Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Brian Aram Pedro Date

Defining the clinical and biological relevance of leader and follower cell mutations in collective cancer invasion

By

Brian Aram Pedro
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Cancer Biology

Adam Marcus, Ph.D.
Advisor

Lawrence Boise, Ph.D.
Committee Member

Shoichiro Ono, Ph.D.
Committee Member

Paula Vertino, Ph.D.
Committee Member

David Yu, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Defining the clinical and biological relevance of leader and follower cell mutations in collective cancer invasion

By

Brian Aram Pedro
B.A., Tufts University, 2014

Advisor:
Adam Marcus, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Sciences
Cancer Biology
2020

Abstract

Defining the clinical and biological relevance of leader and follower cell mutations in collective cancer invasion

By Brian Aram Pedro

Cancer metastasis, the direct cause of 90% of cancer-related deaths, remains a poorly-understood process. Metastatic carcinomas often utilize collective invasion, whereby cohesive packs of cells travel through the microenvironment. Furthermore, *in vitro* studies of lung cancer collective invasion have revealed that specialized leader and follower cells cooperate to facilitate the emergence of invasive chains, and that follower cells cannot invade in the absence of leaders. However, the biology underlying these phenotypes, including the role of gene mutations, has not been fully explored. We discovered novel leader-specific and follower-specific gene mutations, including, notably, a leader-specific mutation in ARP3. Introduction of this mutation into follower cells conferred leader-like behavior, including the ability to lead invasive chains, suggesting it could play an important role in driving leader cell emergence and behavior.

There is currently a lack of predictive biomarkers for determining high-risk patients in a number of cancer types, including lung cancer. Even among patients diagnosed with localized disease, over 40% are not expected to survive beyond five years. We thus investigated whether high-risk patients could be identified by the presence of mutations within a leader-cell derived cluster of genes on chromosome 16q. Using cohorts of lung squamous cell carcinoma and lung adenocarcinoma patients from The Cancer Genome Atlas, we found poorer survival for 16qMC+ patients; furthermore, this correlation was observed among two cohorts of hepatocellular carcinoma patients, another cancer type with high rates of metastasis and disease recurrence.

Finally, we combined SaGA with single-cell RNA-sequencing to further dissect the biology of leader and follower cells during active collective invasion. We discovered that leader and follower mutational profiles are mutually exclusive on the single cell level. These mutations also correlate strongly with leader and follower expression markers including MYO10 and IL13RA2, indicating that these mutations could be utilized as precise genomic markers for individual leaders and followers. We further discovered that leader cells harbor cancer stem cell-like gene expression, and that TGFβ signaling may facilitate leader-follower cooperation. Ultimately, these data demonstrate that leader- and follower-specific mutations can both elucidate the mechanisms of collective invasion and help to better stratifying high-risk cancer patients.

Defining the clinical and biological relevance of leader and follower cell mutations in collective cancer invasion

By

Brian Aram Pedro
B.A., Tufts University, 2014

Advisor:
Adam Marcus, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Sciences
Cancer Biology
2020

Acknowledgements

None of this work would have been possible without the guidance and support of a large number of people. Firstly, it is important to acknowledge that doing good science is really hard, and there is no way I could have done this work without the enormous help I received from our collaborators, cores and shared resources, and my dissertation committee. In particular I would like to thank the Wei Zhou and Melissa Gilbert-Ross labs for their advice and feedback at lab meetings, and the Vertino lab including Elizabeth, my co-first author with whom I learned to navigate the publication process for the first time. I also received tons of help with microscopy from the Integrated Cellular Imaging core, with sequencing and vector design from the Integrated Genomics Core, and with huge amounts of statistics and bioinformatics analysis from the Winship Biostatistics and Bioinformatics Shared Resource. I would also like to thank my committee members, Paula Vertino, Larry Boise, David Yu, and Sho Ono, for their helpful guidance and for helping me navigate the next steps in my research whenever the path forward was not so clear.

Being an MD/PhD student comes with a unique set of circumstances and experiences, both positive and negative, and I am very lucky to have shared those experiences with an awesome cohort of classmates: Josh, Mohib, Yaseen, Nusaiba, Bejan, Danny, Rafi, Stephanie, Sean, Matt, and Ha Eun have all been instrumental in keeping my sanity throughout the years, and I am so thankful for all of you. I also owe a tremendous debt of gratitude to my labmates in the Marcus lab, past and present, who not only helped me navigate the complicated science that we strive to understand, but made it truly enjoyable to come to the lab every day. I consider you all friends as much as colleagues, and I appreciate all of your help over the years.

I cannot give enough thanks to Adam, for agreeing to take me on 4 years ago, for giving me both the guidance and freedom to grow into an independent researcher, and for your unwavering positive attitude and optimism. I was first interested in your lab because of the cool and exciting research, but my mind was made up after meeting your other graduate students and seeing how confident they were as scientists. I knew then that your lab was the kind of environment I wanted to be trained in, and I can absolutely say that I made the right choice. It has been a pleasure working with you and learning from you, and I will continue to consider you a mentor and friend throughout my career.

And finally, I have to thank my entire family, immediate and extended, who have always been unconditionally supportive of my crazy decision to be a (seemingly) perpetual student. An unbelievable amount of credit goes to my wife, Tori, who gives me more encouragement, confidence, and stress relief than I could ever ask for. From the first time I floated the idea of doing an MD/PhD 8 years ago, to agreeing to move to the south with me straight out of college, to taking care of our family whenever I get swamped with work or studying, you have been such a constant bright spot in my life, and I truly could not have done any of this without you. To our sons, Charlie and Owen, who are too young to read this right now but hopefully will someday: thank you for helping me keep everything in perspective and reminding me why I went into this career path in the first place. Watching you grow is such a great source of happiness and a welcome distraction from research whenever I need it. I am also so grateful for my brother, Kevin, my dad, Phil, my mother-in-law, Trish, my father-in-law, Chuck, and my sister-in-law, Beth, who have all been incredibly supportive throughout this process. And lastly, thank you to my mom, Lisa, my inspiration for pursuing cancer research; your endless compassion, courage and perseverance continue to motivate me every single day.

Table of Contents

List of figures and tables

List of Abbreviations

In alphabetical order:

ARP3 – Actin-related protein 3

BMP6 – Bone morphogenic protein 6

CIL – Contact inhibition of locomotion

CMV – cytomegalovirus

CSC – Cancer stem cell

CTC – Circulating tumor cell

DMSO – Dimethyl sulfoxide

EGFR – Epidermal growth factor receptor

EMT – Epithelial to mesenchymal transition

FACS – Fluorescence-activated cell sorting

GSEA – Gene set enrichment analysis

HCC – Hepatocellular carcinoma

ICI – Immune checkpoint inhibitor

IL13RA2 – Interleukin-13 receptor A2

KDM5B – Lysine demethylase 5B

KM – Kaplan-Meier

LUAD – Lung adenocarcinoma

LUSC – Lung squamous cell carcinoma

MET – Mesenchymal to epithelial transition

MYO10 – Myosin 10

NSCLC – Non-small cell lung cancer

OS – Overall survival

PFS – Progression-free survival

RNA-seq – RNA-sequencing

SaGA – Spatiotemporal genomic and cellular analysis

scRNA-seq – Single-cell RNA-sequencing

SNP – Single nucleotide polymorphism

TCGA – The Cancer Genome Atlas

TGFβ – Transforming growth factor beta

TKI – Tyrosine kinase inhibitor

UBC – ubiquitinC

VEGF – Vascular endothelial growth factor

WT – wild-type

Chapter 1: Introduction

1.1. Overview of cancer metastasis

1.1.1. Statistics and clinical implications of cancer metastasis

Cancer encompasses a diverse array of diseases with different drivers, pathological characteristics and clinical courses, but one common thread is the role of metastasis in driving mortality. Metastatic disease – defined by the presence of cancer cells that have broken off from the primary tumor, traveled to a distant site, and ultimately colonized a secondary tumor – is a hallmark of cancer (4) and is responsible for the vast majority of patient deaths across all cancer types. Furthermore, despite its crucial role in cancer mortality, the mechanisms of this complex, multi-step process are still poorly understood. It is clear, however, that when patients present with metastatic disease, treatment becomes increasingly challenging, and overall prognosis drops precipitously (5) (Fig. 1.1).

This is especially evident in lung cancer, which is the second leading most common cancer among men and women in the United States, and is responsible for more cancer deaths than colon, breast, and prostate cancers combined (5). This is driven by the large proportion (over 50%) of patients who already have metastatic disease at the time of diagnosis; among these patients, five-year relative survival is a dismal 4.7%, compared with 56% for patients diagnosed with only localized disease (Fig. 1.1). Similarly, among patients with liver cancer, for the 18% who are diagnosed with metastatic disease, five-year relative survival is just 2.4%, compared with 32.6% for those with localized disease (2).

Even among patients who are diagnosed with earlier stage cancer that has yet to metastasize, those with certain cancers including lung and liver cancer remain at high risk for subsequently developing

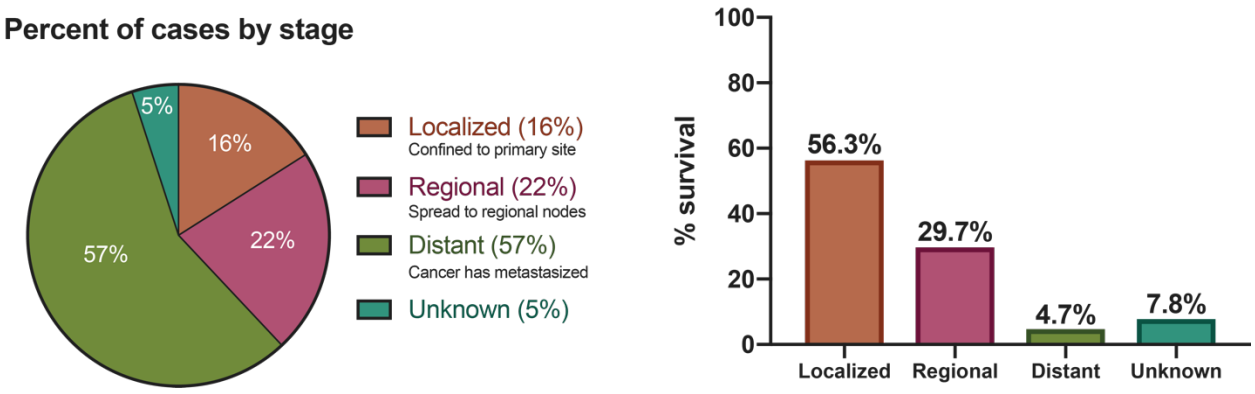


Figure 1.1. Lung cancer survival statistics. Percentages of lung cancer cases by stage at diagnosis, and % five-year survival for each group. Adapted from data from the National Cancer Institute Surveillance, Epidemiology and End Results (SEER) program (2).

metastatic disease. For lung cancer, five-year survival for patients with localized disease is 57.4%, indicating significant mortality among this group (Fig. 1.1). This suggests that certain patients are at higher risk for disease progression and metastasis; however, there remain few options for identifying those higher-risk patient groups at the time of diagnosis. For patients with non-small cell lung cancer (NSCLC), of which squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD) are the major subtypes, those diagnosed at stage I (i.e. localized disease) are typically treated with surgery, and if removal is deemed complete, are moved onto an every-six-month screening regimen without further treatment (6). There are no clinically actionable biomarkers to indicate which early-stage patients are at higher risk for disease recurrence and should therefore be treated with adjuvant chemotherapy, targeted therapy, or radiotherapy in addition to surgical removal (6). Thus, in order to develop markers that could better stratify high-risk patients, deeper knowledge of the mechanisms of metastasis is essential.

1.1.2. Steps of the metastatic cascade

For cancer cells to metastasize, they must carry out a complex series of events, each of which presents unique challenges (Fig. 1.2). Following invasion away from the primary tumor site, cancer cells must intravasate into either the bloodstream (hematogenous route) or the lymphatic system (7). Cells disseminating via the hematogenous route then extravasate out of the bloodstream at a secondary organ site, and must survive and ultimately proliferate there to form a secondary tumor (7). Tumor cells in the lymphatic system can form secondary tumors within lymph nodes and subsequently metastasize to distant organs. Through either route, there are substantial obstacles that must be overcome for cancer cells to metastasize successfully. In order to invade, cells must be able to break away from the primary tumor, break down the surrounding extracellular matrix, and migrate through it. This shift to a more motile phenotype has been well-characterized in the context of epithelial-to-mesenchymal transition (EMT), in which epithelial-derived cancer cells alter the expression of certain genes, including downregulation of E-cadherin and upregulation of N-cadherin, vimentin, and fibronectin (8, 9). To survive after intravasation into the bloodstream, cancer cells must be resistant to anoikis, or detachment-induced apoptosis observed in normal cells (7). After extravasation at the secondary site, cells must survive and subsequently proliferate to form a secondary tumor. Because of the complexity and inefficiency of this process, the vast minority (less than 1%, perhaps as low as 0.02%) of cells successfully form metastases after entering the bloodstream (7). The capacity to survive in low numbers and initiate a secondary tumor is often considered a stem cell-like property that would only be present in a small subset of the tumor population.

1.1.3. Single and collective cell migration and invasion

The majority of research into cancer cell migration and invasion has focused on single-cell mechanisms. In the traditional EMT model, a single call can transition from a non-motile epithelial

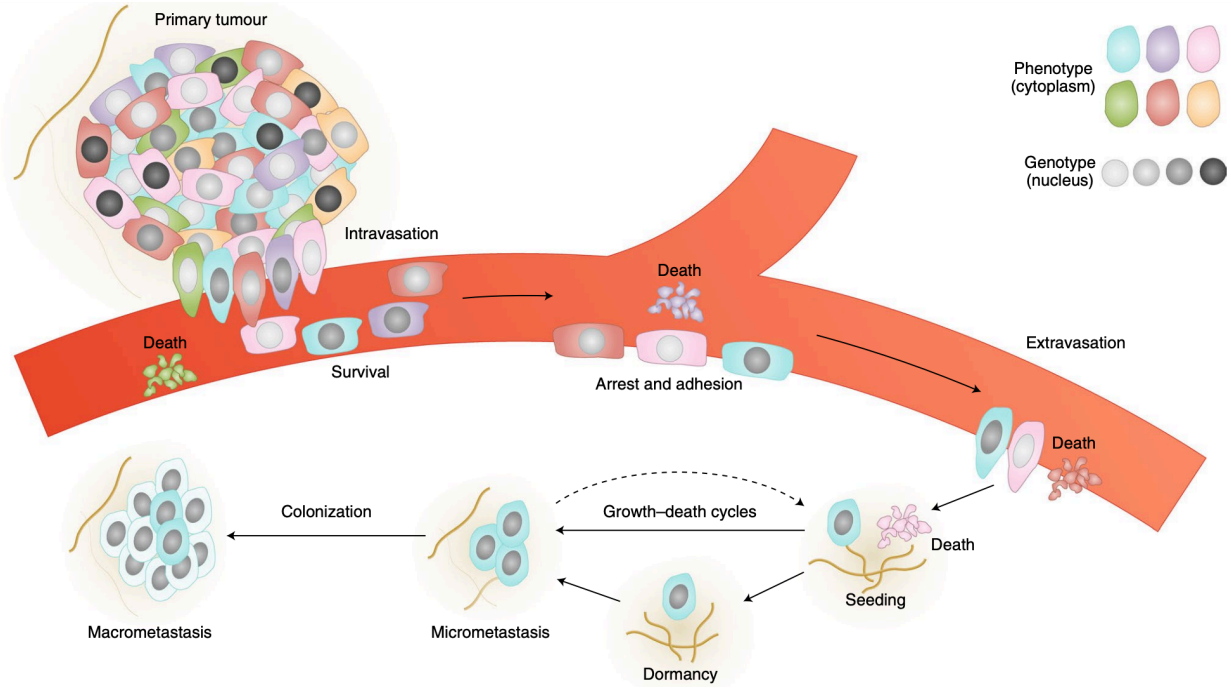


Figure 1.2. Steps of the metastatic cascade. Taken from (1). Schematic of the metastatic cascade from primary tumor to secondary site, with cells of different phenotypes and genotypes cooperating to carry out different individual steps.

phenotype to a mesenchymal morphology with the ability to invade and travel through the microenvironment (8, 9). This is accomplished by downregulation of E-cadherin, resulting in dissolution of cell-cell junctions, and increased expression of proteins such as N-cadherin that allow for enhanced cell-matrix interactions (8, 9). Upon reaching the secondary site, cells can undergo the reverse process, or mesenchymal to epithelial transition (MET), to shift back to a proliferative phenotype. Alternatively, single cells can take on an amoeboid morphology, which also lacks cell-cell junctions but is more reliant on actomyosin dynamics for movement rather than cell-matrix interactions (10). However, recent research has focused on the importance of heterogeneous packs of collectively invading and migrating cancer cells, which could cooperate and confer distinct advantages at each point throughout the process (11-13). Furthermore, studies have shown that groups of multiple cells are more successful in ultimately forming metastases than cells that travel alone, further highlighting the clinical importance of collective invasion (14).

Although only recently linked to cancer invasion, the concept of leader and follower cells in collective migration is well-described in human biology, including during development, wound healing, and sprouting angiogenesis (15). During angiogenesis, tip cells travel ahead and signal to stalk cells that travel behind during the formation of new blood vessels. Tip cells typically express VEGFR2 and respond to gradients of VEGFA during angiogenesis, maintaining their phenotype and leading to secretion of the Notch ligands DLL4 and JAG1; stalk cells, which express Notch, receive these signals from tip cells, thus stabilizing their stalk cell phenotype (15). Maintenance of these roles is crucial for successful angiogenic sprouting, as tip cells are uniquely able to interact with the extracellular matrix (ECM) while stalk cells are needed to form a new basement membrane (15). Similarly, during embryonic development, neural crest cells migrate together in cellular streams, ultimately giving rise to skull bone, facial muscles and nervous tissue, among other structures (16). Much like invasive cancer

cells, induction of neural crest collective migration depends upon signaling pathway associated with EMT, resulting in decreased cell-cell adhesion and increased motility (16). In cancer, these signaling pathways that govern these fundamental biological processes are often "hijacked," allowing cancer cells to become motile to escape pressures such as hypoxia and lack of nutrients. For example, VEGF and Notch have been implicated in lung cancer collective invasion (3), leading to distinct leader and follower phenotypes that cooperate similarly to tip cells and stalk cells during angiogenesis.

Further bolstering the evidence for collective invasion as a major mode of cancer metastasis are studies showing that metastatic tumors in mice arise from multicellular seeds, rather than seeding of single cells, and that cells aggregated into clusters more successfully form metastatic tumors than a single-cell suspension (14, 17). This is important, because it underlines the importance of understanding and ultimately therapeutically targeting collective invasion; although some cancer cells may indeed invade as single cells, they are less likely to succeed in seeding secondary tumors. Collective packs of cells, while more rarely found in the circulation than single circulating tumor cells (CTCs), may more readily result in metastatic tumor formation, and should thus be preferentially targeted when considering anti-metastatic therapy (14, 17).

1.2. Overview of tumor heterogeneity

1.2.1. Clonal evolution of tumor subpopulations

The concept that individual tumors could be genetically distinct from one another originated decades ago (18), and with continued advancement in sequencing technologies, it has become evident that there is heterogeneity between cancer types, within cancer types, and even within subpopulations of individual tumors (19). This heterogeneity is driven by differential epigenetic regulation, gene expression, and most notably, somatic gene mutations. In a Darwinian evolution model, one of the earliest and most common hypotheses for tumor clonal development (20), internal and external selection pressures result in the clonal expansion of tumor cells that have gained certain advantages through spontaneous somatic mutations. These internal pressures can include hypoxia or immune system targeting, while external pressures include cytotoxic chemotherapy, and more recently, targeted therapies or immunotherapies. Cells that have become altered, likely through gene mutations, to acquire survival advantages in the face of these pressures are ultimately able to proliferate while other cells are killed or undergo apoptosis.

Tumors with increased genomic instability are more prone to spontaneous mutations, including chromosomal duplications, deletions, and fusions, as well as point mutations (19). In certain cancer types, including lung, greater genomic instability results in larger somatic mutational burdens (21); this presents unique problems for targeted therapy and treatment resistance, as cells in these tumors are more readily able to acquire resistance-conferring mutations and thus recur as a tumor that is more difficult to target. This is evident in EGFR-targeted therapies, which have recently become widely used in non-small cell lung cancer (22). EGFR activating mutations are among the most common drivers in NSCLC, and EGFR-tyrosine kinase inhibitors (TKIs) have become first-line treatment for

these patients (6, 22, 23). However, resistance to these therapies can be acquired when cells gain additional EGFR mutations, including T790M which occurs in 50% of patients with disease progression after EGFR-TKI therapy (22, 23). Patients can be treated with a third-generation EGFR-TKI, osimertinib, that selectively targets EGFR-T790M; however, resistance to osimertinib can also arise through alternative EGFR mutations or activation of additional TKIs (23). Thus, while improved sequencing technology has helped to illuminate the extent of genetic heterogeneity within tumor populations, and has allowed for the development of novel targeted therapies, it has also become clear that the mutational mechanisms leading to this heterogeneity also present a unique challenge to the success of targeted therapy.

1.2.2. Phenotypic heterogeneity in cancer and the SaGA platform

Tumor heterogeneity is most commonly described in terms of genetic differences between and within tumors. However, it is equally as important to consider the varying phenotypes that result from these genetic alterations, especially in the context of the metastatic cascade. Acquiring the ability to invade and metastasize is a well-characterized hallmark of cancer (4), and recent studies have shown that cells with different specialized phenotypes can work together to carry out this complex, multi-step process (3, 11, 24). This includes collective invasion, the major mode of invasion displayed by carcinomas, which has been shown to include phenotypically distinct leader and follower cells (3, 11, 24).

In order to isolate and further study these cells, we previously developed the SaGA (Spatiotemporal Genomic and Cellular Analysis) platform (3) (Fig. 1.3). Through photoconversion of the Dendra2 protein from green to red fluorescence, SaGA enables precise optical highlighting of any single cell, or group of cells, which can then be isolated via fluorescence-activated cell sorting (FACS). This platform is particularly useful for the isolation of phenotypically rare cells within a population, as they

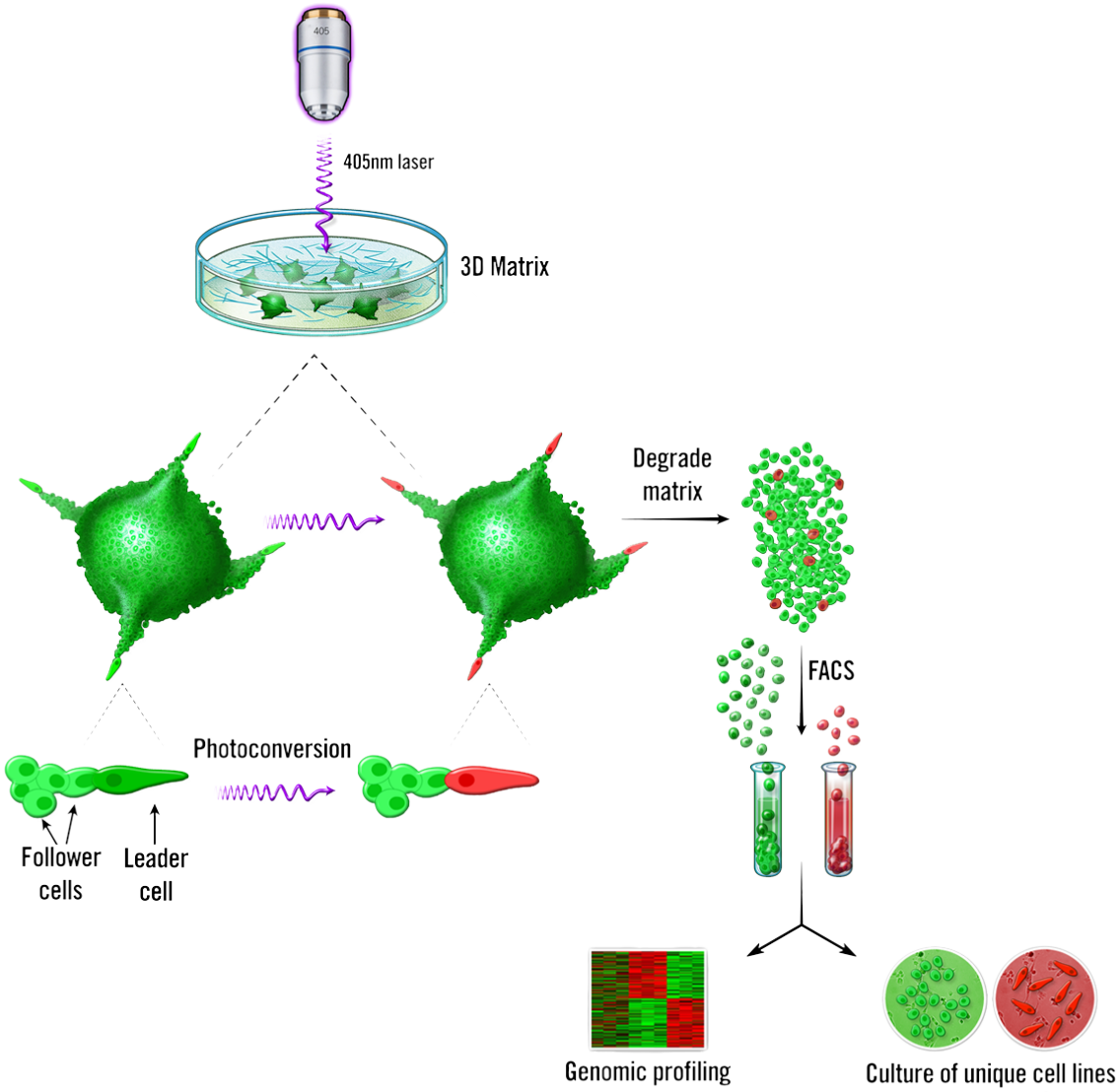


Figure 1.3. Overview of the SaGA platform. Taken from (3). Spheroids comprised of approximately 3,000 Dendra2-expressing cells are embedded in Matrigel and allowed to invade for 24 hours. Dendra2 is selectively photoconverted to red fluorescence in user-defined leader or follower cells via 405nm laser, followed by matrix degradation and separation of photoconverted cells via FACS.

can be selected and expanded to produce purified populations. This allows for detailed experimentation and characterization of these rare cells that would otherwise not be feasible. We previously used SaGA to isolate, culture and analyze leader and follower cells from collectively invading H1299 NSCLC cells (Fig. 1.3). This analysis demonstrated that leader cells maintain a highly invasive phenotype with the unique ability to pioneer invasive chains away from a primary tumor (or spheroid) and through the microenvironment; meanwhile, the bulk of these collective chains are comprised of follower cells, which are poorly invasive unless leader cells are present (3). Thus, while invasion and metastasis may be hallmarks of cancer, it seems that there are subpopulations of cells that acquire the ability to carry out different components of this process. This is likely due to the differing selection pressures that arise throughout the metastatic cascade (Fig. 1.2); as previously discussed, metastatic cancer cells must be able to invade and migrate through the tissue microenvironment, intravasate to enter the bloodstream or lymphatic system, resist anoikis, eventually extravasate at a secondary site, and ultimately colonize a metastatic tumor (7, 25). These requirements specifically select for cells that have acquired abnormal phenotypes, either through mutations or other means such as epigenetic modifications; thus, the phenotypic heterogeneity that exists within a tumor population increases the chances of success at each step along the process (26) (Fig. 1.2).

Given the cooperation between phenotypically and genetically distinct cells during metastasis, it is inherently difficult to develop a single targeted anti-metastatic therapy. As there is no single cell type or genetic marker solely responsible for carrying out metastasis, any successful therapeutic would have to cover multiple targets at once. Further taking into account inter-tumor heterogeneity, with countless genetic events likely producing similar phenotypes in different tumors, highlights the necessity of personalized, precision medicine approaches to identify targetable markers for individual patients. SaGA enables precise correlation between cells' metastatic phenotypes and genomic profile, thus

providing a potential avenue toward development of personalized, anti-metastatic therapeutics (Fig. 1.3).

1.2.3. Leader-follower dynamics in cancer cell invasion

SaGA, in addition to enabling analysis of isolated leader and follower cell populations, has also elucidated the deep complexity of leader-follower interactions during collective invasion and confirmed the difficulty of therapeutically targeting leader cells. Genomic profiling of leader and follower populations has revealed vastly different gene expression profiles as well as epigenetic profiles between the two populations; leader-follower cooperative signaling appears to function at least in part through modified vascular signaling, and the leader cell phenotype is driven by a combination of hypermethylation and upregulation of certain metastasis-promoting genes such as MYO10 and JAG1 (3, 27). Given the particular importance of leaders in the formation of invasive chains (3), therapeutic inhibition of collective invasion would ideally be targeted to leader cells. However, lung cancer leader cells have also been shown to be highly resistant when treated with established chemotherapeutics and other small molecule inhibitors (28), meaning new strategies for targeting leader cells are needed. To accomplish this, it is crucial to establish the biological drivers for leader cell behavior, including invasion, motility, and communication with followers.

Numerous drivers, across multiple cancer types, have been implicated in leader cell motility. Reported leader cell driver genes include keratin-14, a basal epithelial gene, in breast and ovarian cancer (11, 14, 29); cathepsin B, a matrix protease, in salivary carcinoma (30), and VEGFA in non-small cell lung cancer (3). However, there are currently no consensus leader cell drivers identified between or even within cancer types. Thus, in order to develop targeted therapeutics for patients, it may be necessary to use technology that can rapidly isolate and genomically profile leader cells from an individual tumor. While this type of rapid precision medicine is not likely achievable with current techniques, it is important to continue expanding technologies such as the SaGA platform, both to refine the

techniques as well as continue building a database of leader and follower cell drivers from different cancer types and cell lines.

It remains to be determined whether leader and follower cells have important roles in the metastatic process subsequent to collective invasion, including secondary tumor formation. As some of the pathways expressed in leader cells are also implicated in cancer stem cells (CSCs), including JAG1/Notch and VEGF signaling (31), it is possible that leader cells could also include a CSC-like population that ultimately gives rise to eventual metastatic tumors. Expression of another CSC marker, CD44, has been shown to induce leader cell emergence in breast cancer (32), but leader cells have not been widely implicated as CSCs in lung cancer or other cancer types.

1.3. Dissertation goals

The work described in this dissertation is a multi-layered exploration into the ways that phenotype-specific gene mutations can broaden our understanding of leader and follower cell biology in collective invasion. Prior to this dissertation work, the SaGA platform provided a novel method for isolating leader and follower cells, and initial genetic and phenotypic studies had been performed on the two populations. However, these initial studies also made clear that the biological drivers behind these cell types were complex and would require further analysis to tease apart. Thus, the studies described here began with an intriguing discovery – leader and follower cell populations, both derived from the H1299 non-small cell lung cancer cell line, each harbored certain gene mutations that were present in one population but not the other. This was surprising, as it is typically assumed that cells from a single cell line would represent a clonal population with the same mutation profile. However, given this finding, subsequent studies sought to determine how these mutations could give new insights into the mechanisms of leader and follower cell behavior.

Chapter one focuses on determining whether these mutations could be directly involved in driving the invasive leader cell phenotype, by introducing one of the leader-specific mutations, ARP3 K240R, into follower cells and measuring whether it results in leader-like behavior. In chapter two, the clinical utility of these leader-specific mutations is explored, using publicly available lung and liver cancer patient cohorts from The Cancer Genome Atlas (TCGA) to measure whether patients with a certain subset of leader-specific mutations, interestingly all located on a single chromosome arm, experienced poorer survival. Given the high rates of recurrence and metastasis, and poor survival rates of these two cancer types, this work aims to identify a new strategy for identifying higher-risk patients. Finally, chapter 3 combines the SaGA platform with single-cell RNA-sequencing (scRNA-seq) to dissect H1299 leader and follower cell biology on the single-cell level. Utilizing the previously-identified

leader- and follower-specific mutations as precise genomic markers for each single cell allows for robust characterization of these cell types, revealing new insights including TGFβ crosstalk and a potential role for leaders as a tumor-initiating population.

Taken together, these data aim to deepen our understanding of leader and follower cell cooperation during collective invasion, by leveraging the finding of novel leader- and follower-specific mutations to 1) manipulate leader and follower cell phenotypes, 2) develop a clinical tool for identifying higher-risk lung and liver cancer patients, and 3) create precise genomic markers for leaders and followers that enable detailed single-cell analysis of their underlying biology.

Chapter 2: Genetic heterogeneity within collective invasion packs drives leader and follower cell phenotypes

Adapted from the work published in the *Journal of Cell Science* 2019, 132, jcs231514, doi:10.1242/jcs.231514.

Elizabeth L. Zoeller*[1], Brian Pedro*[1], Jessica Konen[1,^], Bhakti Dwivedi[2], Manali Rupji[2], Niveda Sundararaman[3], Lei Wang[4], John R. Horton[5], Chaojie Zhong[9], Benjamin G. Barwick[8,9], Xiaodong Cheng[5], Elisabeth D. Martinez[4,6], Matthew P. Torres[3], Jeanne Kowalski[2,7], Adam I. Marcus[2,8,#], Paula M. Vertino[2,9,#,^^]

*Contributed equally to this work

Affiliations:

[1] Graduate Program in Cancer Biology, Emory University, Atlanta, GA 30322,

[2] Winship Cancer Institute, Emory University, Atlanta, GA 30322, USA,

[3] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332,

[4] Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center, Dallas, TX 75390,

[5] Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030,

[6] Department of Pharmacology, UT Southwestern Medical Center, Dallas, TX 75390,

[7] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322,

[8] Department of Hematology and Medical Oncology, Emory University, Atlanta, GA 30322,

[9] Department of Radiation Oncology, Emory University, Atlanta, GA 30322

^Present address: Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

^^Present address: Department of Biomedical Genetics, University of Rochester Medical Center, Rochester, NY, 14642

Contributions:

The work presented here is the result of a co-first authorship with ELZ. BP performed experiments in figures 2.1, 2.2, and 2.3, and contributed to the writing and editing of the manuscript.

Abstract

Collective invasion, the coordinated movement of cohesive packs of cells, has become recognized as a major mode of metastasis for solid tumors. These packs are phenotypically heterogeneous and include specialized cells that lead the invasive pack and others that follow behind. To better understand how these unique cell types cooperate to facilitate collective invasion, we analyzed transcriptomic sequence variation between leader and follower populations isolated from the H1299 non-small cell lung cancer cell line using an image-guided selection technique. We now identify 14 expressed mutations that are selectively enriched in leader or follower cells, suggesting a novel link between genomic and phenotypic heterogeneity within a collectively invading tumor cell population. Functional characterization of a leader-specific candidate mutation showed that ARP3 enhances collective invasion by promoting the leader cell phenotype. These results demonstrate an important role for distinct genetic variants in establishing leader and follower phenotypes.

2.1. Introduction

Metastatic disease is the cause of 90% of deaths among cancer patients (33). In non-small cell lung cancer (NSCLC), which comprises 80-85% of all lung cancer diagnoses, metastases are commonly observed in the bones, lungs, brain, liver, and adrenal glands. Patients presenting with metastatic disease have significantly worse prognoses than those with early-stage disease; for example, the 5-year survival rate for stage I NSCLC is 55%, while stage IV disease (in which distant metastases are present) carries a mere 4% five-year survival rate (5). Successful colonization of distant sites requires that cells from the primary tumor gain the capacity to invade through the surrounding basement membrane, travel through the bloodstream or lymphatic system, and ultimately expand to establish colonies at the metastatic site (26, 34).

Cells migrate through the microenvironment via multiple mechanisms. A classic example is the epithelial-to-mesenchymal transition, where cells lose expression of epithelial features such as E-cadherin and gain expression of mesenchymal proteins including vimentin and N-cadherin. This shift is thought to promote cellular detachment and enable cancer cells to undergo single-cell invasion. In contrast, collective cell migration refers to the coordinated movement of a group of cohesive cells (15). This phenomenon is well-described in embryonic development and wound healing, and histological evidence from primary patient tumor samples (11, 24, 35-37), mouse models of metastasis (12, 13), and 3-D cultures (3, 11, 38-41) suggest that cells from solid tumors often migrate and invade in cohesive packs as well. These collective invasion packs and streams vary in width, shape, and cell number, as well as in the mechanisms guiding their movement (42-48).

Understanding the mechanisms that underlie the outgrowth of metastatic clones is further complicated by the heterogeneous mix of cell populations within each tumor. This intratumor

heterogeneity arises from cell-to-cell variation in the genetic background each expanding to create a unique subclonal population (19). Superimposed upon this subclonal genomic heterogeneity is the potential for epigenetic heterogeneity reflected in variations in gene expression even among genetically identical cells. Recently, multiregional sequencing of primary lung tumors characterized the intratumor heterogeneity of NSCLC and showed that upwards of 24% to 30% of mutations went undetected in at least one sampling region, demonstrating that almost a third of mutations are occurring in spatially distinct subclonal populations and may have been missed in broad scale data based on single sampling, such as those analyzed as part of the TCGA project (49, 50). These unique subpopulations may be endowed with properties that enhance attributes beneficial to tumor cells, such as resistance to drug therapy or the ability to invade and metastasize (51-54). For instance, the clonal profile of metastatic disease often does not reflect the profile of the primary tumor, but instead includes one or just a few subpopulations from the primary site (14, 53, 55, 56).

One example of phenotypic heterogeneity associated with invasive behavior includes rare, specialized leader cells that lead collective invasive packs, and follower cells that adhere to and follow behind the leaders, both of which cooperate to achieve collective invasion (3, 11, 24, 48, 57). We developed a novel platform (**S**p**a**tiotemporal **G**enomic and Cellular **A**nalysis, or **SaGA**) to isolate specific leader and follower cell populations from collectively invading NSCLC cells (3). Characterization of these cell types revealed that isolated follower cells are highly proliferative but poorly invasive, while isolated leader cells are highly invasive, but poorly proliferative (3). These cellular sub-types cooperate through an atypical angiogenic signaling pathway that is dependent upon VEGF. Previous data suggest a symbiotic relationship, in which both leader and follower cells are necessary for collective invasion to proceed successfully; however, key questions remain as to what drives the biology and emergence of leader and follower cells from a tumor cell population.

In this study, we aimed to elucidate the role of genetic heterogeneity on collective invasion. We analyzed invading leader and follower populations arising from a common H1299 parental NSCLC cell line grown as 3-D spheroids. Strikingly, this revealed mutational landscapes that differ significantly between leader and follower cells, including several expressed mutations that were found exclusively in one cell type or the other. To our knowledge this is the first identification of leader- and follower-specific gene mutations within the same collectively invading tumor cell population.

2.2. Materials and Methods

Cell lines and transfections:

Leader and follower cells isolated via the SaGA technique from the H1299 human NSCLC cell line (3), as well as the parental H1299 cell line (ATCC, Manassas, VA), were cultured in RPMI-1640 media supplemented with 10% fetal bovine serum and 100 units mL^{-1} of penicillin/streptomycin and maintained at 37°C and 5% CO_2. H1299 cells were mycoplasma tested and authenticated using single nucleotide polymorphism analysis through the Emory Genomics Core as previously described (3). H1299 cells had been transfected with the Dendra2 as previously described (3). 293T cells were maintained in Dulbecco's Modified Eagle's Media (DMEM) supplemented with 10% fetal bovine serum and 100 units mL^{-1} of penicillin/streptomycin at 37C and 5% CO_2.

For ARP3 lines, lentivirus was prepared by seeding $2x10^6$ HEK293T cells in a 100 cm dish and co-transfecting with 5 μg transfer vector, 0.5 μg pMD2.G (Plasmid #12259, Addgene), 5 μg psPAX2 (Plasmid #12260, Addgene), and 1 μg of lentiviral vector. After 24 hours, media was replaced with 5 mL fresh complete media, then virus-containing media was collected after 24 hours. Media was centrifuged for 3 minutes at 1000 rpm, 4°C, and then supernatant was filtered through a 0.45 μm low protein-binding filter. Target cells were seeded at 70% confluence in a 6-well plate one day prior to lentivirus collection. After virus collection, target cell media was replaced with 1 mL complete media plus 340 μL virus stock and 1.34 μL polybrene (10 mg mL^{-1} stock), added dropwise. After 24 hours, media was replaced with 2 mL complete media. Selection antibiotics (shRNA: puromycin; ARP3 expression vectors: hygromycin) were added 48 hours after viral infection.

shACTR3 constructs were obtained from Millipore Sigma TRCN0000029383 and TRCN0000380403. mCherry-ARP3 lentiviral constructs were created by cloning ARP3-pmCherryC1 (a gift from Christien Merrifield; Addgene plasmid # 27682 (Taylor et al., 2011)) into the pCDH-UBC-MCS-EF1 Hygro backbone. The UBC promoter was subsequently exchanged for a CMV promoter. The ARP3 K240R mutation was created using site-directed mutagenesis.

RNA-sequencing and variant calling:

RNA-sequencing was performed in triplicate on H1299 parental, leader and follower cells. For parental cells, three different passages were used. For follower cells, three separately-isolated populations were used. For leader cells, two separately-isolated populations were used: one passage of one population, and two passages of the other. RNA library preparation and sequencing were carried out by the Emory Integrated Genomics core and Omega Bio-Tek, Inc. using the TruSeq Stranded mRNA kit, followed by quantification using a Quantus Fluoremeter (Promega, Madison, WI, USA) and integrity assessment using an Agilent 2200 Tapestation instrument. Sequencing was performed using a HiSeq2500 instrument (Illumina, Inc., San Diego, CA, USA), with 50M total sequencing reads generated per sample using the PE100 run format.

Data processing and statistical analyses were performed by the Emory Biostatistics and Bioinformatics Shared Resource. Raw paired-end fastq reads were assessed for quality and contamination using FastQC (Andrews, 2010) and trimmed with Trimmomatic v0.32 (58). Quality filtered reads were mapped against human reference genome hg19 using STAR aligner v2.3.0 (59). Picard tools v1.111 (http://broadinstitute.github.io/picard) was used to assess post-alignment QC and to remove PCR duplicates. With an average yield of 30M post-filtered reads, 88% of the reads mapped uniquely with 78% of reads covered in the coding and UTR region. Genomic variants from RNA-seq were called

using SamTools v0.1.19 mpileup (60) with Varscan v2.3.6 (61) and functionally annotated using ANNOVAR (62). A filtering criterion was applied requiring that reported variant had ≥ 6X read depth coverage, ≥2X supporting alternate reads in all samples. Variants associated with intronic, intergenic or synonymous changes, pseudo genes, non-coding RNAs, or sex chromosomes were excluded. Variants were filtered if they were known in dbSNP but not present in COSMIC database. This resulted in a total of 6240 variants. A pairwise two-sided independent t-test was done to compare cell populations mean variant allele frequencies between the groups.

Western blotting:

Total cellular protein expression was assessed via Western blotting as previously described (63).

Reagents and antibodies:

Primary antibodies for Western blot: ARP3 antibody (Santa Cruz, cat. no. sc-48344 – immunogen: residues 1-110 of human ARP3) was used at 1:1000. GAPDH antibody (Cell Signaling, cat. no. 2118) was used at 1:30,000. Beta-tubulin antibody was used (Sigma, cat. no. T4026) at 1:5000. Horseradish peroxidase-conjugated secondary antibodies (Jackson ImmunoResearch) were used at 1:10,000 for Western blot.

3-D invasion assays, spheroid microscopy and image analysis:

Spheroids were generated as previously described (63) and embedded in 2 mg mL^{-1} Matrigel (BD Biosciences) diluted in complete media. Images were taken at 0, 24, and in some cases 48 hours post-embedding at 4x using either an Olympus IX51 or CKX41 microscope.

For mixed spheroid experiments, cells were plated together in low-adhesion wells in the indicated ratios with 3000 total cells per spheroid and embedded as previously described. After 24 hours, spheroids were imaged using a Leica SP8 inverted confocal microscope. Invasive area and spheroid circularity were measured using ImageJ as previously described (3).

Target validation:

DNA and RNA were isolated from H1299 parental, leader and follower cells using DNeasy Blood & Tissue Kit and the RNeasy Mini Kit (Qiagen Sciences, Germantown, MD, USA), respectively. Isolation of samples occurred in two independent biological replicates. RNA was reverse transcribed with M-MLV Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) top generate cDNA. Primers were designed and used to amplify regions surrounding each locus of interest subject to Sanger sequencing at GENEWIZ (South Plainfield, NJ, USA). Primer sequences are listed in Table S2.1. PCR products were cloned into the pCR4-TOPO TA vector by TOPO-TA cloning (ThermoFisher Scientific 450071, Carlsbad, CA, USA) and transformed into bacteria. Fifty individual colonies for each gene in each cell type were and re-streaked on a new ampicillin plate, and twenty colonies each sent to GENEWIZ for Sanger sequencing.

Cell proliferation assay:

H1299 cells expressing constructs for either overexpression or knockdown of ARP3 were seeded at $1x10^3$ cells per well in six wells each of a 96 well plate. After 24-120hr growth, cells were fixed with 100μL per well cold 10% trichloroacetic acid for 1 hr at 4°C, washed three times in water, and air dried. Cells were then stained with 100μL 0.4% sulforhodamine B in 1% acetic acid for 30 minutes, washed three times in 1% acetic acid, and dried. Dye was reconstituted in 200 μL 10 mM Tris base

(pH 10.5) for 30 min. Optical density (OD) at 510 nm was measured, and the fold-change in OD determined as: ($OD_{X\ hours}$) / ($OD_{24\ hours}$).

Targeted deep resequencing and analysis:

Targeted region of KDM5B surrounding exon 15 was PCR amplified using primer sequences listed in Table S1. Amplified products were purified by SPRI beads (KAPA Biosystems, Cape Town, South Africa). Libraries were then created with custom TruSeq compatible adapters and KAPA Hyper Prep Kit (KAPA Biosystems, Cape Town, South Africa). Quality for each library was checked using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Libraries were sequence on an Illumina HiSeq 4000 (Illumina, Inc., San Diego, CA, USA) using 150 bp paired-end sequencing at NYU Genome Technology Center. Quality trimmed reads were mapped to the human genome (GRCh37) using Bowtie 2 (64). Variant allele frequencies for the KDM5B SNP (rs1141108) and the linked KDM5B L685W mutation were quantified as the fraction of reads (average depth= 348,327 reads per line) exhibiting the *A* or *G* allele at position chr1: 202715284 or the *A* or *C* variant at position chr1:202715414 using R packages Rsamtools, ShortRead, GenomicAlignments, and BSgenome.Hsapiens.UCSC.hg19 (65-69). Differences in KDM5B variant allele frequencies were based on analysis of variance with Tukey's post-hoc correction using the R functions 'aov' and 'TukeyHSD', respectively.

Statistical analysis:

Two-tailed, unpaired Student's t-test was used to assess statistical significance between any two conditions. Ordinary one-way ANOVA with Tukey's multiple comparisons test was used for experiments in which three or more conditions were being compared. Confidence intervals of proportions in the mixed spheroid experiments were calculated via the Wilson/Brown method.

Fisher's exact test was used to test the association of the identified mutations and the phenotypes (e.g. mutant vs. wild-type, leader versus follower) in TOPO-TA cloning experiments.

2.3. Results

Leader and follower cell populations contain distinct mutational profiles.

We previously developed the SaGA technique for isolation of leader and follower cells from collectively invading packs of human NSCLC cancer cells (3). Briefly, cells expressing Dendra2 green-to-red photoconvertible fluorescent protein are formed into multicellular spheroids, embedded in Matrigel and allowed to invade for 24 hours. Leader or follower cells are then selected based upon physical positioning within invasive chains, optically highlighted by photoconversion using a 405nm laser, and separated by fluorescence-activated cell sorting (FACS) (Fig. 2.1A). Using this approach, we isolated three follower populations and two leader populations from H1299 parental cells. Following expansion of each population in 2-D culture, RNA-seq was performed in triplicate, using three separate passages of parental H1299 cells, the three separately isolated populations of follower cells, and the two separately isolated populations of leader cells (including two passages of one of the leader populations). Sequence variants were determined for each population (leader, follower, parental) independently by mapping the RNA-seq profiles to human reference genome Hg19 (GRCh37), resulting in a total of 6240 filtered variants combined in the three populations (see Methods for details). Notably, when comparing variant allele frequencies (VAF) via pairwise t-test analysis, a number of variants were disproportionately present in the leader versus follower populations. We therefore further filtered for those variants that exhibited >20% VAF in either leaders or the followers and <1% in the other (VAF student's t-test p-value <0.01 between leaders and followers). For the purposes of this study, we further excluded those located in 5' or 3' UTRs and known SNPs. Application of these criteria identified fourteen missense mutations – six leader-specific and eight follower-specific (Table 2.1). This represents the first identification of leader- and follower-specific mutations within the collective invasion pack.

Table 2.1. RNA-seq reveals leader- and follower-specific gene mutations

	Gene symbol	Full name	Protein Function	Variant locus (GRCh37)	VAF (%)		
					Parental	Leaders	Followers
Leader–enriched	**ACTR3**	Actin-related protein 3 (ARP3)	Major component of Arp2/3 complex	chr2:114699797; A:G	8.76	23.4	0.08
	MCM5	Minichromosome maintenance complex component 5	Pre-replication complex during DNA replication	chr22:35809920; G:A	7.59	26.9	0.14
	MIPEP	Mitochondrial intermediate peptidase	Oxidative phosphorylation protein maturation	chr13:24413837; A:C	7.75	37.2	0.98
	NAE1	NEDD8 activating enzyme E1 subunit 1	Activation of neddylation pathway	chr16:66852492; T:C	26.2	58.9	0.13
	NUP93	Nucleoporin 93	Component of nuclear pore complex	chr16:56868312; G:A	25.9	57.8	0.29
	ZNF302	Zinc finger protein 302	Function has yet to be determined	chr19:35175335; G:C	4.71	34.6	0.47
Follower–enriched	**CLEC11A**	C-type lectin domain family 11, member A	Growth factor for hematopoietic progenitor cells	chr19:51228679; C:G	8.42	0	22.4
	KDM5B	Lysine Demethylase 5B	Demethylates lysine 4 of histone H3	chr1:202715414; A:C	17.5	0.74	28.1
	NDUFS1	NADH:Ubiquinone oxidoreductase core subunit S1	Core subunit of electron transport chain Complex I	chr2:207012514; C:A	10.4	0.28	22.8
	RERE	Arginine-glutamate dipeptide repeats	Possible role in controlling cell survival	chr1:8416225; G:C	38.7	0	65.4
	RNF115	Ring finger protein 115	E3 ubiquitin ligase	chr1:145686997; T:C	14.7	0.27	21.9
	SKA1	Spindle and kinetochore associated complex subunit 1	Chromosome segregation during mitosis	chr18:47902232; C:G	13.0	0	21.6
	TBP	TATA-Box binding protein	Involved in transcription initiation by RNA pol. II	chr6:170871308; G:T	12.4	0.36	21.7

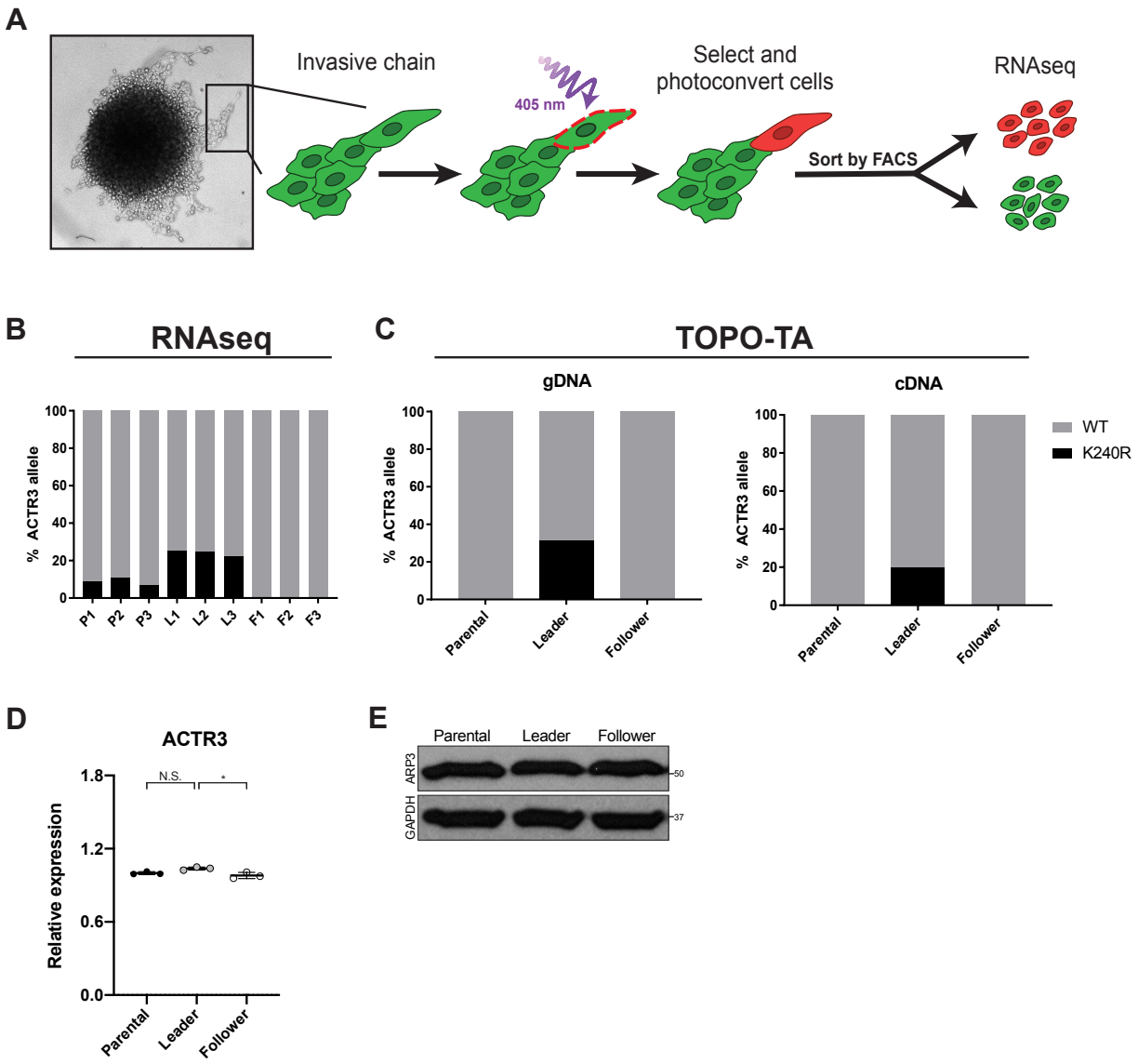


Figure 2.1. ARP3 K240R is a validated mutation in H1299 leader and follower cells. (A) Schematic of the SaGA protocol used to isolate leader and follower cell populations. (B) Variant allele frequencies for ACTR3 from RNA-sequencing of H1299 parental (P), leader (L) and follower (F) cells. *n*=3 separate populations per group. (C) TOPO-TA cloning and subsequent Sanger sequencing confirms the presence of ACTR3 K240R mutation in both cDNA and genomic DNA (gDNA) from parental, leader and follower cells, respectively. *n*=20 colonies (parental, follower gDNA and parental, leader cDNA); 19 colonies (leader gDNA); 18 colonies (follower cDNA) (association between the genotype and cell phenotype was determined by Fisher's exact test as follows: mutant vs. wild-type, leader versus follower ARP3 gDNA p=0.008, ARP3 cDNA *p*=0.11. (D) Relative mRNA expression (via RNA-seq; normalized to parental average) and (E) protein levels (via Western blot) of ACTR3 in H1299 parental, leader, and follower populations. *$p<0.05$ by one-way ANOVA with Tukey's post-test.

Given the cell type specificity, we hypothesized that these mutations could be key contributors to the emergence of leader vs. follower phenotypes from the parental population. To test this, we chose a leader-enriched candidate mutation for further study, *ACTR3* chr2:114699797 A to G, which results in a mutation in ARP3 at lysine 240 (ARP3 K240R) (Fig. 2.1B). We first confirmed this mutation by Sanger sequencing of genomic DNA and cDNA from the parental H1299 population as well as the isolated leader and follower populations (Fig. S2.1). The variant was detectable at subclonal levels in genomic DNA, indicating that it was unlikely to have arisen *de novo* during transcription, but represented a subpopulation of genomic alleles in the parental population. Moreover, the selectivity for the leader population observed at the RNA levels was preserved at the genomic DNA level (Fig. S2.1; Fig. 2.1C). Analysis of ARP3 expression in H1299 parental, leader and follower cells showed that ARP3 mRNA and protein levels were comparable between the populations (Fig. 2.1D,E). While there was some variation in the frequency of the mutation in the parental population between methods and DNA/RNA samples isolated at different times, there was little variation in the allelic balance in leader and follower populations, which maintained a consistent frequency of their respective mutations at both DNA and RNA level, suggesting that there is no allelic bias in the expression of the mutant version in either case. Thus, our selection of leader and follower cells based on phenotypic criteria also selected for subpopulations with distinct allelic balance of expressed mutations.

Predicted functional impact of the leader-enriched ARP3 K240R mutation.

We next sought to characterize the potential impact of the leader-enriched *ACTR3* mutation, which results in a K to R shift in ARP3 (K240R). ARP3 is a key component of the Arp2/3 complex that helps facilitate cellular migration by promoting lamellipodia protrusion (70). Overexpression of ARP3 has been correlated with invasion, metastasis, and poor survival in multiple cancer types, including gastric, colorectal, liver, and gallbladder (71-74). Furthermore, multiple mass spectrometry studies

indicate ARP3 K240 as a post-translational modification (PTM) site, with evidence of both ubiquitylation and acetylation (75-77) (Fig. 2.2A, inset). To evaluate the functional impact of the K240R mutation on the ARP3 protein, we used SAPH-ire (Structural Analysis of PTM Hotspots) (78), which predicts the functional potential of PTMs in protein families that have existing experimental and 3D structure data. SAPH-ire calculates a Function Probability Score (FPS) using a neural network model trained with an array of protein sequence and PTM-specific features extracted from PTMs with established functional impact. Consistent with these data, K240 had among the highest FPS values of all known modified residues within the ARP3 protein family and was among the top 90% of PTMs with well-established functional significance (i.e. 4 or more publications) across all protein families (Fig. 2.2A). SAPH-ire also revealed six experimental ubiquitylation sites in the ARP3 protein family between residues alignment positions 1298 – 1315, four of which correspond to ubiquitylation of ARP3 specifically (K240, K244, K251, and K254). Of these, K240 had the highest mean solvent accessible surface area (SASA) and was also proximal to a protein interaction interface (Fig. 2.2B). The high solvent accessibility, proximity to a protein-protein interface, and predicted functional impact of the ARP3 K240 site suggest that this mutation could indeed be altering the activity of ARP3 in leader cells.

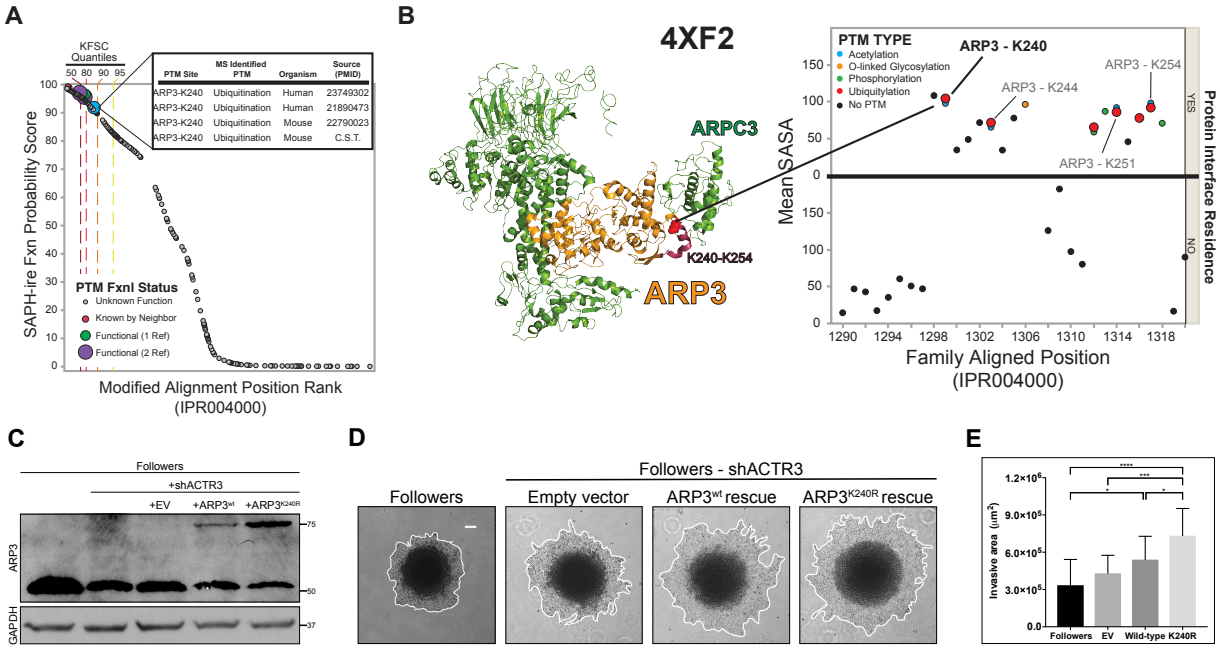


Fig. 2.2. PTM hotspot analysis of ARP3 K240 suggests functional impact of the K240R mutation. (A) Plot of SAPH-ire probability score by rank for all modified alignment positions in the ARP protein family IPR004000. The ARP3 K240 ubiquitylation site is highly ranked along with other MAPs that contain PTMs with well-established function (4 or more supporting references), as indicated by known function source count (KFSC) quantiles. (Inset) Table of ARP3 K240 PTMs identified by mass spectrometry of human and mouse tissues, including literature sources. (B) Local PTM topology of the ARP3 family near ARP3 K240. PTM sites plotted by solvent accessible surface area (SASA) and proximity to the interface of a protein-protein interaction. Human ARP3 PTMs are labeled, revealing multiple ubiquitylation sites between K240-K254. (Left) Structure of Arp2/3 complex (PDB 4XF2) indicating ARP3 K240 (spheres) within the K240-K254 region of ARP3 (red). (C) Western blot showing exogenous (upper band) versus endogenous (lower band) ARP3 expression in unmodified followers, and shACTR3-followers rescued with either empty vector (EV), wild-type ARP3, or ARP3 K240R. Rescue constructs were mCherry-tagged to allow for visualization within invasive chains. (D) Images of invasion in Matrigel at 24hrs of spheroids comprised of unmodified follower cells, or shACTR3 followers transfected with either empty vector, wild-type ARP3, or ARP3 K240R constructs. (E) Quantification of spheroid invasive area (mean±s.d., n=17, 15, 16, and 17 spheroids per group, respectively, across N=3 experiments. **p<0.01, ***p<0.001, ****p<0.0001 by one-way ANOVA with Tukey's post-test). Scale bar: 100 μm.

ARP3 K240R promotes invasion and leader cell behavior.

We sought to determine the impact of ARP3 K240R by introducing this mutation into follower cells and testing for leader cell behavior. Given the relatively low VAF (23.4%) of mutant *ACTR3* in leader cell DNA/RNA, we sought to replicate the leader cell *ACTR3* allelic balance by employing a rescue approach. ARP3 levels were first stably knocked down using two separate short hairpin RNAs (Fig. S2.2A,B), including one (shACTR3 #2) targeted to the 5' UTR of endogenous *ACTR3*. Knockdown of ARP3 significantly reduced 3-D invasion in H1299 parental, leader and follower cells compared to pLKO.1 controls (Fig. S2.2C,D). Using a sulforhodamine B (SRB) assay to measure cell growth, we found little difference upon ARP3 knockdown until 96 hours in the leader and follower populations, and 120 hours in the parental population, when proliferation was decreased (Fig. S2.2E).

To test the functional consequences of ARP3 mutation, follower cells expressing shACTR3 #2 were then 'rescued' by stably expressing empty vector, mCherry-tagged wild-type ARP3, or mCherry-tagged ARP3 K240R, under the control of the moderate activity UBC promoter (79). Higher expression was achieved for ARP3 K240R compared to wild-type ARP3 (Fig. 2.2C, upper bands), suggesting that ARP3 K240R may be more stable than the wild-type protein. When grown as a spheroid, embedded in a Matrigel matrix, and allowed to invade for 24 hours (Fig. 2.2D), ARP3 knockdown follower cells reconstituted with ARP3 K240R, and to a lesser extent those reconstituted with wild-type ARP3, exhibited significantly higher invasive area compared with unmodified followers or those reconstituted with empty vector (Fig. 2.2E). This indicates that ARP3 expression, and especially ARP3 K240R expression, can increase invasive capacity even in normally poorly-invasive follower cells.

Next, we sought to examine whether ARP3 K240R could specifically promote leader cell behavior (i.e. facilitate collective invasion and travel at the front of invasive chains) in a heterogeneous

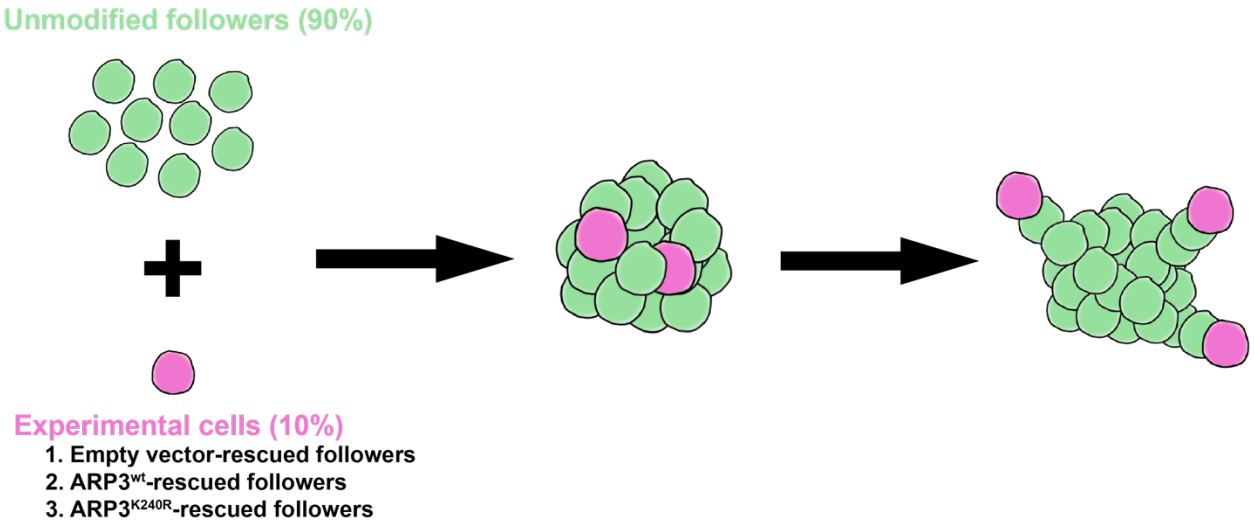


Figure 2.3. Spheroid mixing experimental setup. Spheroids were comprised of either 100% unmodified followers, or 90% unmodified followers plus 10% of shACTR3 followers rescued with either empty vector, wild-type ARP3, or ARP3 K240R.

population. We created 3-D spheroids comprised of 90% unmodified H1299 followers and 10% ARP3 depleted followers rescued with either empty vector, wild-type ARP3, or ARP3 K240R (Fig. 2.3). After 24 hours embedded in Matrigel, we observed a significant increase in invasive area and average number of chains per spheroid, and decreased circularity (indicating more chain-like and less sheet-like invasion) in the mixed spheroids containing 10% ARP3 K240R-rescued followers, as compared with the other three conditions (Fig. 2.4A,B). To determine whether the ARP3 K240R-rescued followers were in fact leading these invasive chains, we used confocal fluorescence imaging to quantify the fraction of chains that exhibited mCherry-ARP3 K240R-rescued followers in the leader position. In spheroids containing 10% wild-type ARP3-rescued followers, we found rescued cells in the leader position in 53.8% of those chains (95% confidence interval 29.1% to 76.8%; Fig. 2.4C). By contrast, in spheroids containing 10% ARP3 K240R-rescued followers, rescue cells were found in the leader position in 87.2% of chains (95% confidence interval 78.0% to 92.9%; Fig. 2.4C). Thus, while both wild-type ARP3-rescued and ARP3 K240R-rescued cells led chains at a higher frequency than expected by random chance, the ARP3 K240R-rescued cells were more efficient in this regard. Together these data indicate that ARP3 K240R confers key leader-like behaviors onto follower cells, including increased invasive capacity, increased numbers of invasive chains, and a greater ability to lead those chains.

As noted above, the K240R-reconstituted ARP3 knockdown cells express higher levels of ARP3 than those reconstituted with the wildtype ARP3. To distinguish the impact of ARP3 dosage from that of ARP3 K240R mutation, we re-created the same rescue cell lines using a CMV promoter (Fig. 2.5A) and repeated the above experiments creating mixed spheroids with 10% rescued cells (Fig. 2.5B). At this higher protein expression level, there was no significant difference in invasive area between the

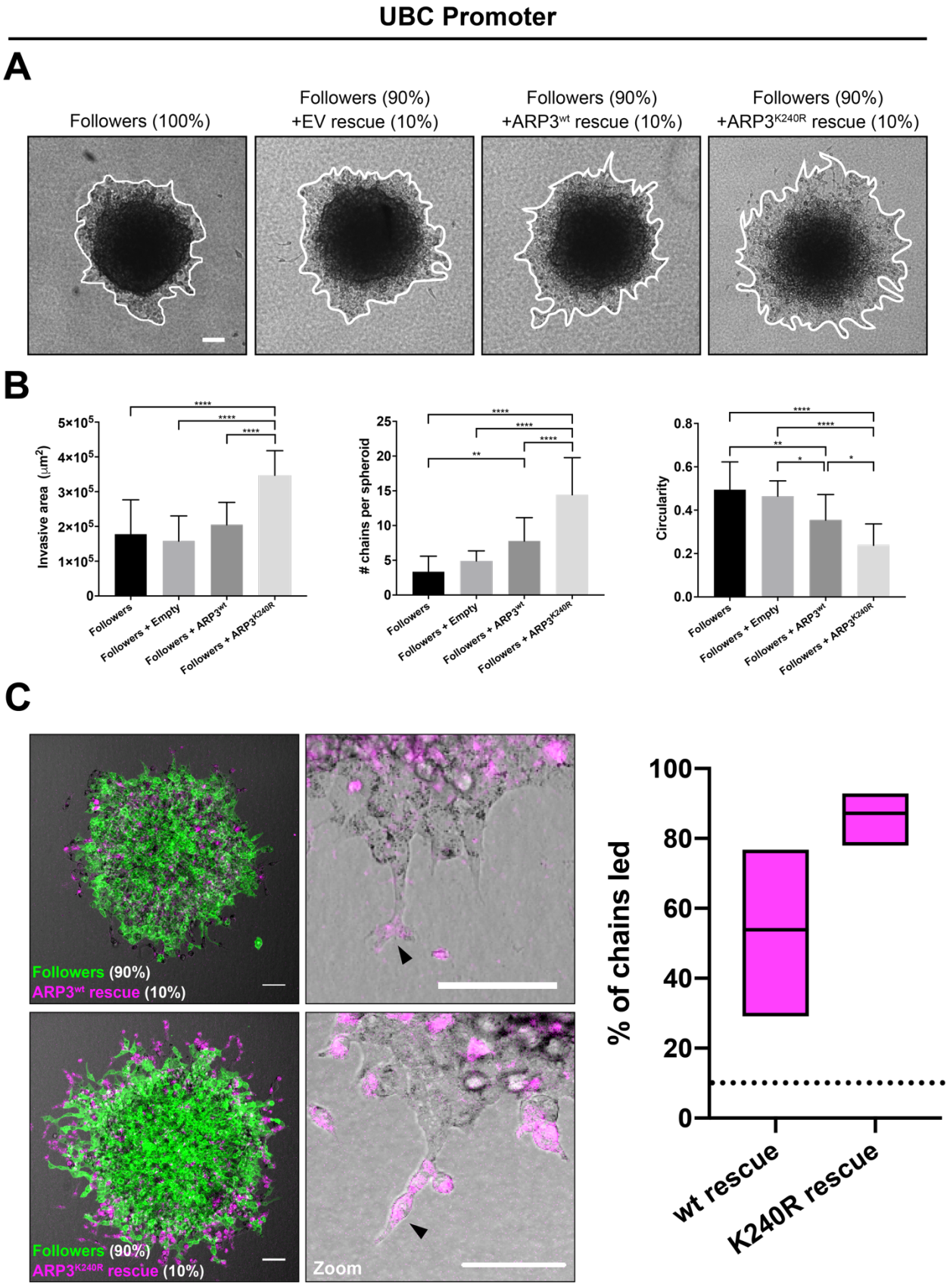

UBC Promoter
A
Followers (100%)
Followers (90%) +EV rescue (10%)
Followers (90%) +ARP3wt rescue (10%)
Followers (90%) +ARP3K240R rescue (10%)
B
Invasive area (μm2)
chains per spheroid
Circularity
Followers
Followers + Empty
Followers + ARP3wt
Followers + ARP3K240R
C
Followers (90%)
ARP3wt rescue (10%)
Followers (90%)
ARP3K240R rescue (10%)
Zoom
% of chains led
wt rescue
K240R rescue

Figure 2.4. ARP3 K240R confers leader-like properties when expressed at low levels. (A) Invasion of mixed spheroids in Matrigel after 24 hr. Rescue constructs were expressed under control of the UBC promoter. Representative images shown for each condition. (B) Quantification of invasive area, circularity, and average number of chains per spheroid for each condition (mean±s.d., n=14, 11, 18, and 16 spheroids for unmodified followers, EV rescue, wildtype ARP3 rescue, and ARP3 K240R rescue, respectively, across N=3 experiments. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001 by one-way ANOVA with Tukey's post-test). (C) Confocal fluorescence imaging of mixed spheroids, with unmodified followers shown in green and mCherry-ARP3 K240R-rescued cells or mCherry-wild-type ARP3-rescued cells shown in magenta. Black arrows indicate invasive chains being led by experimental rescue cells. Graphs show percentage of chains (mean +/- 95% confidence intervals) led by wild-type ARP3-rescued and ARP3 K240R-rescued followers. Dotted line denotes 10% of chains led, corresponding to the proportion of ARP3 rescued cells in the mixed spheroids. Scale bar: 100 μm.

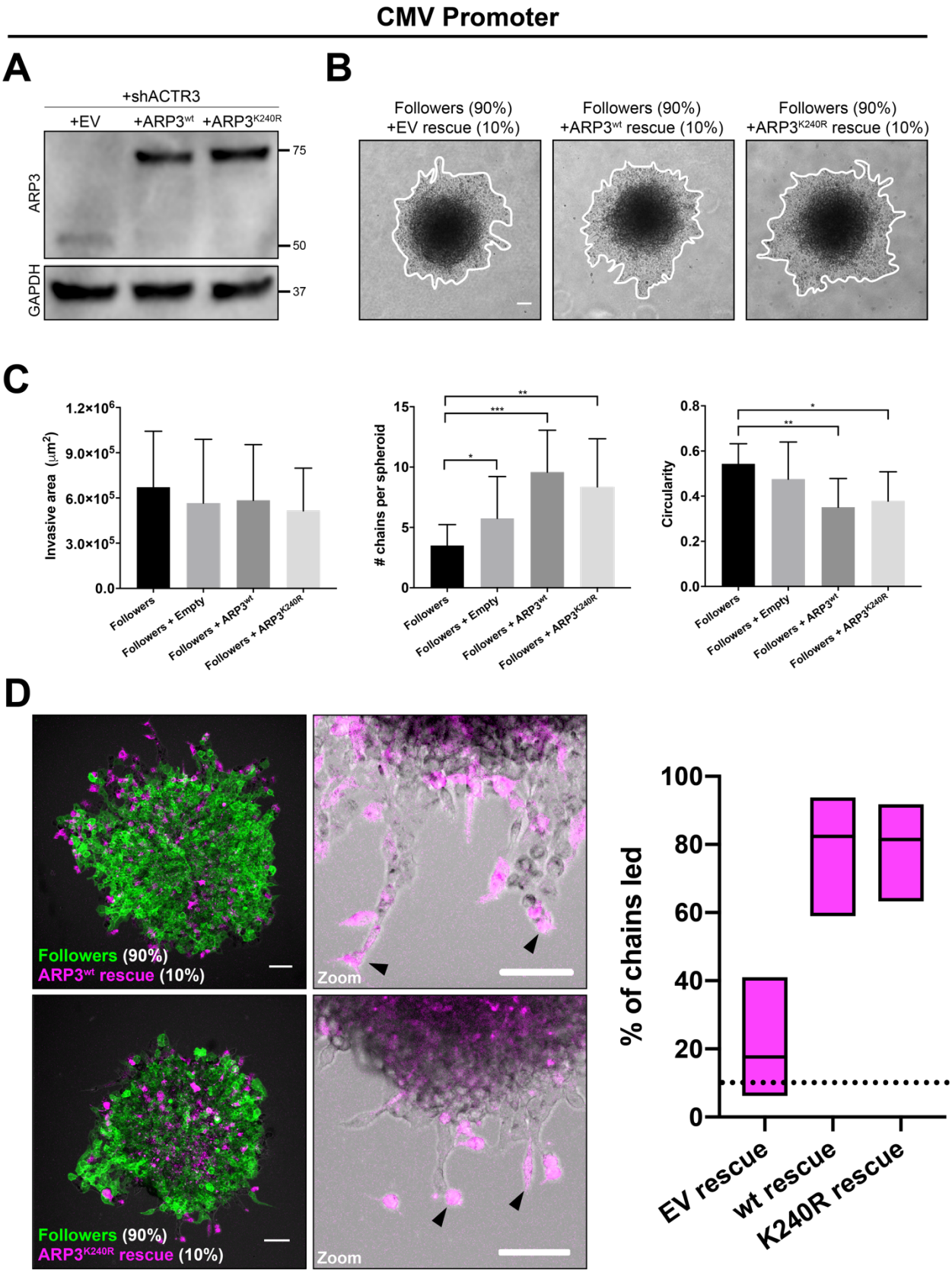

CMV Promoter
A
+shACTR3
+EV
+ARP3wt
+ARP3K240R
ARP3
GAPDH
75
50
37
B
Followers (90%)
+EV rescue (10%)
Followers (90%)
+ARP3wt rescue (10%)
Followers (90%)
+ARP3K240R rescue (10%)
C
Invasive area (μm2)
1.2×10^6
9.0×10^5
6.0×10^5
3.0×10^5
0.0
chains per spheroid
15
10
5
0
Circularity
0.8
0.6
0.4
0.2
0.0
Followers
Followers + Empty
Followers + ARP3wt
Followers + ARP3K240R
D
Followers (90%)
ARP3wt rescue (10%)
Zoom
Followers (90%)
ARP3K240R rescue (10%)
Zoom
% of chains led
100
80
60
40
20
0
EV rescue
wt rescue
K240R rescue

Figure 2.5. Wild-type ARP3 and ARP3 K240R both confer leader-like properties when expressed at high levels. (A) Western blot showing protein expression levels of ARP3 in followers after knockdown of endogenous ARP3 and rescue with empty vector, wild-type ARP3, or ARP3 K240R. Rescue constructs were expressed under control of the CMV promoter. (B-C) 24 hr invasion of mixed spheroids in Matrigel. Representative images (B) and quantification (C) of invasive area, circularity, and average numbers of chains per spheroid shown for each condition (mean±s.d., n=12, 12, 12, and 11 spheroids per group, respectively, across N=2 experiments. *p<0.05, **p<0.01, ***p<0.001, by one-way ANOVA with Tukey's post-test). Scale bars: 100 μm. (D) Confocal fluorescence imaging of mixed spheroids, with unmodified followers shown in green and mCherry-wild-type ARP3-rescued cells or mCherry-ARP3 K240R-rescued cells shown in magenta. Black arrows indicate invasive chains being led by experimental rescue cells. Graphs show the percentage of chains (mean +/- 95% confidence intervals) led by mCherry-empty vector-rescued, wild-type ARP3-rescued and ARP3 K240R-rescued followers. Dotted line denotes the proportion of ARP3 rescued cells in the mixed spheroids thus the 10% of chains expected to be led based on random chance.

groups (Fig. 2.5C). Additionally, while chain number increased and circularity decreased in spheroids containing either wild-type ARP3-rescued or ARP3 K240R-rescued followers (Fig. 2.5C), there was no significant difference between the two, suggesting that the functional difference between ARP3 K240R and wild-type ARP3 is mitigated at higher expression levels. Furthermore, both wild-type ARP3-rescued followers (81.5% of chains led; 95% confidence interval 63.3% to 91.8%) and ARP3 K240R-rescued followers (84.2% of chains led; 95% confidence interval 62.4% to 94.5%) promoted leader cell behavior when mixed with 90% unmodified followers (Fig. 2.5D). The specificity of the effect was further confirmed by mixing experiments with mCherry-empty vector-rescued follower cells, which led only 17.7% of chains (95% confidence interval 6.19% to 41.0%; Fig. 2.5D), suggesting that the enhanced leader ability of the ARP3 K240R cells was not simply due to the different cell types segregating within the spheroid. Based upon these findings, it appears that increased dosage of ARP3 is sufficient to promote leader-like behavior, and that the selective ability of ARP3 K240R to lead invasive chains and drive collective invasion at lower expression levels may arise, in part, from increased effective dose of ARP3 protein.

SNP analysis suggests distinct leader and follower cell lineages.

If high expression of wild-type KDM5B enforces the leader phenotype and/or suppresses the emergence of an alternate phenotype (e.g. followers), and this has an impact on collective behavior (data not shown), then one might predict that there would be a selection against the expression of KDM5B L685W in leader cells and selection for KDM5B L685W in follower cells independently captured from H1299 spheroids relative to the parental population. Fortuitously, a common SNP (rs1141108, chr1: 202715284, *G>A*) is located within the same exon as the KDM5B mutation, and the H1299 cell line was determined to be heterozygous for this SNP in our initial sequence analysis. Targeted deep resequencing of an amplicon including the region containing both the mutation

(chr1:202715414) and the SNP (chr1: 202715284) in genomic DNA thus enabled us to trace the relationship between allelic balance and the frequency of the KDM5B L685W mutation in each cell population (average depth= 348,327 reads).

We first found that approximately one third of alleles in each population carried the *A* variant and two thirds carry the *G* variant at rs1141108, leading us to conclude that there are 3 copies of *KDM5B* and/or chromosome 1 in our strain of H1299 cells and that this ploidy is maintained across the three populations. The *ACTR3* mutation that gives rise to ARP3 K240R on chromosome 2 also occurs in approximately one third of reads from genomic DNA in leader cells (Fig. 2.1B). Focal copy number alterations in the genomic regions surrounding ACTR3 and KDM5B have not been detected in H1299 cells by SNP copy number analyses (COSMIC Cell Lines Project: https://cancer.sanger.ac.uk/cell_lines). These data suggest that our strain of H1299 cells are functionally triploid. We further ascertained that the *KDM5B* variant (chr1:202715414: *A*>*C*) giving rise to the L685W mutation resides exclusively in *cis* with the SNP rs1141108 *G* allele (99.7% concordance across 267,414 total reads containing the mutation). The analysis further confirmed the proportions of *KDM5B* wild-type vs. mutant alleles shown in Fig. 2.1, including the more variable mutation frequency observed among different isolates of the parental population, the complete absence of the KDM5B L685W mutation (<0.00001%) in the leader cells and the very consistent ~30% in the follower cells (Fig. 2.6A). Additionally, the relative proportions of the wild-type and variant expressed at the mRNA level were largely reflective of the proportion at the genomic DNA level across all three populations, again suggesting that there was no allelic bias in expression of the wild-type versus mutant forms of KDM5B (Fig. 2.6A).

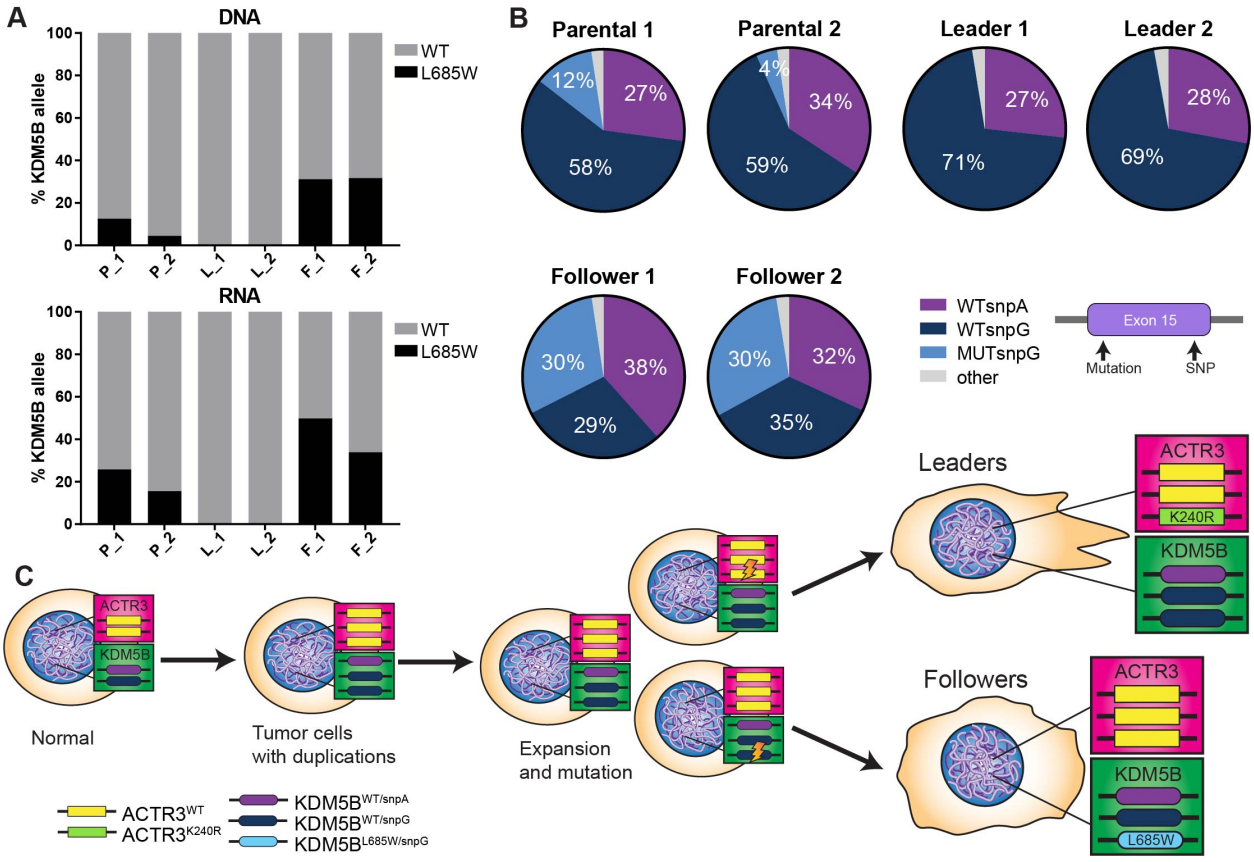


Figure 2.6. Leader and follower cells are derived from two separate populations defined by mutational profile. (A) Deep targeted resequencing across KDM5B exon 15 in genomic DNA (top) and RNA (bottom) isolated from parental, leader, and follower populations (*N*=2 independent isolates of DNA/RNA at separate passages of cells derived from a single phenotypic isolation. Average depth= 348,327 reads per sample). (B) Pie charts depicting proportions of KDM5B genotypes across parental, leader, and follower populations as determined from the deep amplicon sequencing shown in panel A. (C) Model of the potential history of leader and follower populations from parental cell population as inferred from the genetic profiles of KDM5B and ARP3.

The finding that the KDM5B L685W mutation was exclusive to the *G* allele while the *A* allele was exclusively wild-type allowed us to trace the relative proportions of the three genotypes (WT/A, WT/G, mutant/G) across the parental, leader, and follower populations. These data showed that whereas 4-12% of the parental population carries the KDM5B mutant/*G* allele, this allele was excluded from the leader population, which exhibited essentially none of the KDM5B mutant/*G* alleles (Fig. 2.6B), but was nearly 2-fold enriched in the isolated followers (30%) (p=0.008; ANOVA plus Tukey's post-hoc correction) and approached the expected frequency (if each allele is independently sorted). Taken together these data suggest that whereas the parental population varies in the fraction of cells containing the mutant/SNP G allele, growth in 3-D culture, SaGA-based capture and subsequent expansion of leader cells selectively enriches for cells expressing only wild-type KDM5B, whereas that of followers selects for a population in which nearly every cell contains (and expresses) one copy of the mutant KDM5B L685W allele (Fig. 2.6C).

2.4. Discussion

The greatest threat to cancer patient mortality is the metastatic spread of tumor cells from the primary site (33). Collective migration and invasion are major contributors to the dispersion of metastatic cancer cells (15, 57, 80, 81). Collective invasion is typified by the coordinated movement of a group of cohesive cells, often including multiple heterogeneous cell populations with specialized functions. One well-studied example of collective invasion is that of chain-like invasion, in which specialized leader cells lead groups of cells termed follower cells, out of the tumor (3, 57, 81), with both populations playing important roles in the process of invasion. Until now, studies of the distinct populations within invasive chains have been limited by the inability to separate these populations. Development of the SaGA technique (3), enabled us to independently analyze leader and follower cells with different phenotypes to gain insight into population dynamics and the emergence of populations that differ in cell behavior. Our results identify a set of expressed mutations that define leader and follower cells, representing, to our knowledge, the first known instance of distinct mutations as contributors to the leader/follower phenotypes within collectively invading packs.

We confirmed the importance of the leader cell-enriched mutation ARP3 K240R to the invasive leader cell phenotype by introducing it into a population of non-invasive H1299 follower cells. Both in pure spheroids and when mixed with 90% unmodified followers, ARP3 K240R-expressing followers displayed increased ability to invade and lead collective chains at both lower and higher protein expression levels. Rescue with wild-type ARP3 also conferred leader cell behavior, but only when expressed at supra-physiologic levels. One potential explanation is that ARP3 K240R increases the effective dosage of ARP3 protein, essentially recapitulating ARP3 overexpression even at low expression. Indeed, ARP3 K240R accumulated to higher levels than wild-type ARP3 when

exogenously expressed from the same vector. The K240R mutation might interfere with ubiquitylation at K240, resulting in either decreased ARP3 turnover, or enhanced ARP3 activity. Ubiquitylation at K240 has previously been observed by mass spectroscopy in multiple human cell lines as well as mouse tissue, and K240R was predicted by SAPH-ire to have a high likelihood of functional consequence. Further experiments will be necessary to determine whether ARP3 K240R is indeed resistant to ubiquitylation, and whether this impacts the activity of ARP3 in leader cells. ARP3 is a key subunit of the Arp2/3 complex that regulates intracellular actin dynamics in a number of processes, including lamellipodia protrusion during cell motility (70). Indeed, we observed significantly reduced invasion in parental, leader, and follower cells upon ARP3 knockdown, supporting its importance for cell migration and invasion. Overexpression of Arp2/3 complex subunits including ARP2, ARP3, ARPC2, and ARPC5 has been shown to promote invasion in multiple cancer types including lung, colorectal, glioblastoma, and others (71-74, 82-84). Our results now further indicate a role for ARP3 as contributing to tumor collective invasion by promoting the leader cell phenotype.

Emerging research regarding how followers cooperate with leaders to promote their invasive capabilities center on concepts such as contact inhibition of locomotion (CIL) and movement along a chemical gradient (81, 85). CIL occurs when a migrating cell contacts another cell and begins forming protrusions opposite the site of contact to move in the opposite direction (86-88). In this case, follower cells contact leader cells, forcing them to polarize and move in a forward direction. Follower cells also regulate signaling to leader cells, often through creation of chemical gradients that motivate leader cells to move in a particular direction (81, 89-93). Indeed, we previously discovered that follower and leader cells engage in a symbiotic relationship, in which follower cells promote the survival and proliferation of leader cells, which in turn secrete VEGFA to promote the motility of follower cells (3).

Our sequencing data and the ability to discern the clonality of the distinct leader and follower subpopulations relative to that of the parental population from which they were derived allows us to construct a model of the population origins. In a population of tumor cells which are already functionally triploid, *ACTR3* and *KDM5B* undergo mutation, each in separate cells. These mutational events are the beginning of the divergent paths that will develop two separate but cooperative subpopulations. Cells containing mutant *ACTR3* go on to form the highly invasive, slow proliferating leader cells while cells containing the *KDM5B* mutation ultimately become the follower population of invasion-deficient, rapidly proliferating cells. Expressing the leader specific *ACTR3* mutation in follower cells increases chain-like invasion. Likewise, expressing the follower specific *KDM5B* mutation in leader cells increases chain-like invasion (data not shown). Our phenotypic data adds to this model the necessity for cells exhibiting both the leader phenotype and a follower phenotype to the cooperative behavior demonstrated in collective invasion.

The approach used herein, while innovative, does have certain limitations. First, the work focuses on the characterization of leader and followers isolated from a single NSCLC-derived cell line. How universal the identified mutations/genes might be in contributing to leader/follower behavior in other cell lines or cancer types is currently unknown. In addition, while our functional characterization supports a potential role for ACTR3, this was only one of the phenotype-selective mutations identified. It is possible that no single alteration in isolation is sufficient to fully drive either phenotype, but rather the combinatorial effects of multiple genetic and epigenetic alterations contribute to collective behavior. Studies aimed at similar analyses of multiple cell lines, or collective invasion models and across cancer types may converge on critical "drivers" or pathways. Ultimately the identification of such key alterations may help in clinical decision-making by identifying predictive

biomarkers or new therapeutic targets. Indeed, a recently discovered Arp2/3 inhibitor CK-666 (94) has been shown to inhibit cell motility *in vivo* (95). An alternative approach would be targeting other components of the Arp2/3 pathway. One such target, PLK4, has been implicated as a driver of cancer invasion and metastasis in part through its interaction with Arp2/3 subunits (96), and PLK4 inhibitors are currently under clinical investigation for patients with advanced solid tumors. Our identification of a panel of mutations delineating leader and follower cell phenotypes in a non-small cell lung cancer tumor population is an initial step toward elucidation of how heterogeneous genetic mutations contribute to cancer metastasis and how these vulnerabilities can be exploited to circumvent the development of metastasis.

Supplementary Information

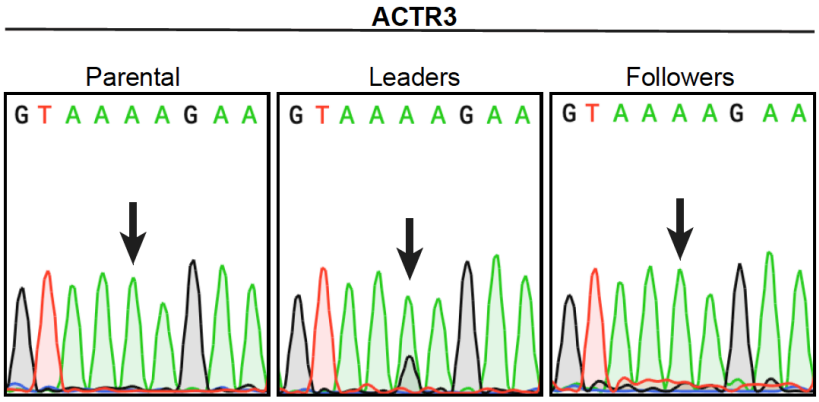


Figure S2.1. Confirmation of leader- and follower-enriched mutations. Sanger sequencing confirming leader-enriched ACTR3 mutation in cDNA (shown) and genomic DNA isolated from H1299 parental, leader and follower populations. Black arrows indicate the bases of interest. Only the wild-type A peak is seen in the parental and follower populations, while the leader population contains both A and G peaks.

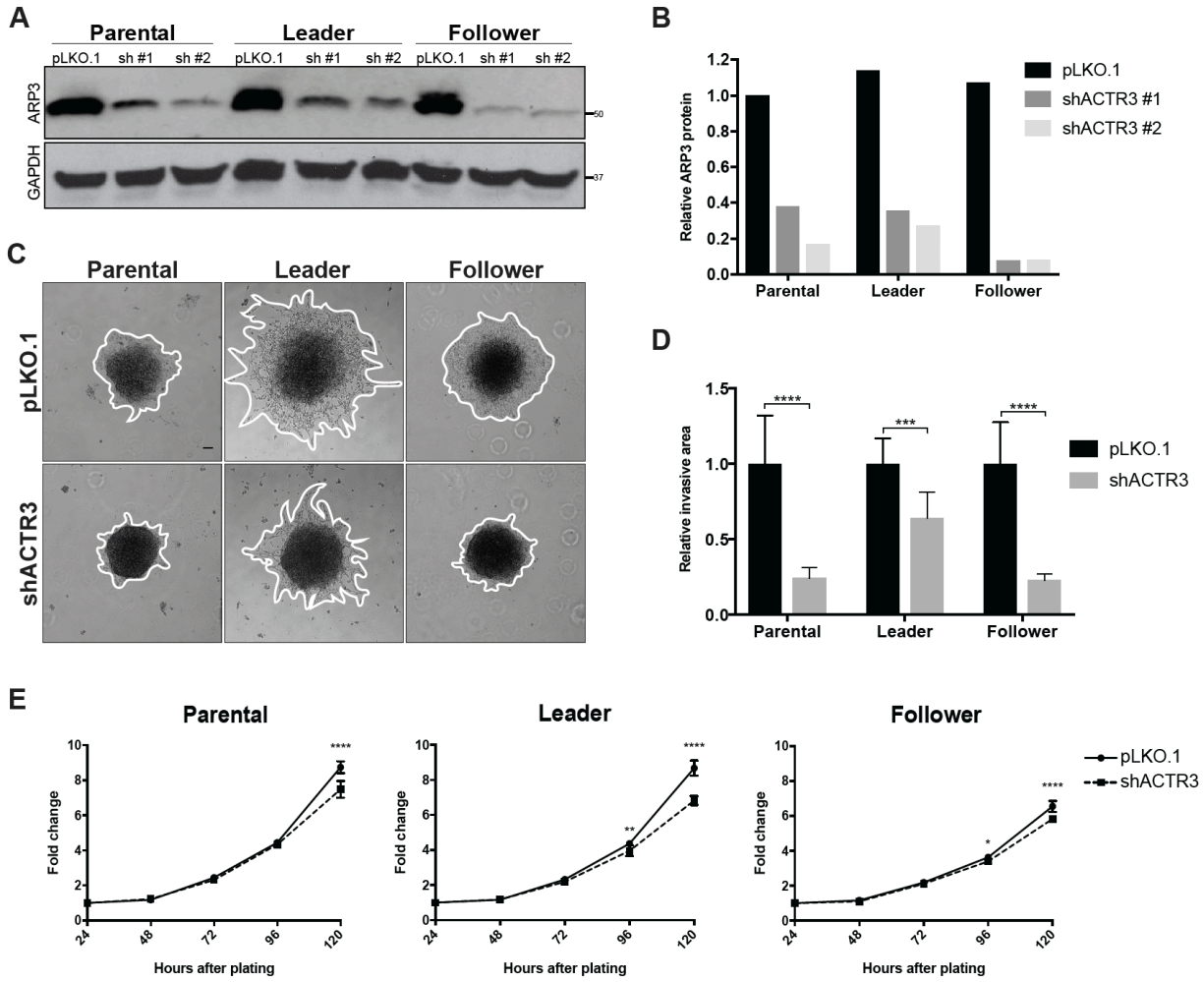


Figure S2.2. ARP3 knockdown inhibits 3-D invasion. (A) Western blot showing ARP3 protein levels in H1299 parental, leader and follower cells upon expression of empty pLKO.1 vector, ARP3 shRNA #1 (Millipore Sigma TRCN0000029383), or ARP3 shRNA #2 (Millipore Sigma TRCN0000380403). (B) Western blot densitometry quantification, indicating 70-90% knockdown of ARP3 protein using either shRNA #1 or shRNA #2. (C) Representative images of 24-hour invasion of H1299 parental, leader, and follower spheroids expressing either empty pLKO.1 or shACTR3 #2. Scale bar = 100μm.(D) Quantification of relative 24-hour invasive area, normalized to pLKO.1 control for each group. (mean±s.d., n=5, 11, and 5 spheroids for parental, leader and follower lines, respectively. ***p<0.001, ****p<0.0001 by two-way ANOVA with Sidak correction). (E) Growth rate of parental, leader, and follower lines expressing either empty pLKO.1 or shACTR3 #2. (mean+s.d., n=5 replicates per time point. *p<0.05, ***p<0.001, ****p<0.0001 by two-way ANOVA with Šidák correction).

Table S2.1. PCR primers for ACTR3

		Primer sequence (5'-3')
gDNA	Forward	GTTACTTTTGTTTCTTTGTTTTTCAG
	Reverse	TTCATATTTGCTGCTGAATACTTTT
cDNA	Forward	TCCCTCCAGAACAATCCTTG
	Reverse	GGTTGTGTAAAGTCTGGATTAGCA

Chapter 3: Prognostic significance of an invasive leader cell-derived mutation cluster on chromosome 16q

Adapted from "Prognostic significance of an invasive leader cell-derived mutation cluster on chromosome 16q." In Press, *Cancer*.

Brian Pedro[1], Manali Rupji[2], Bhakti Dwivedi[2], Jeanne Kowalski[2,3,^], Jessica M. Konen[1,^^], Taofeek K. Owonikoko[2,4], Suresh S. Ramalingam[2,4], Paula M. Vertino[5,6], Adam I. Marcus[2,4,*]

Affiliations:

[1]Graduate Program in Cancer Biology, Emory University, Atlanta, GA 30322

[2]Winship Cancer Institute, Emory University, Atlanta, GA 30322, USA

[3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322

[4]Department of Hematology and Medical Oncology, Emory University, Atlanta, GA 30322

[5]Department of Radiation Oncology, Emory University, Atlanta, GA 30322

[6]Department of Biomedical Genetics and the Wilmot Cancer Institute, University of Rochester Medical Center, Rochester, NY, 14642

^Present address: Department of Oncology, Dell Medical School, The University of Texas at Austin, Austin, TX, 78712

^^Present address: Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract

Background: Intra-tumoral heterogeneity is defined by subpopulations with varying genotypes and phenotypes. Specialized, highly invasive leader cells and less invasive follower cells are phenotypically distinct subpopulations that cooperate during collective cancer invasion. Since leader cells are a rare subpopulation that would be missed by bulk sequencing, a novel image-guided genomics platform was employed to precisely select this subpopulation. We identified a novel leader cell mutation signature and tested its ability to predict prognosis in non-small cell lung cancer (NSCLC) patient cohorts.

Methods: SaGA was used to isolate and perform RNA-sequencing on leader and follower populations from the H1299 NSCLC cell line, revealing a leader-specific mutation cluster on chromosome 16q. Genomic data from lung squamous cell carcinoma (LUSC, n=475) and lung adenocarcinoma (LUAD, n=501) patients from The Cancer Genome Atlas (TCGA) were stratified by 16q mutation cluster status (16qMC+ vs. 16qMC-) and compared for overall survival, progression-free survival, and gene set enrichment analysis (GSEA).

Results: Poorer overall survival and/or progression-free survival was found across all stages and among early-stage patients with 16qMC+ tumors within LUSC and LUAD cohorts. GSEA revealed 16qMC+ tumors to be enriched for expression of metastasis- and survival-associated gene sets.

Conclusion: This represents the first leader cell mutation signature identified in patients and has the potential to better stratify high-risk NSCLC and ultimately improve patient outcomes.

3.1 Introduction

Intra-tumoral heterogeneity stems from internal and external selection pressures (52, 56, 97-100), leading to cellular subpopulations with varying genomes and phenotypes. This heterogeneity is a key contributor to treatment resistance and cancer progression (19, 49, 56, 99-102); however, this heterogeneity may be missed due to bulk sequencing of only a portion of the tumor. Consequently, the complex genetic and phenotypic landscape from the tumor is not fully captured.

Evidence from *in vitro* studies and primary solid tumors suggests that rare cells unwittingly missed from bulk sequencing are important for tumor progression and metastasis (3, 103). Using a 3-D *in vitro* model of lung cancer invasion, we showed that collectively invading packs of tumor cells are heterogeneous, and include rare, specialized leader cells that pioneer invasive chains, and follower cells that adhere to and invade behind leaders (3). Collective invasion is widely observed in carcinomas and increases the overall success of metastasis (12-14). Leader cells promote collective invasion when mixed with poorly invasive follower cells, even when comprising as little as 1 percent of the population.(3) In addition, leader cells are genetically distinct from followers, harboring unique gene expression profiles that may help to facilitate collective invasion (3).

Rare subpopulations such as leader cells could be important for cancer metastasis, yet underrepresented by standard tumor sequencing. We therefore sought to use our imaging-guided genomics platform (Spatiotemporal Genomic and Cellular Analysis, or SaGA) (3) to identify unique leader cell gene mutations and define higher-risk patient groups in non-small cell lung cancer (NSCLC), which includes squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD). Using a novel, leader cell-specific cluster of mutated genes on chromosome 16q, we found that LUSC and

LUAD patients with 1 or more mutation(s) within this cluster have poorer overall and progression-free survival, even among early-stage patients. This represents the first leader cell mutation signature identified in patients and has the potential to better stratify high-risk NSCLC and ultimately improve patient outcomes.

3.2 Methods

Identification of leader- and follower-enriched variants

Isolation via SaGA, RNA-sequencing expression, and variant calling for leader and follower cells from the H1299 cell line were performed as previously described (3, 104). RNA-sequencing data are deposited in the NCBI SRA database under accession number PRJNA542374.

Patient selection and stratification

For TCGA cohorts in cBioPortal, only patients with available mutation data were included. Patients with at least one non-synonymous mutation in at least one 16qMC gene were categorized as "16qMC+". Lollipop plots depicting locations of 16q cluster point mutations in each cohort were constructed using MutationMapper through cBioPortal (105, 106). Patient clinical data were downloaded from cBioPortal.

Enrichment analysis

For GSEA, previously processed versions of TCGA LUAD, LUSC, and LIHC (HCC) RNA-seq data based on human genome build hg19 were downloaded for the included subsets of patients from the GDC legacy archive (https://portal.gdc.cancer.gov/legacy-archive/search/). Raw RSEM expression counts were filtered for lowly expressed genes (average CPM<1.0) and normalized by the TMM method using edgeR (107, 108). Differential expression between 16qMC+ and 16qMC- was calculated for all genes with limma R package (108). Genes were ranked according to -$\log_{10}$(*P* value) multiplied by direction of fold change. GSEAPreranked was performed on the ranked gene list with classic enrichment statistics under default settings and C2 curated gene sets (4762 gene sets) from MSigDB 6.2 release using GSEA Desktop v3.0 (109, 110).

Statistical analysis

Statistical analysis was conducted using SAS Version 9.4 and GraphPad Prism Version 8.2. Ordinary one-way ANOVA with Sidak's multiple comparisons test was used when three or more conditions were being compared. Confidence intervals of percentages were calculated using the Wilson/Brown method. Patient characteristics were reported as counts with percentages for categorical variables and median with range for numeric variables. A chi-square or Fisher's exact test, as appropriate, was conducted to identify associations between categorical demographic characteristics and 16qMC status, and an ANOVA or a Kruskal-Wallis test, as appropriate, was conducted to identify associations between continuous demographic factors and 16qMC status.

OS and PFS were calculated by the Kaplan-Meier method, with *P* values calculated by the log-rank (Mantel-Cox) test. A univariable cox proportional hazards regression analysis was performed to determine any significant association of the demographic factors and OS/PFS. Variables significant at an alpha of 0.2 were used for model selection. A multivariable cox regression analysis using a backward elimination approach was used to select covariates, with removal of covariates of alpha >0.2.

For subgroup survival analysis based on early/late stage or mutation count categories, KM curves were created based upon 16qMC status for both OS and PFS. To account for the small number of events in the strata, Firth's penalized regression approach was used within each subgroup. The multivariable analysis was conducted as described above. Similar subgroup analysis was performed for early-stage (I and II) and late stage (III and IV) patients. For the four-group survival analysis by 16qMC status and *TP53* mutation status, KM curves for each OS and PFS endpoint were created and log-rank *P* values were obtained. Pairwise log-rank *P* values were adjusted using Tukey-Kramer's method for multiple comparisons.

Random sampling analysis

To assess the statistical significance of the 16qMC genes in separating patients by OS and PFS versus a randomly-defined set of genes, a bootstrap approach was used for the TCGA LUSC, LUAD, and HCC cohorts. Specifically, M = 1,000 random samples were taken, with "positive" patients defined by those having any mutations within 8, 9, and 9 genes from 20502, 20502, 20503 total genes in the TCGA LUSC, LUAD, and HCC cohorts, respectively. For each re-sampling, KM plots were constructed for each outcome, and a log-rank test was performed. A Monte Carlo P value was defined by comparing the P value obtained from the log-rank test analysis of 16qMC+ vs. 16qMC- samples to the distribution of P values obtained based on the groups formed using randomly-sampled gene mutations.

3.3 Results

Development of leader cell-specific 16q mutation cluster

We utilized leader and follower cell lines previously derived from the H1299 NSCLC cell line using the SaGA platform (described in (3); schematic in Fig. 3.1A). As leader cells are crucial for collective invasion in 3-D assays (3), we hypothesized that NSCLC patients with genetic evidence of leader cells within the primary tumor could be at higher risk for disease progression and recurrence. Our previous data show that H1299 leader and follower cells contain distinct mutational profiles (104). Additional inclusion of known variants from the dbSNP database (111) resulted in 17 leader-specific and 18 follower-specific mutations (Fig. 3.1B; Table S3.1). Notably, 7 leader-specific mutations were found on chromosome 16q (Table S3.2; Fig. 3.1C, solid lines). We hypothesized that these mutations could help detect leader cell subpopulations; therefore, after confirming comparable mRNA levels of each gene in the leader and follower populations (Fig. S3.1), we used these genes to define a leader cell mutation signature.

Identification of 16q mutation cluster-positive tumors in NSCLC patient cohorts

Gene expression data and clinical outcomes information for LUSC patients (n=475) and LUAD patients (n=501) were extracted from The Cancer Genome Atlas (TCGA) (112). Importantly, 37 of 475 (7.8%) LUSC patients and 30 of 501 (6.2%) LUAD patients had one or more mutations among the seven 16q mutation cluster genes (LUSC: Figs. 3.1D, 3.2A; LUAD: Figs. 3.1D, 3.2D). *NQO1* was mutated in one patient and was excluded from subsequent analyses. Nearly all of the identified point mutations occurred at different loci, suggesting that they could result from a hyper-mutational process rather than being selected due to altered protein function. Among genes directly adjacent to the six leader-derived 16q genes (Fig. 3.1C, dashed lines), the same pattern of randomly-distributed

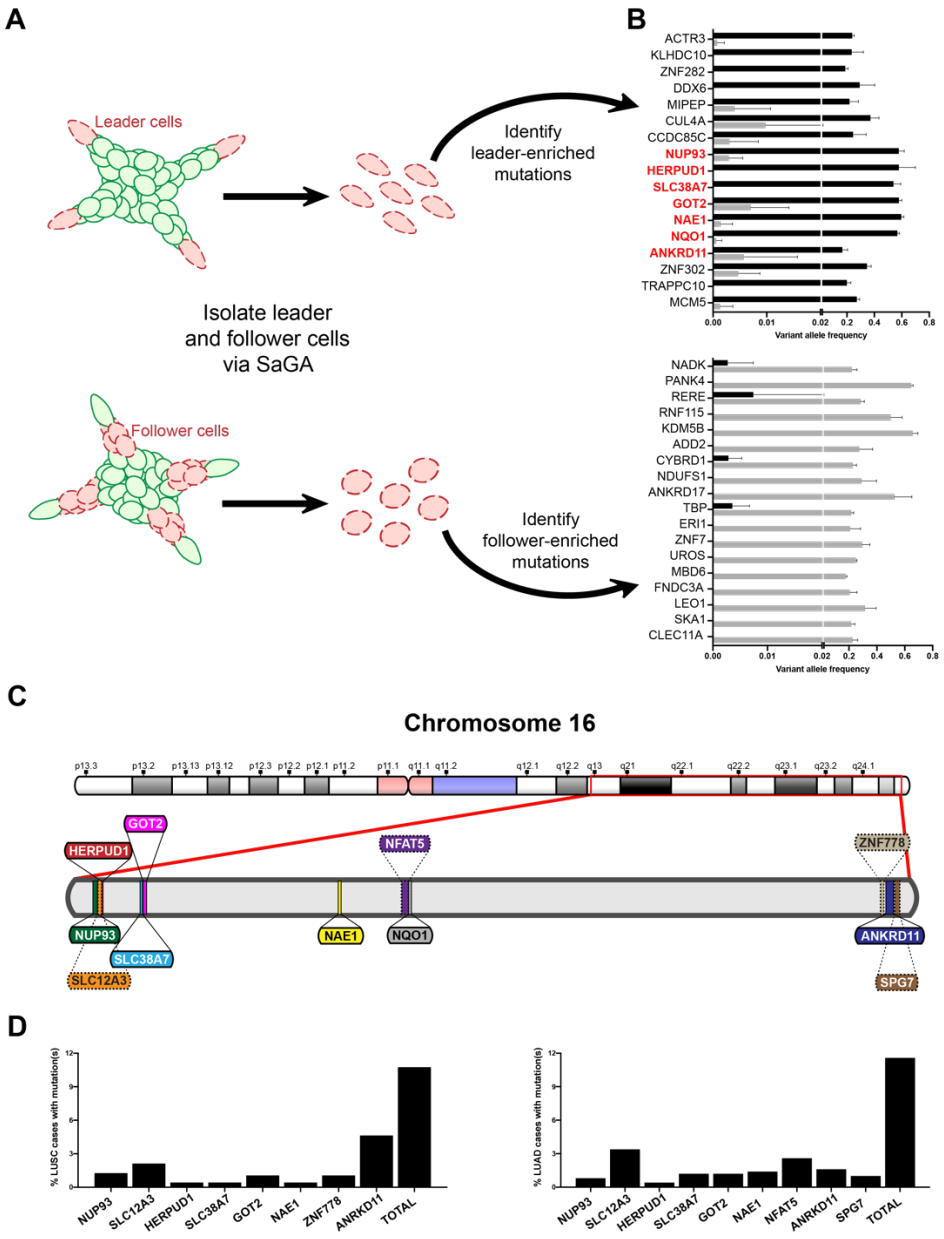


Figure 3.1. Identification of a leader cell-derived mutation cluster on chromosome 16q. (A) Schematic of the SaGA technique for photoconversion, isolation, and downstream analysis of H1299 leader and follower cells. Adapted from (3). (B) Variant allele frequency values from RNA-sequencing of H1299 leader and follower populations for 17 genes identified as containing leader-specific point mutations, and 18 genes identified as containing follower-specific point mutations. (C) Map of chromosome 16q annotated with locations of genes containing leader-specific mutations (solid lines) and adjacent genes subsequently included in the 16q mutation cluster (dotted lines). (D) Percentages of TCGA LUSC and LUAD cases with mutations in each of eight (LUSC) or nine (LUAD) 16q cluster genes.

Table 3.1: Patient characteristics for LUSC and LUAD TCGA cohort

			LUSC			LUAD		
Covariate	**Statistic**	**Group**	**16q cluster status**		**P value[ab]**	**16q cluster status**		**P value**
			16qMC+ (N=51)	16qMC- (N=424)		16qMC+ (N=58)	16qMC- (N=443)	
Gender	N (%)	Female	13 (25.49)	112 (26.54)	0.872	28 (48.28)	240 (54.18)	0.397
	N (%)	Male	38 (74.51)	310 (73.46)		30 (51.72)	203 (45.82)	
Pathologic stage	N (%)	Stage I & II	41 (80.39)	343 (81.47)	0.852	46 (79.31)	346 (78.46)	0.882
	N (%)	Stage III & IV	10 (19.61)	78 (18.53)		12 (20.69)	95 (21.54)	
Smoking history	N (%)	Current/reformed smoker	49 (98)	399 (96.14)	1	55 (96.49)	362 (84.19)	**0.009**
	N (%)	Nonsmoker	1 (2)	16 (3.86)		2 (3.51)	68 (15.81)	
Mutation count[c]	N (%)	High mut. count	43 (84.31)	235 (57.46)	**<0.001**	49 (84.48)	202 (46.12)	**<0.001**
	N (%)	Low mut. count	8 (15.69)	174 (42.54)		9 (15.52)	236 (53.88)	
Age at diagnosis	Median (min-max)		67 (44-83)	68 (39-90)	0.363	65 (40-88)	66 (38-87)	0.494

[a] P-values calculated by ANOVA for numerical covariates and chi-square or Fisher's exact test for categorical covariates.

[b] P-values calculated by Kruskal-Wallis test for numerical covariates.

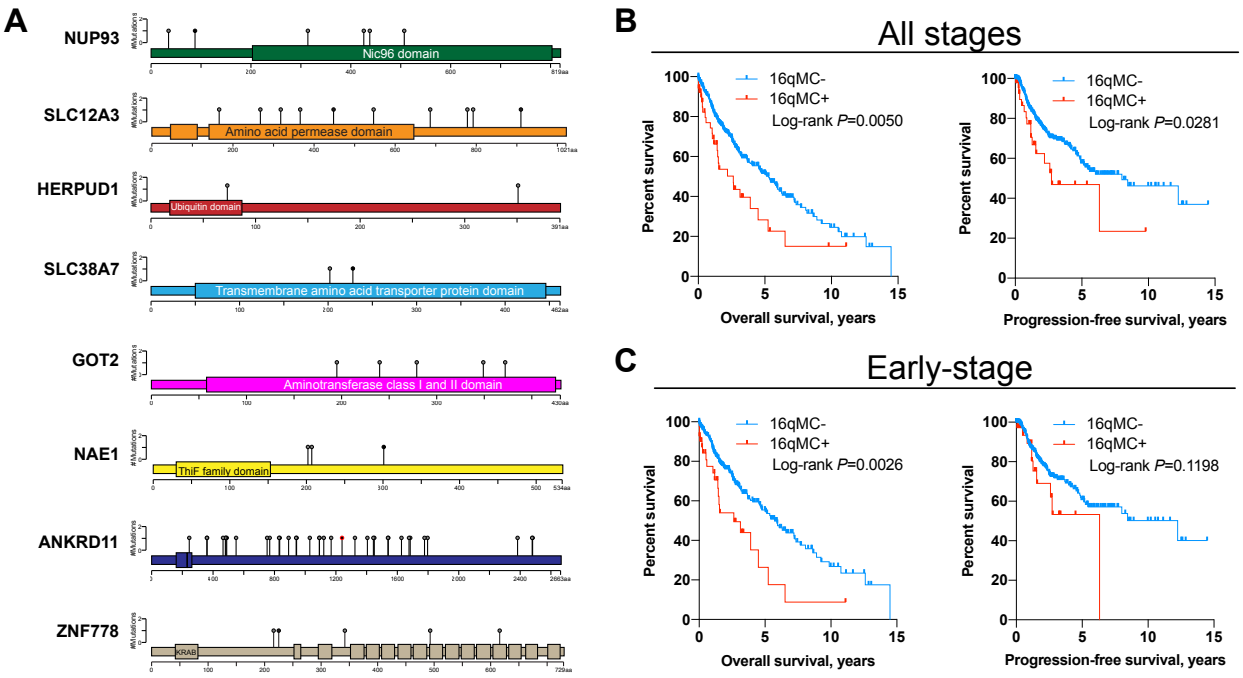
[c] High/low mutation count was defined by a cutoff of 192 mutations, based upon previous mutational burden analysis of TCGA cohorts (113).

mutations was observed in *SLC12A3* and *ZNF778* in the LUSC cohort (Fig. 3.2A), and *SLC12A3*, *NFAT5* and *SPG7* in the LUAD cohort (Fig. 3.2D). Taken together, 10.7% of LUSC patients and 11.6% of LUAD patients had at least one mutation within the respective 8- or 9-gene 16q clusters (Fig. 3.1D); these patients were defined as 16q mutation cluster-positive (16qMC+). The majority of 16qMC+ patients – 94.1% of LUSC and 86.2% of LUAD – had mutations in only one 16q cluster gene (Fig. S3.2). Most mutations were found at variant allele frequencies (VAF) of less than 50%, which likely indicates sub-clonal mutations barring any chromosomal alterations at that locus (Fig. S3.2). Additionally, 16qMC+ tumors had significantly higher mutation counts (Table 3.1). Within the LUAD cohort, the 16qMC+ group contained significantly more smokers (96.5% vs. 84.2%, P=0.009; Table 3.1), *TP53* mutations (77.6% vs. 48.6%, P=0.0005; Table S3.3), and patients who received radiation therapy prior to resection (20.7% vs. 11.0%, P=0.024; Table S3.4).

Prognostic validation of 16qMC in TCGA cohorts

We found that 16qMC+ patients had poorer overall survival (OS) (HR 1.79, 95% CI 1.19-2.71; log-rank P=0.005) and progression-free survival (PFS) (HR 1.78, 95% CI 1.06-3.01; log-rank P=0.028) among all-stage LUSC (Fig. 3.2B; Table 3.2). Notably, early-stage 16qMC+ LUSC patients had poorer OS (HR 2.08, 95% CI 1.27-3.24; log-rank P=0.003) (Fig. 3.2C; Table S3.5). In the LUAD cohort, all-stage 16qMC+ patients experienced poorer OS (HR 1.84, 95% CI 1.73-2.74; log-rank P=0.002; Fig. 2E; Table 3.3) as did early-stage patients (HR 2.06, 95% CI 1.26-3.23; log-rank P=0.003; Fig. 2F; Table S3.6). Multivariable Cox regression analysis indicated 16qMC+ status as a significant predictor of OS (HR 1.71, 95% CI 1.13-2.58; P=0.011) and PFS (HR 1.73, 95% CI 1.00-2.97; P=0.049) among all-stage LUSC patients (Table 3.2), and of OS among all-stage LUAD patients (HR 1.95, 95% CI 1.31-2.91; P=0.001) (Table 3.3). In multivariable analysis among early-stage patients, 16qMC+ status

Lung squamous cell carcinoma



Lung adenocarcinoma

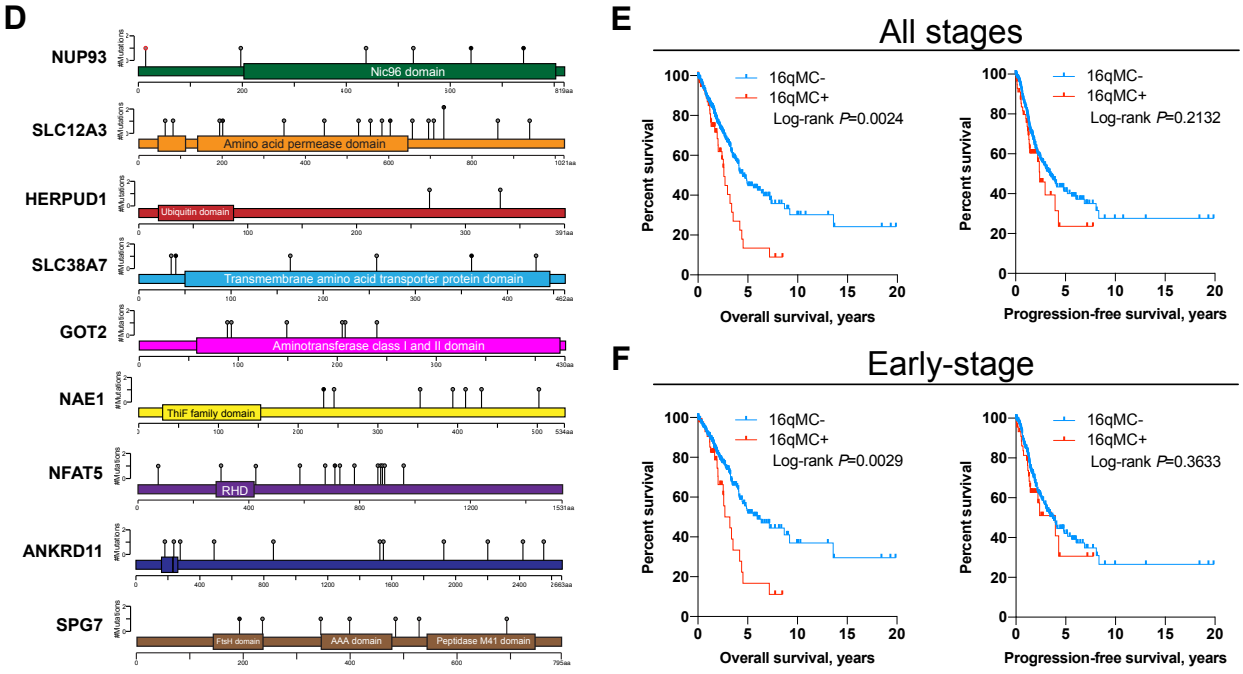


Figure 3.2. 16qMC predicts poor prognosis in non small cell lung cancer cohorts. (A) Lollipop plots illustrating locations of point mutations in 16q cluster genes in TCGA LUSC patients. Black dots depict truncations; gray dots depict missense mutations; red outlines depict driver mutations indicated by OncoKB and/or Cancer Hotspots. (B) Kaplan Meier (KM) curves for OS and PFS of 16qMC+ and 16qMC- TCGA LUSC patients. Median OS: 5.0 years (16qMC-) vs. 2.6 years (16qMC+); median PFS: 8.0 years (16qMC-) vs. 2.7 years (16qMC+). (C) KM curves for OS and PFS of 16qMC+ and 16qMC- stage I and II TCGA LUSC patients. Median OS: 5.4 years (16qMC-) vs. 2.6 years (16qMC+); median PFS: 8.4 years (16qMC-) vs. 6.3 years (16qMC+). (D) Lollipop plots illustrating locations of point mutations in 16q cluster genes in TCGA LUAD patients. Black dots depict truncation mutations; gray dots depict missense mutations; red outlines depict driver mutations indicated by OncoKB and/or Cancer Hotspots. (E) KM curves for OS and PFS of 16qMC+ and 16qMC- TCGA LUAD patients. Median OS: 4.2 years (16qMC-) vs. 2.6 years (16qMC+); median PFS: 3.1 years (16qMC-) vs. 2.4 years (16qMC+). (F) KM curves for OS and PFS of 16qMC+ and 16qMC- stage I and II TCGA LUAD patients. Median OS: 5.6 years (16qMC-) vs. 3.2 years (16qMC+); median PFS: 3.4 years (16qMC-) vs. 4.0 years (16qMC+). *P* values calculated by log-rank test.

Table 3.2. Cox regression analysis for all-stage LUSC TCGA patients

	Univariable analysis				Multivariable analysis			
	Overall survival		Progression-free survival		Overall survival		Progression-free survival	
Covariate	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
16q cluster status (16qMC+ vs. -)	1.79 (1.19-2.71)	**0.006**	1.78 (1.06-3.01)	**0.030**	1.71 (1.13-2.58)	**0.011**	1.73 (1.00-2.97)	**0.049**
Gender (Female vs. Male)	0.83 (0.60-1.16)	0.277	0.89 (0.60-1.32)	0.564				
Pathologic stage (I/II vs. III/IV)	0.61 (0.44-0.84)	**0.003**	0.50 (0.34-0.73)	**<0.001**	0.61 (0.44-0.85)	**0.004**	0.49 (0.33-0.73)	**<0.001**
Smoking history (Smoker vs. non)	0.63 (0.26-1.53)	0.305	0.36 (0.16-0.83)	**0.016**			0.34 (0.15-0.78)	**0.011**
Mutation count (High vs. low)	1.04 (0.78-1.39)	0.765	0.92 (0.65-1.30)	0.625				
Age at diagnosis	1.02 (1.00-1.03)	0.059	1.00 (0.98-1.02)	0.988	1.02 (1.00-1.04)	**0.038**		

Table 3.3. Cox regression analysis for all-stage LUAD TCGA patients

	Univariable analysis				Multivariable analysis			
	Overall survival		Progression-free survival		Overall survival		Progression-free survival	
Covariate	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
16q cluster status (16qMC+ vs. -)	1.84 (1.23-2.74)	**0.003**	1.30 (0.86-1.96)	0.214	1.95 (1.31-2.91)	**0.001**	1.30 (0.86-1.97)	0.207
Gender (Female vs. Male)	0.95 (0.71-1.27)	0.713	0.95 (0.72-1.26)	0.730				
Pathologic stage (I/II vs. III/IV)	0.37 (0.27-0.51)	**<0.001**	0.62 (0.45-0.86)	**0.004**	0.36 (0.27-0.50)	**<0.001**	0.62 (0.45-0.86)	**0.004**
Smoking history (Smoker vs. non)	0.91 (0.60-1.39)	0.676	0.96 (0.65-1.44)	0.859				
Mutation count (High vs. low)	0.98 (0.73-1.31)	0.893	0.94 (0.71-1.24)	0.669				
Age at diagnosis	1.01 (0.99-1.02)	0.349	1.00 (0.98-1.01)	0.695				

remained predictive of poorer OS for LUSC (HR 1.94, 95% CI 1.21-3.12; *P*=0.006; Table S3.5) and LUAD patients (HR 2.02, 95% CI 1.25-3.27; *P*=0.004; Table S3.6).

16qMC+ tumors in both cohorts had increased mutation counts (Fig. 3.3A, B) and 16qMC+ LUSC tumors had more copy number alterations (Fig. 3.3C). To determine whether mutation count was driving the poorer survival among 16qMC+ patients, we stratified by low (≤192) or high (>192) mutation count as previously described for TCGA cohorts (113). We found that 16qMC+ status still correlated with poorer OS among highly-mutated LUSC (Fig. 3.3D), OS and PFS among lowly-mutated LUAD, and OS among highly-mutated LUAD (Fig. 3.3E). Given the higher proportion of *TP53* mutations among 16qMC+ LUAD tumors, we also examined survival by both *TP53* and 16qMC status; although mutated *TP53* was associated with poorer OS, 16qMC+ status further differentiated survival among *TP53* wild-type patients (Fig. S3.3). These data indicate that 16qMC+ status could help identify patients who are at higher risk for disease progression.

As collective invasion is observed in numerous carcinomas (36), we analyzed additional TCGA cohorts to test the prognostic value of leader-cell derived 16qMC in other cancer types. Notably, 16qMC+ patients within a TCGA hepatocellular carcinoma (HCC) cohort (114) also had significantly increased mutation counts, and poorer survival among all-stage and early-stage disease (Tables S3.7-S3.9; Fig. S3.4). As HCC carries poor prognosis and high rates of recurrence, HCC patients could also potentially benefit from 16qMC+ screening.

Given the scattered distribution of mutations in 16qMC+ patients (Figs. 3.2A, 3.2D, S3.4), we determined the prognostic power of the 16qMC genes compared with 1,000 randomly-selected

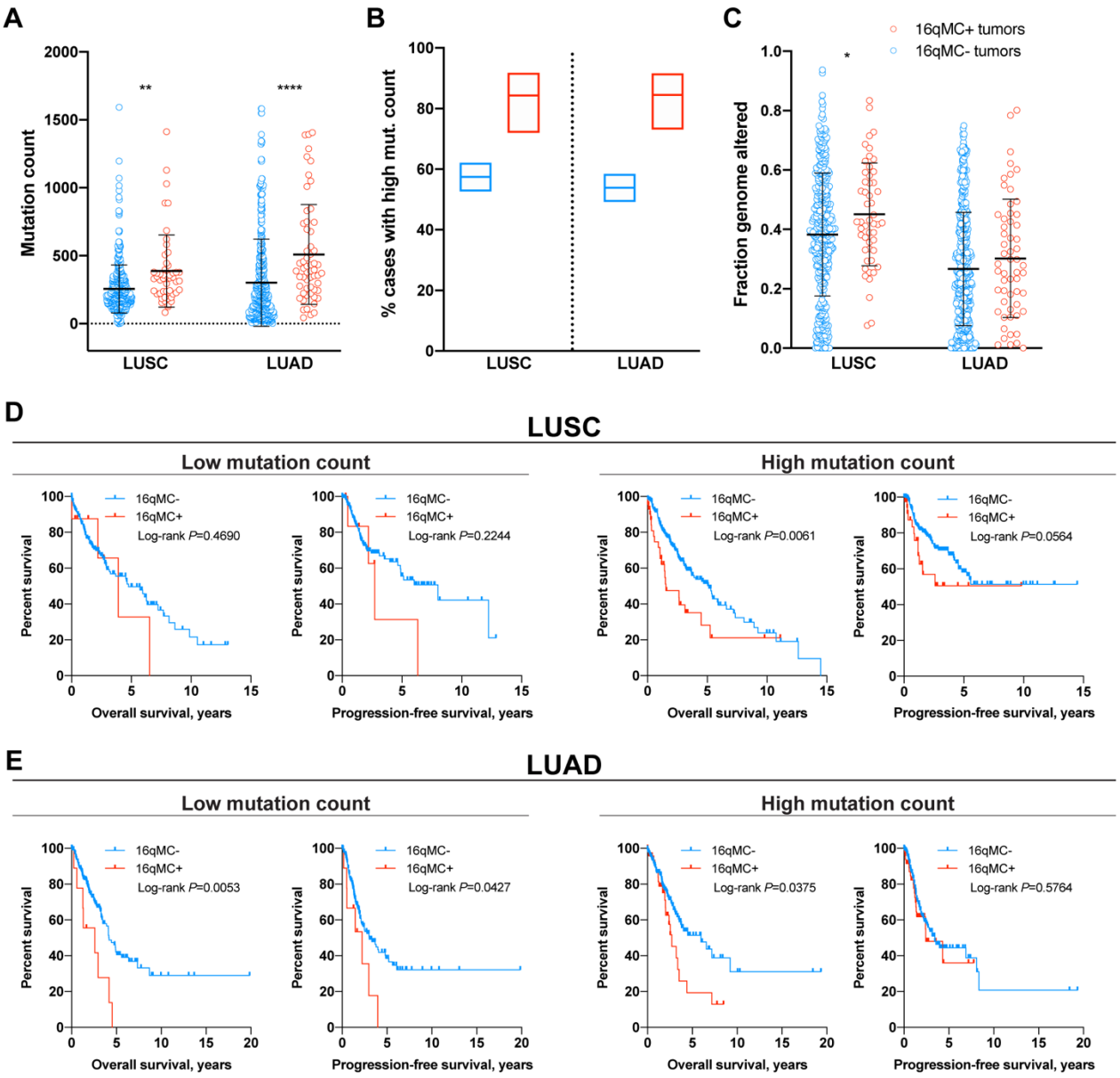


Figure 3.3. 16qMC+ tumors have increased overall mutational burden. (A) Total mutation count for TCGA LUAD and LUSC cohorts. Total mutation count defined as total detected number of non-synonymous mutations. ***P*<0.01, *****P*<0.0001 by ordinary one-way ANOVA with Sidak's multiple comparisons test. Bars show mean+standard deviation. (B) Percentage (with 95% confidence intervals) of tumors with high mutation count (defined as >192 total mutations) among 16qMC- and 16qMC+ patients. Confidence intervals calculated by the Wilson/Brown method. (C) Total fraction of genome altered (FGA), calculated as the percentage of the genome with copy number gains and/or losses, between 16qMC- and 16qMC+ tumors. **P*<0.05 by ordinary one-way ANOVA. Bars show mean+standard deviation. (D-F) KM curves for OS and PFS of 16qMC- and 16qMC+ TCGA LUSC patients (D) and TCGA LUAD patients (E) with either high (>192) or low (<192) mutation counts. Median OS for 16qMC- vs. 16qMC+ LUSC patients: 4.8 vs. 3.9 years (low

mut. count); 5.0 vs. 1.5 years (high mut. count). Median PFS for 16qMC- vs. 16qMC+ LUSC patients: 6.0 vs. 2.7 years (low mut. count). Median OS for 16qMC- vs. 16qMC+ LUAD patients: 4.2 vs. 2.6 years (low mut. count); 6.0 vs. 2.7 years (high mut. count). Median PFS for 16qMC- vs. 16qMC+ LUAD patients: 3.0 vs. 2.2 years (low mut. count); 3.1 vs. 2.4 years (high mut. count). *P* values calculated by log-rank test.

clusters of 8 (LUSC) or 9 (LUAD, HCC) genes. The 16q mutation cluster outperformed the random gene sets in differentiating survival for LUSC (OS: P=0.007; PFS: P=0.025), LUAD (OS: P=0.001), and HCC (OS: P=0.006; PFS: P=0.0290).

Gene set enrichment analysis of 16qMC+ tumors

Next, differentially expressed genes between 16qMC+ and 16qMC- tumors were determined from RNA-sequencing data for the LUSC and LUAD TCGA patient cohorts, and subjected to gene set enrichment analysis (GSEA) (109, 115). Several gene sets related to metastasis, recurrence, relapse, prognosis, or survival were significantly associated with 16qMC+ status (false discovery rate <0.05) (Fig. 4A). In both cohorts, among the most positively-enriched gene sets for 16qMC+ patients was "SHEDDEN LUNG CANCER POOR SURVIVAL A6," a gene set predictive of OS in lung adenocarcinoma patients (116) (LUSC normalized enrichment score (NES)=6.81, LUAD NES=13.75; Fig. 3.4B-D). Conversely, "SHEDDEN LUNG CANCER GOOD SURVIVAL A4," a gene set highly expressed in patients with better survival (116), was depleted in 16qMC+ LUSC (NES=-3.46) and LUAD (NES=-6.47) (Fig. 3.4B, C). Also identified were positive enrichment for "WINNEPENNINCKX_MELANOMA_METASTASIS_UP" in 16qMC+ LUSC (NES=4.87; Fig. 4B) and LUAD (NES=8.11), negative enrichment of "CHANDRAN_METASTASIS_DN" in 16qMC+ LUSC (NES=-3.59) and LUAD (NES=-5.48), and positive enrichment of "BIDUS_METASTASIS_UP" in 16qMC+ LUSC (NES=4.86) and LUAD (NES=5.51) (Fig. 3.4B, C). Together, these results show that the 16q mutation cluster identifies patients with a similar high-risk expression profile as previously established prognostic gene sets, and that 16qMC+ tumors are consistent with more advanced disease, increased likelihood of recurrence, and poorer patient outcomes.

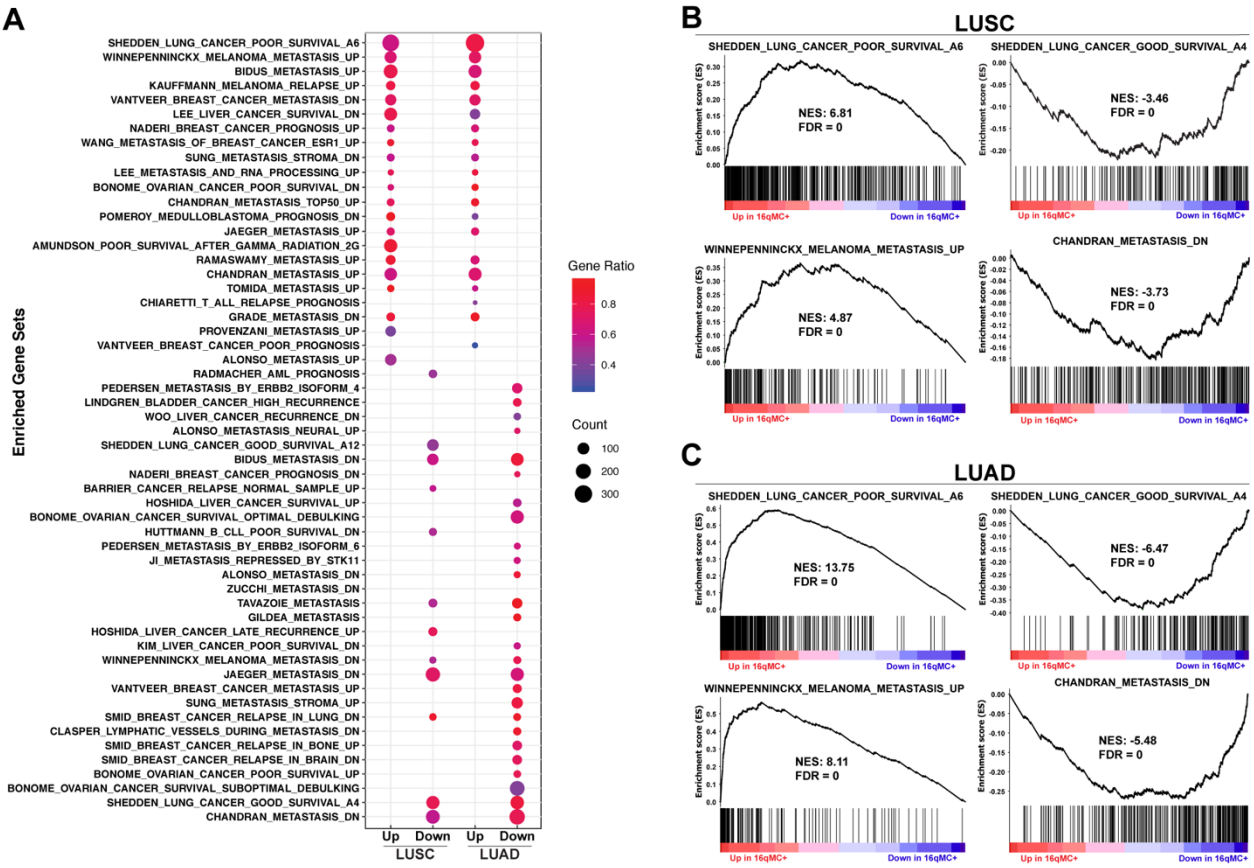


Figure 3.4. Metastasis- and prognosis-related gene sets are enriched in 16qMC+ tumors. (A) Gene sets related to metastasis, recurrence, relapse, prognosis, or survival that were significantly positively- or negatively-enriched (FDR < 0.05) in GSEA of 16qMC+ tumors vs. 16qMC- tumors within the TCGA LUSC and TCGA LUAD cohorts. NES = normalized enrichment score. Dot size indicates the number of core enriched genes, while dot color indicates the proportion of total genes in the given gene set that are enriched in the 16qMC+ population. (B-D) GSEA plots of selected gene sets in LUSC (D) and LUAD (E) cohorts.

3.4 Discussion

Current methods for molecular characterization may not sufficiently capture the full genomic and phenotypic landscape of a tumor population (56, 117), since rare, yet treatment-resistant and invasive cell populations would be missed (1, 3). Our previous work begins to address this problem through the SaGA platform, which was used to isolate specialized, highly invasive leader cells from a larger population of collectively invading packs of NSCLC cells (3).

We identified a novel, leader cell-derived, ten-gene mutation cluster on chromosome 16q. Although 16q deletions have been found in breast, prostate, and other cancers (118-120), 16q alterations in lung cancer have not been widely studied, and co-occurrence of point mutations on 16q have not been reported in any cancer type. In separate cohorts of LUSC and LUAD patients, patients with at least one non-synonymous mutation in any of 8 (LUSC) or 9 (LUAD) of these 16q genes were found to have experienced significantly poorer overall and progression-free survival. These survival differences are maintained in early-stage patients, highlighting the potential clinical utility of this mutation cluster.

The mechanism by which 16qMC+ status differentiates survival requires further study. Although we identified 17 leader-specific mutated genes, only mutations on 16q could differentiate survival, whereas including all 17 genes showed no survival differences in LUSC and LUAD (Log-rank P=0.504 and 0.380, respectively). The majority of 16qMC+ tumors contained only one 16qMC mutation, with no observed effects on expression for the majority of genes (Fig. S3.2). Although this was initially surprising, it is important to consider that the 16qMC was derived from rare and invasive leader cells; therefore the majority of early stage tumors with these mutations may not yet have a detectable effect on genome-wide expression. As the tumor progresses and metastasis occurs, we would predict that the downstream expression consequences of leader specific mutations would become more apparent.

Furthermore, point mutations could impact protein function without affecting gene expression. For example, we previously showed that a leader-specific mutation in ARP3, while not affecting mRNA expression in leader cells, conferred leader cell behavior when introduced into follower cells (104).

Interestingly, in addition to LUSC and LUAD, 16qMC+ tumors also contained significantly elevated mutation counts in TCGA HCC, breast, colorectal, stomach, melanoma, and endometrial cancer cohorts (Fig. S3.4). This indicates that 16qMC+ status could result from a hyper-mutational state, such as microsatellite instability (MSI), in which the 16qMC genes are particularly susceptible to somatic mutations. MSI is observed in lung cancer (121, 122) albeit less frequently than other cancer types such as colorectal. However, our data show that 16qMC+ status correlates with survival even after stratifying patients by high and low mutation counts (Fig. 3.3), and thus additional work is needed to determine whether mutation count is contributing to the poorer survival among 16qMC+ patients.

Using GSEA, genes differentially expressed between 16qMC+ and 16qMC- tumors are enriched in gene sets associated with metastasis and patient prognosis (116, 123-125). These data show that the 16q mutation cluster can stratify high-risk patients through identification of a single point mutation among ten genes. By comparison, other larger-scale, expression-based gene sets, are not as easily translatable to patient care. Targeted sequencing of these ten 16q genes could represent a new strategy for preventing disease recurrence and improving survival in NSCLC, and potentially in HCC as well. Future studies will focus on prospective cohort analysis of early-stage NSCLC patients to better determine how reproducibly 16qMC+ status can differentiate survival. These results are observed across multiple NSCLC cohorts and extend to HCC; however, to better determine the potential clinical utility of 16qMC+ screening, next steps include prospective analyses in additional NSCLC cohorts using primary patient tissue. Additionally, the issue persists that a biopsy could miss rarer

subpopulations of cells (56, 117). Thus, sequencing of circulating tumor DNA through liquid biopsies could provide a more complete picture of the tumor genome (126-128). By using SaGA to identify, isolate, and analyze rare leader cells to discover novel biomarkers, we have laid out an approach that could lead to more effective prognostic strategies.

Supplementary Information

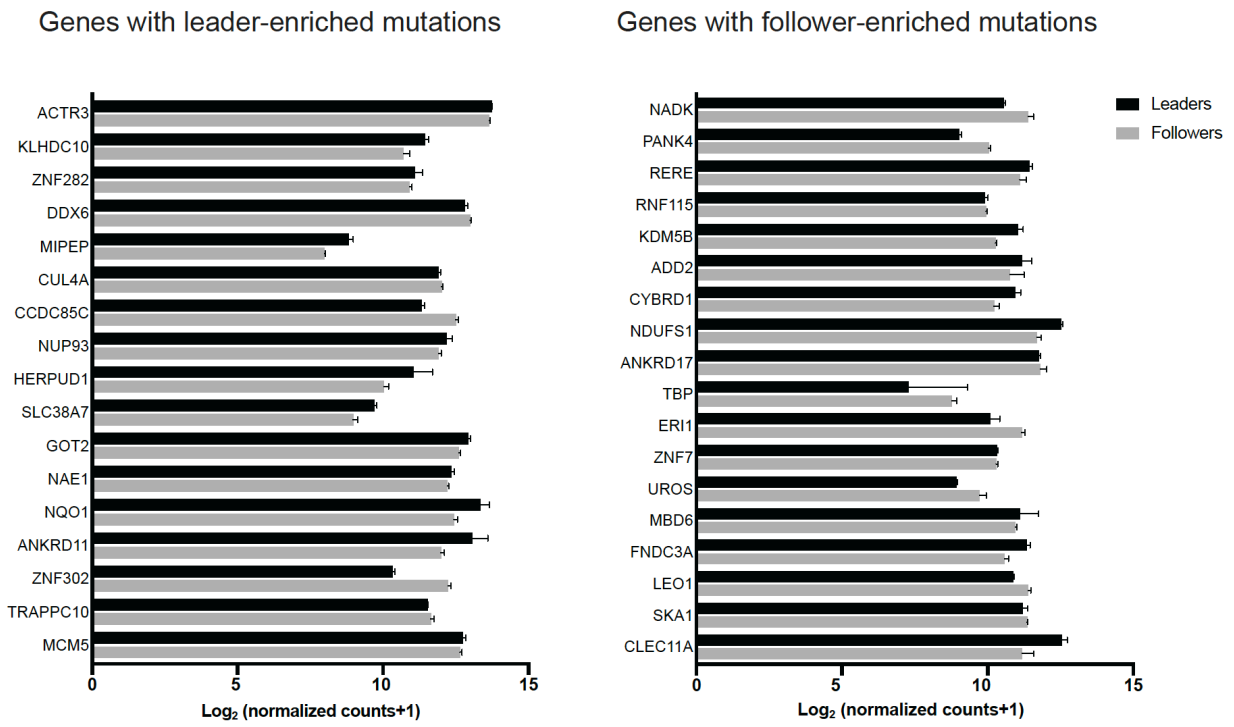


Figure S3.1. Comparison of mRNA levels of genes found to harbor leader- and follower-enriched mutations. mRNA levels shown as log2(normalized counts+1). Expression of all seventeen genes with leader-enriched mutations, and all 18 genes with follower-enriched mutations, was found in both the leader and follower populations.

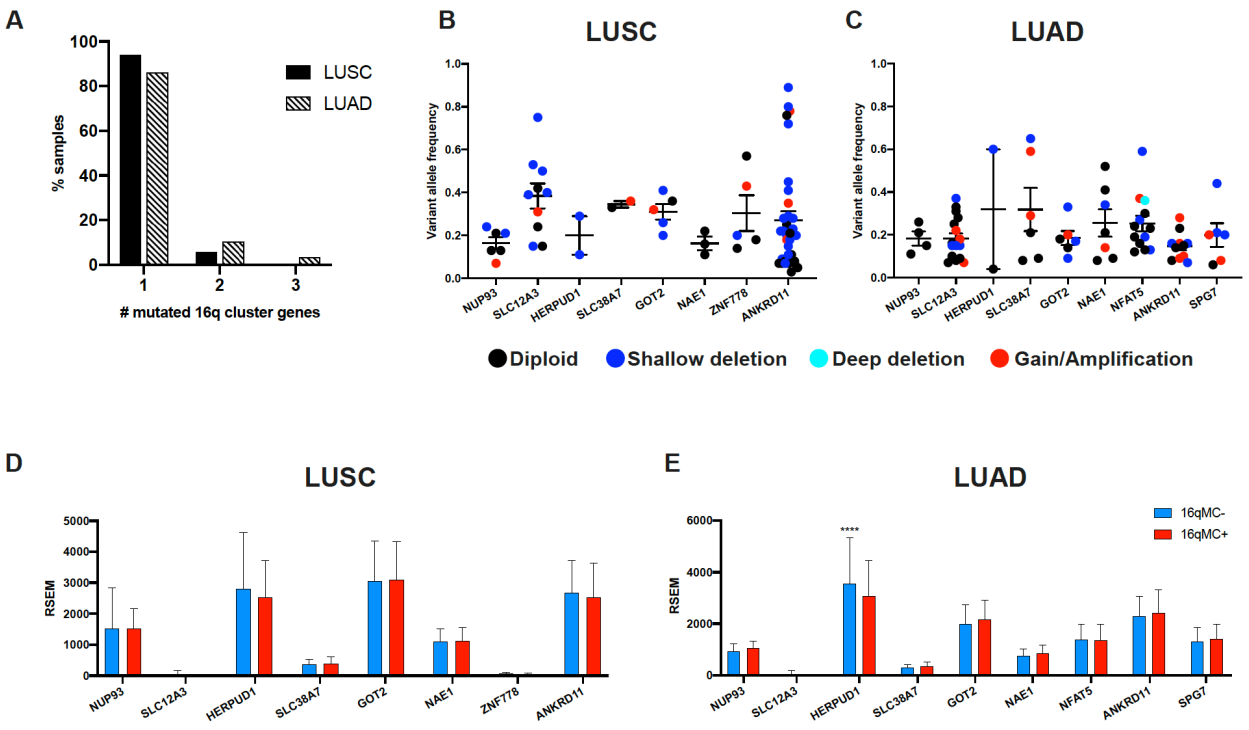


Figure S3.2. Variant allele frequency of 16qMC mutations and mRNA expression of 16qMC genes in TCGA cohorts. (A) Graph of percentages of samples with mutations in 1, 2, or 3 16q cluster genes. LUSC: $n_{1\ mutation}$ = 48 (94.1%), $n_{2\ mutations}$ = 3 (5.9%). LUAD: $n_{1\ mutation}$ = 50 (86.2%), $n_{2\ mutations}$ = 6 (10.3%), $n_{3\ mutations}$ = 2 (3.4%). (B-D) Variant allele frequency (VAF) of each identified 16q cluster mutation in patients from TCGA LUSC (B) and LUAD (C) cohorts. VAF defined as [# variant reads]/[#total reads] at a given locus. Bars show mean + standard deviation. Colors indicate copy-number (diploid, shallow deletion, deep deletion, low-level gain, or amplification) at each mutation locus as reported by cBioPortal. (D-E) mRNA expression of 16qMC genes in 16qMC- and 16qMC+ tumors among LUSC (D) and LUAD (E) TCGA cohorts. ****P<0.0001 by one-way ANOVA with Sidak's multiple comparisons test.

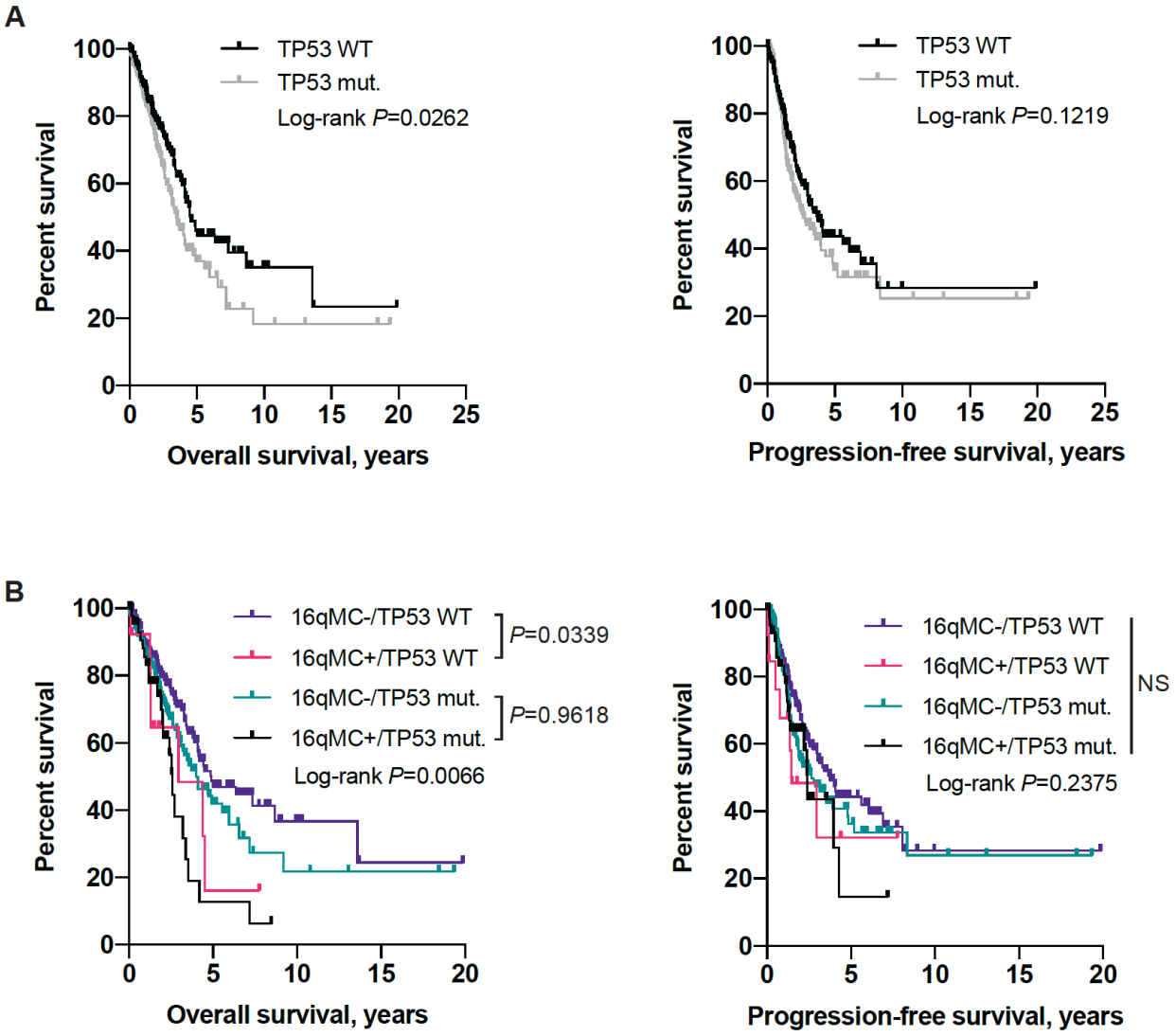


Figure S3.3. Survival analysis by 16qMC+/- and TP53 mutation status in LUAD TCGA patients. (A) KM curves for OS and PFS of patients with TP53 mutant or TP53 wild-type (WT) tumors. Median OS for TP53 mut. vs. TP53 WT: 3.5 vs. 4.5 years. Median PFS for TP53 mut. vs. TP53 WT: 2.6 vs. 3.7 years. P values calculated by log-rank test. (B) KM curves for OS and PFS of patients stratified by both 16qMC+/- and TP53 mutant/WT status. Median OS for each group (years): 16qMC-/TP53 WT: 4.9; 16qMC+/TP53 WT: 3.0; 16qMC-/TP53 mut.: 4.0; 16qMC+/TP53 mut.: 2.6. Median PFS for each group (years): 16qMC-/TP53 WT: 3.8; 16qMC+/TP53 WT: 1.4; 16qMC-/TP53 mut.: 2.6; 16qMC+/TP53 mut.: 2.4. P values calculated by log-rank test with Tukey-Kramer adjustment for multiple comparisons. NS = not significant.

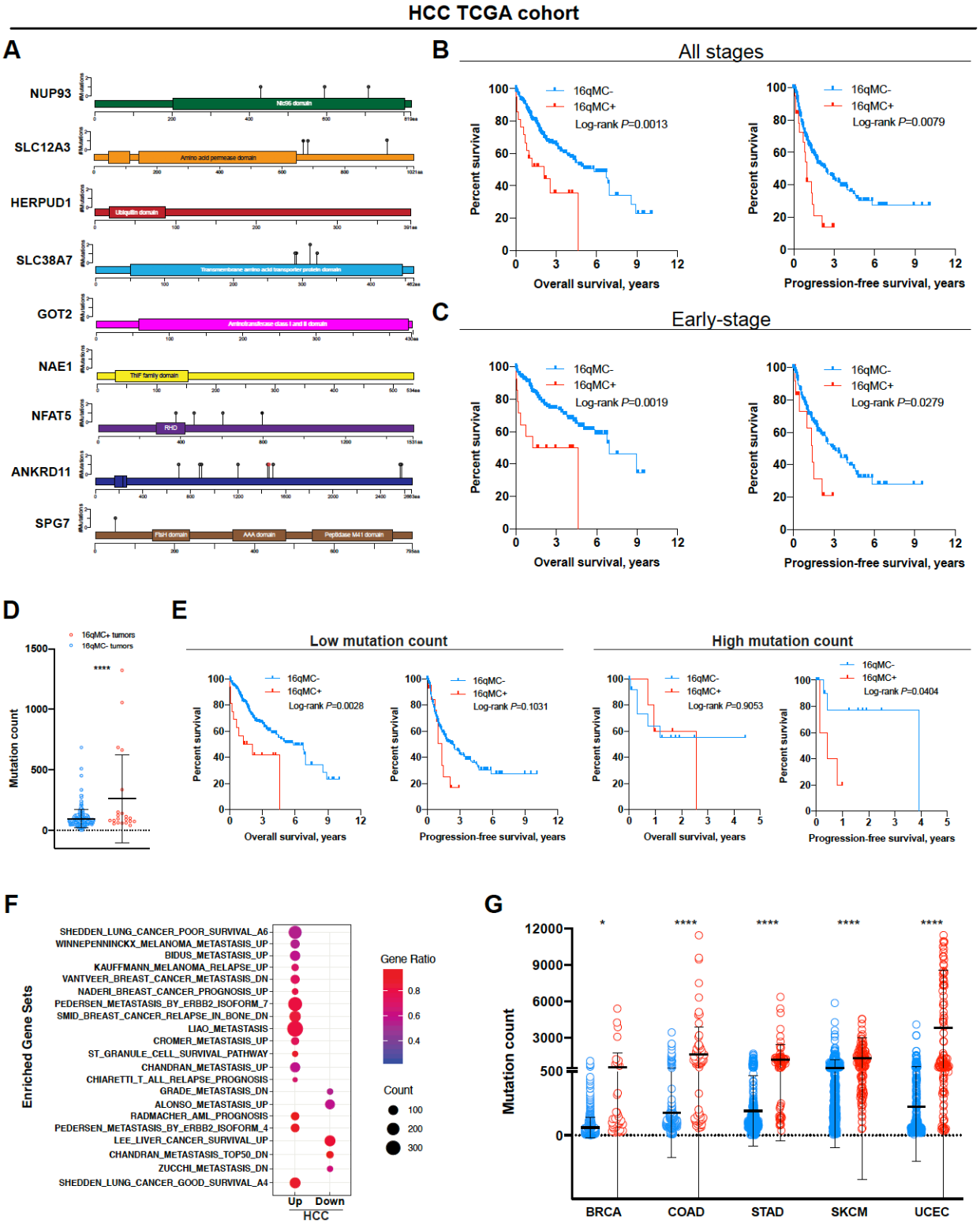

HCC TCGA cohort
A
NUP93
SLC12A3
HERPUD1
SLC38A7
GOT2
NAE1
NFAT5
ANKRD11
SPG7
B
All stages
16qMC-
16qMC+
Log-rank P=0.0013
Log-rank P=0.0079
Percent survival
Overall survival, years
Progression-free survival, years
C
Early-stage
Log-rank P=0.0019
Log-rank P=0.0279
D
16qMC+ tumors
16qMC- tumors
Mutation count
E
Low mutation count
High mutation count
Log-rank P=0.0028
Log-rank P=0.1031
Log-rank P=0.9053
Log-rank P=0.0404
F
Enriched Gene Sets
SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6
WINNEPENNINCKX_MELANOMA_METASTASIS_UP
BIDUS_METASTASIS_UP
KAUFFMANN_MELANOMA_RELAPSE_UP
VANTVEER_BREAST_CANCER_METASTASIS_DN
NADERI_BREAST_CANCER_PROGNOSIS_UP
PEDERSEN_METASTASIS_BY_ERBB2_ISOFORM_7
SMID_BREAST_CANCER_RELAPSE_IN_BONE_DN
LIAO_METASTASIS
CROMER_METASTASIS_UP
ST_GRANULE_CELL_SURVIVAL_PATHWAY
CHANDRAN_METASTASIS_UP
CHIARETTI_T_ALL_RELAPSE_PROGNOSIS
GRADE_METASTASIS_DN
ALONSO_METASTASIS_UP
RADMACHER_AML_PROGNOSIS
PEDERSEN_METASTASIS_BY_ERBB2_ISOFORM_4
LEE_LIVER_CANCER_SURVIVAL_UP
CHANDRAN_METASTASIS_TOP50_DN
ZUCCHI_METASTASIS_DN
SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A4
Up
Down
HCC
Gene Ratio
Count
G
Mutation count
BRCA
COAD
STAD
SKCM
UCEC

Figure S4. 16qMC+ status is associated with poorer survival in HCC and increased mutational burden in multiple cancer types. (A) Lollipop plots illustrating locations of point mutations in 16q cluster genes in TCGA HCC patients. Black dots depict truncations; gray dots depict missense mutations; red outlines depict driver mutations indicated by OncoKB and/or Cancer Hotspots. (B) KM curves for OS and PFS of 16qMC+ and 16qMC- TCGA HCC patients. Median OS: 5.1 years (16qMC-) vs. 2.1 years (16qMC+); median PFS: 3.4 years (16qMC-) vs. 0.4 years (16qMC+). (C) KM curves for OS and PFS of 16qMC+ and 16qMC- stage I and II TCGA HCC patients. Median OS: 6.9 years (16qMC-) vs. 2.9 years (16qMC+); median PFS: 3.0 years (16qMC-) vs. 1.4 years (16qMC+). (D) Quantification of total mutation count for TCGA HCC cohort. Total mutation count defined as total detected number of non-synonymous mutations. ****P<0.0001 by two-tailed, unpaired Student's t-test. Bars show mean+standard deviation. (E) KM curves for OS and PFS of 16qMC- and 16qMC+ TCGA HCC patients with either high (>192) or low (<192) mutation counts. Median OS for 16qMC- vs. 16qMC+ HCC TCGA patients: 5.8 vs. 1.7 years (low mut. count); N/A vs. 2.6 years (high mut. count). Median PFS for 16qMC- vs. 16qMC+ HCC TCGA patients: 1.8 vs. 1.3 years (low mut. count); 3.9 vs. 0.4 years (high mut. count). P values calculated by log-rank test. (F) Gene sets related to metastasis, recurrence, relapse, prognosis, or survival that were significantly positively- or negatively-enriched (FDR < 0.05) in GSEA of 16qMC+ tumors vs. 16qMC- tumors within the TCGA HCC cohort. NES = normalized enrichment score. Dot size indicates the number of core enriched genes, while dot color indicates the proportion of total genes in the given gene set that are enriched in the 16qMC+ population. (G) Quantification of total mutation counts for TCGA breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD), skin cutaneous melanoma (SKCM), and uterine corpus endometrial carcinoma (UCEC) cohorts. *P<0.05, ****P<0.0001 by one-way ANOVA with Sidak's multiple comparisons test. Bars show mean+standard deviation.

Table S3.1. Leader- and follower-enriched gene mutations identified from H1299 cell line

Leader-enriched				Follower-enriched			
	Mutation locus	VAF (Leader)	VAF (Follower)		Mutation locus	VAF (Leader)	VAF (Follower)
ACTR3	chr2:114699797	23.9%	0.1%	NADK	chr1:1686040	0.0%	64.7%
KLHDC10	chr7:129756297	23.5%	0.0%	PANK4	chr1:2441358	0.0%	49.9%
ZNF282	chr7:148909534	18.8%	0.0%	RERE	chr1:8416225	0.0%	65.8%
DDX6	chr11:118650373	29.3%	0.0%	RNF115	chr1:145686997	0.3%	22.0%
MIPEP	chr13:24413837	37.2%	1.0%	KDM5B	chr1:202715414	0.7%	28.1%
CUL4A	chr13:113897295	21.9%	0.4%	ADD2	chr2:70890775	0.0%	29.0%
CCDC85C	chr14:100069569	24.5%	0.3%	CYBRD1	chr2:172379188	0.0%	27.4%
NUP93	chr16:56868312	57.8%	0.3%	NDUFS1	chr2:207012514	0.3%	22.8%
HERPUD1	chr16:56969148	57.9%	0.0%	ANKRD17	chr4:73956729	0.0%	52.5%
SLC38A7	chr16:58713798	53.9%	0.0%	TBP	chr6:170871308	0.4%	21.7%
GOT2	chr16:58768129	57.9%	0.7%	ERI1	chr8:8887402	0.0%	29.6%
NAE1	chr16:66852492	59.6%	0.1%	ZNF7	chr8:146068313	0.0%	20.8%
NQO1	chr16:69745145	56.8%	0.1%	UROS	chr10:127486699	0.0%	25.0%
ANKRD11	chr16:89347145	16.7%	0.6%	MBD6	chr12:57922979	0.0%	17.5%
ZNF302	chr19:35175335	34.6%	0.5%	FNDC3A	chr13:49776097	0.0%	20.6%
TRAPPC10	chr21:45504072	19.9%	0.0%	LEO1	chr15:52258309	0.0%	31.5%
MCM5	chr22:35809920	27.2%	0.1%	SKA1	chr18:47902232	0.0%	21.6%
				CLEC11A	chr19:51228679	0.0%	22.7%

Table S3.2. Chromosome 16q mutations enriched in H1299 leader cells

	Locus (GRCh38)	Nucleotide shift	AA shift	VAF (Leaders)	VAF (Followers)	P value
NUP93	chr16:56834400	G:A	M565I	57.8%	0.29%	0.0015
HERPUD1	chr16:56935236	G:A	R50H	57.9%	0%	0.0142
SLC38A7	chr16:58679894	G:A	T78I	53.9%	0%	0.0033
GOT2	chr16:58734225	C:A	A2S	57.9%	0.70%	0.0001
NAE1	chr16:66818589	T:C	K187R	59.6%	0.13%	0.0003
NQO1	chr16:69711242	G:A	P187S	56.8%	0.06%	0.0002
ANKRD11	chr16:89280737	G:C	D1935E	16.7%	0.06%	0.0130

Table S3.3. Association of common NSCLC driver mutations with 16q mutation cluster

	LUAD			LUSC		
	16qMC+ n=58	**16qMC-** n=447	**P value**	**16qMC+** n=51	**16qMC-** n=427	**P value**
TP53	45 (77.6)	217 (48.6)	**<0.0001**	43 (84.3)	356 (83.4)	>0.99
KRAS	20 (35.5)	135 (30.2)	0.55	1 (2.0)	6 (1.4)	0.55
LKB1	10 (17.2)	66 (14.8)	0.56	1 (2.0)	4 (0.9)	0.43
EGFR	5 (8.6)	60 (13.4)	0.41	0	16 (3.8)	0.40
BRAF	3 (5.1)	37 (8.3)	0.60	2 (3.9)	12 (2.8)	0.65
PDGFRA	6 (10.3)	33 (7.4)	0.43	2 (3.9)	19 (4.5)	>0.99
ALK	8 (13.8)	26 (5.8)	**0.04**	3 (5.9)	15 (3.5)	0.43
ROS1	5 (8.6)	25 (5.6)	0.37	7 (13.7)	29 (6.8)	0.09
PIK3CA	4 (6.9)	23 (5.2)	0.54	12 (23.5)	42 (9.8)	**0.008**
CDKN2A	2 (3.5)	19 (4.3)	>0.99	10 (19.6)	62 (14.5)	0.41
MET	1 (1.7)	20 (4.5)	0.49	2 (3.9)	6 (1.4)	0.21
DDR2	2 (3.5)	16 (3.6)	>0.99	3 (5.9)	12 (2.8)	0.21
RET	4 (6.9)	15 (3.4)	0.26	4 (7.8)	13 (3.0)	0.10
NFE2L2	2 (3.5)	12 (2.7)	0.67	7 (13.7)	66 (15.5)	0.84

Table S3.4: Tumor subtypes and treatment data for TCGA LUSC and LUAD cohorts

			LUSC			LUAD		
Covariate	**Statistic**	**Group**	**16q cluster status**		**P value[d]**	**16q cluster status**		**P value**
			16qMC+ (N=51)	16qMC- (N=424)		16qMC+ (N=58)	16qMC- (N=443)	
Tumor subtype[e]	N (%)	Lung Squamous Cell Carcinoma (NOS)[f]	48 (25.49)	410 (96.7)	0.461			
		Lung Basaloid Squamous Cell Carcinoma	2 (3.9)	11 (2.6)				
		Lung Papillary Squamous Cell Carcinoma	1 (2.0)	5 (1.2)				
		Lung Small Cell Squamous Cell Carcinoma	0	1 (0.2)				
		Lung Adenocarcinoma (NOS)				43 (74.1)	269 (60.2)	0.057
		Lung Adenocarcinoma, Mixed Subtype				11 (19.0)	96 (21.5)	
		Lung Acinar Adenocarcinoma					18 (4.0)	
		Lung Bronchioloalveolar Carcinoma, Non-Mucinous				1 (1.7)	18 (4.0)	
		Mucinous (Colloid) Carcinoma					10 (2.2)	
		Lung Bronchioloalveolar Carcinoma, Mucinous					5 (1.1)	
		Lung Micropapillary Adenocarcinoma				1 (1.7)	2 (0.4)	
		Lung Papillary Adenocarcinoma				1 (1.7)	20 (4.5)	
		Lung Solid Pattern Predominant Adenocarcinoma				1 (1.7)	4 (0.9)	
		Lung Clear Cell Adenocarcinoma					2 (0.4)	
		Lung Mucinous Adenocarcinoma					2 (0.4)	
		Lung Signet Ring Adenocarcinoma					1 (0.2)	
Neoadjuvant therapy	N (%)	No	51 (100.0)	420 (98.4)	>0.99	58 (100.0)	444 (99.3)	>0.99
		Yes	0	5 (1.2)		0	3 (0.7)	
Radiation therapy	N (%)	No	31 (60.8)	331 (77.5)	0.127	37 (63.8)	360 (80.5)	**0.024**
		Yes	8 (15.7)	44 (10.3)		12 (20.7)	49 (11.0)	

[d] P-values calculated by chi-square or Fisher's exact test.
[e] LUSC tumor subtypes 2-4, and LUAD tumor subtypes 3-10, were combined for statistical analysis.
[f] NOS: not otherwise specified

Table S3.5. Cox regression analysis for early-stage LUSC TCGA patients

	Univariable analysis				Multivariable analysis			
	Overall survival		Progression-free survival		Overall survival		Progression-free survival	
Covariate	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
16q cluster status (16qMC+ vs. -)	2.08 (1.27-3.24)	**0.002**	1.75 (0.86-3.18)	0.091	1.94 (1.21-3.12)	**0.006**	1.71 (0.85-3.42)	0.131
Gender (Female vs. Male)	0.90 (0.61-1.29)	0.577	0.98 (0.61-1.51)	0.917				
Smoking history (Smoker vs. non)	0.66 (0.27-2.39)	0.455	0.34 (0.15-1.03)	**0.027**			0.36 (0.13-0.99)	**0.048**
Mutation count (High vs. low)	0.96 (0.69-1.33)	0.787	0.88 (0.59-1.32)	0.526				
Age at diagnosis	1.02 (1.00-1.04)	0.057	1.00 (0.98-1.03)	0.773	1.02 (1.00-1.04)	0.068		

Table S3.6. Cox regression analysis for early-stage LUAD TCGA patients

	Univariable analysis				Multivariable analysis[a]	
	Overall survival		Progression-free survival		Overall survival	
Covariate	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
16q cluster status (16qMC+ vs. -)	2.06 (1.26-3.23)	**0.003**	1.28 (0.77-2.01)	0.315	2.02 (1.25-3.27)	**0.004**
Gender (Female vs. Male)	0.95 (0.66-1.36)	0.769	0.96 (0.70-1.32)	0.787		
Smoking history (Smoker vs. non)	0.93 (0.57-1.60)	0.782	1.01 (0.65-1.66)	0.956		
Mutation count (High vs. low)	1.13 (0.79-1.63)	0.507	1.11 (0.81-1.53)	0.511		
Age at diagnosis	1.02 (1.00-1.04)	0.062	1.00 (0.99-1.02)	0.627	1.02 (1.00-1.04)	0.112

[a]Multivariable analysis for early stage patients for PFS outcome was not reported, as none of the predictors passed the model selection criteria.

Table S3.7: Patient characteristics for HCC TCGA cohort

Covariate	Statistic	Group	16q cluster status		P value[bc]
			16qMC+ (N=21)	16qMC- (N=339)	
Gender	N (%)	Female	5 (23.81)	114 (33.63)	0.353
	N (%)	Male	16 (76.19)	225 (66.37)	
Pathologic stage	N (%)	Stage I & II	14 (70)	237 (73.83)	0.706
	N (%)	Stage III & IV	6 (30)	84 (26.17)	
Mutation count[a]	N (%)	High mut. count	5 (23.81)	13 (3.93)	**<0.001**
	N (%)	Low mut. count	16 (76.19)	318 (96.07)	
Age at diagnosis	Median (min-max)		64 (24-85)	61 (16-90)	0.295
Histologic grade	N (%)	Low Grade (I and II)	14 (66.67)	209 (62.57)	0.707
	N (%)	High Grade (III and IV)	7 (33.33)	125 (37.43)	
Histologic subtype	N (%)	HBV or Combination	5 (26.32)	91 (28.26)	0.442
	N (%)	HCV or Combination	4 (21.05)	43 (13.35)	
	N (%)	NBNC	3 (15.79)	86 (26.71)	
	N (%)	HBV or HCV	1 (5.26)	6 (1.86)	
	N (%)	Other	6 (31.58)	96 (29.81)	

[a]High/low mutation count was defined by a cutoff of 192 mutations.
[b]P-values calculated by ANOVA for numerical covariates and chi-square or Fisher's exact test for categorical covariates.
[c]P-values calculated by Kruskal-Wallis test for numerical covariates

Table S3.8: Cox regression analysis for all-stage HCC TCGA patients

	Univariable analysis				Multivariable analysis			
	Overall survival		Progression-free survival		Overall survival		Progression-free survival	
Covariate	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
16q cluster status (16qMC+ vs. -)	3.21 (1.45-6.28)	**0.002**	2.34 (1.07-4.46)	**0.019**	3.28 (1.10-9.78)	**0.033**	5.27 (1.57-17.67)	**0.007**
Gender (Female vs. Male)	1.61 (0.99-2.60)	0.053	1.19 (0.79-1.77)	0.392	0.79 (0.34-1.83)	0.588		
Mutation count (High vs. low)	2.44 (0.90-5.39)	**0.049**	1.17 (0.39-2.68)	0.749				
Age at diagnosis	1.03 (1.01-1.05)	**0.019**	1.00 (0.98-1.01)	0.637	1.03 (1.00-1.06)	**0.024**		
Histologic grade (I/II vs. III/IV)	1.05 (0.64-1.71)	0.836	1.05 (0.70-1.54)	0.816				

Table S3.9: Cox regression analysis for early-stage HCC TCGA patients

	Univariable analysis				Multivariable analysis			
	Overall survival		Progression-free survival		Overall survival		Progression-free survival	
Covariate	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
16q cluster status (16qMC+ vs. -)	2.50 (1.41-4.46)	**0.002**	2.07 (1.20-3.59)	**0.009**	3.35 (0.70-15.97)	0.129	1.97 (1.11-3.49)	**0.020**
Gender (Female vs. Male)	1.25 (0.87-1.78)	0.23	1.08 (0.79-1.48)	0.640	1.15 (0.51-2.60)	0.736		
Pathologic stage (I/II vs. III/IV)	0.42 (0.29-0.61)	**<0.001**	0.45 (0.32-0.62)	**<0.001**	0.53 (0.26-1.08)	0.079	0.45 (0.33-0.63)	**<0.001**
Mutation count (High vs. low)	2.12 (1.03-4.36)	**0.042**	1.18 (0.55-2.52)	0.668				
Age at diagnosis	1.01 (1.00-1.03)	0.054	1.00 (0.98-1.01)	0.419	1.01 (0.99-1.03)	0.405		
Histologic grade (I/II vs. III/IV)	1.10 (0.77-1.59)	0.598	1.16 (0.85-1.58)	0.337				

Chapter 4: Single-cell RNA-sequencing of lung cancer leader and follower cells reveals distinct mutational profiles and cancer stem cell-like gene expression patterns

Brian Pedro, Manali Rupji, Bhakti Dwivedi, Janna K. Mouw, Jessica Konen, Jeanne Kowalski, Paula M. Vertino, Adam I. Marcus

4.1 Introduction

Metastatic disease accounts for the vast majority of cancer-related deaths, and increasing evidence suggests that carcinomas – tumors of epithelial origin – rely upon collective invasion to successfully metastasize. These collective packs have been demonstrated through *in vitro* experiments, *in vivo* studies, as well as in human tissue samples, and they can contain specialized, phenotypically and genetically distinct leader and follower cells that cooperate to carry out invasion (3, 11, 13, 14). Through the SaGA platform, we have previously demonstrated that leader cells isolated from the H1299 non-small cell lung cancer (NSCLC) cell line are highly invasive and are able to pioneer invasive chains even when present as a rare population among follower cells (3). Furthermore, H1299 leaders and followers distinct in their epigenetics and gene expression profiles, and, importantly, we have shown that leaders and follower populations contain distinct gene mutations that may contribute to their specialized phenotypes (3, 27, 104, 129). Indeed, this phenomenon is not exclusive to the H1299 cell line, as leader and follower cells have been identified across a variety of cellular processes and cancer types including breast, lung, and colorectal (11, 15, 24, 130, 131).

Previous genetic analyses of SaGA-derived leaders and followers were performed via population-based analyses, including microarrays, bulk RNA-seq, and methylation arrays. However, given the high degree of heterogeneity that exists within a single tumor population, these population-based analyses may fail to capture the variations between leader and follower cells. Recently, single-cell sequencing techniques have provided greater insight into the genetic variation and numerous subpopulations that exist within a single parental tumor population. While these studies have provided vast amounts of information detailing the complex genomic events that underly tumor formation and progression, these techniques are inherently lacking in phenotypic information, as a single cell cannot be further

examined subsequent to sequencing. Thus, when performing single-cell analysis on cells undergoing collective invasion, a process that is dependent on cooperation between phenotypically distinct cells, the ability to correlate each cell's phenotype with its genetic profile is crucial.

In this study, we present a novel approach to single cell analysis, adapting the SaGA platform to select and isolate cells actively undergoing collective invasion and immediately subject them to single-cell RNA-sequencing (scRNA-seq). Previously established leader- and follower-specific gene mutations (104) were used precisely label each single cell, enabling for the first time the combination of phenotypic, gene expression, and mutational analysis of single cells during active collective invasion. These analyses found that leader cells display cancer stem cell-like gene expression and tumor initiating capacity, and that TGFβ signaling may be important for leader-follower crosstalk, providing crucial new insights into the roles of these specialized cells in collective invasion and metastasis.

4.2 Materials and Methods

SaGA platform and single-cell RNA-sequencing (scRNA-seq)

Raw expression count data was filtered for low quality samples and genes using seurat R package (132). Samples were filtered by library size (e.g., <0.4M; n=85 and >4M; n=1), number of detected features (e.g., samples that do not contain >7000 expressed genes; n=5), and percentage of mitochondrial reads (>=20%; n=1). Filtering was also applied to remove lowly expressed genes, i.e., genes that are not expressed in all samples (n=10119), and genes that do not have the highest average reads per million (n=23093). The minimum average expression among the genes selected ranges from 0.40-15000 (in RPM). The filtered data was normalized using scran R package (133) with minimum 20 cluster sizes to identify the rescaling factor. Among the two duplicate genes with same exact gene names (MARCH1 and MARCH2), the copy of the gene with lowest fraction of zeroes and/or high variance was selected. The final log2 transformed normalized data included 23091 genes and 190 single cells. The distribution of the single cell populations included: Control (n = 20), Parental (n=65), Leaders (n= 53) and Followers (n= 52). The twenty control cells were excluded from the downstream analysis. Mitochondrial genes were regressed out to improve clustering analysis.

Heterogeneity across all phenotypes was assessed using NOJAH's GWH analysis pipeline (134). The top 6% most variable genes (i.e. the core gene set) were selected from the genome-wide data including all samples, using a combined variance, median absolute deviation (MAD) and IQR statistic. The heatmap of the core gene set was generated using row scaling, maximum distance and ward.D clustering. Consensus clustering was performed on the same core gene set using Canberra distance, complete clustering with 1000 iterations, 80% sample resampling and 100% gene resampling. Core cell set were identified after removing cells with negative silhouette widths based on the silhouette

plots for the identified clusters. The heatmap of the core cell set with core gene set was output using row based scaling, Canberra distance and ward.D2 clustering. Principle component analysis (PCA) and tSNE plots based on the phenotypes and clustering were created using the scater Bioconductor R package (135).

Western blotting:

Total cellular protein and secreted protein expression were assessed via Western blotting as previously described (63).

Immunofluorescence:

After 24-48 hours of spheroid invasion in Matrigel, gels were fixed with 4% paraformaldehyde for 30 minutes at room temperature, followed by washing with 100mM glycine in PBS. Gels were blocked with immunofluorescence (IF) buffer (130mM NaCl, 7mM Na_2HPO_4, 3.5mM NaH_2PO_4, 0.2% Triton X-100, 0.05% Tween-20) with 5% goat serum for 1 hour at room temperature. Primary and secondary antibodies were diluted in IF buffer with 2.5% goat serum. Gels were incubated with primary antibodies at 4°C overnight, washed with IF buffer, incubated with secondary antibodies for 1 hour at room temperature, washed again with IF buffer, and stored in PBS at room temperature.

Reagents and antibodies:

Primary antibodies for Western blotting: TGF-β1 antibody (Abcam, cat. no. ab92486) was used at 1:1000. GAPDH antibody (Cell Signaling, cat. no. 2118) was used at 1:30,000. Horseradish peroxidase-conjugated secondary antibodies (Jackson ImmunoResearch) were used at 1:10,000 for Western blotting.

For immunofluorescence, JAG1 primary antibody (Cell Signaling, cat. no. 70109) was used at 1:500. Goat anti-Rabbit Alexa Fluor 568 secondary antibody (Thermo Fisher Scientific cat. no. A-11032) was used at 1:200. For actin staining, Alexa-Fluor 488 Phalloidin (Thermo Fisher Scientific cat. no. A12379) was used at 1:40.

Recombinant human TGF-β1 (PeproTech cat. no. 100-21) was used at 10g/mL. SB-505124 (Cayman Chemical Company cat. no. 11793) was used at 1.0uM.

3-D invasion assays, spheroid microscopy and image analysis:

Spheroids were generated as previously described (63) and embedded in 2 mg/mL Matrigel (Corning cat. no. 356237) diluted in complete media. Images were taken at 0, 24, and in some cases 48 hours post-embedding at 4x using an Olympus CKX41 microscope.

For immunofluorescence, spheroids were imaged using a Leica SP8 inverted confocal microscope, 10x objective with 0.75x zoom. Invasive area and spheroid circularity were measured using ImageJ as previously described (3).

***In vivo* experiments:** NOD scid gamma (NSG) mice aged 8 weeks were injected in the right flank with either $1x10^6$ cells suspended in 25% Matrigel in PBS. After approximately 12 weeks, mice were sacrificed and primary tumors, lymph nodes, lungs and blood were collected for immunohistochemical, immunofluorescent and RNA analyses.

Immunohistochemistry: Tissues from NSG mice were fixed with neutral buffered formalin, embedded in paraffin and microtome-sectioned at 5um. Hematoxylin and eosin staining was performed on paraffin sections, followed by imaging using an Olympus CKX41 microscope.

4.3 Results

Isolation and single-cell RNA-sequencing (scRNA-seq) of collectively invading H1299 cells

The SaGA technique (3) was modified for single-cell sequencing (Fig. 4.1). As previously described, spheroids comprised of H1299 NSCLC cells were embedded in Matrigel. Either leader cells or follower cells (initially defined by positioning within invasive chains) were highlighted via 405nm laser-driven photoconversion of Dendra2 and, after matrix degradation, isolated via FACS. Single cells were deposited into 10ul RLT buffer in separate wells of 96-well plates (n=84 leader cells and 80 follower cells). It was also important to be able to compare cells located in the spheroid core to those found within invasive chains; therefore, in a third condition, termed parentals, single cells were randomly sorted via FACS into a 96-well plate (n=84 cells). Single H1792 NSCLC cells grown in 2-dimensional culture used as controls (n=10 cells per plate). After library preparation and next-generation sequencing, samples were filtered by library size, number of expressed features, and proportion of mitochondrial reads, resulting in n=52 leader cells, 53 follower cells, and 65 parental cells. The top 6% most variably expressed genes (i.e. the core gene set; n=1,155 genes) across all single cells were defined by a combined variance, median absolute deviation (MAD) and IQR statistic. Consensus clustering was performed on the core gene set using Canberra distance, complete clustering with 1000 iterations, 80% sample resampling and 100% gene resampling. The core cell set was defined by removing cells with negative silhouette widths based on the silhouette plots for the identified clusters, resulting in n=105 cells across three core clusters (n=56 cells, cluster 1; n=18 cells, cluster 2; n=32 cells, cluster 3) (Fig. 4.2.A, B). A tSNE plot was generated for the core cell set based upon the core gene set (Fig. 4.2.C).

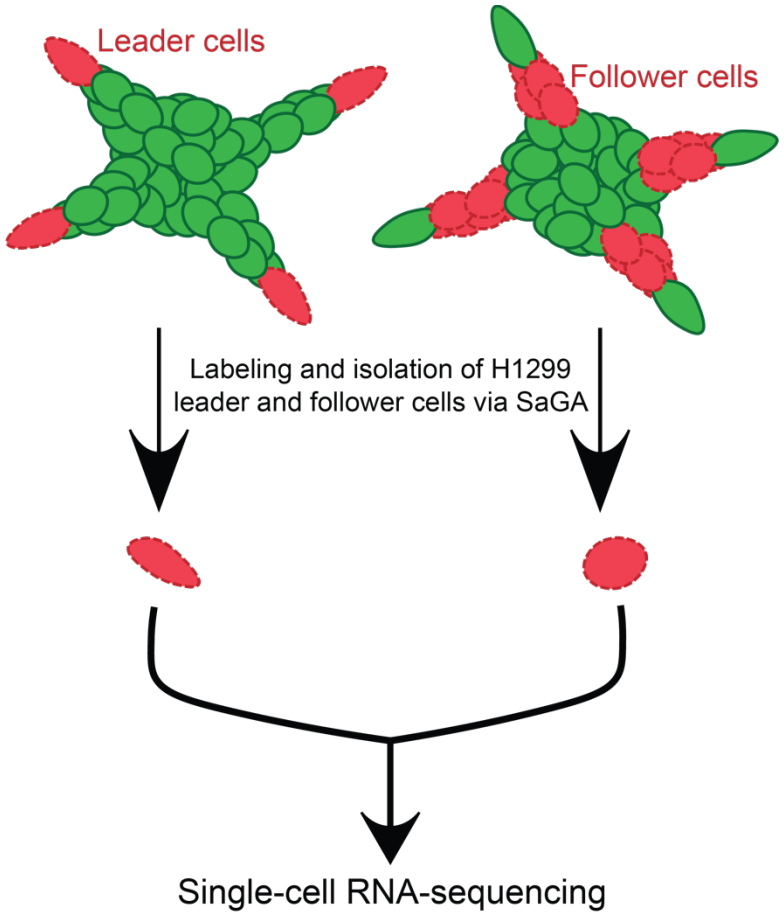


Figure 4.1. Adaptation of the SaGA platform for single-cell RNA-sequencing. Schematic of the SaGA platform (3) adapted for single-cell RNA-sequencing. Spheroids of Dendra2-expressing H1299 NSCLC cells were embedded in Matrigel and allowed to invade for 24 hours. After photoconversion of leader or follower cells and degradation of Matrigel, single red fluorescent cells (or green fluorescent, in the parental condition) were sorted directly into lysis buffer in individual wells of 96-well plates. Plates were frozen at -80C and subsequently processed for single-cell RNA-sequencing.

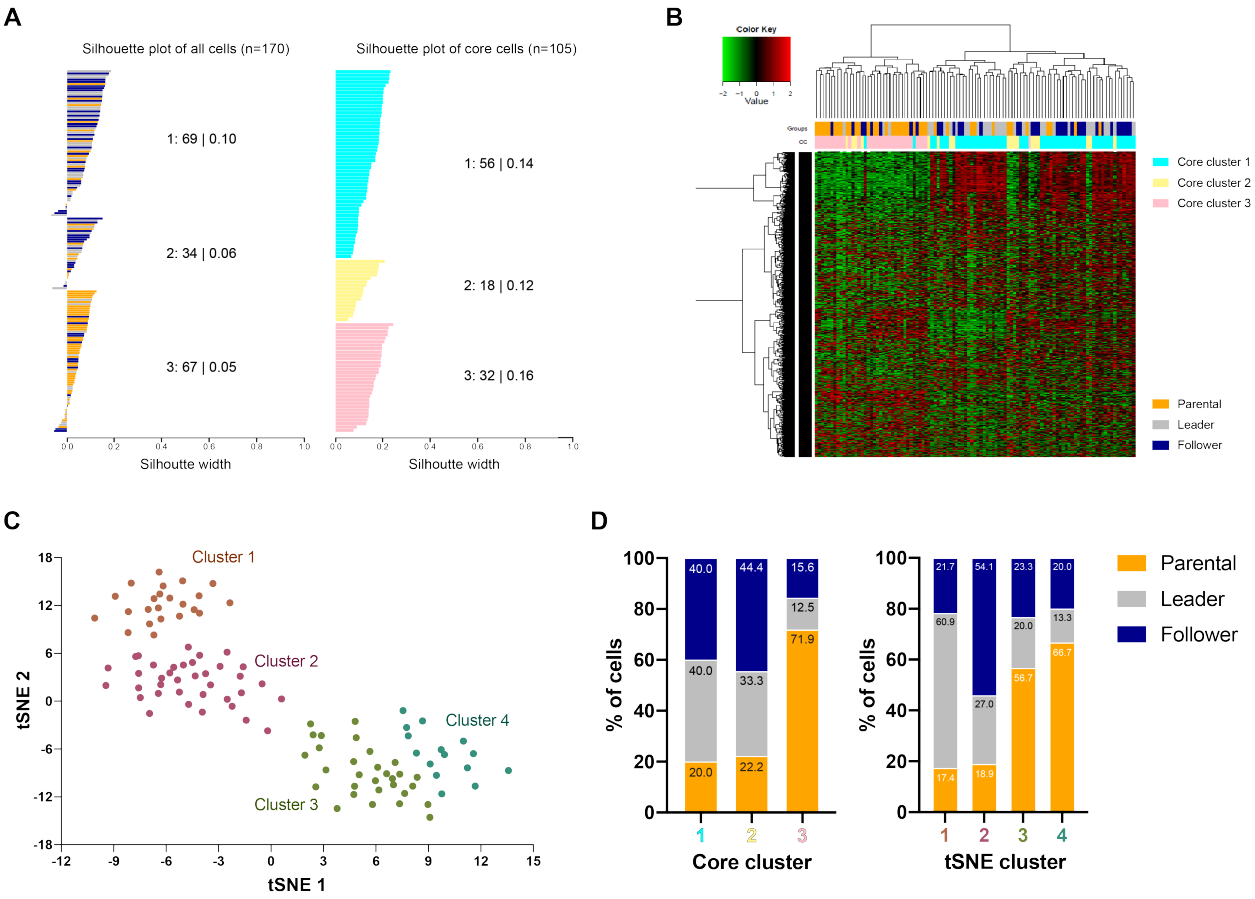


Figure 4.2. Assigned positional phenotypes do not correlate strongly with gene expression profiles. (A) Silhouette plots for all single cells (n=170) and core cells only (n=105) based upon core consensus clustering of the core set of most variably expressed genes (n=1,155). Cells with negative silhouette widths were removed to obtain the core cell set. (B) Heatmap based upon consensus clustering of the 105 core and 1,155 most variably expressed genes. (C) tSNE plot of single H1299 cells based upon expression of the most variably expressed genes. tSNE clusters were determined by positioning of cells within the plot. (D) Composition of each core cluster and each tSNE cluster based upon assigned phenotypes of each single cell prior to isolation.

Given previous work defining the gene expression profiles of leader and follower cells, it was expected that the leader and follower cells, as defined by physical positioning within invasive chains prior to photoconversion, would cluster separately in scRNA-seq analysis, with some parental cells falling into each category. However, the three core clusters of cells contained mixed populations of leaders and followers: Cluster 1 contained 40% each of leader and follower cells, cluster 2 contained 33% leaders and 44% followers, and cluster 3 contained 12.5% leaders and 15.6% followers, with over 70% of cells coming from the parental group (Fig. 4.2.D). Given the lack of correlation between the core clusters and phenotypes, it was determined that a secondary method for defining cell clusters was necessary. The tSNE plot based upon the most variably expressed 1,155 genes was separated into 4 tSNE clusters, which more closely aligned with the assigned phenotypes for each cell: tSNE cluster 1 contained 61% leader cells and 22% follower cells; tSNE cluster 2 contained 27% leader cells and 54% follower cells; and tSNE clusters 3 and 4 were mainly comprised of cells from the parental group (Fig. 4.2.D). However, there was still a number of cells for which the assigned phenotypes did not align fully with the defined clusters, and we therefore sought to determine a more definitive method for labeling each cell.

Labeling of single cells by leader- or follower-mutational profile

One of the major limitations of single-cell sequencing is that the exact analyzed cells cannot be further studied to verify their phenotypes. However, as previously described in (104), H1299 leader and follower cell populations harbor distinct mutational profiles. Therefore, we explored whether these leader- and follower-specific mutations could be detected by scRNA-seq and used to genomically label each cell. We selected mutation loci from the list of leader- and follower-specific genes (104) and verified that they had sufficient coverage in the single-cell sequences. This left 5 leader-specific mutation loci, and 7 follower-specific mutation loci (Fig. 4.3.A). Interestingly, there was 95.9% mutual exclusivity between the mutational profiles on the single-cell level – that is, of the 171 cells with at

least one leader- or follower-specific mutation, only 7 cells (4.1%) had both (Fig. 4.3.A). In each of these 7 cases, the cell had a NADK mutation in addition to one or more leader-specific mutations. Given this finding, we next investigated whether the leader and follower mutation profiles also correlated with expression of leader and follower biomarkers, which have also been previously determined through bulk RNA-sequencing of the populations. We first re-labeled each cell in the most variable gene expression tSNE plot by its mutational profile – leader, follower, or unknown – and found that in tSNE clusters 1 and 4, 100% and 93% of cells, respectively, had a leader mutation profile (4.3.B, C). Among clusters 2 and 3, 89% and 93% of cells, respectively, had a follower mutation profile (Fig. 4.3.C). Thus, classifying each cell by its mutational profile resulted in much clearer correlation with overall gene expression than the positional labels assigned prior to cell collection.

Next, the correlation between mutation profile and specific leader-follower gene expression markers was measured. Among the most differentially expressed leader cell markers from our previous bulk RNA-sequencing was MYO10, an unconventional myosin involved in filopodial elongation. Conversely, IL13RA2, a cell surface protein commonly characterized as a decoy receptor for IL-13, was among the clearest follower cell expression markers from previous analyses (27). Thus, we analyzed the expression of these two genes within each single cell comprising the four tSNE clusters. Clusters 1 and 4 had significantly increased expression of MYO10 compared to clusters 2 and 3, while the converse was true for IL13RA2 (Fig. 4.3.D, E). Thus, it was concluded that the leader and follower mutation profiles correlated strongly with leader and follower cell gene expression profiles, and thus these mutation profiles could be utilized to label each single cell as either a leader or follower in the single cell analysis.

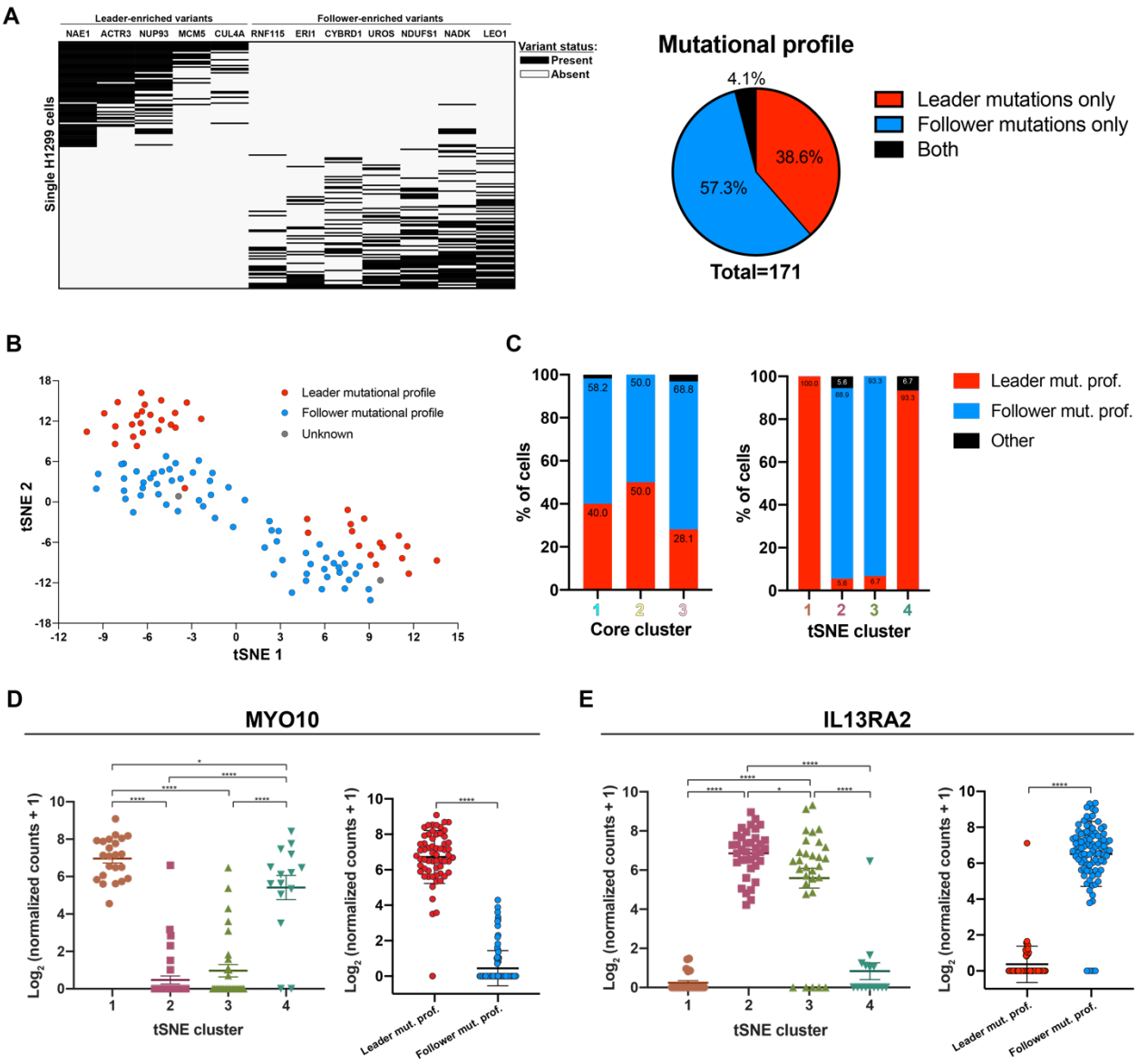


Figure 4.3. Mutational profiles correlate more strongly with gene expression than assigned positional phenotypes. (A) Mutation profile plot for n=171 single cells using previously identified H1299 leader- and follower-specific mutations (Zoeller, Pedro et al. 2019), with quantification of total cells containing exclusively leader mutations, exclusively follower mutations, or both. (B) tSNE plot from 4.XC with each cell labeled by its mutation profile from panel A. Leader mutational profile: ≥1 leader-specific mutation; Follower mutational profile: 0 leader-specific mutations and ≥ 1 follower-specific mutation; Unknown: 0 leader- or follower-specific mutations. (C) Composition of each core cluster and each tSNE cluster based upon mutation profiles. (D-E) Quantification of MYO10 expression (D) and IL13RA2 expression (E) for each tSNE cluster and for cells grouped by mutation profile. **P*<0.05 by one-way ANOVA with Tukey's multiple comparisons test; *****P*<0.0001 by two-tailed, unpaired t-test (for mutation profile groups) or by one-way ANOVA with Tukey's multiple comparisons test (for tSNE clusters).

Leader and follower cells contain cycling and non-cycling populations in 3-D

To determine the genes driving the formation of two separate leader and follower clusters in the gene expression tSNE plot, we performed GSEA for two groups: one comprised of tSNE clusters 1 and 2, and the other comprised of tSNE clusters 3 and 4. Among the mostly highly enriched gene sets in tSNE clusters 1 and 2 were "GO_POSITIVE_REGULATION_OF_CELL_ POPULATION _PROLIFERATION", and "GO_CELL_CYCLE" which included core enrichment of genes including CCNA2/B1/B2, CDK1, and Ki-67, suggesting that tSNE clusters 1 and 2 represented more actively proliferating populations (Fig 4.4.A). By contrast, tSNE clusters 3 and 4 were not enriched for cell cycle or proliferation gene sets; in addition, it is notable that tSNE clusters 3 and 4 contained a higher proportion of cells initially derived from the parental population (Fig. 4.2.D). This was the only population in which cells may have been located in the core of the spheroid, and thus may have been more hypoxic. Indeed, tSNE clusters 3 and 4 showed enrichment of gene sets including "GO_CARBOHYDRATE_TRANSPORT" and "GO_RESPONSE_TO_ENDOPLASMIC_RETICULUM_STRESS" which included core enrichment of HK2, INSR, and DDIT3, genes reported to be hypoxia-inducible (136-138) (Fig. 4.4.B). Furthermore, prolonged hypoxia is known to induce cell cycle arrest (139); thus, tSNE clusters 3 and 4 are consistent with cells that were localized to the core of the spheroid, and were thus more hypoxic and less proliferative. This also suggests that some cells with leader mutation and gene expression profiles are slower to emerge from the core of the spheroid; however, it is possible that these cells would have emerged to lead out invasive chains if invasion were allowed to proceed longer.

Leader cells have a cancer stem cell-like gene expression profile

After labeling each cell by its mutation profile, GSEA was performed on the leader (n=61) and follower (n=95) populations to determine which biological processes may be differentially regulated.

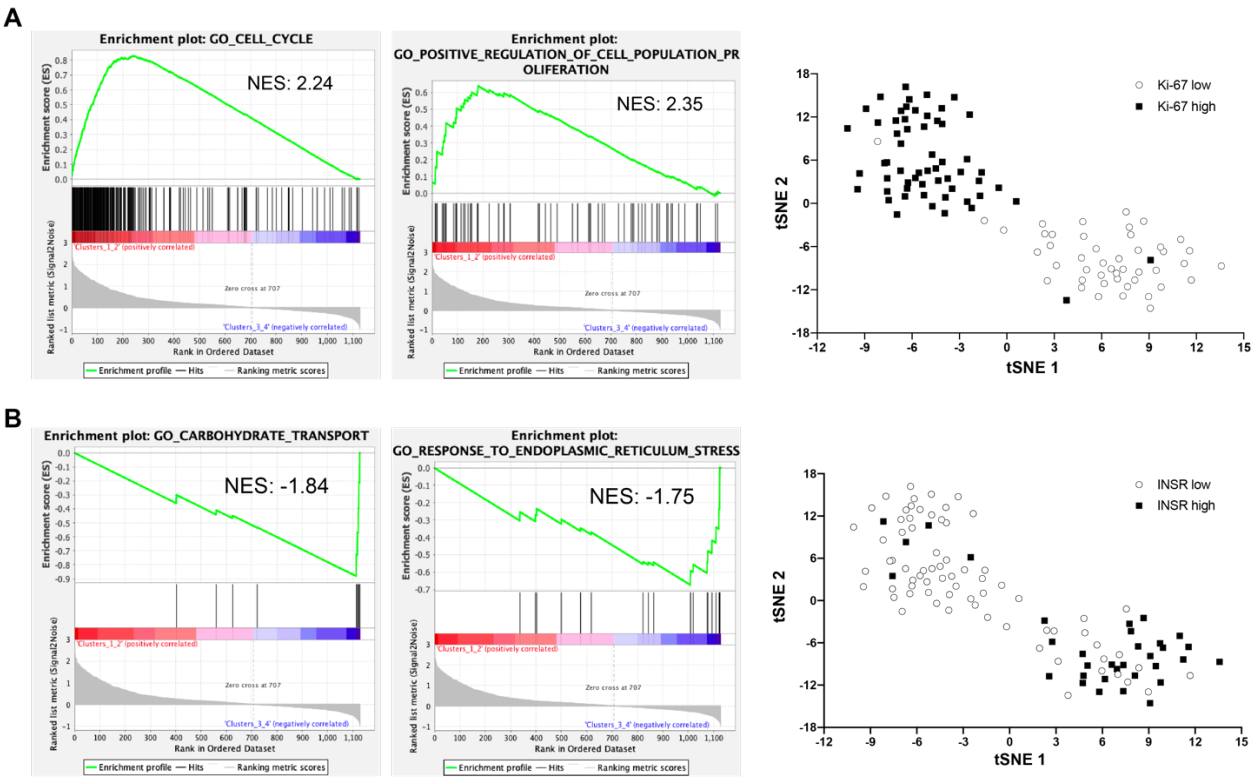


Figure 4.4. Leader and follower cells contain cycling and non-cycling populations in 3-D. (A) GSEA plots for selected gene sets among the topmost enriched in combined tSNE clusters 1 and 2. Shown is tSNE plot from Fig. 4.2C with each cell labeled by high (defined as [log2 (normalized counts+1)]>2) or low (defined as [log2 (normalized counts+1)]<2) expression of Ki-67. (B) GSEA plots for selected gene sets among the topmost enriched in combined tSNE clusters 3 and 4. Shown is tSNE plot from Fig. 4.2C with each cell labeled by high (defined as [log2 (normalized counts+1)]>3.5) or low (defined as [log2 (normalized counts+1)]<3.5) expression of INSR. NES: normalized enrichment score defined by (clusters 1 & 2) vs. (clusters 3 & 4).

Among the top 30 most-enriched gene sets in the leader group were 10 related to regulation of differentiation, development, and morphogenesis – indicative of a more stem cell-like population (Fig. 4.5.A). By contrast, only one gene set related to development was enriched in the follower group, and it indicated positive, rather than negative regulation of development (Fig. 4.5.A). To further probe the possibility that leader cells harbor a more cancer stem cell-like expression profile, we compared expression of 195 stemness-related genes between the two groups. Of those 195 genes, 87 were significantly differentially expressed: 42 were higher leaders, and 45 were higher in followers (Fig. 4.5.B). Notably, among those genes higher in leaders were ALDH1A3, a common lung cancer stem cell (CSC) marker, and JAG1 and NOTCH1, members of the Notch pathway reported to promote CSC survival and self-renewal (Fig. 4.5.B). To determine the potential functional role of these genes in each group, we performed a PANTHER analysis to determine which pathways and biological processes were represented. Significantly overrepresented pathways among the leader stemness-related genes were TGFβ, Wnt, and Notch signaling, as well as angiogenesis, while the most significantly overrepresented biological process was *negative* regulation of differentiation (Fig. 4.5.C, D). Among the follower stemness-related genes there was also significant overrepresentation of TGFβ signaling but a lack of canonical stemness-related pathways, and the most significantly overrepresented biological process was *positive* regulation of differentiation (Fig. 4.5.C, E). Taken together, these data suggest that followers are expressing genes that serve to promote differentiation, while leaders express genes that could help to maintain a more stem-like phenotype.

Leader cells display tumor initiating capacity

To test the hypothesis that leader cells are a cancer stem cell-like population, self-renewal capacity as measured through a 3-D *in vitro* colony formation assay (Fig. 4.6.A). Briefly, single H1299 parental, leader, and follower cells were embedded in 100% Matrigel and colony formation was quantified after

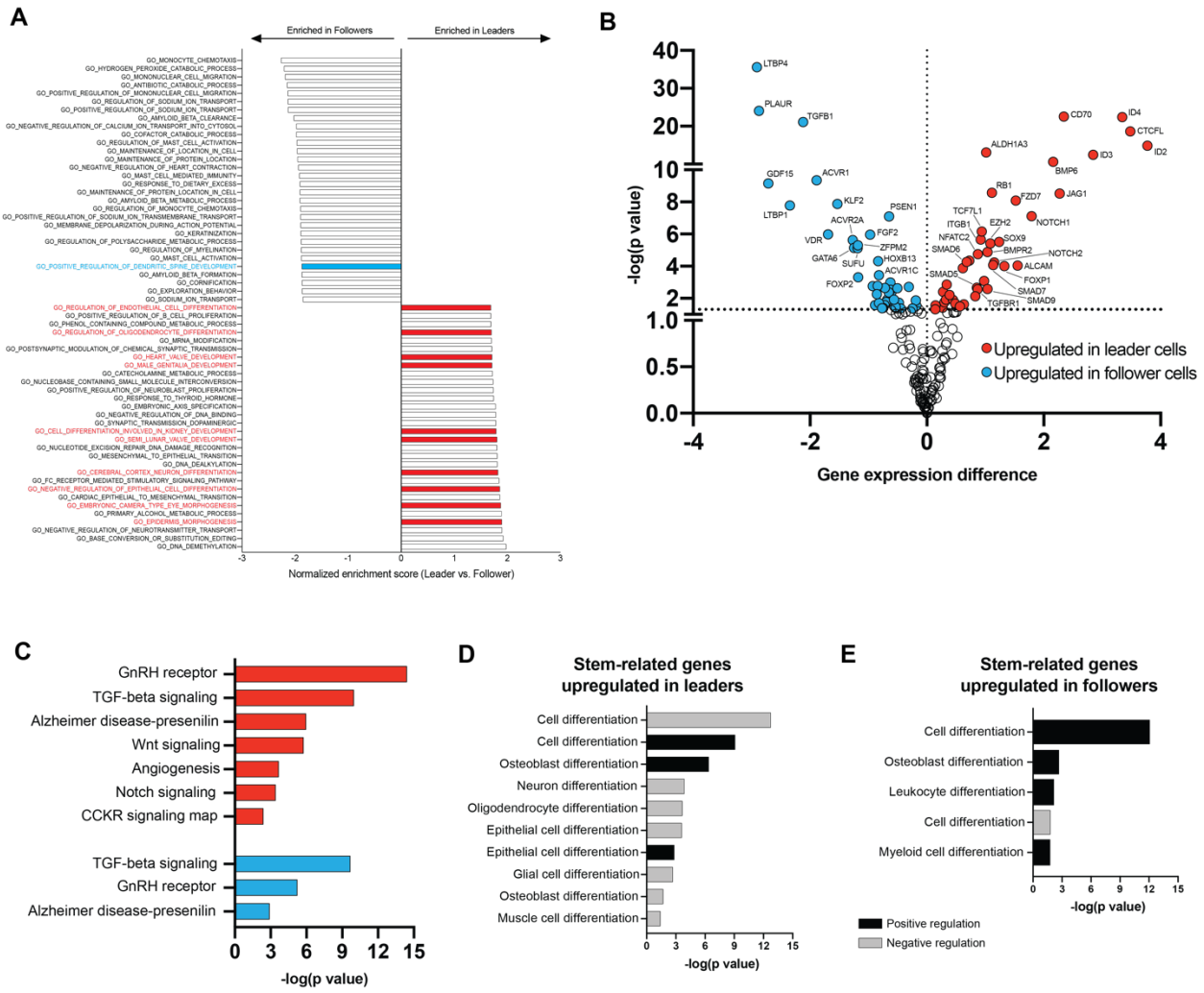


Figure 4.5. Leader cells have stem cell-like gene expression. (A) Graph of top 30 most enriched gene sets in followers and leaders, as defined by each cell's mutation profile. Gene sets were derived from the C5: GO Biological Process Ontology collection from MSigDB. (B) Volcano plot of stemness-related genes (n=195 genes). P values were determined using multiple t-tests with the Holm-Sidak method. (C) PANTHER pathway analysis of stemness-related genes with significantly increased expression in leaders (n=43 genes) and in followers (n=45 genes). (D-E) PANTHER biological process gene ontology analyses of stemness-related genes significantly increased in leaders (D) and followers (E).

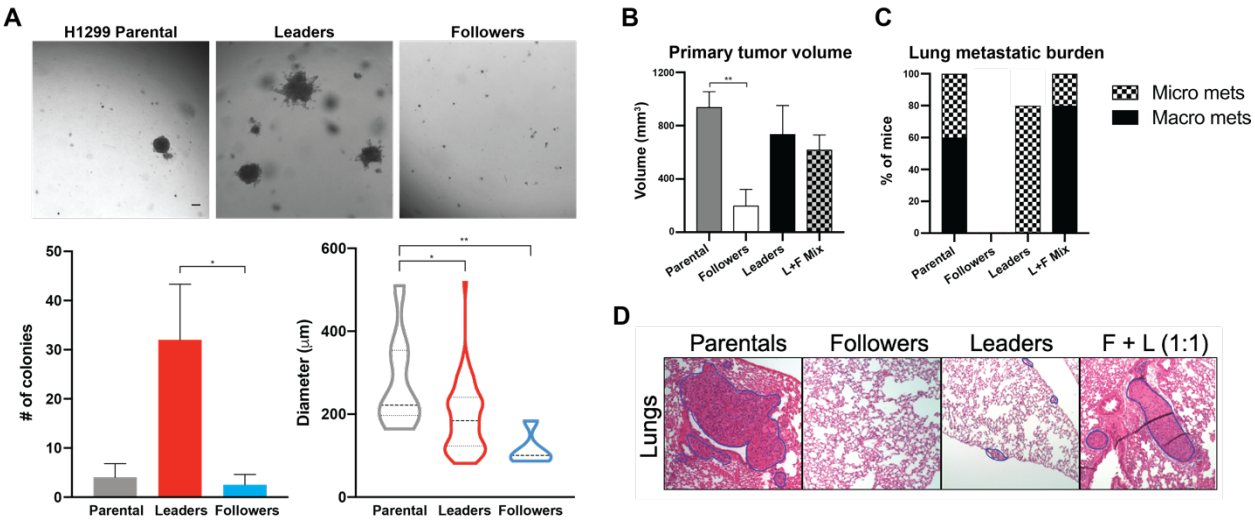


Figure 4.6. Leader cells display tumor-initiating capacity. (A) Representative images and quantification of # of colonies per well (n=3 wells) and average colony diameter from *in vitro* 3-D colony formation assay. **P*<0.05, ***P*<0.01 by one-way ANOVA with Tukey's multiple comparisons test. (B) Quantification of primary tumor volume resulting from flank injections of the listed cell populations in NSG mice. L+F mix includes a 1:1 mix of each cell type. n=5 mice per group. ***P*<0.01 by one-way ANOVA with Tukey's multiple comparisons test. (C) Quantification of metastatic foci in the lungs of mice from panel G; micro metastases: <20 cells; macro metastases: >20 cells. (D) Representative immunohistochemistry images of sections taken from the lungs of mice from panels B-C. Blue outlines indicate micro or macro metastatic lesions.

14 days. Leader cells formed numerous large colonies, indicating they can function as tumor initiating cells. By contrast, follower and parental cells formed very few colonies, suggesting they lack tumor initiating capacity Fig. 4.6.A). To further evaluate the tumor initiating capacity of these cell populations *in vivo*, each separate population, plus a mixed population of leader and follower cells, was injected into the flank of NOD/SCID mice. After 12 weeks, mice were sacrificed and primary tumor volume was measured. The parental, leader and leader/follower mixed populations were able to form substantial primary tumors, while the follower population formed significantly smaller primary tumors (Fig. 4.6.B). Metastatic foci at the lung were also quantified; both macro- and micro-metastases (defined as tumor foci containing >20 or <20 cells, respectively) were observed for the parental and leader/follower mixed populations, while leaders only formed micro-metastases, and followers formed no metastatic foci (Fig. 4.6.C, D). These data suggest that leaders alone may be capable of forming primary tumors, but they may require the presence of follower cells to promote secondary tumor-initiating capacity.

TGFβ crosstalk between leaders and followers regulates tumor initiating capacity

Recent evidence has shown that CSCs exhibit differential primary versus secondary tumor initiating capacity, and that these potentials are modulated by TGFβ signaling (140). Furthermore, TGFβ signaling was enriched among the groups of stemness-related genes expressed in both leaders and followers (Fig. 4.5.C); thus, we determined whether TGFβ crosstalk between leaders and followers could modulate leader cell tumor initiating capacity. We first analyzed expression levels of TGFβ pathway components from scRNA-seq, finding that followers express significantly higher levels of TGFB1, while leaders express higher levels of TGFBR1, one of the main receptors for TGFβ (Fig. 4.7.A). Furthermore, followers showed higher levels of TGFβ1 protein in both whole cell lysates and secreted media compared to leaders (Fig. 4.7.B). Other TGFβ signaling components differentially

expressed between leaders and followers included TGFB2, which was somewhat higher in leaders, and LTBP1/LTBP4, genes encoding latent TGFβ binding proteins, which were significantly higher in followers. Numerous members of the TGFβ-related BMP signaling pathway were differentially expressed, including upregulation of BMP6, BMPR2, SMAD7/8, and ID2/3/4 (which are BMP target genes) in leaders, and upregulation of the BMP receptors ACVR1 and ACVR2A in followers (Fig. S4.1).

To determine whether leader cell self-renewal capacity was affected by TGFβ signaling, the 3-D *in vitro* colony formation assay was repeated using leader cells plus 1) DMSO, 2) 10μg/mL TGFβ1 alone, 3) 10μg/mL TGFβ1 plus 1.0uM SB-505124 (a selective TGFβ-R1 antagonist), or 4) 1.0uM SB-505124 alone in serum-free RPMI 1640 media. At day 15, no significant difference was observed in the colony number for conditions 1-3, but significantly more colonies were observed with SB-505124 alone (Fig. 4.7.C). Additionally, the addition of TGFβ1 led to a modest increase in colony size that was not reduced by the addition of SB-505124; however, addition of SB-505124 alone resulted in significantly smaller colonies than the other three conditions (Fig. 4.7.C). Together, these data indicate that blocking TGFβ1 from binding to TGFβ-R1 on leader cells affects tumor initiating capacity, leading to more numerous, smaller colonies, while addition of extra TGFβ1 enhances tumor initiating capacity as evidenced by the larger average colony size. This data also mirrors the observations at the metastatic site in the previous *in vivo* experiments (Fig. 4.6.D); if followers are considered a source of TGFβ1 for leader cells, then only when this follower-produced TGFβ1 is present at the secondary site alongside leaders are they able to form large metastatic lesions.

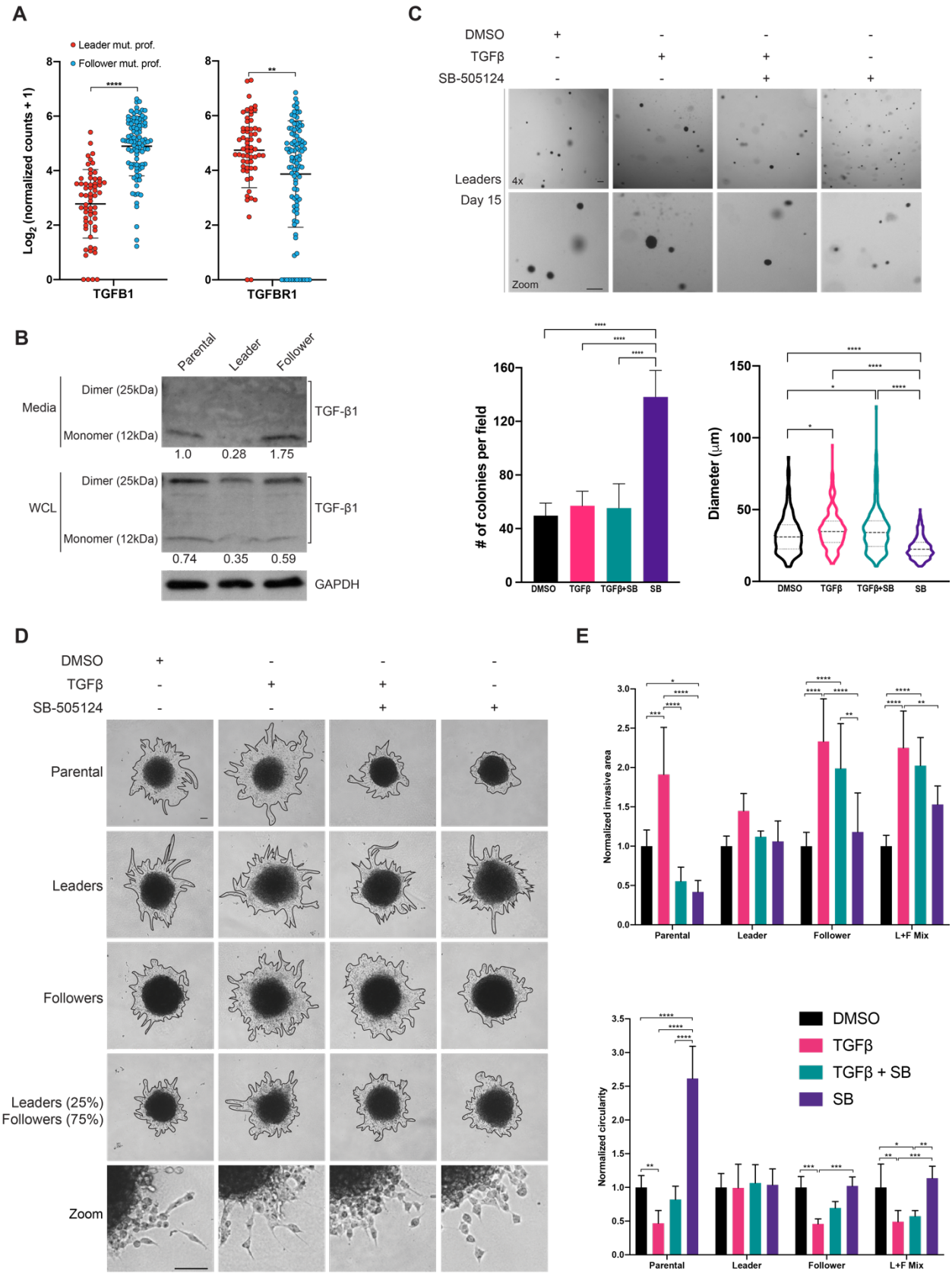

A
Leader mut. prof.
Follower mut. prof.
Log2 (normalized counts + 1)
TGFB1
TGFBR1
B
Parental
Leader
Follower
Media
Dimer (25kDa)
Monomer (12kDa)
TGF-β1
1.0 0.28 1.75
WCL
0.74 0.35 0.59
GAPDH
C
DMSO
TGFβ
SB-505124
Leaders
Day 15
4x
Zoom
# of colonies per field
Diameter (μm)
DMSO TGFβ TGFβ+SB SB
D
Parental
Leaders
Followers
Leaders (25%)
Followers (75%)
Zoom
E
Normalized invasive area
Parental Leader Follower L+F Mix
Normalized circularity
DMSO
TGFβ
TGFβ + SB
SB

Figure 4.7. TGFβ drives leader-follower cooperativity and increases collective invasion. (A) Quantification of TGFB1 and TGFBR1 expression between leaders and followers, defined by each cell's mutational profile. ***P*<0.01, *****P*<0.0001 by unpaired, two-tailed t-test. (B) Western blots showing expression of TGF-β1 protein in both whole-cell lysates (WCL) and conditioned media for H1299 parental, leader, and follower cell populations cultured in 2-D. Densitometry quantification is listed. (C) Representative images and quantification of # of colonies per field (n=2-3 fields across N=wells per condition) and average colony diameter from *in vitro* 3-D colony formation assay. **P*<0.05, *****P*<0.0001 by one-way ANOVA with Tukey's multiple comparisons test. (D) Representative images of 3-D invasion assay for H1299 parental, leader, follower or mixed leader+follower cells plus either DMSO, TGF-β1 (10μg/mL), TGF-β1 (10μg/mL) plus the TGFβR1 inhibitor SB-505124 (1uM), or SB-505124 (1uM) alone. Scale bar: 100um. (E) Quantification of 3-D invasion assays from panel D. Invasive area and circularity normalized to DMSO control for each condition. **P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001 by two-way ANOVA with Tukey's multiple comparisons test. (F) Immunofluorescence of spheroids from panel D stained for actin (Phalloidin, green), and JAG1 (red).

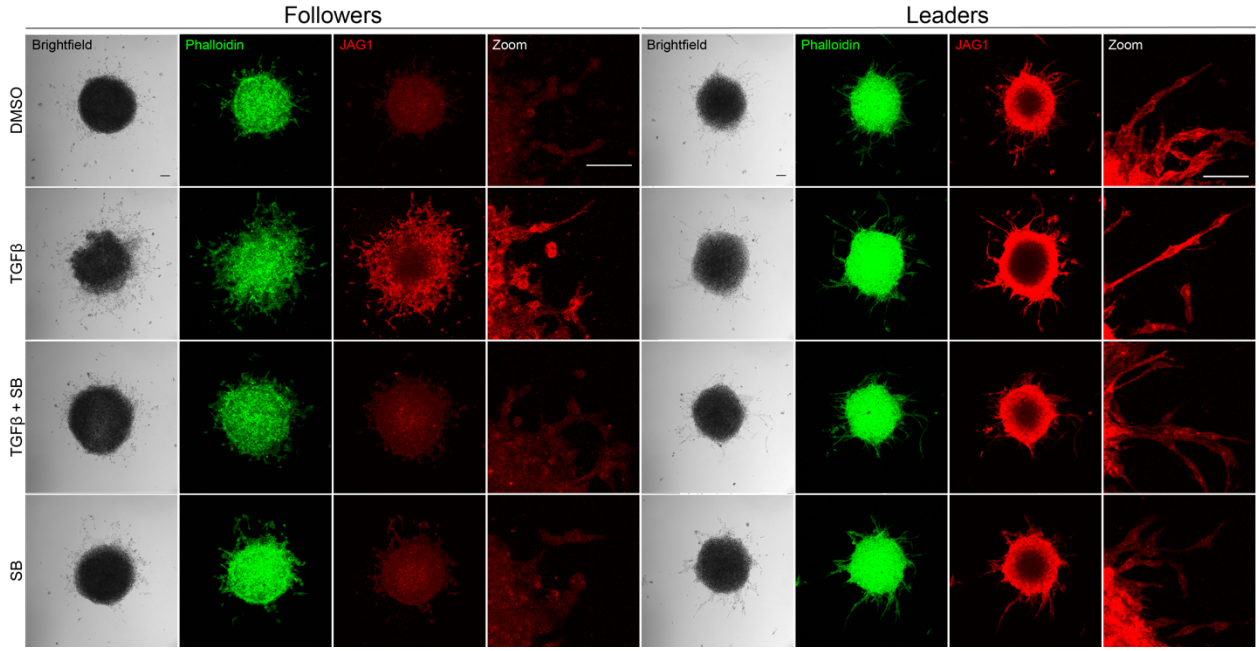


Figure 4.8. TGFβ induces expression of JAG1. Immunofluorescence of spheroids from 4.7.D stained for actin (Phalloidin, green), and JAG1 (red).

TGFβ signaling is important for cooperative invasion

As leader-follower cooperativity is also important during collective invasion, and TGFβ has been reported to promote a more mesenchymal, invasive phenotype in cancer cells, we tested the effects of TGFβ modulation using a 3-D invasion assay (Fig. 4.7.D). Leader cell spheroids showed no change in invasive area or circularity upon addition or inhibition of TGFβ (Fig. 4.7.E). However, follower cell spheroids showed significantly increased invasive area and decreased circularity (an indication of more chain-like vs. sheet-like invasion) upon addition of TGFβ1, and this effect was maintained even with addition of SB-505124 (Fig. 4.7.E). Parental spheroids also become more invasive and more chain-like when TGFβ1 was added; however, only in parental spheroids did SB-505124 alone decrease invasion compared to control (Fig. 4.7.E). This indicates that leaders and followers utilize TGFβ signaling to cooperative during collective invasion, and interruption of this crosstalk thus inhibits invasion. SB-505124 treatment did not affect invasion of leader or follower cells alone, further suggesting that TGFβ signaling is only important for invasion when the two populations are mixed. It is also notable that addition of TGFβ1 to followers is sufficient to induce chain-like invasion, as followers are a poorly invasive population that typically only displays sheet-like invasion with few chains. To test whether TGFβ1 was inducing expression of leader cell genes, we performed immunofluorescence staining for JAG1, a leader cell expression marker identified from both bulk RNA-seq and scRNA-seq. We found that addition of TGFβ1 increased JAG1 expression in followers as well as leaders, despite leaders' already expressing high levels of JAG1 at baseline (Fig. 4.8). SB-505124 blocked the ability of TGFβ1 to induce JAG1 expression, and in leader cells, SB-505124 treatment alone further decreased JAG1 expression from baseline (Fig. 4.8), suggesting that TGFβ may be important for induction and maintenance of JAG1 expression in leader cells.

4.4 Discussion

RNA-sequencing (RNA-seq), including single-cell RNA-seq (scRNA-seq), has traditionally been used to measure gene expression; however, recent studies have begun utilizing the sequence information from RNA-seq to identify genomic variants in a tumor cell population, or in the case of scRNA-seq, within individual cells (141, 142). Combining gene expression and variant analysis can thus provide a multi-dimensional view of the genomic state of a single cell. Indeed, single-cell sequencing has been increasingly employed in the exploration of intra-tumoral genetic heterogeneity across numerous cancer types, often revealing numerous distinct subpopulations that have evolved within a single tumor (143-147). However, a major drawback of single-cell sequencing is that the analyzed cells cannot be subsequently followed to determine correlations between genetic profiles and phenotypes. This is especially important when considering collective cancer invasion, which depends upon cooperation between phenotypically distinct cell types often termed leaders and followers (3, 11). To address this, we adapted the SaGA platform (3) to precisely select, and subject to scRNA-seq, single H1299 leader and follower cells directly from collective invasion packs. Utilizing our previous analyses of H1299 leader and follower populations, we had the unique advantage of previously defined expression and mutational markers that allowed for the to precise labeling of each single cell in our analysis. This genomic labeling enabled, for the first time, a robust examination of the genetic profiles of individual leader and follower cells that were actively participating in collective invasion.

Our results demonstrated that genetically distinct cell populations exist within the H1299 cell line. Importantly, it was found that previously identified leader and follower mutation profiles are mutually exclusive on the single-cell level, and that these mutations correlate strongly with leader and follower cell gene expression markers. Mutations represent more permanent changes in the genome of a cell, and are thus a more robust marker than gene expression, which can vary based upon the context in

which it is measured. Mutational labeling of single cells in this analysis revealed a number of novel findings. Firstly, it must be noted that the user-defined positional phenotypes that were assigned during initial cell selection did not perfectly correlate with leader-follower genomic markers. This is likely due to the dynamic process of collective invasion; a cell isolated from farther back in a chain may appear to be a follower, but actually have the capacity to eventually move forward and lead its own chain. Furthermore, we found that there exist cycling and non-cycling populations of both leaders and followers. The majority of non-cycling leaders and followers were actually isolated from the control condition, which could have included cells from the spheroid core; thus, we hypothesize that the hypoxic environment of the core led to inhibition of the cell cycle in these cells.

When comparing the gene expression profiles mutationally-labeled groups of individual leader and follower cells through GSEA, we discovered that, importantly, leader cells are enriched for gene sets related to processes including development, differentiation, and morphogenesis, suggesting they could be a more cancer stem cell-like population. Indeed, subsequent *in vitro* and *in vivo* experiments showed that leaders display increased tumor initiating capacity. This could represent an important new role for leader cells in the metastatic cascade, where they could drive secondary tumor formation in addition to facilitating collective invasion. Only recently have cancer stem cells been implicated as displaying leader cell behavior (32, 148), and this hypothesis thus warrants further exploration. A potentially important distinction is that of primary versus secondary tumor-initiating capacity. In our *in vivo* experiments, leader cells alone can travel from the flank to the lung and form micro-metastases; however, only in the condition of mixed leader and follower cells do macro-metastases form in the lungs. This raises the possibility that leader cells alone show primary tumor initiating capacity, but require the support of follower cells to form robust secondary tumors. It is also possible that only a subset of leader cells are cancer stem cell-like; for example, *ALDH1A3*, an established cancer stem

cell marker, was completely absent in follower cells and was expressed in 29/61 (47.5%) of leader cells. Further studies, including side-population analysis through flow cytometry, are indicated to better understand stemness potential of different cells within the leader cell population. Unfortunately, there are few consensus markers for lung cancer stem cells (31), and thus there is greater reliance upon phenotypic studies to determine whether leader cells are indeed a stem-like population.

Importantly, recent studies using breast cancer cell populations has shown that CSC populations switch between primary and secondary tumor initiating capacity through modulation of TGFβ signaling, where TGFβ inhibition promotes primary tumor formation but inhibits secondary tumor formation (140). Interestingly, our scRNA-seq analysis also revealed differential expression of TGFβ pathway components between leaders and followers during collective invasion, with leaders expressing higher levels of TGFβ-R1 and followers expressing (and secreting, as shown by Western blotting) higher levels of TGFβ1. In vitro tumor initiation experiments showed that addition of TGFβ1 modestly increased colony size, while TGFβ-R1 inhibition decreased colony size but increased colony number. Thus, we hypothesize that follower cells may act as a source of TGFβ1 for leader cells, allowing for metastatic tumor growth. Further *in vivo* experiments will use fluorescently labeled cells to determine which cell types (leaders, followers, or both) comprise the secondary tumors, and different mixing conditions will explore the lowest percentage of follower cells needed to support secondary tumor growth. Furthermore, scRNA-seq analysis indicated upregulation of other components of the TGFβ and BMP signaling pathways, suggesting there is potentially complex TGFβ crosstalk between leaders and followers.

TGFβ has also been implicated as a driver of EMT and cancer invasion, a characteristic that was further supported by our findings; however, EMT is typically associated with single-cell invasion,

rather than collective invasion. We found that addition of TGFβ1 led to both single cell invasion and collective chain formation in H23 and H1975 NSCLC lines (Fig. S4.2). However, TGFβ stimulation led to increased collective invasion – without increasing single-cell invasion – among H1299 parental and follower cells, and significantly upregulated JAG1, a leader cell marker we previously identified, in both leader and follower cells. Importantly, inhibition of TGFβ signaling through SB-505124 had no effect on invasion of leaders or followers alone, while it significantly decreased collective invasion of parental H1299 cells. This is an indication that leaders and followers may communicate via TGFβ signaling during collective invasion, and that this signaling is important for invasion to proceed. Continued experiments using Western blotting and immunofluorescence will examine the protein expression of TGFβ pathway components within leaders and follower during collective invasion, as well as *in vivo* studies with modulation of TGFβ signaling and subsequent immunohistochemistry to examine effects on collective invasion. Furthermore, the effects of TGFβ modulation on both self-renewal capacity and collective invasion will be explored in additional lung cancer cell lines and patient samples.

There are a number of drawbacks to these studies, including the limitation of studying leader and follower cells from a single cell line. To more fully determine the applicability of these findings to lung cancer and collective invasion in general, the same experiments must be performed using additional cell lines, and ideally patient-derived organoids. Additionally, our analysis was aided by our previous bulk RNA-seq analysis of H1299 leader and follower populations, which provided known mutational and gene expression markers. When applying this type of analysis to patient samples, it may not be cost- or time-effective to perform both bulk and single-cell RNA-seq. Ideally, continued single-cell sequencing studies will result in a database of expression and mutational markers that can be probed in scRNA-seq studies of new patient tumors; however, it remains to be seen if this is feasible. In

addition, our scRNA-seq analysis has implicated a large number of genes as being potentially involved in driving the leader cell phenotype, many of which warrant follow-up studies. One such example is CD70, which is significantly overexpressed in leader cells compared to followers (Fig. 4.5.B) and has previously been implicated in cancer cell stemness and increased invasion (149, 150). Indeed, when we sorted CD70+ and CD70- cells from four different lung cancer cell lines, we found increased invasiveness and tumor initiating capacity for H1299 CD70+ cells; however, the invasive phenotype and tumor-initiating capacity of CD70+ cells were more variable in the other lung cancer cell lines (Fig. S4.3). This highlights the difficulties of identifying broadly applicable biomarkers from a single cell line or primary tumor, as well as the necessity of building a database of biomarkers for rare cell types such as leader cells and cancer stem cells. Overall, by combining the SaGA platform with scRNA-seq, this study provides a novel pathway for precise genomic and phenotypic profiling of leader and follower cells in collective invasion, revealing new insights about the biology of these unique cell types and opening the door for new precision medicine techniques in cancer patient care.

Supplementary information

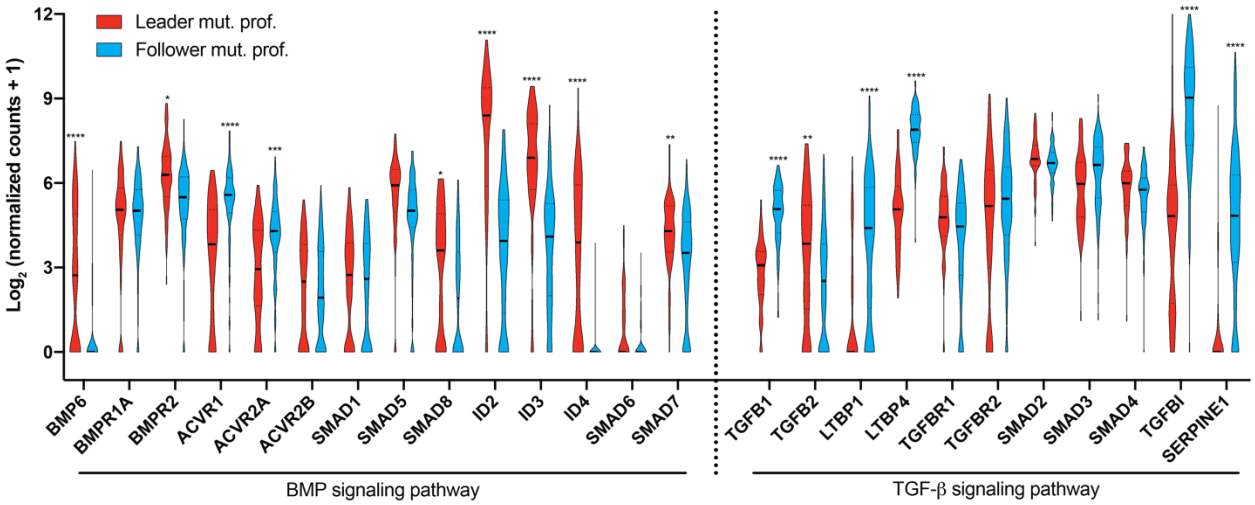


Figure S4.1. Expression of TGFβ family members in H1299 leader and follower cells. Violin plots showing expression of TGFβ and BMP signaling-related genes from scRNA-seq analysis. **P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001 by two-way ANOVA with Tukey's multiple comparisons test.

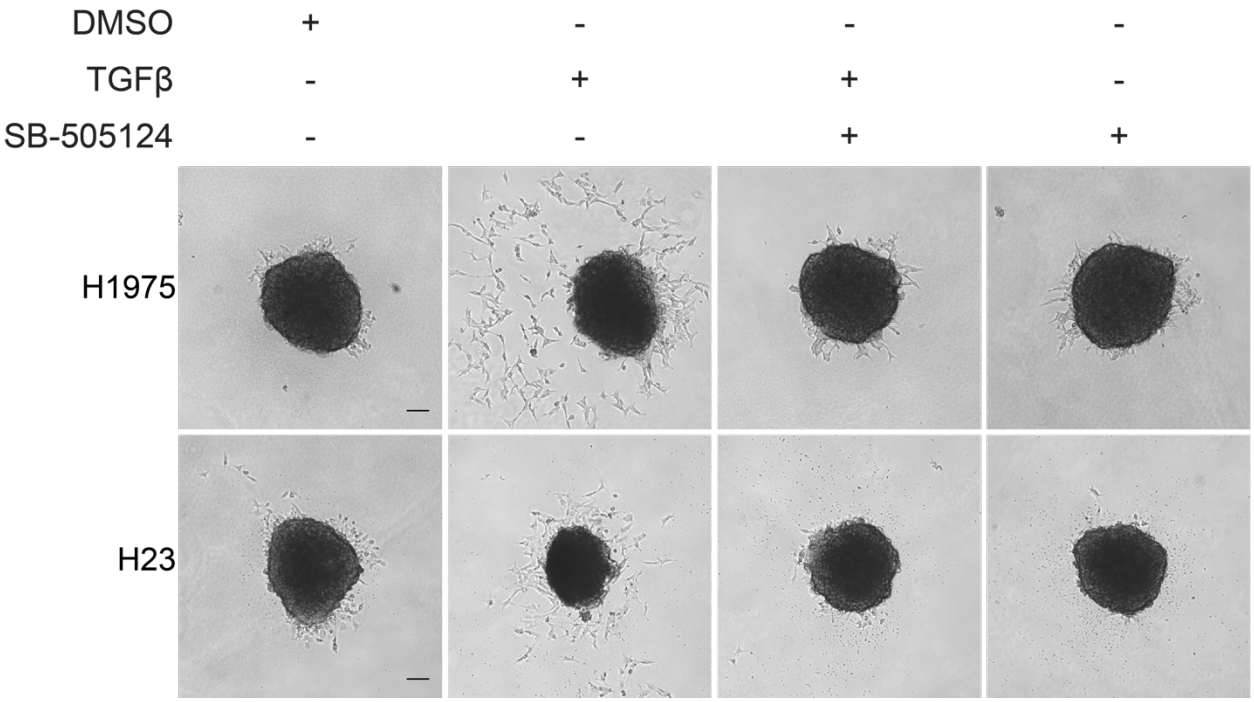


Figure S4.2. TGFβ drives invasion in H1975 and H23 NSCLC cells. Representative images of 3-D invasion assay for H1975 and H23 cells plus either DMSO, TGF-β1 (10μg/mL), TGF-β1 (10μg/mL) plus the TGFβR1 inhibitor SB-505124 (1uM), or SB-505124 (1uM) alone. Scale bar: 100um.

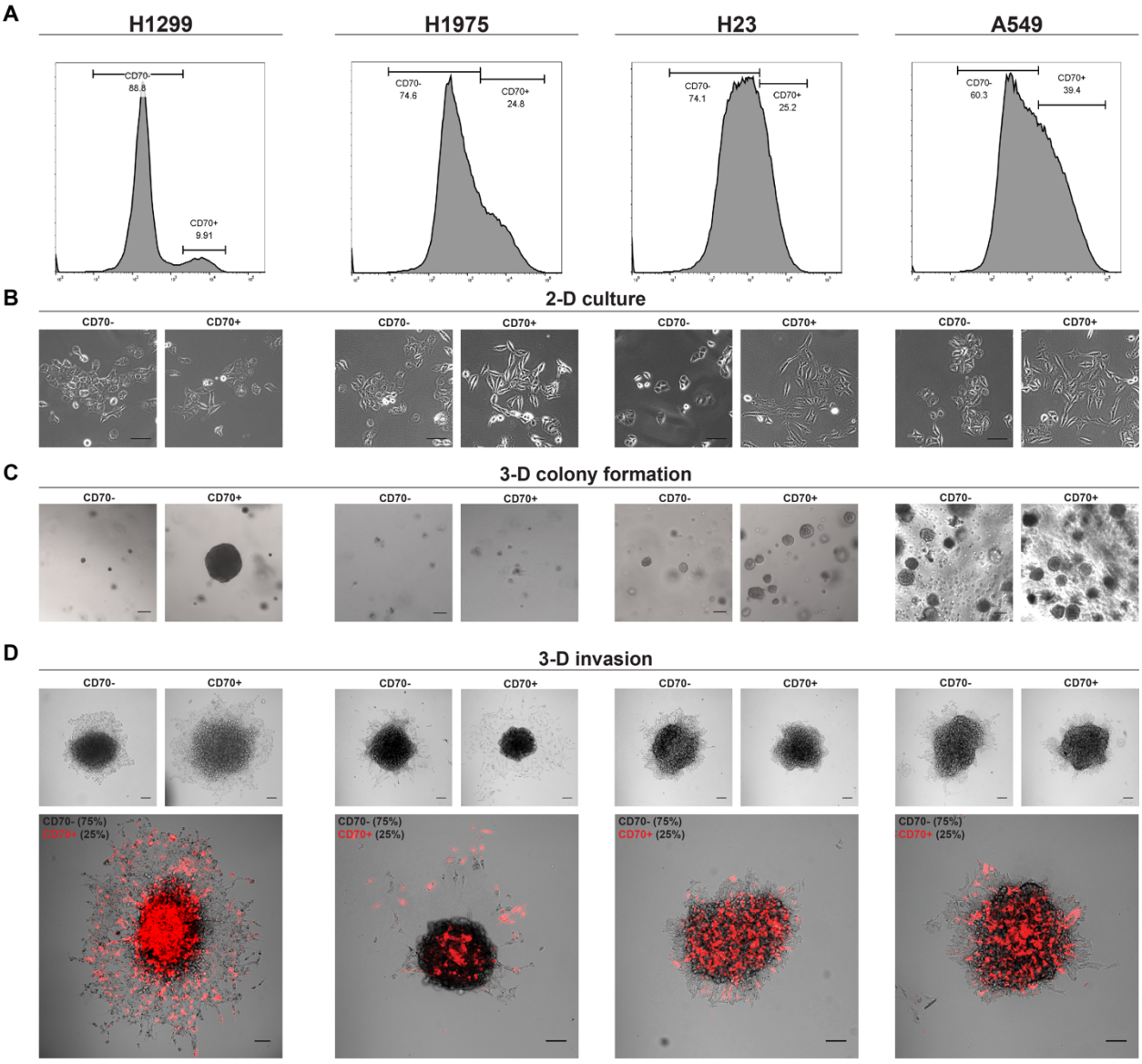


Figure S4.3. Invasiveness and self-renewal capacity of CD70+ cells. (A) CD70+ and CD70- cells were sorted from four different lung cancer cell lines. (B) Phenotypes of CD70+/- cells for each cell line in 2-dimensional culture. (C) Representative images of 3-D colony formation in Matrigel for CD70+/- cells. (D) Representative images of 3-D spheroid invasion in Matrigel for CD70- cells alone, CD70+ cells alone, and a mix of 75% CD70- cells + 25% mCherry-CD70+ cells for each cell line.

Chapter 5: Conclusions and Future Directions

5.1 Role of gene mutations in dissecting leader and follower cell biology

Despite the substantial increase in recent years in studies focusing on leader and follower cells in collective cancer invasion, the vast majority have focused on gene expression or epigenetic changes, with virtually none specifically examining the role of gene mutations in the development of these phenotypes. This is likely due in part to the challenges of linking genetic profiling with a cell's phenotype; however, through the SaGA platform, we have devised a mechanism by which leader and follower cells can be identified, isolated, and genomically profiled, even down to the single-cell level. Thus, we have an unprecedented window into the biological mechanisms underlying the emergence of leader and follower phenotypes during collective invasion, including mutations that could underlie the evolution of these cell types within a tumor population. This led to the discovery that within the H1299 cell line, there exist leader and follower cell populations that contain numerous mutually exclusive gene mutations.

This finding presented a unique opportunity to probe the mechanisms of leader and follower cell behavior. Firstly, we sought to test whether these mutations were merely passengers, or if they could be active drivers of leader and follower cell phenotypes. Indeed, introduction of a single leader-specific mutation in the protein ARP3, a crucial component of the Arp2/3 complex that drives cell motility through actin dynamics, conferred leader-like capabilities upon follower cells, including invasiveness and pioneering chains of unmodified followers. This is a strong indication that gene mutations, and specifically the ARP3 K240R mutation, can in fact drive leader cell behavior; furthermore, these results continue to highlight the potential utility of searching for leader cell mutations in other cell lines,

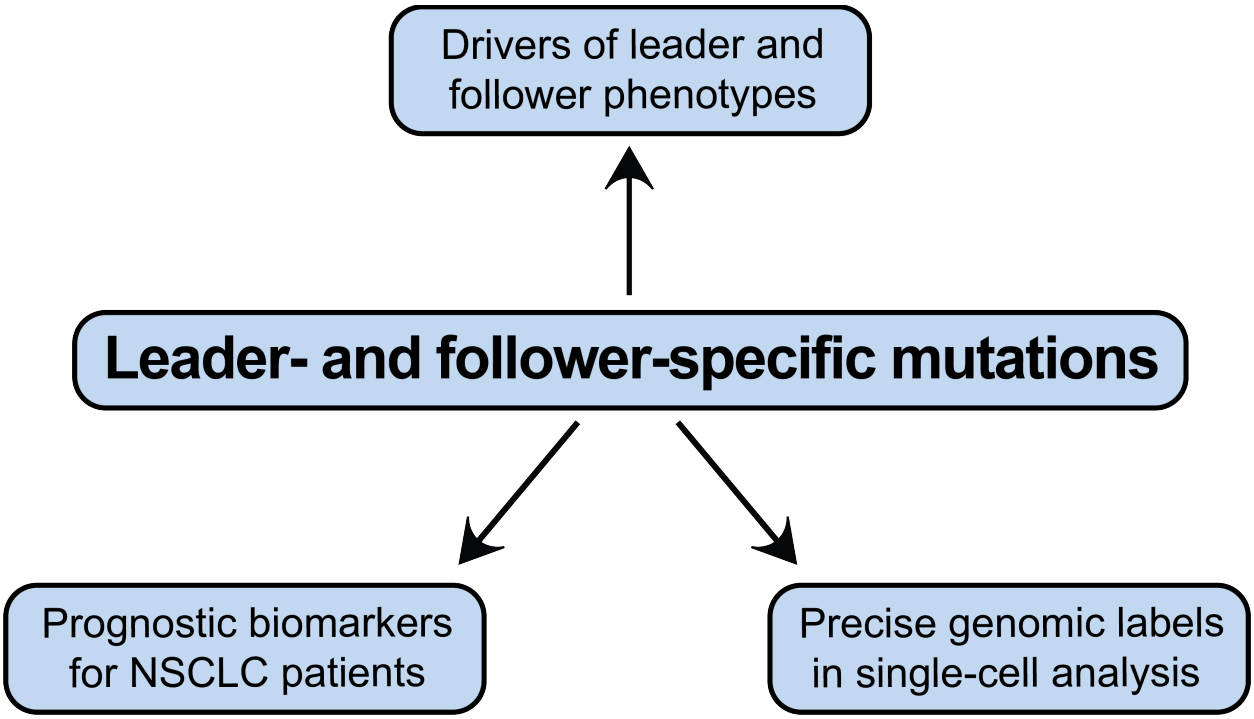


Figure 5.1. Dissertation conclusions regarding leader- and follower-specific mutations. Chapters 2, 3, and 4 demonstrate that leader- and follower-specific mutations have roles as drivers as drivers of leader and follower cell phenotypes in collective invasion, as prognostic biomarkers for identification of high-risk NSCLC patients, and as precise genomic labels for leader and follower cells in single-cell analysis of collectively invading cancer cells.

cancer types, and patient samples as the search for targeted anti-metastatic therapy continues (Fig. 5.1).

The finding that ARP3 K240R promotes leader cell behavior suggests numerous avenues for future investigation. Firstly, as the ultimate goal of these experiments is to identify a strategy for targeting leader cells and thus inhibiting invasion and metastasis, it should be tested whether treatment with the small-molecule Arp2/3 inhibitor CK-666 (95) can inhibit invasion *in vitro* and metastasis *in vivo*, both using the H1299 cell line as well as others. Furthermore, through scRNA-seq, we have shown that leader- and follower-specific mutations are useful genomic markers; therefore, in future *in vivo* experiments, one could sequence for these mutations in both primary and secondary tumors, as well as CSCs and CSC clusters, as a means of tracking leaders and followers throughout the metastatic process. If the mutations identified from the H1299 cell line prove to be found in other cell lines and patient samples, the same mutations could be used in other contexts as well; however, the more likely scenario is that different mutations will be found in different samples; thus, it would be ideal to use the SaGA platform to begin building a database of known leader and follower mutations, thus increasing the likelihood of identifying mutations that are found across samples and tumors.

5.2 Clinical implications of the leader-derived 16q mutation cluster

The vast majority of leader and follower cell studies have thus far focused on dissecting their basic mechanisms and underlying biology. While this is crucial in order to ultimately develop targeted anti-metastatic therapeutics, our data have shown that these markers, including leader-specific mutations, also have direct clinical utility as a prognostic biomarker. Specifically probing for mutations in a group of 10 genes on chromosome 16q identifies patients who experienced poorer overall and progression-

free survival among cohorts of LUSC and LUAD patients (the two major types of NSCLC), as well as a cohort of HCC patients. Importantly, these mutations were predictive of survival even among patients diagnosed with earlier-stage disease. Furthermore, 16qMC+ tumors were enriched for gene expression sets associated with poor prognosis from previous large-scale sequencing studies. These data suggest that a targeted panel for these 10 genes could be feasibly used by clinicians to identify higher-risk, newly diagnosed NSCLC (Fig. 5.1).

Prior to any clinical implementation of a 16qMC panel, additional clinical validation of the mutation cluster will be needed. As the analyses described here were performed retrospectively in publicly available patient cohorts, the first step will be to perform retrospective analyses using banked tumor tissue, and prospective analyses using newly-isolated patient tissue. The prospective analyses will be particularly informative, as they could elucidate both the efficiency with which the 16qMC panel can be applied, and eventually, its accuracy in stratifying patients as higher risk. However, prospective studies, especially those measuring survival, necessarily require long follow-up periods before sufficient data are obtained. Thus, it will be important to simultaneously continue validating this mutation cluster retrospectively, both using patient samples and additional public databases as they become available.

Subsequent steps toward clinical implementation would include a clinical trial, in which patients would be stratified as 16qMC+ or 16qMC- upon diagnosis. Stage I NSCLC patients are typically treated with surgical resection, followed by either regular screening (if the removal is deemed complete), or additional therapy including radiotherapy and/or chemotherapy if removal is incomplete (6). Therefore, following surgical resection, the tumor tissue could be subjected to the 16qMC panel, and if found to be 16qMC+, those patients could be given either more frequent screening or adjuvant

chemotherapy or immunotherapy, followed by monitoring to measure disease progression/recurrence and overall survival. If it is found that 16qMC+ patients who receive additional screening or treatment do indeed have improved outcomes compared to those who only receive standard screening, then the 16qMC panel could be approved as a standard prognostic test in the clinic. Immune checkpoint inhibitor (ICI) therapy could be an attractive adjuvant treatment option for 16qMC+ patients, as studies have shown mutational burden to be a useful biomarker in predicting response to ICI therapy (151-153), and our studies found that 16qMC+ tumors consistently have increased average mutational burden in NSCLC and across numerous other cancer types.

One of the major potential complications with sequencing primary tumor tissue, however, is that there is a high likelihood of missing cell subpopulations, and certain mutations that are either rare, or simply located in other regions of the tumor. Therefore, an alternative approach to 16qMC sequencing could be to employ a liquid biopsy approach, in which blood is taken from the patient and circulating tumor DNA (ctDNA) is sequenced (127). This would also provide the advantage of being minimally invasive, without requiring actual tumor tissue. However, it remains to be seen whether sufficient ctDNA could be obtained to detect 16qMC mutations.

5.3 Characterization of leader cells as a cancer stem cell-like population and the role of TGFβ signaling in leader-follower cooperativity

By applying mutational labels to individual cells in our scRNA-seq analysis, we were able to dissect the gene expression profiles of leader and follower cells to a degree that was not previously possible (Fig. 5.1). Importantly, this led to the discovery that the gene expression found in leader cells is similar to what might typically be seen in a cancer stem cell-like population, with enrichment of gene sets related to development, morphogenesis, and differentiation. This also makes sense when considering

that Notch signaling has been previously associated with H1299 leader cells – Notch signaling, and in particular JAG1, a leader cell expression marker (27), are commonly implicated in regulating cancer cell stemness (154, 155). Furthermore, leader cells have been found to be highly chemo-resistant compared to followers (28), another property typically attributed to stem-like cells (156). Our *in vitro* and preliminary *in vivo* experiments confirmed that leaders possess increased self-renewal and tumor-initiating capacity, which further supports the possibility of leaders being a CSC population. It is also possible that only a subset of leader cells is stem-like, as some stem cell-associated markers such as *ALDH1A3* are only expressed in a fraction of leader cells.

Our data also support the possibility that leaders have differential primary vs. secondary tumor-initiating capacity. Leader cells alone could form robust primary flank tumors and successfully form lung micro-metastases, but macro-metastases were only observed in the conditions with mixed leaders and followers. Given previous studies implicating TGFβ signaling modulation as a switch between primary and secondary tumor-initiating cells, and our data showing that followers express and secrete higher levels of TGFβ, continued experiments will focus on determining whether follower cells can 1) act as a source of TGFβ for leader cells, and 2) promote secondary tumor formation by leader cells. To address this question, *in vivo* experiments will use fluorescently labeled leader cells mixed in different proportions with followers. This will reveal whether secondary tumors are comprised of leader cells alone, followers alone, or a mixed population, and also whether TGFβ crosstalk is occurring at the secondary site.

Previous studies examining clonality of metastatic lung tumors in a mouse model of breast cancer have demonstrated that these metastases typically arise from seeding of polyclonal tumor cell clusters and thus develop into polyclonal tumors (14). Therefore, we could expect to see polyclonal metastases

as well; however, this could arise from expansion of both leader and follower populations, or from differentiation of a stem cell-like leader population. Fluorescent labeling of leader cells should help to answer this question – if there are subpopulations in the metastatic tumor that arise from a leader cell stem-like population, they would be expected to maintain this fluorescence, whereas cells arising from follower cells would not.

The finding that leader cells are potentially a tumor-initiating population has significant clinical relevance. Previous research into leader and follower cells has been focused on their functions during collective invasion; however, if leader cells are also responsible in some cases for colonization and tumor formation at the secondary site, it would mean that leader cells play a crucial role during multiple steps of the metastatic cascade. Development of targeted anti-leader cell therapeutics would also become even more important, as inhibiting leader cells could potentially halt both collective invasion (i.e. early steps metastasis) and secondary tumor formation (i.e. late steps of metastasis). In addition, our *in vivo* data indicate a role for anti-follower cell therapeutics, specifically for inhibiting secondary tumor formation, as leader cells can only form micro-metastases in the absence of follower cells. It is therefore possible that targeting follower cells could prevent them from signaling to leader cells through TGFβ or other mechanisms, resulting in diminished formation of macro-metastases. Alternatively, if TGFβ crosstalk is indeed a mechanism by which leaders and followers communicate, inhibitors of TGFβ signaling could be employed. As TGFβ has been shown to act as a tumor suppressor during early tumorigenesis but promote metastasis at later stages, it has potential as an therapeutic target but must be approached carefully (157). Although there are currently no FDA-approved TGFβ inhibitors, there are a number of agents currently in Phase I, II and III trials (157).

The next phase of this research will also necessarily include expansion of the SaGA platform into other cell lines, cancer types, and patient-derived organoids. The latter is particularly important for the eventual application of SaGA toward developing personalized therapeutics; it needs to be proven that SaGA can be used to isolate phenotypically distinct cells from a patient-derived organoid, and subsequently subject them to genomic analyses to determine whether they express any currently targetable markers. In order to significantly impact patient, care these steps need to be undertaken expeditiously yet in a highly reproducible manner. In addition, it would be highly beneficial in the meantime to construct a database of most common potential leader and follower driver mutations, especially those for which targeted therapies are available. Although metastatic disease still accounts for the vast majority of cancer-related mortality, survival rates across all cancer types, especially including lung cancer, continue to climb (5); continuing to dissect the mechanisms leader and follower cell cooperation during metastasis will help ensure that this upward trend continues.

References

1. Lawson DA, Kessenbrock K, Davis RT, Pervolarakis N, Werb Z. Tumour heterogeneity and metastasis at single-cell resolution. Nat Cell Biol. 2018;20(12):1349-60. Epub 2018/11/30. doi: 10.1038/s41556-018-0236-7. PubMed PMID: 30482943.

2. Howlader N, Noone A, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis D, Chen H, Feuer E, Cronin K. SEER Cancer Statistics Review, 1975-2016. National Cancer Institute Bethesda, MD.

3. Konen J, Summerbell E, Dwivedi B, Galior K, Hou Y, Rusnak L, Chen A, Saltz J, Zhou W, Boise L, Vertino P, Cooper L, Salaita K, Kowalski J, Marcus A. Image-guided genomics of phenotypically heterogeneous populations reveals vascular signaling during symbiotic collective cancer invasion. Nature Communications. 2017;8:15078. doi: 10.1038/ncomms15078. PubMed PMID: 28497793; PMCID: PMC5437311.

4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.

5. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians. 2020;70(1):7-30. doi: 10.3322/caac.21590.

6. Network NCC. NCCN Clinical Practice Guidelines in Oncology - Non-Small Cell Lung Cancer2019; Version 3.2019. Available from: NCCN.org.

7. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147(2):275-92. Epub 2011/10/18. doi: 10.1016/j.cell.2011.09.024. PubMed PMID: 22000009; PMCID: PMC3261217.

8. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009;119(6):1420-8. doi: 10.1172/JCI39104. PubMed PMID: 19487818; PMCID: PMC2689101.

9. Lee JM, Dedhar S, Kalluri R, Thompson EW. The epithelial-mesenchymal transition: new insights in signaling, development, and disease. J Cell Biol. 2006;172(7):973-81. Epub 2006/03/29. doi: 10.1083/jcb.200601018. PubMed PMID: 16567498; PMCID: PMC2063755.

10. Friedl P, Wolf K. Tumour-cell invasion and migration: diversity and escape mechanisms. Nature reviews Cancer. 2003;3(5):362-74. doi: 10.1038/nrc1075. PubMed PMID: 12724734.

11. Cheung KJ, Gabrielson E, Werb Z, Ewald AJ. Collective invasion in breast cancer requires a conserved basal epithelial program. Cell. 2013;155(7):1639-51. doi: 10.1016/j.cell.2013.11.029.

12. Richardson AM, Havel LS, Koyen AE, Konen JM, Shupe J, Wiles WG, Martin WD, Grossniklaus HE, Sica G, Gilbert-Ross M, Marcus AI. Vimentin Is Required for Lung Adenocarcinoma Metastasis via Heterotypic Tumor Cell–Cancer-Associated Fibroblast Interactions during Collective Invasion. Clinical Cancer Research. 2018;24(2):420-32. doi: 10.1158/1078-0432.Ccr-17-1776.

13. Gilbert-Ross M, Konen J, Koo J, Shupe J, Robinson BS, Wiles WGt, Huang C, Martin WD, Behera M, Smith GH, Hill CE, Rossi MR, Sica GL, Rupji M, Chen Z, Kowalski J, Kasinski AL, Ramalingam SS, Fu H, Khuri FR, Zhou W, Marcus AI. Targeting adhesion signaling in KRAS, LKB1 mutant lung adenocarcinoma. JCI Insight. 2017;2(5):e90487. Epub 2017/03/16. doi: 10.1172/jci.insight.90487. PubMed PMID: 28289710; PMCID: PMC5333956 exists.

14. Cheung KJ, Padmanaban V, Silvestri V, Schipper K, Cohen JD, Fairchild AN, Gorin MA, Verdone JE, Pienta KJ, Bader JS, Ewald AJ. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. Proc Natl Acad Sci U S A. 2016;113(7):E854-63. Epub 2016/02/03. doi: 10.1073/pnas.1508541113. PubMed PMID: 26831077; PMCID: PMC4763783.

15. Friedl P, Gilmour D. Collective cell migration in morphogenesis, regeneration and cancer. Nature reviews Molecular cell biology. 2009;10(july):445-57. doi: 10.1038/nrm2720.

16. Szabó A, Mayor R. Mechanisms of Neural Crest Migration. Annual Review of Genetics. 2018;52(1):43-63. doi: 10.1146/annurev-genet-120417-031559.

17. Aceto N, Bardia A, Miyamoto DT, Donaldson MC, Wittner BS, Spencer JA, Yu M, Pely A, Engstrom A, Zhu H, Brannigan BW, Kapur R, Stott SL, Shioda T, Ramaswamy S, Ting DT, Lin CP, Toner M, Haber DA, Maheswaran S. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell. 2014;158(5):1110-22. doi: 10.1016/j.cell.2014.07.013. PubMed PMID: 25171411; PMCID: PMC4149753.

18. Hieger I. BIOLOGICAL ASPECTS OF CANCER. British Medical Journal. 1958;2(5094):494. doi: 10.1136/bmj.2.5094.494-c.

19. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-28. doi: 10.1016/j.cell.2017.01.018. PubMed PMID: 28187284.

20. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23. doi: 10.1126/science.959840.

21. Galuppini F, Dal Pozzo CA, Deckert J, Loupakis F, Fassan M, Baffa R. Tumor mutation burden: from comprehensive mutational screening to the clinic. Cancer Cell International. 2019;19(1):209. doi: 10.1186/s12935-019-0929-4.

22. Wu S-G, Shih J-Y. Management of acquired resistance to EGFR TKI-targeted therapy in advanced non-small cell lung cancer. Molecular cancer. 2018;17(1):38-. doi: 10.1186/s12943-018-0777-1. PubMed PMID: 29455650.

23. Ramalingam SS, Vansteenkiste J, Planchard D, Cho BC, Gray JE, Ohe Y, Zhou C, Reungwetwattana T, Cheng Y, Chewaskulyong B, Shah R, Cobo M, Lee KH, Cheema P, Tiseo M, John T, Lin M-C, Imamura F, Kurata T, Todd A, Hodge R, Saggese M, Rukazenkov Y, Soria J-C, Investigators F. Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC.

The New England journal of medicine. 2020;382(1):41-50. Epub 2019/11/21. doi: 10.1056/NEJMoa1913662. PubMed PMID: 31751012.

24. Westcott JM, Prechtl AM, Maine EA, Dang TT, Esparza MA, Sun H, Zhou Y, Xie Y, Pearson GW. An epigenetically distinct breast cancer cell subpopulation promotes collective invasion. J Clin Invest. 2015;125(5):1927-43. doi: 10.1172/JCI77767. PubMed PMID: 25844900; PMCID: PMC4463195.

25. Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nature Reviews Cancer. 2003;3(6):453-8. doi: 10.1038/nrc1098.

26. Fidler I. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nature Reviews Cancer. 2003;3(6):453-8.

27. Summerbell ER, Mouw JK, Bell JSK, Knippler CM, Pedro B, Arnst JL, Khatib TO, Commander R, Barwick BG, Konen J, Dwivedi B, Seby S, Kowalski J, Vertino PM, Marcus AI. Epigenetically heterogeneous tumor cells direct collective invasion through filopodia-driven fibronectin micropatterning. Science Advances. In Revision.

28. Commander R. Nature Communications. 2020.

29. Bilandzic M, Rainczuk A, Green E, Fairweather N, Jobling TW, Plebanski M, Stephens AN. Keratin-14 (KRT14) Positive Leader Cells Mediate Mesothelial Clearance and Invasion by Ovarian Cancer Cells. Cancers. 2019;11(9):1228. doi: 10.3390/cancers11091228. PubMed PMID: 31443478.

30. Wu JS, Li ZF, Wang HF, Yu XH, Pang X, Wu JB, Wang SS, Zhang M, Yang X, Cao MX, Tang YJ, Liang XH, Zheng M, Tang YL. Cathepsin B defines leader cells during the collective invasion of salivary adenoid cystic carcinoma. Int J Oncol. 2019;54(4):1233-44. Epub 2019/04/11. doi: 10.3892/ijo.2019.4722. PubMed PMID: 30968153; PMCID: PMC6411368.

31. Leon G, MacDonagh L, Finn SP, Cuffe S, Barr MP. Cancer stem cells in drug resistant lung cancer: Targeting cell surface markers and signaling pathways. Pharmacology & Therapeutics. 2016;158:71-90. doi: https://doi.org/10.1016/j.pharmthera.2015.12.001.

32. Yang C, Cao M, Liu Y, He Y, Du Y, Zhang G, Gao F. Inducible formation of leader cells driven by CD44 switching gives rise to collective invasion and metastases in luminal breast carcinomas. Oncogene. 2019;38(46):7113-32. doi: 10.1038/s41388-019-0899-y.

33. Mehlen P, Puisieux A. Metastasis: a question of life or death. Nat Rev Cancer. 2006;6(6):449-58. doi: 10.1038/nrc1886. PubMed PMID: 16723991.

34. Efferth T, Leber M. Molecular principles of cancer invasion and metastasis (Review). International Journal of Oncology. 2009;34(4). doi: 10.3892/ijo_00000214.

35. Kabla AJ. Collective cell migration: leadership, invasion and segregation. J R Soc Interface. 2012;9(77):3268-78. doi: 10.1098/rsif.2012.0448. PubMed PMID: 22832363; PMCID: 3481577.

36. Friedl P, Locker J, Sahai E, Segall JE. Classifying collective cancer cell invasion. Nat Cell Biol. 2012;14(8):777-83.

37. Ewald AJ, Huebner RJ, Palsdottir H, Lee JK, Perez MJ, Jorgens DM, Tauscher AN, Cheung KJ, Werb Z, Auer M. Mammary collective cell migration involves transient loss of epithelial features and individual cell migration within the epithelium. J Cell Sci. 2012;125(Pt 11):2638-54. doi: 10.1242/jcs.096875. PubMed PMID: 22344263; PMCID: 3403234.

38. Friedl P, Noble PB, Walton PA, Laird DW, Chauvin PJ, Tabah RJ, Black M, Zanker KS. Migration of coordinated cell clusters in mesenchymal and epithelial cancer explants in vitro. Cancer Res. 1995;55(20):4557-60. PubMed PMID: 7553628.

39. Hegerfeldt Y, Tusch M, Bro E-b, Friedl P. Collective Cell Movement in Primary Melanoma Explants. Cancer Res. 2002;62:2125-30.

40. Nabeshima K, Inoue T, Shimao Y, Kataoka H, Koono M. Cohort migration of carcinoma cells : Differentiated colorectal carcinoma cells move as coherent cell clusters or sheets. Histol Histopathol. 1999;14:1183-97.

41. Leighton J, Kalla R, Turner J, Fennell R. Pathogenesis of Tumor Invasion II. Aggregate Replication. Cancer Research. 1960;20(575):586.

42. Friedl P. Prespecification and plasticity: shifting mechanisms of cell migration. Curr Opin Cell Biol. 2004;16(1):14-23. Epub 2004/03/24. doi: 10.1016/j.ceb.2003.11.001 S0955067403001571 [pii]. PubMed PMID: 15037300.

43. Alexander S, Koehl GE, Hirschberg M, Geissler EK, Friedl P. Dynamic imaging of cancer growth and invasion: a modified skin-fold chamber model. Histochem Cell Biol. 2008;130(6):1147-54. Epub 2008/11/07. doi: 10.1007/s00418-008-0529-1. PubMed PMID: 18987875.

44. Alexander S, Weigelin B, Winkler F, Friedl P. Preclinical intravital microscopy of the tumour-stroma interface: invasion, metastasis, and therapy response. Curr Opin Cell Biol. 2013;25(5):659-71. doi: 10.1016/j.ceb.2013.07.001. PubMed PMID: 23896198.

45. Friedl P, Wolf K. Plasticity of cell migration: a multiscale tuning model. J Cell Biol. 2010;188(1):11-9. doi: 10.1083/jcb.200909003. PubMed PMID: 19951899; PMCID: 2812848.

46. Haeger A, Krause M, Wolf K, Friedl P. Cell jamming: collective invasion of mesenchymal tumor cells imposed by tissue confinement. Biochim Biophys Acta. 2014;1840(8):2386-95. doi: 10.1016/j.bbagen.2014.03.020. PubMed PMID: 24721714.

47. van Zijl F, Krupitza G, Mikulits W. Initial steps of metastasis: cell invasion and endothelial transmigration. Mutat Res. 2011;728(1-2):23-34. doi: 10.1016/j.mrrev.2011.05.002. PubMed PMID: 21605699; PMCID: 4028085.

48. Pandya P, Orgaz JL, Sanz-Moreno V. Actomyosin contractility and collective migration: may the force be with you. Curr Opin Cell Biol. 2017;48:87-96. doi: 10.1016/j.ceb.2017.06.006. PubMed PMID: 28715714.

49. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, Gronroos E, Muhammad MA, Horswell S, Gerlinger M, Varela I, Jones D, Marshall J, Voet T, Van Loo P, Rassl DM, Rintoul RC, Janes SM, Lee SM, Forster M, Ahmad T, Lawrence D, Falzon M, Capitanio A, Harkins TT, Lee CC, Tom W, Teefe E, Chen SC, Begum S, Rabinowitz A, Phillimore B, Spencer-Dene B, Stamp G, Szallasi Z, Matthews N, Stewart A, Campbell P, Swanton C. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346(6206):251-6. Epub 2014/10/11. doi: 10.1126/science.1253462. PubMed PMID: 25301630; PMCID: PMC4636050.

50. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow CW, Cao Y, Gumbs C, Gold KA, Kalhor N, Little L, Mahadeshwar H, Moran C, Protopopov A, Sun H, Tang J, Wu X, Ye Y, William WN, Lee JJ, Heymach JV, Hong WK, Swisher S, Wistuba, II, Futreal PA. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014;346(6206):256-9. Epub 2014/10/11. doi: 10.1126/science.1256930. PubMed PMID: 25301631; PMCID: PMC4354858.

51. Waclaw B, Bozic I, Pittman ME, Hruban RH, Vogelstein B, Nowak MA. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature. 2015;525(7568):261-4. Epub 2015/08/27. doi: 10.1038/nature14971. PubMed PMID: 26308893; PMCID: PMC4782800.

52. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. Trends Cancer. 2016;2(1):49-63. Epub 2016/03/08. doi: 10.1016/j.trecan.2015.11.003. PubMed PMID: 26949746; PMCID: PMC4756277.

53. McPherson A, Roth A, Laks E, Masud T, Bashashati A, Zhang AW, Ha G, Biele J, Yap D, Wan A, Prentice LM, Khattra J, Smith MA, Nielsen CB, Mullaly SC, Kalloger S, Karnezis A, Shumansky K, Siu C, Rosner J, Chan HL, Ho J, Melnyk N, Senz J, Yang W, Moore R, Mungall AJ, Marra MA, Bouchard-Cote A, Gilks CB, Huntsman DG, McAlpine JN, Aparicio S, Shah SP. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nat Genet. 2016;48(7):758-67. Epub 2016/05/18. doi: 10.1038/ng.3573. PubMed PMID: 27182968.

54. Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal SA, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Griffin CA, Burton J, Swerdlow H, Quail MA, Stratton MR, Iacobuzio-Donahue C, Futreal PA. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010;467(7319):1109-13. Epub 2010/10/29. doi: 10.1038/nature09460. PubMed PMID: 20981101; PMCID: PMC3137369.

55. Rinner B, Galle B, Trajanoski S, Fischer C, Hatz M, Maierhofer T, Michelitsch G, Moinfar F, Stelzer I, Pfragner R, Guelly C. Molecular evidence for the bi-clonal origin of neuroendocrine tumor derived metastases. BMC Genomics. 2012;13:594. Epub 2012/11/07. doi: 10.1186/1471-2164-13-594. PubMed PMID: 23127113; PMCID: PMC3500212.

56. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, Begum S, McDonald NQ, Butler A, Jones D, Raine K, Latimer C, Santos CR, Nohadani M, Eklund AC, Spencer-Dene B, Clark G, Pickering L, Stamp G, Gore M, Szallasi Z, Downward J, Futreal PA, Swanton C. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883-92. Epub 2012/03/09. doi: 10.1056/NEJMoa1113205. PubMed PMID: 22397650; PMCID: PMC4878653.

57. Haeger A, Wolf K, Zegers MM, Friedl P. Collective cell migration: guidance principles and hierarchies. Trends Cell Biol. 2015;25(9):556-66. doi: 10.1016/j.tcb.2015.06.003. PubMed PMID: 26137890.

58. Bolger AM, Usadel B, Lohse M. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.

59. Dobin A, Davis CA, Zaleski C, Schlesinger F, Drenkow J, Chaisson M, Batut P, Jha S, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29(1):15-21. doi: 10.1093/bioinformatics/bts635.

60. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PMCID: PMC2723002.

61. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012;22(3):568-76.

62. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. Epub 2010/07/06. doi: 10.1093/nar/gkq603. PubMed PMID: 20601685; PMCID: PMC2938201.

63. Konen J, Wilkinson S, Lee B, Fu H, Zhou W, Jiang Y, Marcus AI. LKB1 kinase-dependent and -independent defects disrupt polarity and adhesion signaling to drive collagen remodeling during invasion. Molecular biology of the cell. 2016;27(7):mbc.E15-08-0569-1084. doi: 10.1091/mbc.E15-08-0569.

64. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PMCID: PMC3322381.

65. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118. Epub 2013/08/21. doi: 10.1371/journal.pcbi.1003118. PubMed PMID: 23950696; PMCID: PMC3738458.

66. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics. 2009;25(19):2607-8. Epub 2009/08/06. doi: 10.1093/bioinformatics/btp450. PubMed PMID: 19654119; PMCID: PMC2752612.

67. Morgan M, Pagès H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. 2017.

68. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria2017.

69. Team TBD. BSgenome.Hsapiens.UCSC.hg19: Full genome sequences for Homo sapiens (UCSC version hg19). 2014.

70. Goley ED, Welch MD. The ARP2/3 complex: an actin nucleator comes of age. Nature reviews Molecular cell biology. 2006;7(10):713-26. doi: 10.1038/nrm2026.

71. Zheng HC, Zheng YS, Li XH, Takahashi H, Hara T, Masuda S, Yang XH, Guan YF, Takano Y. Arp2/3 overexpression contributed to pathogenesis, growth and invasion of gastric carcinoma. Anticancer Research. 2008;28(4 B):2225-32.

72. Iwaya K, Oikawa K, Semba S, Tsuchiya B, Mukai Y, Otsubo T, Nagao T, Izumi M, Kuroda M, Domoto H, Mukai K. Correlation between liver metastasis of the colocalization of actin-related

protein 2 and 3 complex and WAVE2 in colorectal carcinoma. Cancer Sci. 2007;98(7):992-9. Epub 2007/04/27. doi: 10.1111/j.1349-7006.2007.00488.x. PubMed PMID: 17459058.

73. Lv J, Liu J, Xiao M, Xu H, Xu C, Zhang X, Tang L, Jiang F, Zhou Y, Zhang Z, Qu L, Lu C. ARP3 promotes tumor metastasis and predicts a poor prognosis in hepatocellular carcinoma. Pathol Res Pract. 2018;214(9):1356-61. Epub 2018/07/28. doi: 10.1016/j.prp.2018.05.028. PubMed PMID: 30049513.

74. Yang Z-L, Miao X, Xiong L, Zou Q, Yuan Y, Li J, Liang L, Chen M, Chen S. CFL1 and Arp3 are Biomarkers for Metastasis and Poor Prognosis of Squamous Cell/Adenosquamous Carcinomas and Adenocarcinomas of Gallbladder. Cancer Investigation. 2013;31(2):132-9. doi: 10.3109/07357907.2012.756113.

75. Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, Mani DR, Burgess MW, Gillette MA, Jaffe JD, Carr SA. Integrated proteomic analysis of post-translational modifications by serial enrichment. Nat Methods. 2013;10(7):634-7. doi: 10.1038/nmeth.2518. PubMed PMID: 23749302; PMCID: PMC3943163.

76. Wagner Sa, Beli P, Weinert BT, Nielsen ML, Cox J, Mann M, Choudhary C. A Proteome-wide, Quantitative Survey of In Vivo Ubiquitylation Sites Reveals Widespread Regulatory Roles. Molecular & Cellular Proteomics. 2011;10(10):M111.013284-M111. doi: 10.1074/mcp.M111.013284.

77. Wagner Sa, Beli P, Weinert BT, Scholz C, Kelstrup CD, Young C, Nielsen ML, Olsen JV, Brakebusch C, Choudhary C. Proteomic analyses reveal divergent ubiquitylation site patterns in murinetissues. Molecular & Cellular Proteomics. 2012:1578-85. doi: 10.1074/mcp.M112.017905.

78. Torres MP, Dewhurst H, Sundararaman N. Proteome-wide Structural Analysis of PTM Hotspots Reveals Regulatory Elements Predicted to Impact Biological Function and Disease. Molecular & Cellular Proteomics. 2016;15(11):3513. doi: 10.1074/mcp.M116.062331.

79. Qin JY, Zhang L, Clift KL, Hulur I, Xiang AP, Ren BZ, Lahn BT. Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. PLoS One. 2010;5(5):e10611. Epub 2010/05/21. doi: 10.1371/journal.pone.0010611. PubMed PMID: 20485554; PMCID: PMC2868906.

80. Bronsert P, Enderle-Ammour K, Bader M, Timme S, Kuehs M, Csanadi A, Kayser G, Kohler I, Bausch D, Hoeppner J, Hopt UT, Keck T, Stickeler E, Passlick B, Schilling O, Reiss CP, Vashist Y, Brabletz T, Berger J, Lotz J, Olesch J, Werner M, Wellner UF. Cancer cell invasion and EMT marker expression: a three-dimensional study of the human cancer-host interface. J Pathol. 2014;234(3):410-22. doi: 10.1002/path.4416. PubMed PMID: 25081610.

81. Mayor R, Etienne-Manneville S. The front and rear of collective cell migration. Nat Rev Mol Cell Biol. 2016;17(2):97-109. Epub 2016/01/05. doi: 10.1038/nrm.2015.14. PubMed PMID: 26726037.

82. Liu Z, Yang X, Chen C, Liu B, Ren B, Wang L, Zhao K, Yu S, Ming H. Expression of the Arp2/3 complex in human gliomas and its role in the migration and invasion of glioma cells. Oncol Rep. 2013;30(9):2127-36. doi: 10.3892/or.2013.2669.

83. Semba S, Iwaya K, Matsubayashi J, Serizawa H, Kataba H, Hirano T, Kato H, Matsuoka T, Mukai K. Coexpression of actin-related protein 2 and Wiskott-Aldrich syndrome family verproline-homologous protein 2 in adenocarcinoma of the lung. Clin Cancer Res. 2006;12(8):2449-54. Epub 2006/04/28. doi: 10.1158/1078-0432.CCR-05-2566. PubMed PMID: 16638851.

84. Kinoshita T, Nohata N, Watanabe-Takano H, Yoshino H, Hidaka H, Fujimura L, Fuse M, Yamasaki T, Enokida H, Nakagawa M, Hanazawa T, Okamoto Y, Seki N. Actin-related protein 2/3 complex subunit 5 (ARPC5) contributes to cell migration and invasion and is directly regulated by tumor-suppressive microRNA-133a in head and neck squamous cell carcinoma. Int J Oncol. 2012;40(6):1770-8. Epub 2012/03/02. doi: 10.3892/ijo.2012.1390. PubMed PMID: 22378351.

85. Jolly MK, Somarelli JA, Sheth M, Biddle A, Tripathi SC, Armstrong AJ, Hanash SM, Bapat SA, Rangarajan A, Levine H. Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. Pharmacol Ther. 2018. Epub 2018/10/01. doi: 10.1016/j.pharmthera.2018.09.007. PubMed PMID: 30268772.

86. Abercrombie M. Contact inhibition in tissue culture. In Vitro. 1970;6(2):128-42. Epub 1970/09/01. PubMed PMID: 4943054.

87. Abercrombie M, Heaysman JE. Observations on the social behaviour of cells in tissue culture. I. Speed of movement of chick heart fibroblasts in relation to their mutual contacts. Exp Cell Res. 1953;5(1):111-31. Epub 1953/09/01. PubMed PMID: 13083622.

88. Mayor R, Carmona-Fontaine C. Keeping in touch with contact inhibition of locomotion. Trends Cell Biol. 2010;20(6):319-28. Epub 2010/04/20. doi: 10.1016/j.tcb.2010.03.005. PubMed PMID: 20399659; PMCID: PMC2927909.

89. Theveneau E, Marchant L, Kuriyama S, Gull M, Moepps B, Parsons M, Mayor R. Collective chemotaxis requires contact-dependent cell polarity. Dev Cell. 2010;19(1):39-53. Epub 2010/07/21. doi: 10.1016/j.devcel.2010.06.012. PubMed PMID: 20643349; PMCID: PMC2913244.

90. Malet-Engra G, Yu W, Oldani A, Rey-Barroso J, Gov NS, Scita G, Dupre L. Collective cell motility promotes chemotactic prowess and resistance to chemorepulsion. Curr Biol. 2015;25(2):242-50. Epub 2015/01/13. doi: 10.1016/j.cub.2014.11.030. PubMed PMID: 25578904.

91. Dona E, Barry JD, Valentin G, Quirin C, Khmelinskii A, Kunze A, Durdu S, Newton LR, Fernandez-Minan A, Huber W, Knop M, Gilmour D. Directional tissue migration through a self-generated chemokine gradient. Nature. 2013;503(7475):285-9. Epub 2013/09/27. doi: 10.1038/nature12635. PubMed PMID: 24067609.

92. Valentin G, Haas P, Gilmour D. The chemokine SDF1a coordinates tissue migration through the spatially restricted activation of Cxcr7 and Cxcr4b. Curr Biol. 2007;17(12):1026-31. Epub 2007/06/16. doi: 10.1016/j.cub.2007.05.020. PubMed PMID: 17570670.

93. Muinonen-Martin AJ, Susanto O, Zhang Q, Smethurst E, Faller WJ, Veltman DM, Kalna G, Lindsay C, Bennett DC, Sansom OJ, Herd R, Jones R, Machesky LM, Wakelam MJ, Knecht DA, Insall RH. Melanoma cells break down LPA to establish local gradients that drive chemotactic dispersal. PLoS Biol. 2014;12(10):e1001966. Epub 2014/10/15. doi: 10.1371/journal.pbio.1001966. PubMed PMID: 25313567; PMCID: PMC4196730.

94. Nolen BJ, Tomasevic N, Russell A, Pierce DW, Jia Z, McCormick CD, Hartman J, Sakowicz R, Pollard TD. Characterization of two classes of small molecule inhibitors of Arp2/3 complex. Nature. 2009;460(7258):1031-4. Epub 2009/08/04. doi: 10.1038/nature08231. PubMed PMID: 19648907; PMCID: PMC2780427.

95. Ilatovskaya DV, Chubinskiy-Nadezhdin V, Pavlov TS, Shuyskiy LS, Tomilin V, Palygin O, Staruschenko A, Negulyaev YA. Arp2/3 complex inhibitors adversely affect actin cytoskeleton remodeling in the cultured murine kidney collecting duct M-1 cells. Cell Tissue Res. 2013;354(3):783-92. Epub 2013/09/17. doi: 10.1007/s00441-013-1710-y. PubMed PMID: 24036843; PMCID: PMC3850072.

96. Kazazian K, Go C, Wu H, Brashavitskaya O, Xu R, Dennis JW, Gingras A-C, Swallow CJ. Plk4 Promotes Cancer Invasion and Metastasis through Arp2/3 Complex Regulation of the Actin Cytoskeleton. Cancer Res. 2017;77(2):1-14. doi: 10.1158/0008-5472.CAN-16-2060.

97. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(37):E5528-E37. doi: 10.1073/pnas.1522203113 %J Proceedings of the National Academy of Sciences.

98. Sun R, Hu Z, Sottoriva A, Graham TA, Harpak A, Ma Z, Fischer JM, Shibata D, Curtis C. Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nat Genet. 2017;49(7):1015-24. Epub 2017/06/06. doi: 10.1038/ng.3891. PubMed PMID: 28581503; PMCID: PMC5643198.

99. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell. 2015;27(1):15-26. doi: 10.1016/j.ccell.2014.12.001.

100. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338-45. Epub 2013/09/21. doi: 10.1038/nature12625. PubMed PMID: 24048066.

101. Gerlinger M, McGranahan N, Dewhurst SM, Burrell RA, Tomlinson I, Swanton C. Cancer: evolution within a lifetime. Annu Rev Genet. 2014;48:215-36. Epub 2014/10/09. doi: 10.1146/annurev-genet-120213-092314. PubMed PMID: 25292359.

102. Caswell DR, Swanton C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. BMC Med. 2017;15(1):133. Epub 2017/07/19. doi: 10.1186/s12916-017-0900-y. PubMed PMID: 28716075; PMCID: PMC5514532.

103. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, Salm M, Horswell S, Escudero M, Matthews N, Rowan A, Chambers T, Moore DA, Turajlic S, Xu H, Lee SM, Forster MD, Ahmad T, Hiley CT, Abbosh C, Falzon M, Borg E, Marafioti T, Lawrence D, Hayward M, Kolvekar S, Panagiotopoulos N, Janes SM, Thakrar R, Ahmed A, Blackhall F, Summers Y, Shah R, Joseph L, Quinn AM, Crosbie PA, Naidu B, Middleton G, Langman G, Trotter S, Nicolson M, Remmen H, Kerr K, Chetty M, Gomersall L, Fennell DA, Nakas A, Rathinam S, Anand G, Khan S, Russell P, Ezhil V, Ismail B, Irvin-Sellers M, Prakash V, Lester JF, Kornaszewska M, Attanoos R, Adams H, Davies H, Dentro S, Taniere P, O'Sullivan B, Lowe HL, Hartley JA, Iles N, Bell H, Ngai Y, Shaw JA, Herrero J, Szallasi Z, Schwarz

RF, Stewart A, Quezada SA, Le Quesne J, Van Loo P, Dive C, Hackshaw A, Swanton C, Consortium TR. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med. 2017;376(22):2109-21. Epub 2017/04/27. doi: 10.1056/NEJMoa1616288. PubMed PMID: 28445112.

104. Zoeller EL, Pedro B, Konen J, Dwivedi B, Rupji M, Sundararaman N, Wang L, Horton JR, Zhong C, Barwick BG, Cheng X, Martinez ED, Torres MP, Kowalski J, Marcus AI, Vertino PM. Genetic heterogeneity within collective invasion packs drives leader and follower cell phenotypes. Journal of Cell Science. 2019;132(19):jcs231514. doi: 10.1242/jcs.231514.

105. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. 2012;2(5):401. doi: 10.1158/2159-8290.CD-12-0095.

106. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science Signaling. 2013;6(269):pl1. doi: 10.1126/scisignal.2004088.

107. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40. Epub 2009/11/17. doi: 10.1093/bioinformatics/btp616. PubMed PMID: 19910308; PMCID: PMC2796818.

108. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. Epub 2015/01/22. doi: 10.1093/nar/gkv007. PubMed PMID: 25605792; PMCID: PMC4402510.

109. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based

approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545. doi: 10.1073/pnas.0506580102.

110. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2017;46(D1):D649-D55. doi: 10.1093/nar/gkx1132.

111. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29(1):308-11. doi: 10.1093/nar/29.1.308.

112. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, Akbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta TM, Cancer Genome Atlas N, Stuart JM, Benz CC, Laird PW. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173(2):291-304 e6. Epub 2018/04/07. doi: 10.1016/j.cell.2018.03.022. PubMed PMID: 29625048; PMCID: PMC5957518.

113. Colli LM, Machiela MJ, Myers TA, Jessop L, Yu K, Chanock SJ. Burden of Nonsynonymous Mutations among TCGA Cancers and Candidate Immune Checkpoint Inhibitor Responses. Cancer Res. 2016;76(13):3767-72. Epub 2016/05/20. doi: 10.1158/0008-5472.CAN-16-0170. PubMed PMID: 27197178; PMCID: PMC4930685.

114. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20. Epub 2013/09/28. doi: 10.1038/ng.2764. PubMed PMID: 24071849; PMCID: PMC3919969.

115. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267-73. doi: 10.1038/ng1180.

116. Shedden K, Taylor JMG, Enkemann SA, Tsao M-S, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature Medicine. 2008;14:822. doi: 10.1038/nm.1790.

117. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, Martinez P, Phillimore B, Begum S, Rabinowitz A, Spencer-Dene B, Gulati S, Bates PA, Stamp G, Pickering L, Gore M, Nicol DL, Hazell S, Futreal PA, Stewart A, Swanton C. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet. 2014;46(3):225-33. Epub 2014/02/04. doi: 10.1038/ng.2891. PubMed PMID: 24487277; PMCID: PMC4636053.

118. Fabris VT. From chromosomal abnormalities to the identification of target genes in mouse models of breast cancer. Cancer Genet. 2014;207(6):233-46. Epub 2014/09/02. doi: 10.1016/j.cancergen.2014.06.025. PubMed PMID: 25176624.

119. Cleton-Jansen AM, van Eijk R, Lombaerts M, Schmidt MK, Van't Veer LJ, Philippo K, Zimmerman RM, Peterse JL, Smit VT, van Wezel T, Cornelisse CJ. ATBF1 and NQO1 as candidate targets for allelic loss at chromosome arm 16q in breast cancer: absence of somatic ATBF1 mutations

and no role for the C609T NQO1 polymorphism. BMC cancer. 2008;8:105. Epub 2008/04/18. doi: 10.1186/1471-2407-8-105. PubMed PMID: 18416817; PMCID: PMC2377272.

120. Kluth M, Jung S, Habib O, Eshagzaiy M, Heinl A, Masser S, Mader M, Runte F, Barow P, Korbel J, Steurer S, Krech T, Huland H. Deletion lengthening at chromosomes 6q and 16q targets multiple tumor suppressor genes and is associated with an increasingly poor prognosis in prostate cancer2017;8(65):108923-35.

121. Shen C, Wang X, Tian L, Che G. Microsatellite alteration in multiple primary lung cancer. Journal of thoracic disease. 2014;6(10):1499-505. doi: 10.3978/j.issn.2072-1439.2014.09.14. PubMed PMID: 25364529.

122. Fong KM, Zimmerman PV, Smith PJ. Microsatellite instability and other molecular abnormalities in non-small cell lung cancer. Cancer research. 1995;55(1):28-30. PubMed PMID: 7805035.

123. Woo HG, Park ES, Cheon JH, Kim JH, Lee J-S, Park BJ, Kim W, Park SC, Chung YJ, Kim BG, Yoon J-H, Lee H-S, Kim CY, Yi N-J, Suh K-S, Lee KU, Chu I-S, Roskams T, Thorgeirsson SS, Kim YJ. Gene Expression–Based Recurrence Prediction of Hepatitis B Virus–Related Human Hepatocellular Carcinoma. Clin Cancer Res. 2008;14(7):2056. doi: 10.1158/1078-0432.CCR-07-1473.

124. Villanueva A, Hoshida Y, Battiston C, Tovar V, Sia D, Alsinet C, Cornella H, Liberzon A, Kobayashi M, Kumada H, Thung SN, Bruix J, Newell P, April C, Fan JB, Roayaie S, Mazzaferro V, Schwartz ME, Llovet JM. Combining Clinical, Pathology, and Gene Expression Data to Predict Recurrence of Hepatocellular Carcinoma. Gastroenterology. 2011;140(5):1501-12.e2. doi: 10.1053/j.gastro.2011.02.006.

125. Lee J-S, Chu I-S, Heo J, Calvisi DF, Sun Z, Roskams T, Durnez A, Demetris AJ, Thorgeirsson SS. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. Hepatology. 2004;40(3):667-76. doi: 10.1002/hep.20375.

126. Matikas A, Syrigos KN, Agelaki S. Circulating Biomarkers in Non-Small-Cell Lung Cancer: Current Status and Future Challenges. Clin Lung Cancer. 2016;17(6):507-16. Epub 2016/10/23. doi: 10.1016/j.cllc.2016.05.021. PubMed PMID: 27373516.

127. Zhang YC, Zhou Q, Wu YL. The emerging roles of NGS-based liquid biopsy in non-small cell lung cancer. J Hematol Oncol. 2017;10(1):167. Epub 2017/10/25. doi: 10.1186/s13045-017-0536-6. PubMed PMID: 29061113; PMCID: PMC5654124.

128. Sorber L, Zwaenepoel K, Deschoolmeester V, Van Schil PE, Van Meerbeeck J, Lardon F, Rolfo C, Pauwels P. Circulating cell-free nucleic acids and platelets as a liquid biopsy in the provision of personalized therapy for lung cancer patients. Lung Cancer. 2017;107:100-7. Epub 2016/05/18. doi: 10.1016/j.lungcan.2016.04.026. PubMed PMID: 27180141.

129. Pedro B, Rupji M, Dwivedi B, Kowalski J, Konen J, Owonikoko TK, Ramalingam SS, Vertino PM, Marcus AI. Prognostic significance of an invasive leader cell-derived mutation cluster on chromosome 16q. Cancer. In Press.

130. Zhang J, Goliwas KF, Wang W, Taufalele PV, Bordeleau F, Reinhart-King CA. Energetic regulation of coordinated leader-follower dynamics during collective invasion of breast cancer cells. Proceedings of the National Academy of Sciences of the United States of America. 2019;116(16):7867-72. Epub 2019/03/28. doi: 10.1073/pnas.1809964116. PubMed PMID: 30923113.

131. Libanje F, Raingeaud J, Luan R, Thomas Z, Zajac O, Veiga J, Marisa L, Adam J, Boige V, Malka D, Goéré D, Hall A, Soazec J-Y, Prall F, Gelli M, Dartigues P, Jaulin F. ROCK2 inhibition triggers the collective invasion of colorectal adenocarcinomas. The EMBO journal. 2019;38(14):e99299-e. Epub 2019/06/18. doi: 10.15252/embj.201899299. PubMed PMID: 31304629.

132. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018;36(5):411-20. doi: 10.1038/nbt.4096.

133. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5:2122. Epub 2016/12/06. doi: 10.12688/f1000research.9501.2. PubMed PMID: 27909575; PMCID: PMC5112579.

134. Rupji M, Dwivedi B, Kowalski J. NOJAH: NOt Just Another Heatmap for genome-wide cluster analysis. PLOS ONE. 2019;14(3):e0204542. doi: 10.1371/journal.pone.0204542.

135. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179-86. doi: 10.1093/bioinformatics/btw777.

136. Menendez MT, Teygong C, Wade K, Florimond C, Blader IJ. siRNA Screening Identifies the Host Hexokinase 2 (HK2) Gene as an Important Hypoxia-Inducible Transcription Factor 1 (HIF-1) Target Gene in Toxoplasma gondii-Infected Cells. mBio. 2015;6(3):e00462-15. doi: 10.1128/mBio.00462-15.

137. Roudnicky F, Dieterich LC, Poyet C, Buser L, Wild P, Tang D, Camenzind P, Ho CH, Otto VI, Detmar M. High expression of insulin receptor on tumour-associated blood vessels in invasive bladder cancer predicts poor overall and progression-free survival. The Journal of Pathology. 2017;242(2):193-205. doi: 10.1002/path.4892.

138. Liu T, Laurell C, Selivanova G, Lundeberg J, Nilsson P, Wiman KG. Hypoxia induces p53-dependent transactivation and Fas/CD95-dependent apoptosis. Cell Death & Differentiation. 2007;14(3):411-21. doi: 10.1038/sj.cdd.4402022.

139. Krtolica A, Krucher NA, Ludlow JW. Hypoxia-induced pRB hypophosphorylation results from downregulation of CDK and upregulation of PP1 activities. Oncogene. 1998;17(18):2295-304. doi: 10.1038/sj.onc.1202159.

140. Fico F, Bousquenaud M, Ruegg C, Santamaria-Martinez A. Breast Cancer Stem Cells with Tumor- versus Metastasis-Initiating Capacities Are Modulated by TGFBR1 Inhibition. Stem Cell

Reports. 2019;13(1):1-9. Epub 2019/07/02. doi: 10.1016/j.stemcr.2019.05.026. PubMed PMID: 31257133; PMCID: PMC6626885.

141. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93(4):641-51. Epub 2013/10/01. doi: 10.1016/j.ajhg.2013.08.008. PubMed PMID: 24075185; PMCID: PMC3791257.

142. Fasterius E, Uhlen M, Al-Khalili Szigyarto C. Single-cell RNA-seq variant analysis for exploration of genetic heterogeneity in cancer. Sci Rep. 2019;9(1):9524. Epub 2019/07/04. doi: 10.1038/s41598-019-45934-1. PubMed PMID: 31267007; PMCID: PMC6606766.

143. Davis-Marcisak EF, Sherman TD, Orugunta P, Stein-O'Brien GL, Puram SV, Roussos Torres ET, Hopkins AC, Jaffee EM, Favorov AV, Afsari B, Goff LA, Fertig EJ. Differential Variation Analysis Enables Detection of Tumor Heterogeneity Using Single-Cell RNA-Sequencing Data. Cancer Res. 2019;79(19):5102-12. Epub 2019/07/25. doi: 10.1158/0008-5472.CAN-18-3882. PubMed PMID: 31337651; PMCID: PMC6844448.

144. Zhang C, He H, Hu X, Liu A, Huang D, Xu Y, Chen L, Xu D. Development and validation of a metastasis-associated prognostic signature based on single-cell RNA-seq in clear cell renal cell carcinoma. Aging. 2019;11(22):10183-202. Epub 2019/11/20. doi: 10.18632/aging.102434. PubMed PMID: 31747386.

145. Ocasio J, Babcock B, Malawsky D, Weir SJ, Loo L, Simon JM, Zylka MJ, Hwang D, Dismuke T, Sokolsky M, Rosen EP, Vibhakar R, Zhang J, Saulnier O, Vladoiu M, El-Hamamy I, Stein LD, Taylor MD, Smith KS, Northcott PA, Colaneri A, Wilhelmsen K, Gershon TR. scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy. Nature communications. 2019;10(1):5829-. doi: 10.1038/s41467-019-13657-6. PubMed PMID: 31863004.

146. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nature communications. 2018;9(1):3588-. doi: 10.1038/s41467-018-06052-0. PubMed PMID: 30181541.

147. Peng J, Sun B-F, Chen C-Y, Zhou J-Y, Chen Y-S, Chen H, Liu L, Huang D, Jiang J, Cui G-S, Yang Y, Wang W, Guo D, Dai M, Guo J, Zhang T, Liao Q, Liu Y, Zhao Y-L, Han D-L, Zhao Y, Yang Y-G, Wu W. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell research. 2019;29(9):725-38. Epub 2019/07/04. doi: 10.1038/s41422-019-0195-y. PubMed PMID: 31273297.

148. Quan Q, Wang X, Lu C, Ma W, Wang Y, Xia G, Wang C, Yang G. Cancer stem-like cells with hybrid epithelial/mesenchymal phenotype leading the collective invasion. Cancer Science. 2019;n/a(n/a). doi: 10.1111/cas.14285.

149. Riether C, Schürch CM, Bührer ED, Hinterbrandner M, Huguenin A-L, Hoepner S, Zlobec I, Pabst T, Radpour R, Ochsenbein AF. CD70/CD27 signaling promotes blast stemness and is a viable therapeutic target in acute myeloid leukemia. The Journal of experimental medicine. 2017;214(2):359-80. Epub 2016/12/28. doi: 10.1084/jem.20152008. PubMed PMID: 28031480.

150. Liu L, Yin B, Yi Z, Liu X, Hu Z, Gao W, Yu H, Li Q. Breast cancer stem cells characterized by CD70 expression preferentially metastasize to the lungs. Breast cancer (Tokyo, Japan). 2018;25(6):706-16. Epub 2018/06/14. doi: 10.1007/s12282-018-0880-6. PubMed PMID: 29948958.

151. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, Sher X, Liu XQ, Lu H, Nebozhyn M, Zhang C, Lunceford JK, Joe A, Cheng J, Webber AL, Ibrahim N, Plimack ER, Ott PA, Seiwert TY, Ribas A, McClanahan TK, Tomassini JE, Loboda A, Kaufman D. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science (New York, NY). 2018;362(6411):eaar3593. doi: 10.1126/science.aar3593. PubMed PMID: 30309915.

152. Gandara DR, Paul SM, Kowanetz M, Schleifman E, Zou W, Li Y, Rittmeyer A, Fehrenbacher L, Otto G, Malboeuf C, Lieber DS, Lipson D, Silterra J, Amler L, Riehl T, Cummings CA, Hegde PS, Sandler A, Ballinger M, Fabrizio D, Mok T, Shames DS. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. Nature medicine. 2018;24(9):1441-8. Epub 2018/08/06. doi: 10.1038/s41591-018-0134-3. PubMed PMID: 30082870.

153. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, Chan TA. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science. 2015;348(6230):124. doi: 10.1126/science.aaa1348.

154. Xiao W, Gao Z, Duan Y, Yuan W, Ke Y. Notch signaling plays a crucial role in cancer stem-like cells maintaining stemness and mediating chemotaxis in renal cell carcinoma. Journal of experimental & clinical cancer research : CR. 2017;36(1):41-. doi: 10.1186/s13046-017-0507-3. PubMed PMID: 28279221.

155. Hassan KA, Wang L, Korkaya H, Chen G, Maillard I, Beer DG, Kalemkerian GP, Wicha MS. Notch pathway activity identifies cells with cancer stem cell-like properties and correlates with worse survival in lung adenocarcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research. 2013;19(8):1972-80. Epub 2013/02/26. doi: 10.1158/1078-0432.CCR-12-0370. PubMed PMID: 23444212.

156. Zhao J. Cancer stem cells and chemoresistance: The smartest survives the raid. Pharmacology & therapeutics. 2016;160:145-58. Epub 2016/02/17. doi: 10.1016/j.pharmthera.2016.02.008. PubMed PMID: 26899500.

157. Haque S, Morris JC. Transforming growth factor-β: A therapeutic target for cancer. Human vaccines & immunotherapeutics. 2017;13(8):1741-50. Epub 2017/06/02. doi: 10.1080/21645515.2017.1327107. PubMed PMID: 28575585.