interconnected genes critical to the gene network) associated with each metastatic path and to compute protein-protein interactions (PPIs). The hub genes associated with path 1 (RNA LVI clone) included those related to cell cycle (*FZR*), cell division (*CDC20*), and nucleosome structure

(*HISTIH2BG*) (see Figure 5B and Table S2.3). This is consistent with chromosome segregation being a key biological process associated with path 1, implying that CIN caused by abnormal chromosome segregation may have led to metastasis by LVI-associated clonal populations. Path 2 (RNA LN2 clone) was associated more with inflammation and invasion-related matrix metalloprotease genes (*MMP9, MMP2*), as well as genes for cell adhesion *(VCAN*, *ITGB3*) (see Figure 5C and Table S2.3). Elevated expression of these particular genes is consistent with Toll-like receptor signaling and antigen receptor−mediated signaling being enriched biological processes, suggesting that path 2 may be related to the epithelial-mesenchymal transition (EMT)-inflammation axis of cancer cell metastasis. Finally, almost all of the hub genes (48 of the top 50) associated with path 3 (RNA LN1 clone) were ribosomal, and all of these were downregulated, with the exception of *RPL10L* (see Figure 5D and Table S4.1). This reduction in the expression of transcripts encoding ribosomal proteins, combined with enrichment for biological processes antigen processing and presentation via MHC class I, TAP-dependent (GO [Gene Ontology]: 0002479), and translation termination, suggests that path 3 is trying to limit protein synthesis and neo-antigen presentation on cell surfaces.

Extensive downregulation of transcripts encoding ribosomal proteins is a potential marker of poor prognosis in breast cancer

Finally, we aimed to determine whether the hub gene expression profiles associated with each metastatic path correlated with differences in breast cancer survival outcomes. To extrapolate our findings beyond this one TNBC patient, we used KM-plotter,[33] which incorporates gene expression and survival data from publicly accessible databases to generate Kaplan-Meier overall survival (OS) curves. By comparing the survival outcomes of patients exhibiting high versus low mean expression of the top 20 hub genes from each metastatic path, we inferred that path 1 (RNA LVI clone) and path 3 (RNA LN1 clone) are correlated with significantly worse prognoses for breast cancer patients (Figure 5E; $p = 0.0037$, hazard ratio [HR] 1.63 [1.17–2.28] and $p = 0.0045$, HR 0.58 [0.39–0.85], respectively). By contrast, path 2 showed little impact on breast cancer OS ($p = 0.099$, HR 0.76 [0.55–1.05]). Moreover, of the 1,089 patient samples we included in this OS analysis, a larger number were found to exhibit a path 3 hub gene expression signature ($N = 792$), in comparison to both path 1 ($N = 273$) and path 2 ($N = 473$). Knowing that increased ribosome biogenesis is a hallmark of many cancers and a recognized marker of poor disease prognosis,[34] our finding that the extensive downregulation of both S ribosomal proteins (RPS) and L ribosomal proteins (RPL) associated with path 3 correlated with worse clinical outcomes was unexpected, and it warranted further investigation. Of our 80 ribosomal protein genes, 78 were found to be downregulated in path 3 (Table S4.1), and 64 of these had a log rank $p < 0.05$ in KM-plotter, correlating with a poor prognosis when downregulated. Only *RPL10L*, which exhibited an increased expression in path 3, was associated with improved prognosis ($p = 3.2e{-}5$).

To determine whether such ribosomal protein gene downregulation occurs universally in breast cancer, or potentially, even in other cancer types, we performed a pan-cancer differential gene expression analysis of 80 RPL and RPS ribosomal protein genes across 17 different cancer types using the expression data from The Cancer Genome Atlas (TCGA) with integration by ENCORI (Encyclopedia of RNA Interactomes[35]) (Table S4.2). We found that ribosomal protein genes are broadly downregulated in breast cancer, with 51 of 80 genes having significantly reduced expression in comparison to normal tissue. Furthermore, breast cancer had the lowest median fold change (<0.9) in ribosomal protein expression, when compared to other cancer types (Figure 5F; Table S4.3). By contrast, 9 of 17 cancer types demonstrated extensive ribosomal protein gene upregulation, with >1.1 median fold change. This indicates that the reduced expression of ribosomal proteins is a potential marker for poor survival outcome in breast cancer.

DISCUSSION

Lymphovascular invasive tumor cells have received little attention, despite their strong correlation with clinical outcomes in breast and other cancers. We used a multi-omics approach to study pathologist-defined cell clusters sampled from multiple sites in one TNBC patient. Combining laser capture microdissection with simultaneous genome and transcriptome sequencing, we were able to explore the genetic mechanisms promoting the metastatic spread of breast-derived cells from the primary tumor to the axillary LNs. Our analyses revealed a highly heterogeneous genetic profile, with a vast mutational landscape, typified by extensive chromosome aberrations and single-nucleotide mutations. These mutations affected many key oncogenes such as *BRCA1*, *TP53*, *CHEK2*, *APC*, *MYC*, *CDK6*, and *MCL1*. They were present not only in cancerous cells but also histologically normal breast epithelial cells. Our analysis found that LVI-associated breast cancer cells exhibited common chromosomal and transcriptional features such as CIN. Moreover, we discovered evidence of polyclonal metastasis in this patient, with three transcriptionally distinct metastatic pathways identified.

It is important to note that the cell clusters analyzed in these experiments were dissected directly from H&E-stained pathological slides prepared from frozen tissues. This enabled the histological identification and dissection of specific cells based on their association with specific morphological structures. Importantly, we were able to recover clusters of lymphovascular invasive cells directly from lymphatic vessels, which, due to the unidirectionality of lymph flow, represent disseminated tumor cells categorically *en route* to the axillary LNs from the primary tumor. Furthermore, because previous breast cancer studies found LVI and LN metastasis to be mediated by collective cellular invasion,[36] these methods allowed us to characterize what we believe are more physiologically representative multicellular invasive units.

Performing unsupervised hierarchal clustering, we found LVI-associated cells to be genomically and transcriptionally similar, exhibiting an exceptionally large number of chromosome aberrations and expressing high levels of a number of genes related to invasion and EMT regulation such as *FOXC1*, *ART3*, *BIRC7*, *RAB40B*, *PTP4A3*, and *NOTCH1*. Our data also suggest that LVI cells are genomically unstable, consistent with *BCRA1* being a founding driver mutation in our patient, and spindle checkpoint/chromosome

maintenance genes being identified as key biological processes. Chromosome aberrations have been observed in many human cancers,[37] and recent work[38] has linked CIN to metastatic spread, whereby the rupture of CIN-induced micronuclei elicits a cytosolic DNA response, which ultimately converge onto noncanonical nuclear factor κB (NF-κB) activation concomitant with increased metastasis in animal models. Although the number of NF-κB and inflammatory genes directly associated with the LVI clones in our study is limited, we believe that a similar mechanism may still be at play, as RNA trajectory analysis placed LVI cells in a state of differentiation preceding that of the LN cancer RNA clone D (RNA LN2 clone), which had a clear inflammatory and invasive phenotype. Moreover, because the LVI-associated cells in our study were isolated from within lymphatic vessels, their transcriptional profiles may be suppressed, as a method to circumvent immune surveillance and physical stresses of lymphatic transit. It is also worth mentioning that a number of primary tumor cell clusters were found to exhibit a CNV/SNV profile similar to that of the LVI cells, suggesting that these invasive cells may have arisen directly from cancer cells positioned along the stromal interface of the primary tumor. While spatial mapping of the cell clusters collected would be required to confirm this association, similar correlations have been described by other groups,[39] suggesting that these primary cancer cells could be therapeutically targeted to inhibit the metastatic spread of disease via the lymphatic system.

Our study also found evidence of polyclonal metastasis, with two genetically and transcriptionally distinct cancer cell populations identified in the LNs of this patient. Based on evolutionary evidence, these two populations were likely established by asynchronous metastatic events that occurred at both early (RNA LN2 clone) and late (RNA LN1 clone) time points. This is consistent in part with the theory that early dissemination can seed metastatic breast cancer.[40,41] Moreover, the two LN clones were associated with transcriptionally distinct paths, and disparate OS outcomes when extrapolated to a cohort of 1,089 breast cancer patients. Path 2 (RNA LN2 clone) exhibited a relatively benign gene signature compared to path 1 (RNA LVI clone) and path 3 (RNA LN clone 1). This is consistent with path 2 exhibiting a more pro-inflammatory phenotype that may encourage immune detection, and points to a role for the EMT-inflammation axis in promoting disease progression. The possibilities are not mutually exclusive, since immune surveillance mechanisms restricting tumor growth can be overcome by tumor-induced immunosuppressive changes, as has been extensively reviewed.[42–46]

Aggressive cancers are generally believed to have a high rate of protein synthesis. Ribosomal protein expression is often increased in most cancers and regarded as a marker of poor disease prognosis.[34] Accordingly, we were very surprised that path 3 (RNA LN1 clones) showed a reduced expression of transcripts encoding numerous ribosomal proteins, and that this was associated with worse survival outcomes in breast cancer patients. We hypothesize that path 3 is trying to limit protein synthesis and neo-antigen presentation on cell surfaces. This could serve as both a coping mechanism in response to proteostatic stress[34,47] and an immune evasion strategy to prevent cytotoxic T cell responses with immune cell infiltration of the tumor.[48] The extensive downregulation in ribosomal protein expression and its effect on survival outcomes appears to be unique to breast cancer, with most other cancer types exhibiting increased expression of these genes. The mechanism behind this phenomenon, and its significance to breast cancer, remains unclear. Because these experiments are a temporal assessment of an advanced heterogeneous disease, it is uncertain whether this gene expression pattern is stable, or is a temporary phenotype in the spectrum of tumor cell plasticity. Recent studies have shown that energetics and protein synthesis rates are highly plastic, and alterations in these phenomena may underpin resistance to therapy.[49] Further investigations into the role of ribosomal protein expression in breast cancer progression must be explored.

Although there are obvious limitations to any publication based on just one patient, there are also questions that are difficult (if not impossible) to answer by the sequencing of bulk samples, especially when taken from just one body site. For example, simpler experimental designs may not have detected multiple exits from the primary tumor. Unfortunately, we did not have access to distant metastasis tissues from this one patient, so we could not determine which (if any) of our hypothesized LVI exits was their precursor. Previous studies have attributed 25%–35% of distant metastasis to LN metastasis, albeit by studying heterogenous bulk samples.[50,51] Future studies using multiregional sampling and simultaneous DNA/RNA sequencing (i.e., similar to our experiment but on a much larger number of patients) are likely required for definitive answers to such questions.

Limitations of study

Beyond the inherent limitations of a case study based on one patient, the deeper problem is that we could not sample more time points and more body sites. Hence, we had to infer the existence of multiple clones. We do not see any obvious way to avoid making such inferences since ethical issues will always make it difficult (if not impossible) to collect the requisite samples. The number of clones for the CNV and SNV analyses was robust at three, but the number of clones for the RNA analysis could conceivably have been five. We chose four to simplify the interpretation. The computed gene networks for the different paths were based on a comparison of RNA clones B, D, and C, for paths 1, 2, and 3, respectively, against RNA clone A. Since clone B came mostly from breast tissue and lymph vessels, while clones C and D came mostly from LNs, the microenvironments are different, and that too can influence gene expression. Experimental validation of the computed gene networks is also lacking, so the specific genes that are identified must be interpreted with caution. However, the survival outcome (Figure 5E) and ribosomal protein (Figure 5F) results used public data for over a thousand patients, which should be more robust.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Patient basic information
- METHOD DETAILS
  - Laser capture microdissection (LCM) and cell cluster lysis
  - Oligo-dT30VN bead labeling; mRNA and gDNA separation
  - Reverse transcription, amplification, purification of cDNA
  - Tn5 cDNA library preparation and sequencing
  - Purification and amplification of gDNA
  - Housekeeping test for MDA products
  - WGS library preparation and sequencing
  - WES library preparation and sequencing
  - WGS data processing and CNV calling
  - CNV heatmap construction and clone identification
  - Maximum parsimony tree construction
  - WES data processing, SNV clone calling, heatmap construction, and clone identification
  - WTS data processing and quality control
  - RNA clone calling for cell clusters
  - RNA trajectory reconstruction and gene set enrichment
  - DNA and RNA-clone comparison
  - Gene set enrichment and hub gene identification
  - Hub gene identification and Kaplan-Meier analysis
  - Pan-cancer ribosome protein gene analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - DNA and RNA amplification and library qualification
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2021.100404.

ACKNOWLEDGMENTS

This research was supported by the Science, Technology and Innovation Commission of Shenzhen Municipality under grant no. GJHZ20170314152701465 and the Joint Fund of the National Natural Science Foundation of China and the Natural Science Foundation of Guangdong Province (U1601224). This research was initiated with funding by Alberta Innovates, in the form of an AITF/iCORE Strategic Chair (RES0010334) to G.K.-S.W. We would also like to thank Dr. Ivan Topisirovic for his input and comments on the manuscript.

AUTHOR CONTRIBUTIONS

G.K.-S.W., J.R.M., B.L., Y.H., and X.Z. conceived the study. G.K.-S.W., W.W., and J.R.M. designed the experiments. A.A.J., G.L., and S.S. collected clinical samples. W.W., F.L., Q.Z., and T.J. performed the experiments. Z.Z., S.Q., C.C., and X.Z. analyzed the sequencing data. D.L.H., G.K.-S.W., Z.Z., J.R.M., and L.M.P. wrote the manuscript, with all of the other authors providing feedback.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 15, 2020
Revised: March 30, 2021
Accepted: August 25, 2021
Published: October 19, 2021

REFERENCES

1. Leong, A.S., and Zhuang, Z. (2011). The changing role of pathology in breast cancer diagnosis and treatment. Pathobiology *78*, 99–114.
2. Vogelstein, B., and Kinzler, K.W. (2015). The Path to Cancer–Three Strikes and You're Out. N. Engl. J. Med. *373*, 1895–1898.
3. Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.
4. Vogelstein, B., Fearon, E.R., Hamilton, S.R., Kern, S.E., Preisinger, A.C., Leppert, M., Nakamura, Y., White, R., Smits, A.M., and Bos, J.L. (1988). Genetic alterations during colorectal-tumor development. N. Engl. J. Med. *319*, 525–532.
5. Brouwer, A., De Laere, B., Peeters, D., Peeters, M., Salgado, R., Dirix, L., and Van Laere, S. (2016). Evaluation and consequences of heterogeneity in the circulating tumor cell compartment. Oncotarget *7*, 48625–48643.
6. Ellsworth, D.L., Blackburn, H.L., Shriver, C.D., Rabizadeh, S., Soon-Shiong, P., and Ellsworth, R.E. (2017). Single-cell sequencing and tumorigenesis: improved understanding of tumor evolution and metastasis. Clin. Transl. Med. *6*, 15.
7. Lawson, D.A., Bhakta, N.R., Kessenbrock, K., Prummel, K.D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C.-Y., et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature *526*, 131–135.
8. Karaayvaz, M., Cristea, S., Gillespie, S.M., Patel, A.P., Mylvaganam, R., Luo, C.C., Specht, M.C., Bernstein, B.E., Michor, F., and Ellisen, L.W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat. Commun. *9*, 3588.
9. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.
10. Bao, L., Qian, Z., Lyng, M.B., Wang, L., Yu, Y., Wang, T., Zhang, X., Yang, H., Brünner, N., Wang, J., and Ditzel, H.J. (2018). Coexisting genomic aberrations associated with lymph node metastasis in breast cancer. J. Clin. Invest. *128*, 2310–2324.
11. Aceto, N., Bardia, A., Miyamoto, D.T., Donaldson, M.C., Wittner, B.S., Spencer, J.A., Yu, M., Pely, A., Engstrom, A., Zhu, H., et al. (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell *158*, 1110–1122.
12. Cheung, K.J., Padmanaban, V., Silvestri, V., Schipper, K., Cohen, J.D., Fairchild, A.N., Gorin, M.A., Verdone, J.E., Pienta, K.J., Bader, J.S., and Ewald, A.J. (2016). Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. Proc. Natl. Acad. Sci. USA *113*, E854–E863.
13. Cheung, K.J., and Ewald, A.J. (2016). A collective route to metastasis: seeding by tumor cell clusters. Science *352*, 167–169.
14. Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods *12*, 519–522.
15. Yizhak, K., Aguet, F., Kim, J., Hess, J.M., Kübler, K., Grimsby, J., Frazer, R., Zhang, H., Haradhvala, N.J., Rosebrock, D., et al. (2019). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science *364*, eaaw0726.
16. Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P., Dentro, S.C., Dawson, K.J., Butler, T., Rahbari, R., Mitchell, T.J., et al. (2020). The mutational landscape of normal human endometrial epithelium. Nature *580*, 640–646.

17. Au, S.H., Storey, B.D., Moore, J.C., Tang, Q., Chen, Y.L., Javaid, S., Sarioglu, A.F., Sullivan, R., Madden, M.W., O'Keefe, R., et al. (2016). Clusters of circulating tumor cells traverse capillary-sized vessels. Proc. Natl. Acad. Sci. USA *113*, 4947–4952.

18. Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., et al. (2012). Genome-wide copy number analysis of single cells. Nat. Protoc. *7*, 1024–1041.

19. Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155–160.

20. Castilla, L.H., Couch, F.J., Erdos, M.R., Hoskins, K.F., Calzone, K., Garber, J.E., Boyd, J., Lubin, M.B., Deshano, M.L., Brody, L.C., et al. (1994). Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer. Nat. Genet. *8*, 387–391.

21. Cancello, G., Maisonneuve, P., Rotmensz, N., Viale, G., Mastropasqua, M.G., Pruneri, G., Veronesi, P., Torrisi, R., Montagna, E., Luini, A., et al. (2010). Prognosis and adjuvant treatment effects in selected breast cancer subtypes of very young women (<35 years) with operable breast cancer. Ann. Oncol. *21*, 1974–1981.

22. Criscitiello, C., Azim, H.A., Jr., Schouten, P.C., Linn, S.C., and Sotiriou, C. (2012). Understanding the biology of triple-negative breast cancer. Ann. Oncol. *23* (*Suppl 6*), vi13–vi18.

23. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods *14*, 483–486.

24. Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., Zhang, H., et al. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat. Genet. *48*, 1119–1130.

25. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

26. Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315.

27. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523.

28. Blick, T., Widodo, E., Hugo, H., Waltham, M., Lenburg, M.E., Neve, R.M., and Thompson, E.W. (2008). Epithelial mesenchymal transition traits in human breast cancer cell lines. Clin. Exp. Metastasis *25*, 629–642.

29. Wenes, M., Shang, M., Di Matteo, M., Goveia, J., Martín-Pérez, R., Serneels, J., Prenen, H., Ghesquière, B., Carmeliet, P., and Mazzone, M. (2016). Macrophage Metabolism Controls Tumor Blood Vessel Morphogenesis and Metastasis. Cell Metab. *24*, 701–715.

30. Uribesalgo, I., Hoffmann, D., Zhang, Y., Kavirayani, A., Lazovic, J., Berta, J., Novatchkova, M., Pai, T.P., Wimmer, R.A., László, V., et al. (2019). Apelin inhibition prevents resistance and metastasis associated with anti-angiogenic therapy. EMBO Mol. Med. *11*, e9266.

31. Primac, I., Maquoi, E., Blacher, S., Heljasvaara, R., Van Deun, J., Smeland, H.Y.H., Canale, A., Louis, T., Stuhr, L., Sounni, N.E., et al. (2019). Stromal integrin α11 regulates PDGFR-β signaling and promotes breast cancer progression. J. Clin. Invest. *129*, 4609–4628.

32. Muetze, T., Goenawan, I.H., Wiencko, H.L., Bernal-Llinares, M., Bryan, K., and Lynn, D.J. (2016). Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks. F1000Res. *5*, 1745.

33. Nagy, Á., Lánczky, A., Menyhárt, O., and Győrffy, B. (2018). Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. Sci. Rep. *8*, 9227.

34. Pelletier, J., Thomas, G., and Volarević, S. (2018). Ribosome biogenesis in cancer: new players and therapeutic avenues. Nat. Rev. Cancer *18*, 51–63.

35. Li, J.H., Liu, S., Zhou, H., Qu, L.H., and Yang, J.H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. *42*, D92–D97.

36. Giampieri, S., Manning, C., Hooper, S., Jones, L., Hill, C.S., and Sahai, E. (2009). Localized and reversible TGFbeta signalling switches breast cancer cells from cohesive to single cell motility. Nat. Cell Biol. *11*, 1287–1296.

37. Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

38. Bakhoum, S.F., Ngo, B., Laughney, A.M., Cavallo, J.A., Murphy, C.J., Ly, P., Shah, P., Sriram, R.K., Watkins, T.B.K., Taunk, N.K., et al. (2018). Chromosomal instability drives metastasis through a cytosolic DNA response. Nature *553*, 467–472.

39. Hoadley, K.A., Siegel, M.B., Kanchi, K.L., Miller, C.A., Ding, L., Zhao, W., He, X., Parker, J.S., Wendl, M.C., Fulton, R.S., et al. (2016). Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases. PLoS Med. *13*, e1002174.

40. Hosseini, H., Obradović, M.M.S., Hoffmann, M., Harper, K.L., Sosa, M.S., Werner-Klein, M., Nanduri, L.K., Werno, C., Ehrl, C., Maneck, M., et al. (2016). Early dissemination seeds metastasis in breast cancer. Nature *540*, 552–558.

41. Schwartz, R.S., and Erban, J.K. (2017). Timing of Metastasis in Breast Cancer. N. Engl. J. Med. *376*, 2486–2488.

42. Schreiber, R.D., Old, L.J., and Smyth, M.J. (2011). Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science *331*, 1565–1570.

43. Quail, D.F., and Joyce, J.A. (2013). Microenvironmental regulation of tumor progression and metastasis. Nat. Med. *19*, 1423–1437.

44. Liu, Y., and Cao, X. (2016). Characteristics and Significance of the Premetastatic Niche. Cancer Cell *30*, 668–681.

45. Suarez-Carmona, M., Lesage, J., Cataldo, D., and Gilles, C. (2017). EMT and inflammation: inseparable actors of cancer progression. Mol. Oncol. *11*, 805–823.

46. Dominguez, C., David, J.M., and Palena, C. (2017). Epithelial-mesenchymal transition and inflammation at the site of the primary tumor. Semin. Cancer Biol. *47*, 177–184.

47. Bublik, D.R., Bursać, S., Sheffer, M., Oršolić, I., Shalit, T., Tarcic, O., Kotler, E., Mouhadeb, O., Hoffman, Y., Fuchs, G., et al. (2017). Regulatory module involving FGF13, miR-504, and p53 regulates ribosomal biogenesis and supports cancer cell survival. Proc. Natl. Acad. Sci. USA *114*, E496–E505.

48. Bonaventura, P., Shekarian, T., Alcazer, V., Valladeau-Guilemond, J., Valsesia-Wittmann, S., Amigorena, S., Caux, C., and Depil, S. (2019). Cold Tumors: A Therapeutic Challenge for Immunotherapy. Front. Immunol. *10*, 168.

49. Uchenunu, O., Pollak, M., Topisirovic, I., and Hulea, L. (2019). Oncogenic kinases and perturbations in protein synthesis machinery and energetics in neoplasia. J. Mol. Endocrinol. *62*, R83–R103.

50. Naxerova, K., Reiter, J.G., Brachtel, E., Lennerz, J.K., van de Wetering, M., Rowan, A., Cai, T., Clevers, H., Swanton, C., Nowak, M.A., et al. (2017). Origins of lymphatic and distant metastases in human colorectal cancer. Science *357*, 55–60.

51. Venet, D., Fimereli, D., Rothé, F., Boeckx, B., Maetens, M., Majjaj, S., Rouas, G., Capra, M., Bonizzi, G., Contaldo, F., et al. (2020). Phylogenetic reconstruction of breast cancer reveals two routes of metastatic

dissemination associated with distinct clinical outcome. EBioMedicine *56*, 102793.

52. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
53. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164.
54. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.
55. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
56. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498–2504.
57. Isserlin, R., Merico, D., Voisin, V., and Bader, G.D. (2014). Enrichment Map - a Cytoscape app to visualize and explore OMICs pathway enrichment results. F1000Res. *3*, 141.
58. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545–15550.
59. Zafar, H., Wang, Y., Nakhleh, L., Navin, N., and Chen, K. (2016). Monovar: single-nucleotide variant detection in single cells. Nat. Methods *13*, 505–507. https://doi.org/10.1038/nmeth.3835.
60. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323. https://doi.org/10.1186/1471-2105-12-32.
61. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
62. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience *7*, 1–6.
63. Baslan, T., Kendall, J., Ward, B., Cox, H., Leotta, A., Rodgers, L., Riggs, M., D'Italia, S., Sun, G., Yong, M., et al. (2015). Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. Genome Res. *25*, 714–724.
64. Shah, S.P., Xuan, X., DeLeeuw, R.J., Khojasteh, M., Lam, W.L., Ng, R., and Murphy, K.P. (2006). Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics *22*, e431–e439.
65. Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. Stat. Med. *9*, 811–818.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human tumor biopsy tissue	Cross Cancer Institute, Edmonton, Alberta, Canada, T6G 1Z2	HREBA.CC-18-0470
Chemicals, peptides, and recombinant proteins		
Clear Frozen Section Compound	VWR	95057-838
Hematoxylin 7211	Thermo Fisher	7211
Eosin-Y	Thermo Fisher	7111
Lithium Carbonate (Bluing reagent)	Thermo Fisher	7301
Xylene	Fisher Chemicals	X 5-1
Ethyl alcohol anhydrous	Commercial alcohols	P016EAAN
Ethyl alcohol 95%	Commercial alcohols	P016EA95
Buffer RLT Plus	QIAGEN	Cat# 1053393
ERCC ExFold RNA Spike-In Mixes	Invitrogen	Cat# 4456739
M-280 Streptavidin Dynabeads®	Invitrogen	Cat #11205D
RNase inhibitor	NEB	Cat# M0314L
Nuclease-free water	Ambion	Cat# AM9938
MgCl2 1M	Invitrogen	Cat# 20-303
Betaine solution 5M	Sigma-Aldrich	Cat# B0300-1VL
SuperScript® II Reverse Transcriptase	Invitrogen	Cat# 18064-014
dNTPs (10mM)	NEB	Cat# N0447
2 × KAPA HiFi HotStart ReadyMix	KAPA BIOSYSTEMS	Cat# KK2602
Agencourt AMPure XP	AGENCOURT	Cat# A63881
Transposase	BGI	Cat # BGE005
10% SDS	Ambion	Cat# AM9822
ATP Solution, Tris buffered	Thermo Fisher	Cat# R1441
T4 DNA ligase (600U/μL)	Enzymatics	Cat# L6030
Exonuclease I (20U/μL)	NEB	Cat# M0293L
Exonuclease III (100U/μL)	NEB	Cat# M0206L
dNTPs (2.5mM)	Invitrogen	Cat# R72501
BSA	NEB	Cat# B9000S
rTaq (5U/μL)	Thermo Fisher	Cat# EP0402
Phosphate buffer saline (pH 7.4)	GIBCO	Cat# 10010-031
Critical commercial assays		
REPLI-g Single Cell Kit	QIAGEN	Cat# 150345
MGIEasy DNA Rapid Library Prep Kit	MGI	Cat# 200033-00
MGIEasy Exome Capture V4 Probe Set	MGI	Cat# 1000007745
Qubit dsDNA HS Assay kit	Invitrogen	Cat# Q32854
KAPA HiFi HotStart ReadyMixPCR Kit	KAPA BIOSYSTEMS	Cat# KR0370
Deposited data		
WGS sequencing data of 97 LCM cell cluster	This paper	https://db.cngb.org/cnsa (accession number CNP0000440)
WES sequencing data of 97 LCM cell cluster	This paper	https://db.cngb.org/cnsa (accession number CNP0000440)
RNA sequencing data of 97 LCM cell cluster	This paper	https://db.cngb.org/cnsa (accession number CNP0000440)

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oligonucleotides		
Biotinylated-Oligo-dT30VN primer	Sangon Biotech	5′-Bio-AAGCAGTGGTATCAACGCA GAGTACT30VN-3′
TSO	Sangon Biotech	5′-AAGCAGTGGTATCAACGCAGAGTACr GrG+G-3′
IS PCR Primer	Sangon Biotech	5′-AAGCAGTGGTATCAACGCAGAGT AC-3′
Splint oligo	Sangon Biotech	5′-GCCATGTCGTTCTGTGAGCCAAGG-3′
PhoAd153 F primer	Sangon Biotech	5′-phoGAACGACATGGCTACGATCCGA CTT-3′
Ad153 R primer	Sangon Biotech	5′- TGTGAGCCAAGGAGTTGTTGTC TTC-3′
Ad153-F-tag	Sangon Biotech	5′-phosGAACGACATGGCTACGATC CGACTTTCGTCGGCAGCGTC-3′
Ad153-R-tag	Sangon Biotech	5′-TGTGAGCCAAGGAGTTGTTGT CTTCN$_{10}$GTCTCGTGGGCTCGG-3′
CYB5A Forward	Sangon Biotech	5′-GGCAACGCTTAGACTCTGTGTG-3′
CYB5A Reverse	Sangon Biotech	5’-CTGCCCTTGGCCTAACTAACCT-3′
Software and algorithms		
BWA (Version: 0.7.17)	Li and Durbin, 2009[52]	http://bio-bwa.sourceforge.net/
ANNOVAR (v2017-07-17)	Wang et al., 2010[53]	https://annovar.openbioinformatics.org/en/latest/
bowtie2 (Version: 2.3.1)	Langmead et al., 2009[54]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Picard (Version:1.9)	GATK	https://broadinstitute.github.io/picard/
BEDTools (v2.17.0)	Quinlan and Hall, 2010[55]	https://bedtools.readthedocs.io/en/latest/
CHAT (version:1.0.5)	Muetze et al., 2016[32]	https://apps.cytoscape.org/apps/chat
Cytoscape (Version:3.6.1)	Shannon et al., 2003[56]	https://cytoscape.org
EnrichmentMap (version:3.1.0)	Isserlin et al., 2014[57]	https://apps.cytoscape.org/apps/enrichmentmap
GSEA (v06-Apr-2017)	Subramanian et al., 2005[58]	https://gsea-msigdb.org/gsea/index.jsp
MonoVar (no version available)	Zafar et al., 2016[59]	https://bitbucket.org/hamimzafar/monovar
RSEM (v1.2.29)	Li and Dewey et al., 2011[60]	https://deweylab.github.io/RSEM/
SAMtools (Version: 0.1.19)	Li et al., 2009[61]	http://samtools.sourceforge.net/
SOAPnuke (Version: 1.5.6)	Chen et al., 2018[62]	https://github.com/BGI-flexlab/SOAPnuke
ape (Version: 5.1)	Bioconductor	https://cran.r-project.org/web/packages/ape/index.html
copynumber (Version:1.22.0)	Bioconductor	https://bioconductor.org/packages/release/bioc/html/copynumber.html
ggtree (Version: 1.14.6)	Bioconductor	http://bioconductor.org/packages/release/bioc/html/ggtree.html
hcluster (Version: 1.1.25)	Bioconductor	https://www.rdocumentation.org/packages/fastcluster/versions/1.1.25/topics/hclust
pheatmap (version: 1.0.12)	Bioconductor	https://www.rdocumentation.org/packages/pheatmap/versions/1.0.10/topics/pheatmap
DNAcopy (Version: 1.22.0)	Bioconductor	http://www.bioconductor.org/packages/release/bioc/html/DNAcopy.html
Monocle2 (Version: 2.0)	Qiu et al., 2017[26]	http://cole-trapnell-lab.github.io/monocle-release/docs/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SC3 (Version:1.10.1)	Bioconductor	http://bioconductor.org/packages/release/bioc/htmlSC3.html
dendextend (version: 1.8.0)	Bioconductor	https://bioconductor.org/packages/release/bioc/html/DECIPHER.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Gane Ka-Shu Wong (gane@ualberta.ca).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All raw and processed sequencing data generated in this study are deposition on the publicly accessible database CNGB Nucleotide Sequence Archive (CASA: CNP0000440, CNSA: https://db.cngb.org). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study and analysis were approved by Health Research Ethics Board of Alberta, Cancer Care committee (Approval # HREBA.CC-18-0470), and the BGI institutional review board on bioethics and biosafety (Approval No. BGI-IRB 15148). Patient informed consent was obtained prior to the performance of experimental procedures.

Patient basic information

Tissue samples were provided by a 24-year-old Caucasian female with T3 N1 M1, ER(-), PR(-), HER-2(-), grade 3, invasive ductal breast carcinoma following palliative mastectomy and full axillary lymph node dissection. This patient had received previous treatment with two cycles of docetaxel and four cycles of doxorubicin/cyclophosphamide chemotherapy with little or no clinical response. At the time of sample collection, metastasis has occurred to the patient's local regional axillary lymph nodes (3 of 15 nodes were positive for metastatic carcinoma), as well as the liver. Two tissue samples were collected from the tumor-stromal interface of the primary tumor, and two additional tissue samples were collected from carcinoma positive axillary lymph nodes. Samples were obtained within 10min from resection and immediately immersed in liquid nitrogen for flash freezing. The patient survived ~18 months following initial diagnosis, and passed away ~9 months following sample donation.

METHOD DETAILS

Laser capture microdissection (LCM) and cell cluster lysis

Frozen tissues were embedded in Clear Frozen Section compound (VWR cat no. 95057-838), sectioned on a cryostat (Leica, CM3050S) and stained with hematoxolin and eosin (H&E) to identify areas for laser capture microdissection. LCM was performed with a Leica CRT6000 laser capture microdissection microscope (Concord, ON, Canada) within 15 min of sectioning to avoid RNA degradation. All of our cell clusters were subject to pathologist identification. From the primary cancer tissue, we obtained cell clusters labeled epithelial, cancer, and LVI. Only when the tumor or normal proportion is over 95% was a cell cluster labeled as cancer or epithelial, respectively. For lymph node, the categories were lymphocyte and cancer. Each cell cluster determination was made by two pathologists, based on cellular morphology, and we have included as Figure S1 representative microscope images used for this purpose. Cell clusters were classified as groups of cells over 10μm in diameter. Each cell cluster contained an estimated 50-200 cells. Extracted cell clusters were incubated with a lysis mixture (20μL RLT Plus buffer (QIAGEN 1053393), plus 1μL of spike-in RNA (1:250,000), and immediately placed on ice for subsequent steps.

Oligo-dT30VN bead labeling; mRNA and gDNA separation

M-280 Streptavidin Dynabeads® (Invitrogen, catalog no. 11205D) were washed according to the manufacturer's recommendations, mixed 1:1 with biotinylated-Oligo-dT30VN primer (100 μM), and incubated for 20min at room temperature. Oligo-dT30VN-labeled beads were then washed and resuspended in bead resuspension buffer (Superscript II first-strand buffer, RNase inhibitor, nuclease-free water (NF-H2O)). For mRNA and gDNA separation, 10 μL of Oligo-dT30VN labeled beads was added to tubes containing

lysed cell clusters. 2.5μL of RLT Plus buffer was added to each tube and bead/cell suspensions were incubated for 20 min at room temperature. Samples were then placed on a magnet for 1min for bead separation. Supernatant containing gDNA was transferred to a new tube. Oligo-dT30VN labeled beads were then washed twice with 10 μL of G&T seq wash buffer (50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM MgCl2, 10 mM DTT, 0.5% Tween-20, 0.2 × RNase inhibitor), and supernatant containing gDNA was transferred into new tubes and stored at −80°C until further processing. The remaining Oligo-dT30VN-labeled beads coupled to cell-cluster extracted mRNA were prepared for reverse transcription.

Reverse transcription, amplification, purification of cDNA

For reverse transcription (RT), 10μL of RT Mastermix (5 × SuperScript II First-Strand Buffer, 5M Betaine, 100mM MgCl2, 100mM DTT, 100uM TSO, RNase inhibitor, SSII, dNTPs (10mM)) was added to tubes containing the Oligo-dT30VN-labeled beads coupled to the mRNA from each cell cluster. Samples were placed in a Veriti 96-Well Thermal Cycler (Applied Biosystems, Catalog no. 4375786) for RT using the following program settings: 42°C 2min, 42°C 60min, 50°C 30min, and 60°C 10min. 12.5 μL of PCR Mastermix (2 × KAPA HiFi HotStart ReadyMix, IS PCR Primer) was then added to each RT reaction tube. Tubes were placed in the thermal cycler for cDNA amplification at the following program settings: 98°C 3min, 98°C 20 s, 67°C 15 s, 72°C 6min, 98°C 20 s for 22 cycles and 72°C 5min. Amplified cDNA was stored at −20°C until purification. To purify amplified cDNA, room temperature AMPure XP beads (Agencourt, catalog no A63881) were added at a 0.8:1 ratio to cDNA containing tubes and incubated at room temperature for 5min. The supernatant was removed, beads were washed twice with 100μL 80% ethanol and allowed to dry. Beads were then resuspended in 21μL of NF-H2O.

Tn5 cDNA library preparation and sequencing

1.0ng of cDNA from each sample was mixed with a fragmentation mixture containing the BGI enzyme Tn5 Transposase (BGI, catalog no. BGE005) embedded with adaptors, and heated to 55°C for 7 min. The reaction was stopped by adding 5 μL of 0.1% SDS to each sample. 25 μL of PCR reaction mix (5 x KAPA Fidelity Buffer, 10 mM each dNTP, PhoAd153 F primer (10 μM), Ad153 R primer (10 μM), Ad153-F-tag (0.5 μM), Ad153-R-tag (0.5 μM), KAPA HiFi DNA polymerase), was added to each fragmented cDNA sample. Samples were transferred to a thermal cycler for amplification using the following program settings: 72°C 5 min, 95°C 3 min, 98°C 20 s, 60°C 15 s, 72°C 25 s for 15 cycles, 72°C 5min. After amplification, 0.6X and 0.2X AMPure XP beads were used to select 300bp ± 100bp size fragments. These were then pooled for a total of 520ng cDNA per sample. cDNA in was cyclized by adding 20 μM Splint oligo (Invitrogen, Shanghai, China), and NF-H2O to each sample for a final volume of 70 μL and heating samples to 95°C for 3min. 10x TA buffer (100mM ATP, T4 DNA ligase (600U/μL), NF-H2O), was added to each reaction for a final volume of 120 μL. Samples were incubated at 37°C for 1hr. EXO digestion was then performed by adding 10x TA buffer mixed with EXO I (20U/μL), EXO III (100U/μL) and NF-H2O to each sample for a final volume of 128 μL. Samples were incubated at 37°C for 30 min. Reaction products were then purified by adding 320 μL of AMPure XP beads to obtain the cDNA library. Rolling circle amplification (RCA) was performed to produce DNA Nanoballs (DNBs) that were loaded on to the BGISEQ-500 sequencing platform (BGI, Shenzhen, China). Qualified cDNA libraries were sequencing with 100bp paired-end reads.

Purification and amplification of gDNA

To enriched cell cluster gDNA, AmPure XP Beads were added to gDNA containing supernatants at a 0.6:1 gDNA to bead ratio, and incubated for 8 min at room temperature. The supernatants were then discarded, and the remaining beads were washed twice with 100μL 80% ethanol, followed by the addition of 5μL NF-H2O. gDNA was amplified using a REPLI-g Single Cell Kit (QIAGEN, Catalog no. 150345). Briefly, 3.5μL of Buffer D2 (denaturation buffer) was added to the beads and incubated for 10 min at 65°C. The reaction was stopped by adding 3 μL of Stop Solution to each sample. 40μL of Master Mix (29μL REPLI-g sc Reaction Buffer, 2μL REPLI-g sc DNA Polymerase, NF-H2O sc), was added to each denatured DNA sample for amplification. Samples were then incubated at 30°C for 8 h and the reaction was stopped by heating samples to 65°C for 3 min.

Housekeeping test for MDA products

Prior to the WES/WGS sequencing, the quality of the amplified DNA products was assessed using a multiplex PCR based method that evaluated the presence of eight genes (CYB5A, PRPH, GABARAPL2, ACTG1, NDUFA7, UQCRC1, MYC, MIF) from different chromosomes. 1μL of PCR mix (3.0 μL 10x Buffer, dNTP (2.5mM) 3.2 μL, 3.0 μL Primer Mix (10 μM) 0.2 μL 100 x BSA, 0.4 μL rTaq (5U/μL)), was added to each amplified gDNA sample. These were then placed in a thermal cycler for amplification of the above gene products using the program settings: 95°C 4min; 95°C 30 s; 56°C 50 s and 72°C 1min for 35 cycles; 72°C 10min. Agarose gel electrophoresis was then performed on the PCR amplification products. Samples in which 4 or more bands were detected were subjected to downstream library preparation.

WGS library preparation and sequencing

Whole genome sequencing (WGS) libraries were constructed from quantified and amplified gDNA from each cell cluster using an MGIEasy DNA Rapid Library Prep Kit (BGI, catalog no, 940-200033-00), and the BGISEQ-500 sequencing platform. Briefly, high-quality gDNA was randomly fragmented using a Covaris LE220 ultrasonicator (Covaris, Woburn, MA, USA). AMPure XP magnetic bead-based cleanup was conducted to select fragments ranging from 100-700 base pairs (main band 200-300bp). Selected

fragments were tailing-end repaired by adding Adaptor Mix after which the ligated was DNA purified. Purified DNA samples were transferred to a thermal cycler and amplified with the following program settings: 95°C 3 min; 8 cycles of 98°C 20 s, 60°C 15 s, 72°C 30 s; then 72°C 10 min. The PCR products were purified with AMPure XP magnetic beads. Samples were mixed with different barcodes and NF-H_2O was added to each sample for a final volume of 48 μL. The homogenized PCR products were then denatured by heating them to 95°C for 3 min in a thermal cycler. 11.8 μL of Reaction Mixture was added to each sample. Denatured DNA was circularized by placing samples in a thermal cycler for 30 min at 37°C. Circularized DNA was then digested by added 4μL of Digestion Reaction Solution to each sample for 30 min at 37°C. This reaction was stopped by added 7.5 μL of Digestion Stop Buffer. Single stranded circular DNA was then purification using AMPure XP magnetic beads. RCA was performed to produce DNBs that were loaded on to the BGISEQ-500 sequencing platform. The qualified WGS libraries were sequenced with an average coverage of 0.5∼1X with 100bp single-end reads.

WES library preparation and sequencing

Whole exome sequencing (WES) libraries were constructed from quantified and amplified cDNA from each cell clusters using MGIEasy Exome Capture V4 Probe set (BGI, Shenzhen, China). cDNA pre-hybridization was performed by heating samples to 95°C for 5min followed by hybridization at 65°C for 24h. After elution of hybridized cDNA products, a post-PCR reaction mixture (2X KAPA HiFi HotStart Ready Mix, Ad-153-F (20 μM) and 4 NF-H2O) was added to each sample. Samples were divided in half and amplified on a thermal cycler with the following program settings: 95°C 3 min;13 cycles of 98°C 20 s, 60°C 15 s, 72°C 15sec; then 72°C 10 min. PCR products were then purified using AMPure XP magnetic beads. PCR products totaling 330ng were pooled together. Samples were processed for splint circulation and made into a single strand circular DNA for WES library construction. RCA was performed to produce DNBs that were loaded onto the BGISEQ-500 sequencing platform and sequenced for 1 lane with 100bp paired-end reads.

WGS data processing and CNV calling

Deconvoluted sequencing FASTQ data corresponding to each cell cluster sample was aligned to HG19/NCBI37 using BWA-MEM algorithms (BWA, Version: 0.7.17). SAMtools (Version: 0.1.19) was used to sort BAM files, mark and removed PCR duplicates, and calculate each chromosome's depth and coverage. BAM files produced by alignment were counted in 5k,10k, 20k, 50k genomic bins using a ''non-overlapping'' ''variable binning'' strategy as previously described.[18,63] '''Variable binning'' results in each bin having variable start and end coordinates. Since the reference length is fixed, a smaller genome cutting bin number will require a larger bin length, i.e., for 5k, 10k, 20k, 50k bins, the median genomic length spanned by each bin is 554kb, 220kb, 136kb and 54kb, respectively. The variable start and end coordinates were determined by mapping back 200 million simulated sequence reads with 100nt length to the HG19/NCBI37 reference to determine bins for further calculation. 'Non-overlapping bins' means the boundary of each bin did not overlap with the genome coordinates, enabling us to clearly identity the copy number variation value of each chromosome segments and annotate the CNV affected genes. Unique normalized read counts of each variable bin was calculated using the Circular Binary Segmentation (CBS) method from R Bioconductor 'DNAcopy' package.[64] The parameters used for CBS segmentation were alpha = 0.05, nperm = 1000, undo.SD = 1.0, min.width = 5. Default parameters were used for MergeLevels which removed erroneous chromosome breakpoints. The median absolute pairwise difference (MAPD) was calculated to quantify the copy number noise of each cell cluster. We choose the 5k bin for further clustering analysis as it had a higher average MAPD value compared with 10k, 20k and 50k bins. Next, we filtered out cell clusters with coverage lower than 10% as calculated using BEDTools (v2.17.0), and with a MAPD greater than 1.00. This accounted for approximately 18% (17 of 97 cell clusters) of the total cell clusters for this patient.

CNV heatmap construction and clone identification

To construct a clustered CNV heatmap, we calculated Euclidean distances from the copy number data matrix. Each column represents one cell cluster, and each row represents the relative copy number ratio of diploid cells from each segment. Ward.D2 hierarchical clustering algorithm was performed in R using the pheatmap (version: 1.0.12) package available on CRAN. Columns representing a single cell were hierarchically clustered using Ward.D2 linkage on the basis of pairwise Euclidean distances, and the x axis was ordered by chromosome coordinates of genome position. To estimate the reliability of each clade defined by the Ward.D2 clustering algorithm, we used the bootstrap method by R package pvclust (Version: 2.0.0) to calculate the AU (Approximately Unbiased) p value and BP (Bootstrap Probability) value. An AU p value > 0.95 is seen as stable clade to define a CNV clone.

Maximum parsimony tree construction

To detect common chromosome breakpoints and segments that were shared by cell cluster samples in each identified DNA-CNV clone, we applied a multiple-sample population segmentation algorithm using a Bioconductor R package copynumber (Version: 1.22.0), with parameter gamma = 1. Piecewise constant curves were fitted to the cell cluster CNV data by minimizing the distance between the curve and the observed multi-cell data, and returning multi-cell segments with a fitting clonal CNV result. The Maximum Parsimony tree was calculated from the CNV-clone matrix using the parsimony ratchet algorithm with R package phangorn (Version: 2.4.0). Homozygous deletion, heterozygous deletion, neutral, or amplification, were treated as characters, and missing values were treated as ambiguous items. Hamming distance was calculated for branch lengths with R package ape (Version: 5.1). Phylogenetic trees were exported in Newick format, and R package ggtree (Version: 1.14.6) was introduced for visualization.

WES data processing, SNV clone calling, heatmap construction, and clone identification

On target sequencing reads in the FASTQ files of each cell cluster were cleaned by SOAPnuke (Version: 1.5.6), aligned by BWA (Version: 0.7.17), sorted by SAMTools (Version: 0.1.19), PCR duplications removed by Picard, and realigned by GATK Realigner. The putative SNVs for each cell cluster were called by Monovar with the parameter setting: p = 0.002, a = 0.2, t = 0.05, m = 4, c = 1 and annotated by ANNOVAR (humandb:20170901). Putative SNVs were then filtered using the following parameters: 1000G_ALL < 0.5%, ExAC_ALL < 0.5%, ESP6500siv2_ALL < 0.5%, genomicSuperDups score < 0.9. Only mutations with nonsynonymous, stop/gain and stop/loss in exonic and splicing regions were kept as final SNVs for each cell cluster. The SNV matrix containing the final SNVs of each cell cluster was generated, where 0 represent not mutated or unidentified SNV sites, 1 represent heterozygous SNV sites and 2 represent homozygous SNV sites. SNV heatmaps were constructed using the Ward.D2 clustering method with Euclidean distances of SNV matrix construed by Monovar. DNA-SNV clones were identified using the same method as described above for the DNA-CNV clones.

WTS data processing and quality control

Sequencing data was first processed to filter out low quality reads which were defined as: 1) ''N'' bases accounting for 5% read length; 2) Bases with quality < 15 accounting for 50% read length; 3) Containing the adaptor sequence; 4) Duplicate reads. The reads that passed were then aligned to ribosomal RNA sequences downloaded from NCBI Reference Sequence Database using SOAPaligner (soap2 V2.21t). The unmapped reads were aligned to human genome assembly GRCh37 (hg19). Gene expression TPM was calculated using bowtie2 plus RSEM with default parameters. Saturation curves were then calculated for each cell cluster, and curve densities were compared between each saturation curve. Cell clusters showing an unsaturated curve and an obviously skewed density plot were considered unqualified samples.

RNA clone calling for cell clusters

For each cell cluster, TPM was calculated for each given gene in Refseq. TPM matrices for genes were supplied to SC3,[23] a single-cell consensus clustering pipeline, with the following parameters: pct_dropout_min = 2 and pct_dropout_max = 90. After the consensus matrix was built by SC3, the average silhouette width and stability index values were calculated. These were combined with cell type and the best empirically performing clustering were determined. Once the stable clusters were determined to identify RNA clones, genes exhibiting the highest variability among each LCM cell cluster was calculated. The resultant marker genes were identified by a ROC curve (AUROC) > 0.85 and p value < 0.01. The top ten maker genes of each cluster were shown in the heatmap with a log2(TPM+1) value.

RNA trajectory reconstruction and gene set enrichment

TPM marker genes identified using SC3 were supplied to Monocle2 to generate pseudotime plots that reflect cell fate decisions and differentiation trajectories. Genes were identified as being differentially expressed between trajectories using a cut-off q value of q < 0.01. We choose the top 800 qualifying genes and defined these as ''significant changing genes'' for each RNA trajectory. By further analyzing the branches of each RNA trajectory, we found statically significant (q < 0.001) branch-dependent genes. We used the previously defined ''significant changing genes'' and the branch-dependent genes associated with each trajectory to do GSEA, and to determine the related GO BPs (gene ontology biological processes). We defined significant biological processes as those with a q < 0.01 as calculated by GSEA.

DNA and RNA-clone comparison

The cell cluster DNA-CNV-tree was constructed using the Euclidean distance of the CNV data matrix, clustered with the hclust function using WARD.D2 linkage in R, then outputted as Newick format. Each clade of the tree reflects a DNA-CNV-clone. The RNA-tree was constructed by SC3 and the tree clade was outputted as Newick format. We mapped the DNA-CNV-tree and DNA-SNV-tree using a Sankey diagrams generated by an online tool at http://sankey-diagram-generator.acquireprocure.com. The DNA-CNV-clones was mapped to the RNA-clones to compare the consistency between DNA and RNA clones.

Gene set enrichment and hub gene identification

Genes exhibiting a significant change in expression (Table S2.2,Top 800 with q < 0.01), from each metastasis paths identified following Monocle 2.0 analysis was supplied to Metascape[27] to independently perform biological pathway and process enrichment analysis using the following ontology sources: GO Biological Processes and GO Molecular Functions (Database Last Update Date: 2019-06-11). All genes associated with each metastatic path were used for this enrichment. Enrichment terms with a p value < 0.01, a minimum count of 3, and an enrichment factor > 1.5, were collected and grouped into clusters based on membership similarities calculated by Metascape. Specifically, accumulative hypergeometric distributions and Benjamini-Hochberg statistics were applied to calculate the p values and q-values of each term.[65] Kappa scores were used as a similarity metric when performing hierarchical clustering on the enriched terms, and sub-trees with a similarity of > 0.3 are considered a cluster. The most statistically significant terms within a cluster are chosen to represent each cluster. To further capture the relationships between terms, a subset of enriched terms was selected and used to generate a network map whereby terms with a similarity > 0.3 are connected by edges. Networks were visualized using Cytoscape with a layout generated by employing the yFiles Radial method, whereby each node represents an enriched term and is colored by its cluster ID. We further grouped each term into five subgroups (''shared by three paths,'''' Shared by

Primary Cancer Clones," "Shared by CIN Clones," "Shared by Lymph node Clones" and "Path Specific"), and marked these subgroupings on each Cytoscape network based upon cell cluster location. We then choose the 20 most respective terms combined across all metastatic paths to generate a heatmap in which the coloring is representative of the log p values associated with each term for each metastatic path. Terms in the heatmap were further defined as belonging to 1 of 3 three categories: 1) Metastasis related terms"; 2) "Immune related terms"; and 3) "Other terms."

Hub gene identification and Kaplan-Meier analysis

Genes exhibiting a significant change in expression, as well as their fold-change value (Table S2.2), were supplied to the Cytoscape application CHAT (Contextual hub analysis tool, version:1.0.5) to identify hub genes related to each metastatic path, and generate a PPI network from the BioGRID database of Human/*Homo sapiens* Taxonomy. The mean expression of the top 20 hub genes for each metastatic path were also supplied to Kaplan-Meier Plotter[33] to generated Kaplan-Meier survival curves for these signatures, based upon the survival data of 1089 breast cancer patients sourced from GEO (Gene Expression Omnibus), EGA (The European Genome-phenome Archive), and TCGA (The Cancer Genome Atlas) databases. The top 20 genes were selected based upon the significance of prognosis prediction. 80 of the most characterized ribosome protein genes (RPS and RPL) were also supplied to Kaplan-Meier Plotter to generate Kaplan-Meier overall survival p value across 21 types of cancer available in Kaplan-Meier Plotter database (Table S4.4).

Pan-cancer ribosome protein gene analysis

80 of the most characterized ribosome protein genes (RPS and RPL) were individually supplied to ENCORI (The Encyclopedia of RNA Interactomes; http://starbase.sysu.edu.cn/index.php), a Pan-Cancer Analysis Platform which enables differential gene expression analysis between tumor and normal tissues using available mRNA data from TCGA (The Cancer Genome Atlas; https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). The differential expression data of these 80 ribosomal proteins was available in ENCORI for 17 types of cancers with 7086 tumor and 704 normal samples. Results of this analysis are shown in Table S4.2.

QUANTIFICATION AND STATISTICAL ANALYSIS

DNA and RNA amplification and library qualification

Amplification DNA was quantified using housekeeping genes as described using a Qubit dsDNA High Sensitivity kit (Invitrogen,USA; catalog number: Q32854)as per the manufacturer's instructions. Amplified RNA was quantified using an Agilent's 2100 Bioanalyzer (Agilent Technologies, CITY, STATE) and the library was quantified using Qubit dsDNA High Sensitivity kit (Invitrogen,USA;catalog number: Q32854).

Statistical analysis

The statistic of AU (Approximately Unbiased) p value is calculated by multiscale bootstrap resampling by pvcluster package (v2.0.0). P values of chi-square test were based on asymptotic theory. Silhouette width and stability index statistics were calculated using the SC3 package. All p values were two-sided and $q < 0.01$ was considered significant. Maker genes of each RNA clonal were defined by a q-value < 0.01 and a ROC value > 0.85. A Significant change in gene expression was defined as $q < 0.01$ with top800 genes of three paths. The hub genes for each path were defined using an adjust P value < 0.01. All other statistical analyses were carried out as described in the text using the R statistical environment (v3.4.4 and v3.5.0). Pathway network graphs were generated using Cytoscape (v3.6.1).