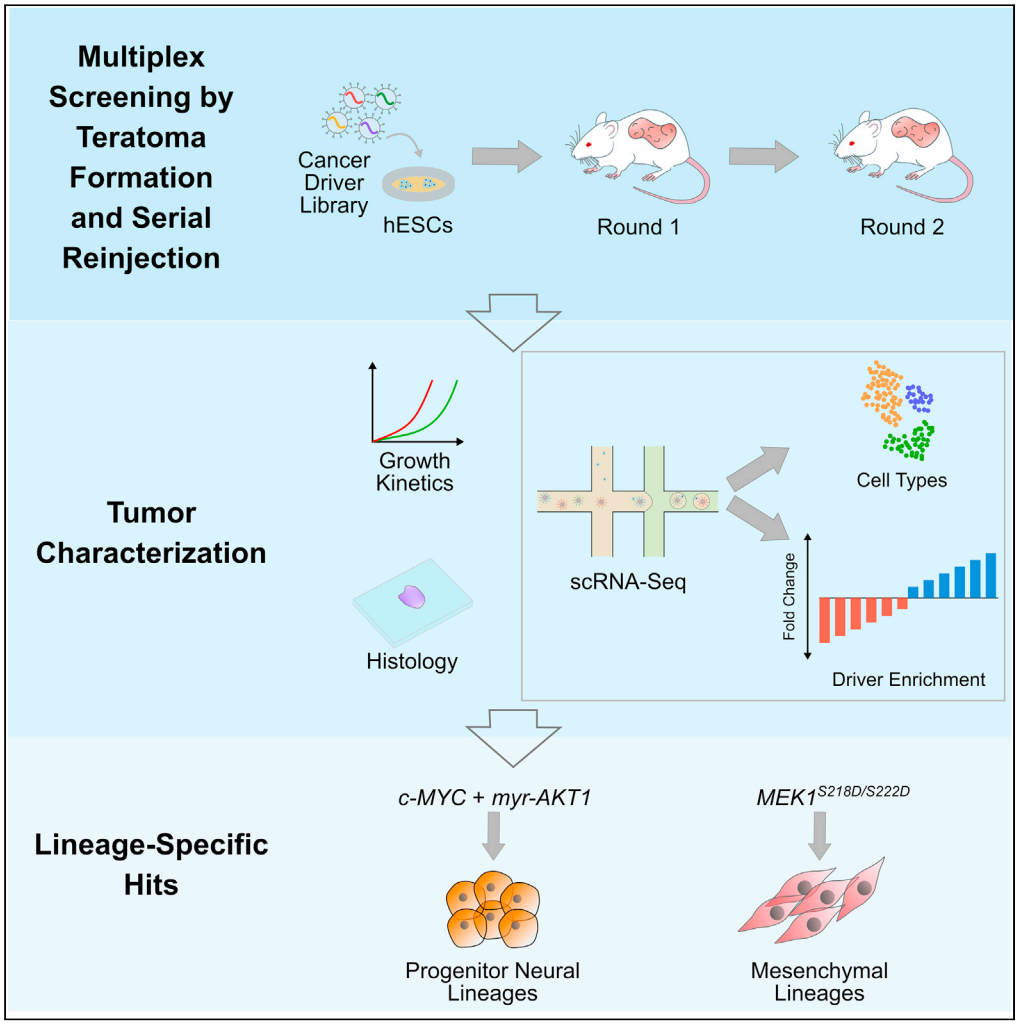
iScience

CellPress
OPEN ACCESS

Article

Charting oncogenicity of genes and variants across lineages via multiplexed screens in teratomas



Udit Parekh, Daniella McDonald, Amir Dailamy, ..., Ann Tipps, Christian Metallo, Prashant Mali

pmali@ucsd.edu

Highlights

Developed multiplex in vivo screens of cancer driver genes across multiple lineages

Couples teratoma differentiation, scRNA-seq readout and tumor serial injection

c-MYC alone or with myristoylated *AKT1* drives neural progenitor proliferation

Mutant *MEK1*$^{S218D/S222D}$ enhances fitness of mesenchymal lineages like fibroblasts

Parekh et al., iScience *24*, 103149
October 22, 2021
https://doi.org/10.1016/j.isci.2021.103149

iScience

CellPress
OPEN ACCESS

Article

Charting oncogenicity of genes and variants across lineages via multiplexed screens in teratomas

Udit Parekh,[1] Daniella McDonald,[2,3] Amir Dailamy,[2] Yan Wu,[2] Thekla Cordes,[2] Kun Zhang,[2] Ann Tipps,[4] Christian Metallo,[2,5] and Prashant Mali[2,6,*]

SUMMARY

Deconstructing tissue-specific effects of genes and variants on proliferation is critical to understanding cellular transformation and systematically selecting cancer therapeutics. This requires scalable methods for multiplexed genetic screens tracking fitness across time, across lineages, and in a suitable niche, since physiological cues influence functional differences. Towards this, we present an approach, coupling single-cell cancer driver screens in teratomas with hit enrichment by serial teratoma reinjection, to simultaneously screen drivers across multiple lineages in vivo. Using this system, we analyzed population shifts and lineage-specific enrichment for 51 cancer associated genes and variants, profiling over 100,000 cells spanning over 20 lineages, across two rounds of serial reinjection. We confirmed that *c-MYC* alone or combined with myristoylated *AKT1* potently drives proliferation in progenitor neural lineages, demonstrating signatures of malignancy. Additionally, mutant *MEK1*$^{S218D/S222D}$ provides a proliferative advantage in mesenchymal lineages like fibroblasts. Our method provides a powerful platform for multi-lineage longitudinal study of oncogenesis.

INTRODUCTION

The onset of cancer is often posited to be an evolutionary process enabled by the acquisition of somatic mutations and genetic lesions over time (Greaves and Maley, 2012; Merlo et al., 2006). Yet, these mutations lead to survival advantages and cancer only if accumulated in the relevant types of cells at the appropriate stage of differentiation, at the appropriate time and in the appropriate cell state (Hagis et al., 2019; Schneider et al., 2017; Sottoriva et al., 2015). Understanding this process of tumorigenesis and dissecting the tissue-specific molecular mechanisms which govern neoplastic transformation is a longstanding goal in cancer biology. In particular, with the explosion in cancer genome sequencing data (Alexandrov et al., 2013; Bailey et al., 2018; Beroukhim et al., 2010; Forbes et al., 2017; Sanchez-Vega et al., 2018; Zack et al., 2013), understanding the tissue-specific oncogenicity of the growing list of genetic variants of unknown significance may lead to significant advances in building early detection systems, which would improve patient outcomes (Etzioni et al., 2003), as well as inform the development and application of therapeutic strategies (Balani et al., 2017).

In this regard, xenograft and genetically engineered animal models (Cheon and Orsulic, 2011; Gould et al., 2015; Richmond and Yingjun, 2008) have been especially useful to recapitulate the process by which healthy cells undergo transformation, but such models often do not completely capture human biology, transformation or tumorigenesis (Cheon and Orsulic, 2011; Fischer, 2021; Gould et al., 2015; Rangarajan and Weinberg, 2003; Richmond and Yingjun, 2008). On the other hand, immortalized cell lines and primary cells have been important workhorses of cancer research (Gillet et al., 2013; Wilding and Bodmer, 2014), aiding in mechanistic studies, the uncovering of therapeutic vulnerabilities and understanding resistance mechanisms (Martz et al., 2014). Yet, elucidating the wide spectrum of tissue-specific programs governing transformation (Boehm et al., 2005; Clark et al., 1988; Daley et al., 1987; Elenbaas et al., 2001; Geder et al., 1976; Hahn et al., 1999; Park et al., 2018; Rangarajan et al., 2004; Sasaki et al., 2009) remains a challenge (Sack et al., 2018), especially as access to and culture of diverse types of primary cells is challenging. Such in vitro systems also exclude the environmental and physiological context which are key modulators of driver-specificity, and often lack the ability to perturb cells along their differentiation trajectories or in distinct states, an important factor in driver-specific transformation (Puisieux et al., 2018).

[1]Department of Electrical and Computer Engineering, University of California San Diego, San Diego, USA

[2]Department of Bioengineering, University of California San Diego, San Diego, USA

[3]Biomedical Sciences Graduate Program, University of California San Diego, San Diego, USA

[4]School of Medicine, University of California San Diego, San Diego, USA

[5]Salk Institute of Biological Studies, La Jolla, USA

[6]Lead contact

*Correspondence: pmali@ucsd.edu

https://doi.org/10.1016/j.isci.2021.103149

Systems which allow us to ethically gain access to developing human tissue models, recapitulating the architecture and signaling programs of native tissue, in a suitable physiological niche could be an invaluable tool to deconstruct tissue-specific drivers of neoplastic transformation. In this regard, there has been significant work toward the use of directed differentiation of stem cells (Smith and Tabar, 2019) in 2-dimensional monolayer culture as well as into organoids (Bian et al., 2018; Drost et al., 2015, 2017; Fumagalli et al., 2017; Lannagan et al., 2019; Li et al., 2014b; Matano et al., 2015; Rosenbluth et al., 2020) to model cancer. These models capture various cell states along the developmental trajectory, while organoids also capture the diversity of cells present in specific tissue, organized in native-like architecture. However, these systems are cultured in specialized conditions, which may not represent in vivo settings, and lack vasculature which is a central characteristic of cancers (Hanahan and Weinberg, 2011). In vivo engraftment of suitably differentiated human stem cells provides an avenue to study the human-specific dynamics of oncogenesis and cancer evolution in an appropriate milieu (Duan et al., 2015; Koga et al., 2020; Pei et al., 2012, 2016). But, each of these systems provides access to a single or a few lineages at a time, thus, to test drivers in panoply of lineages to assess their oncogenic potential in each is a laborious and slow process.

Recently, our group demonstrated that human pluripotent stem cell (hPSC) derived teratomas, which are typically benign tumors with differentiated cells from all three germ layers and regions of organized tissue-like architecture (Lensch et al., 2007), can enable high throughput genetic screens simultaneously across cell types of all germ layers (McDonald et al., 2020). Leveraging this, we propose here a method that uses the teratoma model, in combination with single cell RNA-sequencing (scRNA-seq) based open reading frame (ORF) overexpression screens (Parekh et al., 2018) and serial tumor propagation, to massively assess the tissue-specific oncogenicity of genes and gene variants in parallel across a diverse set of lineages. Specifically, to modify and adapt the teratoma platform for screens of oncogenicity, we added the critical element of serial tumor propagation coupled to single cell profiling of the serially reinjected tumors. This enabled rich, longitudinal assessment of tumor evolution. In addition, the serial injection process, not only enabled the enrichment of top hits, but also allowed for the development of fully transformed states with the hallmarks of malignancy and allowed the multi-parameter assessment of oncogenicity and tumor evolution via fitness, transcriptomic profiling, growth kinetics, and histology. Furthermore, the open reading frame based vectors allowed us to explore the space of mutants and variants in a facile manner and also enabled extension to other important driver categories such as fusions. Taken together, this versatile, modular methodology enabled us to adapt the teratoma platform to oncogenic screens enabling de novo creation of lineage-restricted transformed phenotypes from previously normal cells in an in vivo niche.

RESULTS

Design of a multiplex in vivo screening platform

To enable a screening platform which would simultaneously allow the determination of lineages along with detection of perturbations, we implemented pooled overexpression screens in teratomas with scRNA-seq readout, using a lentiviral overexpression vector we previously developed for compatibility with scRNA-seq based pooled genetic screens (Parekh et al., 2018) (Figure 1A). The vector contains a unique 20 bp barcode for each library element, located 200 bp upstream of the lentiviral 3′ long terminal repeat (LTR). This yields a polyadenylated transcript with the barcode proximal to the 3′ end, thus allowing detection in droplet based scRNA-seq systems which rely on poly-A capture. Each cancer driver (Figure 1B and Table S1) was cloned into the lentiviral backbone vector, packaged individually into lentivirus particles and this individually packaged lentivirus then combined, to avoid barcode shuffling due to the recombination driven template switching inherent in pooled lentiviral packaging (Hill et al., 2018; Parekh et al., 2018; Sack et al., 2016; Xie et al., 2018). To allow for combinatorial driver transduction, these driver libraries were then transduced into H1 human embryonic stem cells (hESCs), which we had previously characterized to be of normal karyotype (McDonald et al., 2020). The hESCs were transduced at the highest titer where we did not see visible morphological changes, and maintained under antibiotic selection for 4 days after transduction (Figures 1C and S1, STAR Methods). In addition, for cells profiled prior to injection by scRNA-seq, for the majority of drivers we observed no significant changes per cluster in the proportion of cells expressing drivers as compared to those expressing the internal negative control (Figure S1), confirming that the driver library was not significantly perturbing the hESCs from their basal state.

To form teratomas with these library-transduced hESCs, we subcutaneously injected 6–8 million of these cells suspended in a 1:1 mixture of Matrigel and the pluripotent stem cell medium mTeSR1 in the right flank

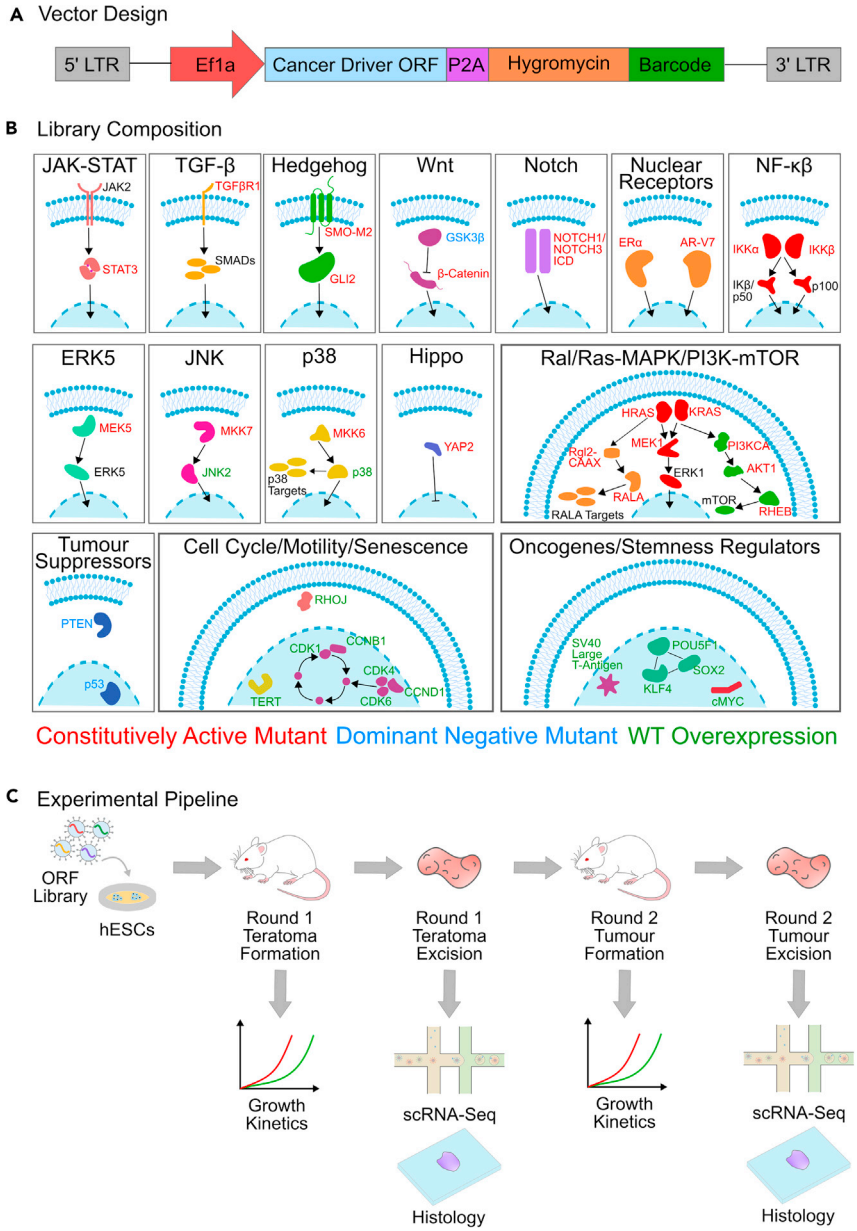


Figure 1. Overview of experimental design and library construction

(A) Schematic of lentiviral overexpression vector.

(B) Composition of the cancer driver ORF library, encompassing major signaling pathways, oncogenes and stemness regulators involved in oncogenesis and cancer progression.

(C) Schematic of experimental framework for evaluation of effects of cancer driver overexpression in developing teratomas and reinjected tumors: Individual cancer driver ORFs are cloned into the barcoded ORF overexpression vector, packaged into lentivirus then pooled for transduction of hESCs. Transduced cells are then injected subcutaneously into $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice to form the round 1 teratomas. Teratoma growth is assayed by caliper-based measurement of approximate elliptical area of the tumor. Once the teratoma size reaches a threshold value, teratomas are excised and assayed by single-cell RNA-seq and histology, and a fraction of cells are reinjected in $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice to form round 2 tumors for further enrichment of drivers. Round 2 tumor growth is also monitored via caliper-based measurements and at the endpoint tumors are assayed via scRNA-seq and histology.

of anesthetized $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice (Figure 1C, STAR Methods). As growth controls, wild type, unmodified H1 hESCs were similarly injected in a separate group of $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice. The growth of all teratomas was monitored via weekly caliper-based measurements of elliptical area

(Figure 1C, STAR Methods). Once the teratomas reached a threshold size for excision, they were extracted for downstream processing. After extraction, tumors were weighed and measured, and representative pieces were frozen for cryosectioning and histological analysis. The remaining pieces were dissociated into single cell suspensions. A part of this suspension was processed for single cell RNA-sequencing (scRNA-seq) using the droplet based 10X Genomics Chromium system, while a second part consisting of 6–8 million cells was reinjected into $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice for serial proliferation to further enrich transformed lineages and drivers of proliferative advantage. These reinjected round 2 tumors were monitored and processed similarly to the initially formed round 1 teratomas (Figure 1C, STAR Methods), to obtain histology information as well as transcriptomic profiles via scRNA-seq.

Deconstructing fitness effects across lineages

Using this method, we screened 51 genes and gene variants representing major signaling pathways and oncogenes (Figure 1B and Table S1) (Martz et al., 2014) involved in oncogenesis, tumor proliferation, survival, cell cycle regulation, and stemness regulation (Figure 1C).

To screen these drivers, we generated four teratomas from the perturbed hPSCs. In these first-round driver library teratomas, we observed palpable and measurable tumors between 20 and 27 days after injection, which grew to a size sufficient for extraction between 41 and 60 days after injection. In comparison, out of eight teratomas formed from unperturbed wild type cells, one did not form a tumor in 90 days of monitoring, whereas the seven remaining teratomas were palpable and measurable between 27 and 39 days after injection and six of them grew to a size sufficient for extraction between 55 and 78 days after injection (Figure 2A).

Once these round 1 driver library teratomas were excised and processed via scRNA-seq, gene expression matrices were generated from the resultant data using 10X Genomics cellranger. We then used these expression matrices with the Seurat pipeline (Butler et al., 2018) to integrate data from all four driver library teratomas and compensate for batch effects. Using this integrated data matrix, cells were then clustered using a shared nearest neighbor algorithm. Cell types were classified by the Seurat label transfer process using a previously classified set of teratomas generated from wild type H1 hESCs (McDonald et al., 2020) as a reference. Clusters were projected as a uniform manifold approximation and projection (UMAP) scatterplot (Figure 2B, STAR Methods), with cell types well distributed across all the teratomas (Figure S2A). 17 out of 23 cell types detected in wild type teratomas were detected in the driver library teratomas.

Drivers were assigned to individual cells by amplifying paired scRNA-seq cell barcodes and library barcodes from the unfragmented scRNA-seq cDNA, enabling genotyping of each cell (Parekh et al., 2018) (STAR Methods). Barcodes were detectably expressed in nearly 45% of cells in this round, and we hypothesize that stochastic silencing of the lentiviral cassette and potentially sparse capture during scRNA-seq may be leading to barcode association for a fraction of the captured cells. 54% of these genotyped cells expressed a single detectable barcode while the remaining genotyped cells expressed two or more barcodes (Figures S2B–S2D).

A subset of the dissociated cells from each of the round 1 teratomas were reinjected to grow round 2 tumors (Figure 1B, STAR Methods), to further enrich dominant drivers and lineages. From these four reinjected tumors, two tumors were processed via scRNAseq and histology. Three of the fastest growing wild type teratomas were also dissociated into single cell suspensions and reinjected to form round 2 control tumors.

For the two driver library tumors which were processed, tumors were palpable and measurable by 21 days after injection and reached the size threshold for excision by 35–39 days after injection. This was a markedly faster growth rate than the control round 2 tumors, which in contrast reached a detectable and measurable size at least 42 days after injection and did not reach the size threshold for excision in up to 120 days of monitoring (Figure 2C). Owing to the hardness and density of the tissue, none of the control round 2 tumors could be dissociated into single cell suspensions using the standard dissociation protocols.

Post-excision, the round 2 driver library tumors were processed in a manner similar to the round 1 teratomas. Mapping clusters to cell types in wild type teratomas revealed only 7 out of the 23 cell types detected in the wild type teratomas, a sign of potential fitness advantage in these lineages leading to

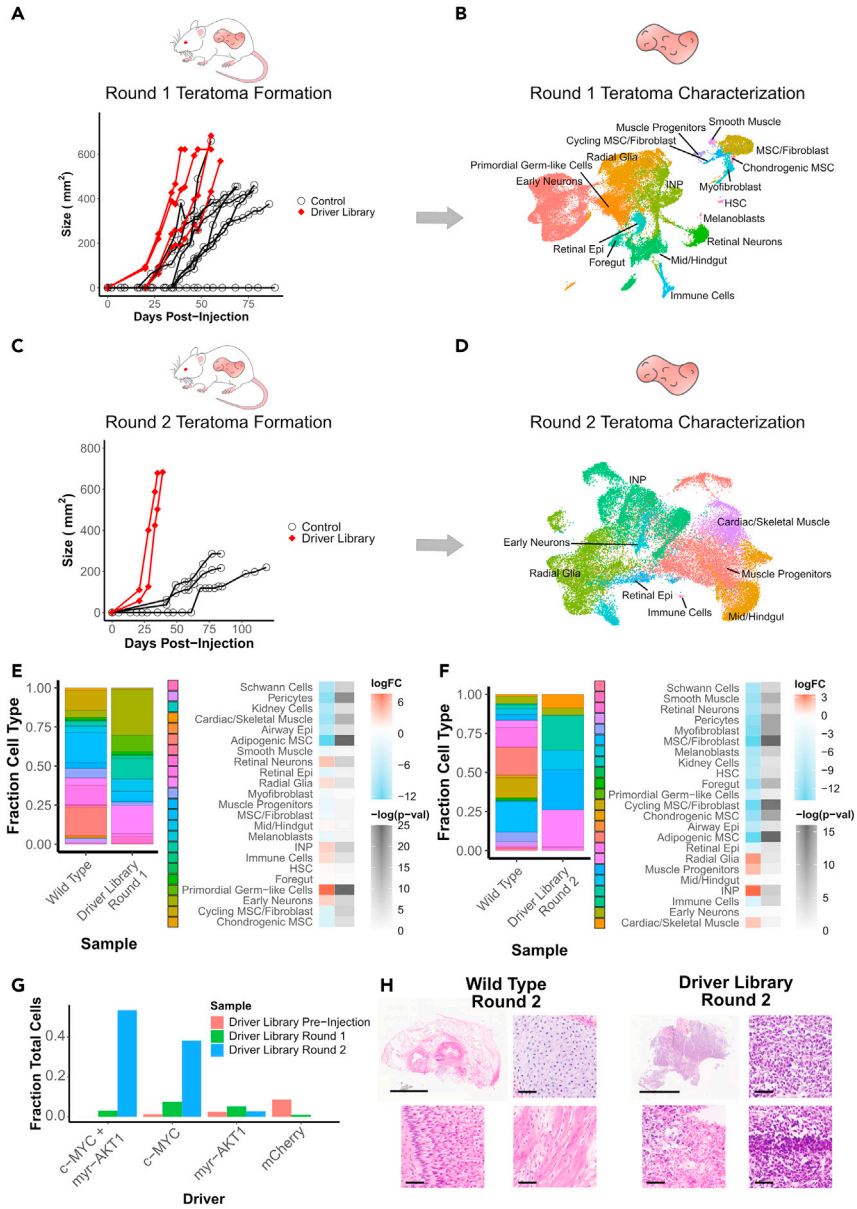


Figure 2. Identification of significantly enriched drivers and cell types

(A) Growth kinetics of round 1 teratoma formation for injections with driver library transduced hESCs vs control WT hESCs.

(B) UMAP visualization of cell types from round 1 teratomas formed by driver library transduced hESCs.

(C) Growth kinetics of round 2 tumors formed from re-injected cells from round 1 teratomas formed by driver library transduced hESCs vs WT hESCs. Control measurements are from a common set of tumors grown from the parent WT hESC line, which were used as growth controls for all experiments in this study.

(D) UMAP visualization of cell types from round 2 tumors formed by re-injected cells from round 1 teratomas of driver library transduced hESCs.

(E) Relative fraction of each cell type in round 1 teratomas formed from library transduced hESCs and WT hESCs, and log fold change with associated -log(p-value) of each cell type for driver library teratomas vs WT teratomas.

(F) Relative fraction of each cell type in round 2 tumors formed from library transduced hESCs and WT hESCs, and log fold change with associated -log(p-value) of each cell type for driver library tumors vs WT tumors.

(G) Relative fraction of top enriched drivers prior to injection and in each round of tumor formation.

(H) H&E stained sections of round 2 tumors formed from WT and driver library transduced hESCs. WT tumors display mature cell types from all three germ layers, such as cartilage (top right), muscle (bottom right) and dermis-like epithelium (bottom left). Driver library tumors display disorganized and more homogeneous composition along with markers of transformation such as nuclear pleomorphism (top right), areas of high mitotic rates (bottom right) and areas of necrosis (bottom left). Scale bars for full sections are 5 mm, scale bars for magnified images are 50 μm.

them dominating the tumor composition (Figure 2D). A significant difference was observed between the two tumor samples even after integrating the data with batch correction via the Seurat integration pipeline (Figure S2E). This challenge in integrating scRNA-seq datasets is also observed in clinical tumor samples, which display patient- or tumor-specific batch effects (Couturier et al., 2020; Fan et al., 2020; Yuan et al., 2018) because of inter-tumor and intra-tumor heterogeneity. Barcodes were detectably expressed in 95% of cells in this round (Figures S2E and S2F), suggesting that cells with a survival and proliferation advantage were those where the lentiviral cassette was not silenced and was robustly expressed.

We then examined the cell type populations which were present in these driver library tumors, compared to teratomas derived from wild type hESCs. In the round 1 teratomas, immature neural lineages and neural progenitors dominated the composition of the tumor with early neurons, radial glia and intermediate neural progenitors (INPs) making up 60% of the teratoma, and primordial germ-like cells another 10% (Figure 2E). Compared to the control teratomas this represented a 4–6 fold increase in proportion of the neural cell types, and a nearly 100 fold increase in proportion of the primordial germ-like cells. In contrast, the mesenchymal lineages were significantly reduced as a proportion of cells in the tumors (Figure 2E). In the round 2 tumors, cell type populations were further shifted. The majority of these round 2 tumors consisted of neural progenitor-like cells, primarily radial glia and INPs, which made up 45% of these tumors, and muscle progenitor-like cells, which made up 25% of the tumors (Figure 2F). Here we observed significant sample-specific effects with tumor 1 composed of a majority of neural-like cells, whereas tumor 2 was composed primarily of muscle, muscle progenitor-like, and gut-like lineages (Figures S2E and S2G).

There was a dramatic redistribution of driver populations detected over the course of these serial injections. Prior to injection, *c-MYC* and myristoylated *AKT1* (*myr-AKT1*) constituted 1.3% and 2.4%, respectively, of total cells whereas a combinatorial transduction of the two, *c-MYC* + *myr-AKT1*, was undetectable in 5949 genotyped cells. In the first round of teratomas *c-MYC*, *myr-AKT1*, and *c-MYC* + *myr-AKT1* constituted 16.7%, 11.6%, and 6.9% of cells, respectively, whereas strikingly in the second round of tumors *c-MYC* + *myr-AKT1* and *c-MYC* made up 56% and 40% of cells, respectively (Figure 2G). These observations are consistent with existing knowledge, for instance, the observation of elevated expression or amplification of *c-MYC*, or *c-MYC* dependent survival and proliferation, in subsets of embryonal tumors such as medulloblastoma (Hovestadt et al., 2019; Pei et al., 2012, 2016; Vladoiu et al., 2019) and atypical rhabdoid/teratoid tumors (Alimova et al., 2019; Ho et al., 2020), and pediatric soft tissue tumors like rhabdomyosarcoma (Gravina et al., 2016; Kouraklis et al., 1999; Zhang et al., 2017). The observed enhanced combinatorial effect of the drivers is also in line with the cooperative action of *c-MYC* and *AKT1*, where the action of *c-MYC* is negatively regulated by the *FOXO* group of transcription factors, which in turn are regulated by the *PI3K/AKT* pathway (Bouchard et al., 2004). A constitutively active form of *AKT1*, such as *myr-AKT1*, phosphorylates *FOXO* transcription factors and abrogates their function, thus allowing for uninhibited *c-MYC* activity (Bouchard et al., 2004; Pei et al., 2016). In our observations with the teratoma based system, *c-MYC* seemed to also drive a muscle-like phenotype which was reduced as a fraction of the population when *myr-AKT1* was expressed in combination. This may also explain the sample-specific clustering observed between the two tumors which had different drivers as the top enriched hits leading to fitness advantages (Figures S2H and S2I).

Histology displayed clear differences between the control and driver library tumors. Although control tumors had a diversity of cell types from all three germ layers in well-organized architectures (Figure 2H), the driver library tumors had far more homogeneous composition and displayed distinct signs of malignancy such as nuclear pleomorphism, areas of high mitotic rate and areas of necrosis (Figure 2H), suggesting the de novo creation of a transformed phenotype driven by *c-MYC* overexpression with or without the accompanying dysregulation of *AKT1* signaling.

Screening less-dominant drivers by removing dominant hits

In the previously described screens, the proliferative advantage conferred by the top hits caused cells expressing those drivers to overwhelm all other cells, such that other drivers were not significantly detected in the second round of tumors formed by serial reinjection. To assay the fitness effects of other drivers, we hypothesized that removing the top hits from the library would allow the detection of enrichment of other drivers.

Towards this we repackaged a lentiviral driver sub-library, removing *c-MYC* and *myr-AKT1* from the pool. Using this sub-library we conducted the screening process similarly to that for the full driver library. We

injected and monitored six round 1 teratomas, out of which we characterized the three fastest growing ones via scRNA-seq. Similar to before, round 1 teratomas were reinjected for driver enrichment, and tumors were monitored for 75 days, with the two fastest growing tumors assayed via scRNA-seq.

The round 1 driver sub-library teratomas were measurable between 18 and 25 days after injection and reached a size sufficient for extraction between 40 and 60 days after injection (Figure 3A). We again assessed the cell type populations in these sub-library screens by mapping the cells to the wild type teratomas as a reference. Using this, we determined that the round 1 teratomas contained 13 major cell types, a majority of which were neural cell types, in particular early neurons, radial glia and INPs, distributed across all three samples (Figures 3B and S3A). In these samples, barcodes were detectable in 27% of cells (Figures S3B–S3D).

In contrast to the full driver library, the round 2 driver sub-library teratomas showed a slower growth rate, reaching a measurable size between 27 and 38 days after injection. Although one of the round 2 tumors reached a sufficient size for extraction at 75 days after injection (Figure 3C), with the fastest growing tumors extracted and assayed via scRNA-seq. In these round 2 sub-library tumors, we again detected 13 major cell types, but as opposed to neural or muscle progenitor lineages, the majority were mesenchymal cell types with adipogenic mesenchymal stromal cells (MSCs), MSC/fibroblasts and chondrogenic MSCs (Figures 3D and S3E). Barcodes were detectable in 46% of cells in this round of tumors (Figure S3F), with the increase in the fraction of genotyped cells suggesting that surviving cells which express barcodes may have a fitness or survival advantage conferred by the expressed driver. We again subsetted the drivers to visualize only those detected in 25 cells or more (Figure S3G and S3H), the majority of which were driven wholly or in part by *MEK1*$^{S218D/S222D}$, whereas the remaining were driven by *RHOJ*, a small GTP-binding protein known to play a role in cell migration, which has recently been found to confer proliferative advantage in multiple lineages (Sack et al., 2018).

We then evaluated the cell type populations in comparison to teratomas formed from wild type H1 hESCs. In the round 1 sub-library teratomas, immature and progenitor neural cell types constituted 70% of the total population of the teratoma, and fibroblast-like mesenchymal lineages making up a further 20% of the teratoma (Figure 3E). However, in the round 2 tumors, in a striking difference compared to the full library screens, the tumors were composed primarily of fibroblast and MSC phenotypes, which constituted 65% of the tumors (Figure 3F). Along with the fibroblast lineages which constituted a majority of these tumors, a small neural component persisted via an expansion in Schwann cells and melanoblasts, which may be derived from Schwann cell precursors (Adameyko et al., 2009). The neural progenitor-like lineages present in the *c-MYC* and *myr-AKT1* driven tumors were not present in these tumors. This change in composition of the tumors was accompanied by an enrichment of the constitutively active mutant *MEK1*$^{S218D/S222D}$ which was present in 3.3% of cells prior to injection but constituted 20% of all cells in the round 2 tumors. In comparison, the fraction of cells expressing the internal negative control, mCherry, fell from 16% of cells prior to injection to 2% of cells in the round 2 tumors (Figure 3G). This role of *MEK1*$^{S218D/S222D}$ in supporting proliferation and survival of fibroblasts is consistent with previously reported results in literature where expression of constitutively active versions of *MEK1* were sufficient to trigger proliferative states and even transformation in fibroblasts in vitro (Brunet et al., 1994; Cowley et al., 1994; Mansour et al., 1994).

Histology images from the two rounds of tumors further confirmed the cellular composition. In round 1 teratomas we observed a majority of neuroectoderm-like cell types, whereas in round 2 tumors the majority of cells were mesenchymal in nature (Figure 3H). In these sub-library screens, we did not observe clear indications of malignancy as we observed in the *c-MYC* and *myr-AKT1* driven full library screens, suggesting that although *MEK1*$^{S218D/S222D}$ may drive proliferative advantage, it may not be a driver of cellular transformation on its own.

Validating enriched driver hits

To validate the hits obtained from the two sets of screens we individually tested the effects of *c-MYC*, *myr-AKT1*, *c-MYC + myr-AKT1*, and *MEK1*$^{S218D/S222D}$. hPSCs were transduced with either one of these driver vectors or a negative control vector expressing mCherry. Before injection, the driver and the control transduced cells were mixed in a 1:1 ratio, with a portion of these mixed cells pelleted and genomic DNA extracted to assess barcode distribution, and the remaining cells injected for teratoma formation in three Rag2$^{-/-}$;γc$^{-/-}$ immunodeficient mice for each driver to be validated. Teratomas were formed from these

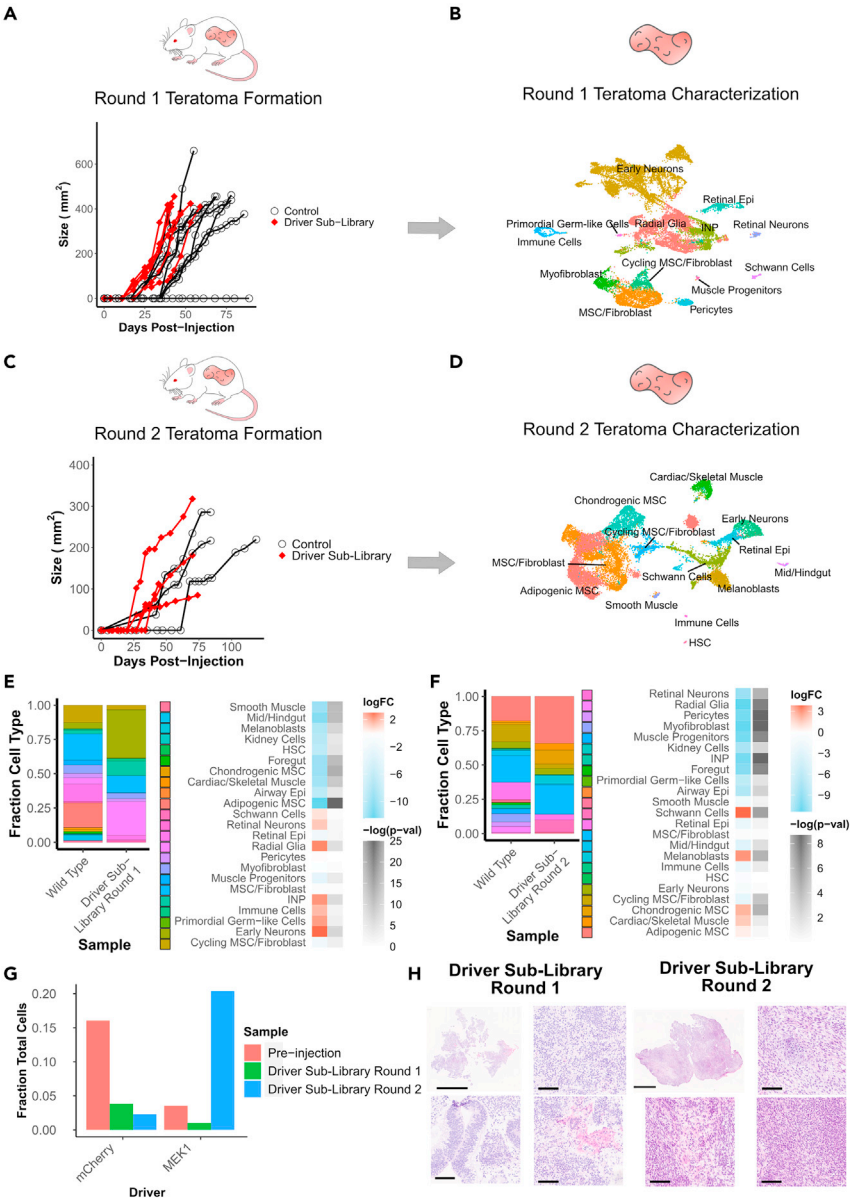


Figure 3. Identification of significantly enriched drivers and cell types for tumors formed with driver libraries without *c-MYC* and *myr-AKT1*

(A) Growth kinetics of round 1 teratoma formation for injections with driver library transduced hESCs vs WT hESCs.

(B) UMAP visualization of cell types from round 1 teratomas formed by driver library transduced hESCs.

(C) Growth kinetics of round 2 tumors formed from re-injected cells from round 1 teratomas formed by driver library transduced hESCs vs WT hESCs. Control measurements are from the common set of tumors grown from the parent WT hESC line, which were used as growth controls for all experiments in this study.

(D) UMAP visualization of cell types from round 2 tumors formed by re-injected cells from round 1 teratomas of driver library transduced hESCs.

(E) Relative fraction of each cell type in round 1 teratomas formed from library transduced hESCs and WT hESCs, and log fold change with associated -log(p-value) of each cell type for driver library teratomas vs WT teratomas.

(F) Relative fraction of each cell type in round 2 tumors formed from library transduced hESCs and WT hESCs, and log fold change with associated -log(p-value) of each cell type for driver library tumors vs WT tumors.

(G) Relative fraction of top enriched drivers prior to injection and in each round of tumor formation.

(H) H&E stained sections of round 1 and round 2 tumors formed from hESCs transduced with driver libraries without *c-MYC* or *myr-AKT1*. Round 1 tumors contain neuroectodermal and epithelial cell types as the majority of cells, while round 2 tumors contain mesenchymal cell types as the majority of cells. Scale bars for full sections are 5 mm, scale bars for magnified images are 100 μm.

injected cells, excised when they reached sufficient size, representative pieces of the excised tumors were immediately flash frozen in liquid nitrogen and some representative pieces preserved for cryosectioning, whereas the remaining tissue was dissociated. Dissociated cells were divided to be stored for genomic DNA extraction, RNA extraction and for serial reinjection to form round 2 tumors. Round 2 tumors were also then allowed to grow to a size sufficient for extraction and then excised and dissociated as for round 1. Dissociated cells were stored for genomic DNA extraction and RNA extraction.

Consistent with the observations during the screens, teratomas formed by hESCs transduced with *c-MYC* and *c-MYC + myr-AKT1* grew at the highest rate compared to all other tumors. In the round 1 tumors, the *c-MYC*, *myr-AKT1*, and *c-MYC + myr-AKT1* were measurable between 27 and 32 days after injection and were at an extractable size between 42 and 48 days after injection. However, for $MEK1^{S218D/S222D}$ round 1 teratomas, tumors were measurable 33–40 days after injection and at an extractable size 59–78 days after injection (Figure 4A). In the round 2 tumors, those driven by *c-MYC + myr-AKT1* grew at the fastest rate, followed by *c-MYC*, both of which showed a significantly enhanced growth rate compared to the control round 2 tumors. Round 2 tumors driven by *myr-AKT1* alone or those driven by $MEK1^{S218D/S222D}$ did not show a significantly enhanced growth rate compared to the control tumors (Figure 4A), although tumors driven by $MEK1^{S218D/S222D}$ clustered toward the higher end of the control tumor growth rate.

To assess the barcode distribution from cells prior to injection through both rounds of tumor formation, we PCR amplified barcodes from genomic DNA isolated from stored cell pellets and quantified them via deep sequencing. We find that negative control barcodes decrease from 21 to 35% of total mapped reads prior to injection, to less than 2% of reads for $MEK1^{S218D/S222D}$, and less than 0.1% of reads for all other drivers by the second round of tumors (Figure 4B).

We then performed bulk RNA-sequencing on these validation tumors to assess their composition. In agreement with the results of the screens, we found that the *c-MYC* driven tumors were composed of a mix of neural and muscle lineages displaying elevated expression of neural-related genes *TUBB3*, *NEUROD1*, *SYP*, *CAMKV*, and *NEFH* (Figure 4C). In comparison, tumors driven by a combination of *c-MYC* and *myr-AKT1* were primarily composed of a neural progenitor like phenotype, expressing *NEFH* and *CAMKV* and elevated levels of *DANCR*, a non-coding RNA which acts to suppress differentiation and is present in many cancers (Jin et al., 2019; Zhang et al., 2018) (Figure 4C). On the other hand, tumors driven by $MEK1^{S218D/S222D}$ displayed elevated expression of mesenchymal markers *VIM*, *ITM2A*, and *CD44* (Figure 4C), which is again consistent with the results from the sub-library screens. We also compared the *c-MYC* and $MEK1^{S218D/S222D}$ driven tumors to pediatric cancers from the TARGET initiative. Using the 2000 most variable genes across the TARGET tumors we performed a principal component analysis (PCA) (Figure S4A). Plotting the two PCs capturing the majority of variation, we observed that *c-MYC* driven tumors clustered toward the TARGET neuroblastoma tumors, whereas those driven by $MEK1^{S218D/S222D}$ clustered toward the kidney tumors (Wilm's tumor and clear cell sarcoma of the kidney) and osteosarcoma.

c-MYC and *myr-AKT1* also contribute to metabolic reprogramming in tumors by promoting nucleotide biosynthesis(Hoxhaj and Manning, 2020; Stine et al., 2015). To further characterize *c-MYC* and *c-MYC + myr-AKT1* driven tumors, we quantified nucleobase levels via mass spectrometry. Compared to WT teratomas, *c-MYC* and *c-MYC + myr-AKT1* driven tumors increased the abundance of purine nucleobases, especially guanine (Figures S4B and S4D). Further, this was supported by broad enrichment of nucleobase synthesis terms, and particularly purine synthesis related Gene Ontology terms in genes which were upregulated compared to WT teratomas (Figures S4C and S4E). This is consistent with previous studies which have demonstrated the regulation of nucleotide metabolism by *c-MYC* (Liu et al., 2008; Stine et al., 2015), with de novo purine synthesis especially implicated in tumor maintenance (Wang et al., 2017) and response to therapy (Barfeld et al., 2015).

Histology showed that round 2 tumors driven by *myr-AKT1* alone showed poorly differentiated neural lineage cells interspersed with mesenchyme (Figure 4D). Round 2 tumors driven by *c-MYC* (Figure 4E) and those driven by *c-MYC + myr-AKT1* (Figure 4F) were composed of poorly differentiated cells with signs of malignancy, including necrosis. Finally, as observed in screen results, round 2 tumors driven by $MEK1^{S218D/S222D}$ were composed primarily of mesenchymal, fibroblast-like cells interspersed with cartilage (Figure 4G). In addition, staining with the proliferation marker Ki-67 demonstrated a distinctly larger population of proliferating cells in the round 2 *c-MYC* and *c-MYC + myr-AKT1* driven tumors followed by round

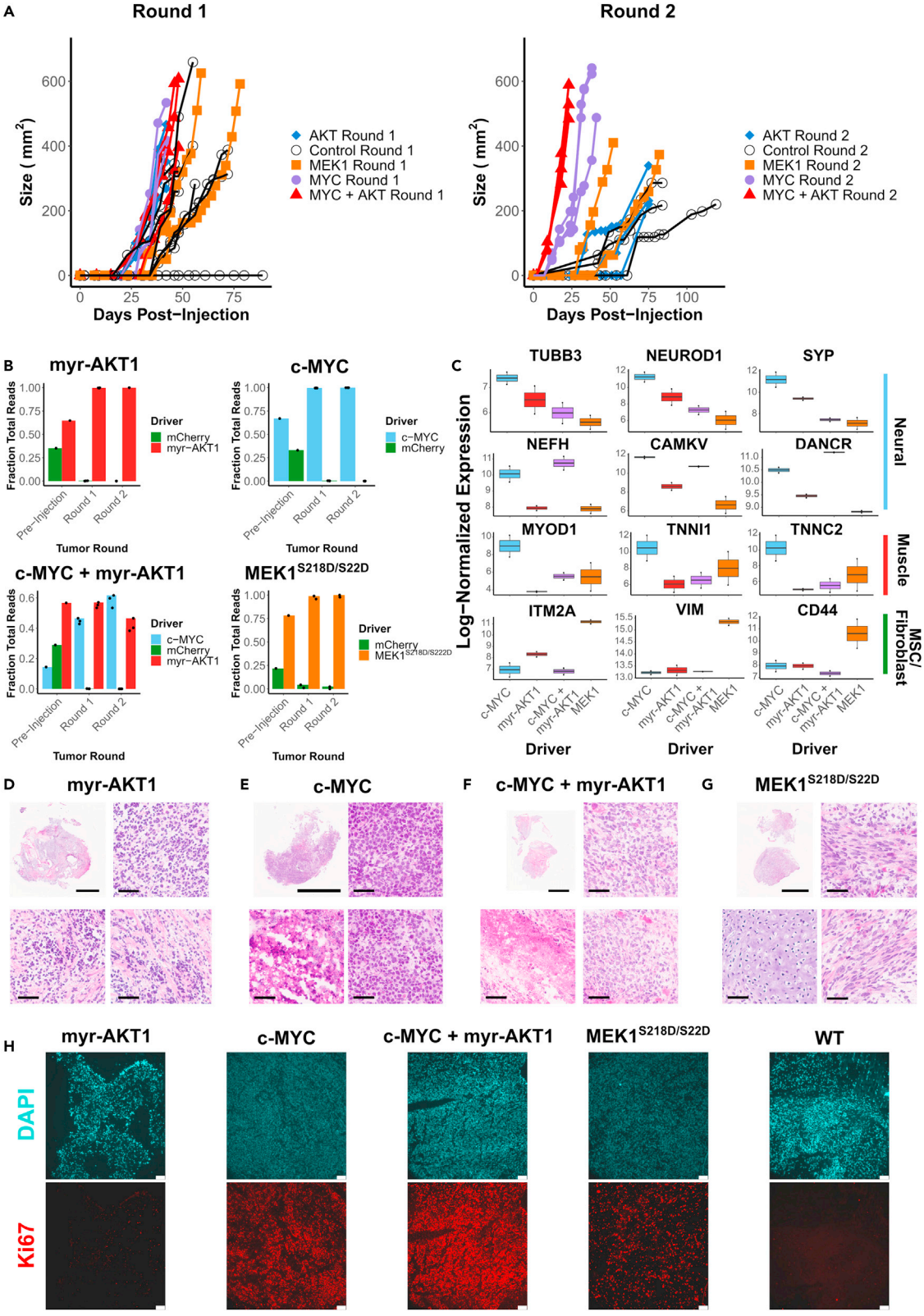

A
Round 1
Round 2
Size (mm2)
Days Post−Injection
AKT Round 1
Control Round 1
MEK1 Round 1
MYC Round 1
MYC + AKT Round 1
AKT Round 2
Control Round 2
MEK1 Round 2
MYC Round 2
MYC + AKT Round 2
B
myr-AKT1
c-MYC
c-MYC + myr-AKT1
MEK1S218D/S22D
Fraction Total Reads
Tumor Round
Pre−Injection
Round 1
Round 2
Driver
mCherry
myr−AKT1
c−MYC
MEK1S218D/S222D
C
Log−Normalized Expression
TUBB3
NEUROD1
SYP
NEFH
CAMKV
DANCR
MYOD1
TNNI1
TNNC2
ITM2A
VIM
CD44
Neural
Muscle
MSC/
Fibroblast
D
myr-AKT1
E
c-MYC
F
c-MYC + myr-AKT1
G
MEK1S218D/S22D
H
myr-AKT1
c-MYC
c-MYC + myr-AKT1
MEK1S218D/S22D
WT
DAPI
Ki67

Figure 4. Validation of tumor formation and proliferative advantages of significantly enriched drivers from library screens

(A) Growth kinetics of round 1 and round 2 tumors formed from a mixture of hESCs transduced with driver hits (c-MYC, myr-AKT1, c-MYC + myr-AKT1 or MEK1 (S218D, S222D) and hESCs transduced with a negative control (mCherry). Control measurements are from the common set of tumors grown from the parent WT hESC line, which were used as growth controls for all experiments in this study.

(B) Fraction of reads detecting either the driver or negative control barcodes at each stage: pre-injection, round 1 tumor formation and round 2 tumor formation. Barcodes were amplified from genomic DNA.

(C) Gene expression of lineage-specific markers for round 2 tumors driven by individual hits. Expression values are normalized and log transformed.

(D) H&E stain of round 2 tumor driven by myr-AKT1 showing regions of poor differentiation (top right) and necrosis (bottom right) interspersed with regions of organized tissue (bottom left).

(E) H&E stain of round 2 tumor driven by c-MYC showing regions of poor differentiation (top and bottom right) and regions of necrosis (bottom left).

(F) H&E stain of round 2 tumor driven by c-MYC + myr-AKT1 showing regions of poor differentiation (top and bottom right) and regions of necrosis (bottom left).

(G) H&E stain of round 2 tumor driven by MEK1 (S218D, S222D) showing primarily regions of mature mesenchymal fibroblast-like tissue (top and bottom right) and cartilage-like (bottom left). (D–G) Scale bars for full sections are 5 mm, scale bars for magnified images are 50 μm.

(H) Immunofluorescence micrographs of DAPI and Ki-67 stained sections. Scale bars are 75 μm.

2 *MEK1*$^{S218D/S222D}$ driven tumors, as compared to the WT teratomas (Figure 4H). While round 2 tumors driven by *myr-AKT1* alone showed no higher proliferative capacity than WT tumors (Figure 4H), confirming that *myr-AKT1* conferred proliferative advantage only in concert with *c-MYC*. These validation results strongly support the observations from the multiplexed screens and confirm the significant fitness advantage conferred by *c-MYC* and *c-MYC + myr-AKT1* on neural lineages, and of *MEK1*$^{S218D/S222D}$ on fibroblasts.

DISCUSSION

To study the important problem of what drives oncogenic transformation in human tissue, investigators have relied on in vitro and animal model systems. While immense progress has been made using these, limitations still remain. Animal models retain significant differences with human biology, while in vitro systems lack vasculature and the physiological cues present in the in vivo niche which are involved in regulating lineage-specific transformation. In addition, currently prevalent models often necessitate the investigation of a single or a few lineages at a time, raising an impediment in studying tissue-specificity across multiple cell types.

In this study, we have developed a flexible, highly multiplexed platform to study the effects of cancer drivers across lineages which harnesses hPSC-derived teratomas to access a diverse set of lineages, ORF overexpression libraries to express cancer drivers, scRNA-seq to read out transcriptomic profiles which determine the cell type and detect the perturbing driver, and serial proliferation of the tumors to enrich drivers and lineages enjoying fitness advantages. Using this platform, we initially screened key cancer drivers across more than 20 cell types and de novo generated a transformed phenotype in a neural lineage via the overexpression of *c-MYC* and *myr-AKT1*, a constitutively active form of *AKT1*, which dominated the serially reinjected tumors and displayed the hallmarks of malignancy. To screen less-dominant drivers, we repeated the screens with *c-MYC* and *myr-AKT1* removed from the library, and captured the proliferative advantage of other drivers, such as the one conferred on mesenchymal lineages like fibroblasts via the overexpression of constitutively active *MEK1*$^{S218D/S222D}$. These results were then validated by individually overexpressing these top hits during teratoma formation and serial reinjection, to confirm the enrichment of lineages detected in the screens.

Although the current demonstration serves as a proof of concept of the methodology, we envision subsequent studies building on this initial demonstration of the platform in several ways. One, although we have focused here primarily on deciphering the role of individual cancer drivers on oncogenic potential across lineages, exploring combinatorial perturbations will be crucial to dissecting both tumorigenicity as observed in native tumors, and also systematically studying variants-of-unknown significance, many of which individually may have only subtle phenotypes. Two, in this study, we focused on cell-intrinsic changes to the transcriptome only. In combination with techniques to map epigenetic characteristics, such as ATAC-seq, this may lead to a more detailed understanding of the determinants of tumor formation given that epigenetic factors such as chromatin state are critical to tumorigenesis, cancer evolution and progression. Three, newly developed spatial technologies may be combined with the platform to uncover cell-cell communication and paracrine signaling driven by cancer drivers, enabling us to tease apart cell endogenous versus exogenous effects. Finally, four, the xenografted mouse models used here for teratoma formation may be improved upon in two ways. In this study, we used a subcutaneous site of injection for teratoma

formation, but the site of injection impacts teratoma differentiation (Chan et al., 2018) as well as the lineage-specificity of cancer drivers. Orthotopic injections for specific tissues may provide a more appropriate niche. In addition, the mice used in this study are immune deficient, thus precluding the study of any effects of immune system interaction. A possible route to addressing this may be through the use of humanized mouse models (Walsh et al., 2017).

Taken together, we have demonstrated a proof of concept for a scalable, versatile platform which can screen multiple lineages and drivers in a single experiment, with a rich transcriptomic readout, thus providing a systematic path to studying the determinants and tissue-specificity of neoplastic transformation in human cells. We envision that refinements to this platform, coupled with the expanding array of available omics technologies will enable the comprehensive characterization of the trajectory of cells from normal to malignant states.

Limitations of the study

Although offering a powerful new approach in studying transformation, certain limitations and challenges merit consideration both in terms of interpretation of the resulting data, as well as in design of future implementations of the same.

Firstly, in its current form, the overexpression vector is designed to be constitutively on. This may confound results since some drivers and lineages could be enriched due to their biased differentiation or inability to escape a state being governed by driver expression itself. In future versions of this platform, that effect may be mitigated by using an inducible expression system or one with recombinase-based control, both of which can be controllably turned on at specific time points, to ensure differentiation is not affected by driver expression.

Secondly, although we demonstrate this proof of concept in a hESC line, background mutations and genetic alterations already present in hPSC lines may bias transformation phenotypes. Third, cell fate engineering methods (McDonald et al., 2020; Parekh et al., 2018) may be paired with the screening platform to repeatedly and predictably derive specific lineages of interest and avoid cell line biases or poor repeatability due to heterogeneity. Such lineage engineering methods could also allow for targeted study of specific tumor types which can be perturbed with well-designed, clinically relevant libraries. Four, drivers that improve proliferation will progressively dominate the cell population over the several weeks' long teratoma formation. Consequently, the system is highly susceptible to being overwhelmed by the top hits, thus leaving only small numbers of internal control cells or non-hit drivers to compare against or analyze separately. To sensitively profile the various drivers and controls, it may be necessary to redesign the system, profile more cells or profile samples at multiple time points to gain an understanding of the temporal evolution of the driver population distribution by cell type. Newer computational methods (Ji et al., 2021) enabling sophisticated analysis of perturbations may also aid in these aspects. Five, the predominantly embryonic versus adult state of the cell types might bias their innate transformation potential. In particular, we anticipate, due to the embryonic origin of the starting cells, this system might be especially applicable to modeling of pediatric tumors which are susceptible to transformation via single genetic alterations. To apply this system to adult tumors, it may require the application of methods for lineage engineering, methods to accelerate maturation, as well as the ability to introduce combinatorial genetic alterations which are required to transform adult cells. This may be done by overlaying libraries such as libraries to knockout tumor suppressors combined with the ORF library outlined here or by systematically applying targeted libraries on cells with an appropriate background genetic alteration such as a mutant *TP53* or *TERT*. For each case, libraries may be designed by targeted and deliberate choice of clinically relevant genetic alterations. Finally, dissociation and sample processing methods can have a large impact on scRNA-seq results (Denisenko et al., 2020; O'Flanagan et al., 2019; Van Den Brink et al., 2017). Improved protocols may allow for more fine-grained dissection of transcriptomes to assess cell state shifts within lineages.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103149.

ACKNOWLEDGMENTS

This work was generously supported by UCSD Institutional Funds and NIH grants (R01HG009285, RO1CA222826, RO1GM123313). This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929).

AUTHOR CONTRIBUTIONS

Conceptualization and Design: U.P., P.M.; Methodology: U.P., D.M., Y.W., K.Z., P.M.; Experiments: U.P., D.M., A.D., T.C.; Computational analyses: U.P., Y.W.; Writing: A.D., U.P., P.M. with input from all authors.

DECLARATION OF INTERESTS

P.M. is a scientific co-founder of Shape Therapeutics, Boundless Biosciences, Navega Therapeutics, and Engine Biosciences, which have no commercial interests related to this study. K.Z. is a cofounder, equity holder, and paid consultant of Singlera Genomics, which has no commercial interests related to this study. The terms of these arrangements have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Received: May 20, 2021
Revised: August 27, 2021
Accepted: September 15, 2021
Published: October 22, 2021

REFERENCES

Adameyko, I., Lallemend, F., Aquino, J.B., Pereira, J.A., Topilko, P., Müller, T., Fritz, N., Beljajeva, A., Mochii, M., Liste, I., et al. (2009). Schwann cell precursors from nerve innervation are a cellular origin of melanocytes in skin. Cell *139*, 366–379.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alimova, I., Pierce, A., Danis, E., Donson, A., Birks, D.K., Griesinger, A., Foreman, N.K., Santi, M., Soucek, L., Venkataraman, S., et al. (2019). Inhibition of MYC attenuates tumor cell self-renewal and promotes senescence in SMARCB1-deficient Group 2 atypical teratoid

rhabdoid tumors to suppress tumor growth in vivo. Int. J. Cancer *144*, 1983–1995.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18.

Balani, S., Nguyen, L.V., and Eaves, C.J. (2017). Modeling the process of human tumorigenesis. Nat. Commun. *8*, 1–10.

Barfeld, S.J., Fazli, L., Persson, M., Marjavaara, L., Urbanucci, A., Kaukoniemi, K.M., Rennie, P.S., Ceder, Y., Chabes, A., Visakorpi, T., et al. (2015). Myc-dependent purine biosynthesis affects nucleolar stress and therapy response in prostate cancer. Oncotarget *6*, 12587–12602.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Bian, S., Repic, M., Guo, Z., Kavirayani, A., Burkard, T., Bagley, J.A., Krauditsch, C., and Knoblich, J.A. (2018). Genetically engineered cerebral organoids model brain tumor formation. Nat. Methods *15*, 631–639.

Boehm, J.S., Hession, M.T., Bulmer, S.E., and Hahn, W.C. (2005). Transformation of human and murine fibroblasts without viral oncoproteins. Mol. Cell. Biol. *25*, 6464–6474.

Bouchard, C., Marquardt, J., Brás, A., Medema, R.H., and Eilers, M. (2004). Myc-induced proliferation and transformation require Akt-mediated phosphorylation of FoxO proteins. EMBO J. *23*, 2830–2840.

Brunet, A., Pagès, G., and Pouysségur, J. (1994). Constitutively active mutants of MAP kinase kinase (MEK1) induce growth factor-relaxation and oncogenicity when expressed in fibroblasts. Oncogene *9*, 3379–3387.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Chan, S.S.K., Arpke, R.W., Filareto, A., Xie, N., Pappas, M.P., Penaloza, J.S., Perlingeiro, R.C.R., and Kyba, M. (2018). Skeletal muscle stem cells from PSC-derived teratomas have functional regenerative capacity. Cell Stem Cell *23*, 74–85.e6.

Cheon, D.-J., and Orsulic, S. (2011). Mouse models of cancer. Annu. Rev. Pathol. Mech. Dis. *6*, 95–119.

Clark, R., Stampfer, M.R., Milley, R., O'rourke, E., Walen, K.H., Kriegler, M., Kopplin, J., and Mccormick, F. (1988). Transformation of human mammary epithelial cells by oncogenic retroviruses. Cancer Res. *48*, 4689–4694.

Cordes, T., and Metallo, C.M. (2019). Quantifying intermediary metabolism and lipogenesis in cultured mammalian cells using stable isotope tracing and mass spectrometry. Methods Mol. Biol. *1978*, 219–241.

Couturier, C.P., Ayyadhury, S., Le, P.U., Nadaf, J., Monlong, J., Riva, G., Allache, R., Baig, S., Yan, X., Bourgey, M., et al. (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nat. Commun. *11*, 3406.

Cowley, S., Paterson, H., Kemp, P., and Marshall, C.J. (1994). Activation of MAP kinase kinase is necessary and sufficient for PC12 differentiation and for transformation of NIH 3T3 cells. Cell *77*, 841–852.

Daley, G.Q., Mclaughlin, J., Witte, O.N., and Baltimore, D. (1987). The CML-specific P210 bcr/abl protein, unlike v-abl, does not transform NIH/3T3 fibroblasts. Science *237*, 532–535.

Denisenko, E., Guo, B.B., Jones, M., Hou, R., De Kock, L., Lassmann, T., Poppe, D., Poppe, D., Clément, O., Simmons, R.K., et al. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. *21*, 130.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Drost, J., Van Jaarsveld, R.H., Ponsioen, B., Zimberlin, C., Van Boxtel, R., Buijs, A., Sachs, N., Overmeer, R.M., Offerhaus, G.J., Begthel, H., et al. (2015). Sequential cancer mutations in cultured human intestinal stem cells. Nature *521*, 43–47.

Drost, J., van Boxtel, R., Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., de Ligt, J., Behjati, S., Grolleman, J.E., van Wezel, T., et al. (2017). Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science *358*, 234–238.

Duan, S., Yuan, G., Liu, X., Ren, R., Li, J., Zhang, W., Wu, J., Xu, X., Fu, L., Li, Y., et al. (2015). PTEN deficiency reprogrammes human neural stem cells towards a glioblastoma stem cell-like phenotype. Nat. Commun. *6*, 10068.

Elenbaas, B., Spirio, L., Koerner, F., Fleming, M.D., Zimonjic, D.B., Donaher, J.L., Popescu, N.C., Hahn, W.C., and Weinberg, R.A. (2001). Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Development *15*, 50–65.

Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G., and Hartwell, L. (2003). The case for early detection. Nat. Rev. Cancer *3*, 243–252.

Fan, J., Slowikowski, K., and Zhang, F. (2020). Single-cell transcriptomics in cancer: computational challenges and opportunities. Exp. Mol. Med. *52*, 1452–1465.

Fischer, M. (2021). Mice are not humans: the case of p53. Trends Cancer *7*, 12–14.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Fumagalli, A., Drost, J., Suijkerbuijk, S.J.E., van Boxtel, R., de Ligt, J., Offerhaus, G.J., Begthel, H., Beerling, E., Tan, E.H., Sansom, O.J., et al. (2017). Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. Proc. Natl. Acad. Sci. U. S. A. *114*, E2357–E2364.

Geder, L., Lausch, R., O'Neill, F., and Rapp, F. (1976). Oncogenic transformation of human embryo lung cells by human cytomegalovirus. Science *192*, 1134–1137.

Gillet, J.-P., Varma, S., and Gottesman, M.M. (2013). The clinical relevance of cancer cell lines. J. Natl. Cancer Inst. *105*, 452–458.

Gould, S.E., Junttila, M.R., and De Sauvage, F.J. (2015). Translational value of mouse models in oncology drug development. Nat. Med. *21*, 431–439.

Gravina, G.L., Festuccia, C., Popov, V.M., Di Rocco, A., Colapietro, A., Sanità, P., Delle Monache, S., Musio, D., De Felice, F., Di Cesare, E., et al. (2016). C-myc sustains transformed phenotype and promotes radioresistance of embryonal rhabdomyosarcoma cell lines. Radiat. Res. *185*, 411–422.

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

Hagis, K.M., Cichowski, K., and Elledge, S.J. (2019). Tissue-specificity in cancer: the rule, not the exception. Science *363*, 1150–1151.

Hagting, A., Karlsson, C., Clute, P., Jackman, M., and Pines, J. (1998). MPF localization is controlled by nuclear export. EMBO J. *17*, 4127–4138.

Hahn, W.C., Counter, C.M., Lundberg, A.S., Beijersbergen, R.L., Brooks, M.W., and Weinberg, R.A. (1999). Creation of human tumour cells with defined genetic elements. Nature *400*, 464–468.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hayer, A., Shao, L., Chung, M., Joubert, L.-M., Yang, H.W., Tsai, F.-C., Bisaria, A., Betzig, E., and Meyer, T. (2016). Engulfed cadherin fingers are polarized junctional structures between collectively migrating endothelial cells. Nat. Cell Biol. *18*, 1311–1323.

Hill, A.J., McFaline-Figueroa, J.L., Starita, L.M., Gasperini, M.J., Matreyek, K.A., Packer, J., Jackson, D., Shendure, J., and Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. Nat. Methods *15*, 271–274.

Ho, B., Johann, P.D., Grabovska, Y., De Dieu Andrianteranagna, M.J., Yao, F., Frühwald, M., Hasselblatt, M., Bourdeaut, F., Williamson, D., Huang, A., et al. (2020). Molecular subgrouping of atypical teratoid/rhabdoid tumors—a reinvestigation and current consensus. Neuro. Oncol. *22*, 613–624.

Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.K.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R., et al. (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. Nature *572*, 74–79.

Hoxhaj, G., and Manning, B.D. (2020). The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. Nat. Rev. Cancer *20*, 74–88.

Ji, Y., Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2021). Machine learning for perturbational single-cell omics. Cell Syst *12*, 522–537.

Jin, S.J., Jin, M.Z., Xia, B.R., and Jin, W.L. (2019). Long non-coding RNA DANCR as an emerging therapeutic target in human cancers. Front. Oncol. *9*, 1225.

Johannessen, C.M., Boehm, J.S., Kim, S.Y., Thomas, S.R., Wardwell, L., Johnson, L.A., Emery, C.M., Stransky, N., Cogdill, A.P., Barretina, J., et al. (2010). COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. Nature *468*, 968–972.

Kim, E., Ilic, N., Shrestha, Y., Zou, L., Kamburov, A., Zhu, C., Yang, X., Lubonja, R., Tran, N., Nguyen, C., et al. (2016). Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. Cancer Discov. *6*, 714–726.

Koga, T., Chaim, I.A., Benitez, J.A., Markmiller, S., Parisian, A.D., Hevner, R.F., Turner, K.M., Hessenauer, F.M., D'Antonio, M., Nguyen, N.-P.D., et al. (2020). Longitudinal assessment of tumor development using cancer avatars derived from genetically engineered pluripotent stem cells. Nat. Commun. *11*, 550.

Kouraklis, G., Triche, T.J., Tsokos, M., and Wesley, R.D. (1999). Myc oncogene expression and nude mouse tumorigenicity and metastasis formation are higher in alveolar than embryonal rhabdomyosarcoma cell lines. Pediatr. Res. *45*, 552–558.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lannagan, T.R.M., Lee, Y.K., Wang, T., Roper, J., Bettington, M.L., Fennell, L., Vrbanac, L., Jonavicius, L., Somashekar, R., Gieniec, K., et al. (2019). Genetic editing of colonic organoids provides a molecularly distinct and orthotopic preclinical model of serrated carcinogenesis. Gut *68*, 684–692.

Lensch, M.W., Schlaeger, T.M., Zon, L.I., and Daley, G.Q. (2007). teratoma formation assays with human embryonic stem cells: a rationale for one type of human-animal chimera. Cell Stem Cell *1*, 253–258.

Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu, X.S. (2014a). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. *15*, 554.

Li, X., Nadauld, L., Ootani, A., Corney, D.C., Pai, R.K., Gevaert, O., Cantrell, M.A., Rack, P.G., Neal, J.T., Chan, C.W.-M., et al. (2014b). Oncogenic transformation of diverse gastrointestinal tissues in primary organoid culture. Nat. Med. *20*, 769–777.

Liu, Y.-C., Li, F., Handler, J., Huang, C.R.L., Xiang, Y., Neretti, N., Sedivy, J.M., Zeller, K.I., and Dang, C.V. (2008). Global regulation of nucleotide biosynthetic genes by c-Myc. PLoS One *3*, e2722.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Mansour, S.J., Matten, W.T., Hermann, A.S., Candia, J.M., Rong, S., Fukasawa, K., Vande Woude, G.F., and Ahn, N.G. (1994). Transformation of mammalian cells by constitutively active MAP kinase kinase. Science *265*, 966–970.

Martz, C.A., Ottina, K.A., Singleton, K.R., Jasper, J.S., Wardell, S.E., Peraza-penton, A., Anderson, G.R., Winter, P.S., Wang, T., Alley, H.M., et al. (2014). Systematic identification of signaling pathways with potential to confer anticancer drug resistance. Sci. Signal. *7*, 1–14.

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., Watanabe, T., Kanai, T., and Sato, T. (2015). Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. Nat. Med. *21*, 256–262.

McDonald, D., Wu, Y., Dailamy, A., Tat, J., Parekh, U., Zhao, D., Hu, M., Tipps, A., Zhang, K., and Mali, P. (2020). Defining the teratoma as a model for multi-lineage human development. Cell *183*, 1402–1419.e18.

McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. *8*, 329–337.e4.

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

O'Flanagan, C.H., Campbell, K.R., Zhang, A.W., Kabeer, F., Lim, J.L.P., Biele, J., Eirew, P., Lai, D., McPherson, A., Kong, E., et al. (2019). Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. Genome Biol. *20*, 210.

Parekh, U., Wu, Y., Zhao, D., Worlikar, A., Shah, N., Zhang, K., and Mali, P. (2018). Mapping cellular reprogramming via pooled overexpression screens with paired fitness and single-cell RNA-sequencing readout. Cell Syst. *7*, 548–555.e8.

Park, J.W., Lee, J.K., Sheu, K.M., Wang, L., Balanis, N.G., Nguyen, K., Smith, B.A., Cheng, C., Tsai, B.L., Cheng, D., et al. (2018). Reprogramming normal human epithelial tissues to a common, lethal neuroendocrine cancer lineage. Science *362*, 91–95.

Pei, Y., Moore, C.E., Wang, J., Tewari, A.K., Eroshkin, A., Cho, Y.J., Witt, H., Korshunov, A., Read, T.A., Sun, J.L., et al. (2012). An animal model of MYC-driven medulloblastoma. Cancer Cell *21*, 155–167.

Pei, Y., Liu, K.W., Wang, J., Garancher, A., Tao, R., Esparza, L.A., Maier, D.L., Udaka, Y.T., Murad, N., Morrissy, S., et al. (2016). HDAC and PI3K antagonists cooperate to inhibit growth of MYC-driven medulloblastoma. Cancer Cell *29*, 311–323.

Puisieux, A., Pommier, R.M., Morel, A.P., and Lavial, F. (2018). Cellular pliancy and the multistep process of tumorigenesis. Cancer Cell *33*, 164–172.

Ramaswamy, S., Nakamura, N., Vazquez, F., Batt, D.B., Perera, S., Roberts, T.M., and Sellers, W.R. (1999). Regulation of G1 progression by the PTEN tumor suppressor protein is linked to inhibition of the phosphatidylinositol 3-kinase/Akt pathway. Proc. Natl. Acad. Sci. U. S. A. *96*, 2110–2115.

Rangarajan, A., and Weinberg, R.A. (2003). Comparative biology of mouse versus human cells: modelling human cancer in mice. Nat. Rev. Cancer *3*, 952–959.

Rangarajan, A., Hong, S.J., Gifford, A., and Weinberg, R.A. (2004). Species- and cell type-specific requirements for cellular transformation. Cancer Cell *6*, 171–183.

Richmond, A., and Yingjun, S. (2008). Mouse xenograft models vs GEM models for human cancer therapeutics. DMM Dis. Models Mech. *1*, 78–82.

Ritchie, M.E., Dai, Z., Sheridan, J.M., Gearing, L.J., Moore, D.L., Su, S., Wormald, S., Wilcox, S., O'Connor, L., Dickins, R.A., et al. (2014). edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. F1000Res *3*, 95.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Roberts, P.J., Mitin, N., Keller, P.J., Chenette, E.J., Madigan, J.P., Currin, R.O., Cox, A.D., Wilson, O., Kirschmeier, P., and Der, C.J. (2008). Rho Family GTPase modification and dependence on CAAX motif-signaled posttranslational modification. J. Biol. Chem. *283*, 25150–25163.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rosenbluth, J.M., Schackmann, R.C.J., Gray, G.K., Selfors, L.M., Li, C.M.-C., Boedicker, M., Kuiken, H.J., Richardson, A., Brock, J., Garber, J., et al. (2020). Organoid cultures from normal and cancer-prone human breast tissues preserve complex epithelial lineages. Nat. Commun. *11*, 1711.

Sack, L.M., Davoli, T., Xu, Q., Li, M.Z., and Elledge, S.J. (2016). Sources of error in mammalian genetic screens. G *6*, 2781–2790.

Sack, L.M., Davoli, T., Li, M.Z., Li, Y., Xu, Q., Naxerova, K., Wooten, E.C., Bernardi, R.J., Martin, T.D., Chen, T., et al. (2018). Profound tissue specificity in proliferation control underlies cancer drivers and aneuploidy patterns. Cell *0*, 1–16.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. Cell *173*, 321–337.e10.

Sasaki, R., Narisawa-Saito, M., Yugawa, T., Fujita, M., Tashiro, H., Katabuchi, H., and Kiyono, T. (2009). Oncogenic transformation of human ovarian surface epithelial cells with defined cellular oncogenes. Carcinogenesis *30*, 423–431.

Schneider, G., Schmidt-Supprian, M., Rad, R., and Saur, D. (2017). Tissue-specific tumorigenesis: context matters. Nat. Rev. Cancer *17*, 239–253.

Smith, R.C., and Tabar, V. (2019). Constructing and deconstructing cancers using human pluripotent stem cells and organoids. Cell Stem Cell *24*, 12–24.

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., et al. (2015). A Big Bang model of human colorectal tumor growth. Nat. Genet. *47*, 209–216.

Stine, Z.E., Walton, Z.E., Altman, B.J., Hsieh, A.L., and Dang, C.V. (2015). MYC, metabolism, and cancer. Cancer Discov. *5*, 1024–1039.

Van Den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and Van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods *14*, 935–936.

Vladoiu, M.C., El-Hamamy, I., Donovan, L.K., Farooq, H., Holgado, B.L., Sundaravadanam, Y., Ramaswamy, V., Hendrikse, L.D., Kumar, S., Mack, S.C., et al. (2019). Childhood cerebellar tumours mirror conserved fetal transcriptional programs. Nature *572*, 67–73.

Walsh, N.C., Kenney, L.L., Jangalwe, S., Aryee, K.E., Greiner, D.L., Brehm, M.A., and Shultz, L.D. (2017). Humanized mouse models of clinical disease. Annu. Rev. Pathol. Mech. Dis. *12*, 187–215.

Wang, X., Yang, K., Xie, Q., Wu, Q., Mack, S.C., Shi, Y., Kim, L.J.Y., Prager, B.C., Flavahan, W.A., Liu, X., et al. (2017). Purine synthesis promotes maintenance of brain tumor initiating cells in glioma. Nat. Neurosci. *20*, 661–673.

Wilding, J.L., and Bodmer, W.F. (2014). Cancer cell lines for drug discovery and development. Cancer Res. *74*, 2377–2384.

Xie, S., Cooley, A., Armendariz, D., Zhou, P., and Hon, G.C. (2018). Frequent sgRNA-barcode recombination in single-cell perturbation assays. PLoS One *13*, e0198635.

Yuan, J., Levitin, H.M., Frattini, V., Bush, E.C., Boyett, D.M., Samanamud, J., Ceccarelli, M., Dovas, A., Zanazzi, G., Canoll, P., et al. (2018). Single-cell transcriptome analysis of lineage diversity in high-grade glioma. Genome Med. *10*, 57.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

Zhang, J., Song, N., Zang, D., Yu, J., Li, J., Di, W., Guo, R., Zhao, W., and Wang, H. (2017). C-Myc promotes tumor proliferation and anti-apoptosis by repressing p21 in rhabdomyosarcomas. Mol. Med. Rep. *16*, 4089–4094.

Zhang, J., Tao, Z., and Wang, Y. (2018). Long non-coding RNA DANCR regulates the proliferation and osteogenic differentiation of human bone-derived marrow mesenchymal stem cells via the p38 MAPK pathway. Int. J. Mol. Med. *41*, 213–219.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 1–12.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal anti-Ki-67	Cell Signaling Technology	9027 (D2H10), RRID: AB_2636984
Bacterial and virus strains		
One Shot Stbl3 Chemically Competent *E. coli*	Thermo Fisher Scientific	C737303
Chemicals, peptides, and recombinant proteins		
BamHI-HF	New England Biolabs	R3136L
HpaI	New England Biolabs	R0105L
Critical commercial assays		
KAPA HiFi Hotstart Ready Mix	Kapa Biosystems	KK2602
Gibson Assembly Master Mix	New England Biolabs	E2611L
Lipofectamine 2000	Thermo Fisher Scientific	11668019
QIAquick PCR Purification Kit	Qiagen	28106
QIAprep Spin Miniprep Kit	Qiagen	27106
RNEasy Mini Kit	Qiagen	74104
DNEasy Blood and Tissue Kit	Qiagen	69506
NEBNext Ultra RNA Library Prep Kit for Illumina	New England Biolabs	E7530
NEBNext Ultra II Directional RNA Library Prep Kit	New England Biolabs	E7775
Chromium Single Cell B Chip Kit	10X Genomics	1000074
Chromium Single Cell 3' Library and Gel Bead Kit v3	10X Genomics	1000075
Deposited data		
Raw and processed data	This paper	GEO: GSE169114
Wild type teratoma characterization data	McDonald et al., 2020	GEO: GSE156170
TARGET Gene Expression data	UCSC Xena	https://xenabrowser.net
Experimental models: cell lines		
Human: H1 ES Cells	WiCell	WA01
Human: HEK293T	ATCC	N/A
Experimental models: organisms/strains		
Male NOD-scid IL2Rgamma^{null} Mice	UC San Diego ACP	RRID: BCBC_4142
Oligonucleotides		
Primer for amplification of ORF barcodes from scRNA-seq cDNA – forward, NEB_EC2H_Barcode_F: GACTGGAGTTCAGACGTGTGCTCTTCCGATCTAG AACTATTTCCTGGCTGTTACGCG	This paper	N/A
Primer for amplification of ORF barcodes from genomic DNA – forward, EC2H_gDNA_Barcode_F: ACACTCTTTCCCTACACGACGCTCTTCCGATCTA CTGTCGGGCGTACACAAATC	This paper	N/A
Primer for amplification of ORF barcodes from genomic DNA – reverse, EC2H_gDNA_Barcode_R: GACTGGAGTTCAGACGTGTGCTCTTCCGATCTC ACTGTTTAACAAGCCCGTCAGTAG	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NEBNext Multiplex Oligos for Illumina	New England Biolabs	E7335S
NEBNext Multiplex Oligos for Illumina (Dual Index)	New England Biolabs	E7600S
Recombinant DNA		
pMDG.2	Addgene	12259
pCMVR8.2	Addgene	12263
Plasmid: Ef1a_mCherry_P2A_Hygro	Addgene	135003
Software and algorithms		
Code for analyzing data	This paper	https://github.com/udit-parekh/teratoma_transformation
CellRanger	10X Genomics	https://support.10xgenomics.com
MAGeCK	Li et al., 2014a, b	https://sourceforge.net/p/mageck/wiki/Home/
Seurat	Butler et al., 2018	https://satijalab.org/seurat/
R/RStudio	R/RStudio	https://www.rstudio.com/
genotyping-matrices	Parekh et al., 2018	https://github.com/yanwu2014/genotyping-matrices
perturbLM	Parekh et al., 2018	https://github.com/yanwu2014/perturbLM
PicardTools	Broad Institute	https://broadinstitute.github.io/picard/
DoubletFinder	McGinnis et al., 2019	https://github.com/chris-mcginnis-ucsf/DoubletFinder
STAR Aligner	Dobin et al., 2013	https://github.com/alexdobin/STAR
DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
edgeR	Robinson et al., 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
limma	Ritchie et al., 2015	https://bioconductor.org/packages/release/bioc/html/limma.html
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
ImageJ	ImageJ Team	https://imagej.nih.gov/ij/
PRISM	Graphpad	https://www.graphpad.com/scientific-software/prism/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Prashant Mali (pmali@ucsd.edu).

Materials availability

Plasmids generated in this study have been deposited to Addgene.

Data and code availability

- All sequencing data have been deposited at GEO and are publicly available. Data can be accessed using accession number GSE169114.
- Code used for analysis is available at this github repository: udit-parekh/teratoma_transformation
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture

H1 male hESC cell line was maintained under feeder-free conditions in mTeSR1 medium (Stem Cell Technologies) supplemented with 1% antibiotic-antimycotic (Thermo Fisher Scientific). Prior to passaging,

tissue-culture plates were coated with growth factor-reduced Matrigel (Corning) diluted in DMEM/F-12 medium (Thermo Fisher Scientific) and incubated for 30 minutes at 37°C, 5% CO_2. Cells were dissociated and passaged using the dissociation reagent Versene (Thermo Fisher Scientific). Prior to lentiviral transduction, hESCs were passaged using Accutase (Innovative Cell Technologies) and plated as dissociated cells to achieve higher transduction efficiency. When passaged with Accutase, cells were plated in mTeSR1 containing Y27632 (10 μM, Tocris). H1 hESCs used started at P30 and were passaged a maximum of 4 passages before injection.

HEK 293T cells were maintained in high glucose DMEM supplemented with 10% fetal bovine serum (FBS) and 1 % antibiotic-antimycotic. HEK 293T cells were passaged using 0.05% Trypsin (Thermo Fisher Scientific).

Animals

Housing, husbandry and all procedures involving animals used in this study were performed in compliance with protocols (#S16003) approved by the University of California San Diego Institutional Animal Care and Use Committee (UCSD IACUC). Mice were group housed (up to 4 animals per cage) on a 12:12 hr light-dark cycle, with free access tso food and water in individually ventilated specific pathogen free (SPF) autoclaved cages. All mice used were healthy and were not involved in any previous procedures nor drug treatment unless indicated otherwise. Animals used in this study were male NOD-scid IL2Rgamma^{null} mice 6–8 weeks of age.

METHOD DETAILS

Library preparation

The lentiviral backbone plasmid (Addgene #135003), compatible with detection in scRNA-seq was constructed as previously described (Parekh et al., 2018) containing the EF1α promoter, mCherry transgene flanked by BamHI restriction sites, followed by a P2A peptide and hygromycin resistance enzyme gene immediately downstream. Each driver ORF in the library was individually inserted in place of the mCherry transgene. A barcode sequence was introduced to allow for identification of the ectopically expressed transcription factor. The backbone plasmid (Addgene #135003) was digested with HpaI, and a pool of 20 bp long barcodes with flanking sequences compatible with the HpaI site, was inserted immediately downstream of the hygromycin resistance gene by Gibson assembly. The vector was constructed such that the barcodes were located only 200 bp upstream of the 3′-LTR region. This design enabled the barcodes to be transcribed near the poly-adenylation tail of the transcripts and a high fraction of barcodes to be captured during sample processing for scRNA-seq.

To create the driver ORF library, individual drivers were PCR amplified out of the Cancer Pathways kit (Addgene #1000000072) (Martz et al., 2014), individual plasmids (Addgene #9053, #85140, #82262, #82297, #82175, #61852, #39872, #23776, #23688, #10745, #23231) (Hagting et al., 1998; Hayer et al., 2016; Johannessen et al., 2010; Kim et al., 2016; Ramaswamy et al., 1999; Roberts et al., 2008), a human cDNA pool (Promega Corporation), or obtained as synthesized double-stranded DNA fragments (gBlocks, IDT Inc) with flanking sequences compatible with the BamHI restriction sites. The barcoded lentiviral backbone was digested with BamHI HF (New England Biolabs) at 37°C for 3 hours in a reaction consisting of: lentiviral backbone, 4 μg, CutSmart buffer, 5 μl, BamHI, 0.625 μl, H_20 up to 50 μl. After digestion, the vector was purified using a QIAquick PCR Purification Kit (Qiagen). Each transcription factor vector was then individually assembled via Gibson assembly. The Gibson assembly reactions were set up as follows: 100 ng digested lentiviral backbone, 3:10 molar ratio of transcription factor insert, 2X Gibson assembly master mix (New England Biolabs), H_20 up to 20 μl. After incubation at 50°C for 1 h, the product was transformed into One Shot Stbl3 chemically competent *Escherichia coli* (Invitrogen). A fraction (150 μL) of cultures was spread on carbenicillin (50 μg/ml) LB plates and incubated overnight at 37°C. Individual colonies were picked, introduced into 5 ml of carbenicillin (50 μg/ml) LB medium and incubated overnight in a shaker at 37°C. The plasmid DNA was then extracted with a QIAprep Spin Miniprep Kit (Qiagen), and Sanger sequenced to verify correct assembly of the vector and to extract barcode sequences. One overexpression vector was created for each ORF, thus a single unique barcode was associated with each ORF.

Viral production

HEK 293T cells were maintained in high glucose DMEM supplemented with 10% fetal bovine serum (FBS).

To prepare lentivirus for the full library, to avoid barcode shuffling (Hill et al., 2018; Parekh et al., 2018; Sack et al., 2016; Xie et al., 2018) lentivirus for each ORF was packaged independently. Cells were seeded in a 6-well plate 1 day prior to transfection, such that they were 60–70% confluent at the time of transfection. For each well of a 6-well plate, 2.25 μl of Lipofectamine 2000 (Life Technologies) was added to 125 μl of Opti-MEM (Life Technologies). Separately 187.5 ng of pMD2.G (Addgene #12259), 750 ng of pCMV delta R8.2 (Addgene #12263) and 562.5 ng of an individual vector was added to 125 μl of Opti-MEM. After 5 minutes of incubation at room temperature, the Lipofectamine 2000 and DNA solutions were mixed and incubated at room temperature for 30 minutes. During the incubation period, medium in each plate was replaced with 2 ml of fresh, pre-warmed medium per well. After the incubation period, the mixture was added dropwise to each well of HEK 293T cells. Supernatant containing the viral particles was harvested after 48 and 72 hours, filtered with 0.45 μm filters (Steriflip, Millipore), and further concentrated using Amicon Ultra-15 centrifugal ultrafilters with a 100,000 NMWL cutoff (Millipore). For library lentiviral production, supernatant of all ORF wells was mixed together and concentrated to a final volume of 600–800 μl, divided into aliquots and frozen at −80°C. To prepare lentivirus for individual vectors, the production protocol as described above was scaled up to a 15 cm dish.

Viral transduction

For viral transduction, on day −1, H1 cells were dissociated to a single cell suspension using Accutase and seeded into Matrigel-coated 6-well plates at a density of $3x10^5$ cells per well in mTeSR containing ROCK inhibitor, Y27632 (10 μM, Tocris). The next day, day 0, cells were approximately 20% confluent. Medium containing Y27632 was replaced with mTeSR1 within 16 hours after plating and cells were allowed to recover for at least 8 hours prior to addition of virus.

Recovered cells were then transduced with lentivirus added to fresh mTeSR containing polybrene (5 μg/ml, Millipore). On day 1, medium was replaced with fresh mTeSR1. Hygromycin (Thermo Fisher Scientific) selection was started from day 2 onward at a selection dose of 50 μg/ml, medium containing hygromycin was replaced daily.

Teratoma formation

A subcutaneous injection of 6–8 million PSCs in a slurry of growth factor reduced Matrigel (Corning) and mTeSR medium (1:1) was made in the right flank of anesthetized, 6–8 week old, male $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice. Weekly monitoring of teratoma growth was made by quantifying approximate elliptical area (mm^2) with the use of calipers measuring outward width and height.

Teratoma processing

Mice were euthanized by slow release of CO_2 followed by secondary means via cervical dislocation. Tumor area was shaved, sprayed with 70% ethanol, and then extracted via surgical excision using scissors and forceps. Tumor was rinsed with PBS, weighed, and photographed. Tumor was then cut into small pieces in a semi-random fashion and frozen in OCT for sectioning and H&E staining courtesy of the Moore's Cancer Center Histology Core. Remaining tumor was cut into small pieces 1–2mm in diameter and subjected to standard GentleMACS (Miltenyi) protocols: Human Tumor Dissociation Kit (medium tumor settings) and Red Blood Cell Lysis Kit. From round 1 teratomas, 6–8 million dissociated live cells were resuspended in growth factor reduced Matrigel (Corning) and subcutaneously injected in the right flank of anesthetized $Rag2^{-/-};\gamma c^{-/-}$ immunodeficient mice to form round 2 tumors. For single cell RNA-seq, samples were also processed with Dead Cell Removal Kit (Miltenyi). Single cells were then resuspended in PBS + 0.04% BSA for processing on the 10X Genomics Chromium (Zheng et al., 2017) platform and downstream sequencing on an Illumina NovaSeq platform. For bulk processing, after red blood cell removal, cells were divided into multiple tubes and pelleted for 5 minutes at 300g. Supernatant was then removed and cells were either directly frozen at −80°C or resuspended in RNALater (Thermo Fisher Scientific), incubated overnight at 4°C, RNALater removed and then frozen at −80°C.

Barcode amplification

Barcodes were amplified from cDNA generated by the single cell system, and gDNA from validation tumors and prepared for deep sequencing through a two-step PCR process.

For amplification of barcodes from cDNA, the first step was performed as four separate 50 µl reactions for each sample. 2.5 µl of the cDNA was input per reaction with Kapa Hifi Hotstart ReadyMix (Kapa Biosystems). The PCR primers used were, NEB_EC2H_Barcode_F: GACTGGAGTTCAGACGTGTGCTCTTCCGATCTA GAACTATTTCCTGGCTGTTACGCG and NEBNext Universal PCR Primer for Illumina (New England Biolabs). The thermocycling parameters were 95°C for 3 min; 20–26 cycles of (98°C for 20 s; 65°C for 15 s; and 72°C for 30 s); and a final extension of 72°C for 5 min. The numbers of cycles were tested to ensure that they fell within the linear phase of amplification. Amplicons (~500 bp) of 4 reactions for each sample were pooled, size-selected and purified with Agencourt AMPure XP beads at a 0.8 ratio. The second step of PCR was performed with two separate 50 µl reactions with 50 ng of first step purified PCR product per reaction. NEBNext Multiplex Oligos for Illumina (Dual Index Primers) were used to attach Illumina adapters and indices to the samples. The thermocycling parameters were: 95°C for 3 min; 6 cycles of (98°C for 20 s; 65°C for 15 s; 72°C for 30 s); and 72°C for 5 min. The amplicons from these two reactions for each sample were pooled, size-selected and purified with Agencourt AMPure XP beads at an 0.8 ratio. The purified second-step PCR library was quantified by Qubit dsDNA HS assay (Thermo Fisher Scientific) and used for downstream sequencing on an Illumina HiSeq platform.

For amplification of barcodes from genomic DNA, genomic DNA was extracted from stored cell pellets with a DNeasy Blood and Tissue Kit (Qiagen). The first step PCR was performed as two separate 50 µl reactions for each sample. 2 µg of genomic DNA was input per reaction with Kapa Hifi Hotstart ReadyMix. The PCR primers used were, EC2H_gDNA_Barcode_F: ACACTCTTTCCCTACACGACGCTCTTCCGATC TACTGTCGGGCGTACACAAATC and EC2H_gDNA_Barcode_R: GACTGGAGTTCAGACGTGTGCTCTT CCGATCTCACTGTTTAACAAGCCCGTCAGTAG. The thermocycling parameters were: 95°C for 3 min; 24–32 cycles of (98°C for 20 s; 60°C for 15 s; and 72°C for 30 s); and a final extension of 72°C for 5 min. The numbers of cycles were tested to ensure that they fell within the linear phase of amplification. Amplicons (200 bp) of the two reactions for each sample were pooled, size-selected with Agencourt AMPure XP beads (Beckman Coulter, Inc.) at a ratio of 1.6. The second step of PCR was performed as two separate 50 µl reactions with 25 ng of first step purified PCR product per reaction. NEBNext Multiplex Oligos for Illumina (Dual Index Primers) were used to attach Illumina adapters and indices to the samples. The thermocycling parameters were: 95°C for 3 min; 6–8 cycles of (98°C for 20 s; 65°C for 20 s; 72°C for 30 s); and 72°C for 2 min. The amplicons from these two reactions for each sample were pooled, size-selected with Agencourt AMPure XP beads at a ratio of 1.6. The purified second-step PCR library was quantified by Qubit dsDNA HS assay (Thermo Fisher Scientific) and used for downstream sequencing on an Illumina NovaSeq platform.

Bulk RNA extraction and RNA-seq library preparation

RNA was extracted from cells using the RNeasy Mini Kit (Qiagen) as per the manufacturer's instructions. The quality and concentration of the RNA samples was measured using a spectrophotometer (Nanodrop 2000, Thermo Fisher Scientific).

Bulk RNA-seq libraries were prepared from 1000 ng of RNA using the NEBNext Ultra RNA Library Prep kit for Illumina or the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) as per the manufacturer's instructions. Libraries were sequenced on an Illumina NovaSeq platform.

Histology

Sectioning and H&E staining was performed by the Moore's Cancer Center Histology Core. In brief, Optimal Cutting Temperature (O.C.T.) blocks were sectioned with a cryostat into 10 micron sections onto a positively charged glass slide. The slide was then stained with Harris hematoxylin and then rinsed in tap water and treated with an alkaline solution. The slide was then de-stained to remove non-specific background staining with a weak acid alcohol. The section was then stained with an aqueous solution of eosin and passed through several changes of alcohol, then rinsed in several baths of xylene. A thin layer of polystyrene mountant was applied, followed by a glass coverslip.

Immunostaining

Fresh frozen sections were rinsed once with PBS before fixation at room temperature for 15 min with 4% paraformaldehyde. Three consecutive washes were then performed with PBS 5 min each before addition of blocking buffer (5% normal goat serum, 0.2% triton x-100 in PBS) for 1 hr. Primary antibody (anti-Ki-67 rabbit [Cell Signaling] diluted 1:200 in blocking buffer) was added overnight (12 hr) at 4C. Three consecutive washes were then performed with PBS 10 min each with gentle agitation before addition of secondary

antibody (Anti-Rabbit Dylight 550 (Abcam) diluted 1:200 in blocking buffer) for 1 hr at 37C shielded from light. Three consecutive washes were then performed with PBS 5 min each with gentle agitation before addition of DAPI (1:10,000 dilution in PBS) for 10 min. This was finally followed by three consecutive washes with PBS 10 min each with gentle agitation before imaging.

Gas chromatograph-mass spectrometry (GC-MS) sample preparation and analysis

Metabolites were extracted, analyzed, and quantified, as previously described (Cordes and Metallo, 2019). Briefly, frozen tissue was pulverized using a cellcrusher and ten to fifteen mg of frozen tissue were then homogenized with a ball mill (Retsch Mixer Mill MM 400) at 30 Hz for 5 minutes in 500 µL −20°C methanol and 200 µL of ice-cold MiliQ water. The mixture was then transferred into a 2 mL Eppendorf tube containing in 500 µL of chloroform, vortexed for 5 min followed by centrifugation with 16,000 × *g* for 5 min at 4°C. To determine the abundances of nucleobases the interface was centrifuged with 1 mL −20°C methanol at 16,000×*g* for 5 min at 4°C, the pellet was hydrolyzed with 1 mL 6N HCL for 2h at 80°C, centrifuged at 16,000×*g* for 5 min and the supernatant was dried at 60°C under airflow.

Metabolite derivatization was performed using a Gerstel MPS. Dried metabolites were dissolved in 15 µl of 2% (w/v) methoxyamine hydrochloride (Thermo Scientific) in pyridine and incubated for 60 min at 45°C. An equal volume of 2,2,2-trifluoro-N-methyl-N-trimethylsilyl-acetamide (MSTFA) (nucleobases) was added and incubated further for 30 min at 45°C. Derivatized samples were analyzed by GC-MS using a DB-35MS column (30 m × 0.25 mm i.d. x 0.25 µm, Agilent J&W Scientific) installed in an Agilent 7890A gas chromatograph (GC) interfaced with an Agilent 5975C mass spectrometer (MS) operating under electron impact ionization at 70 eV. The MS source was held at 230°C and the quadrupole at 150°C and helium was used as a carrier gas at a flow rate of 1 mL/min. The GC oven temperature was held at 80°C for 6 min, increased to 300°C at 6°C/min and after 10 min increased to 325°C at 10°C/min for 4 min.

QUANTIFICATION AND STATISTICAL ANALYSIS

Single cell RNA-seq processing

Fastq files were aligned to a combined hg19 and mm10 reference and expression matrices generated using the count command in cellranger v3.0.1 (10X Genomics). cellranger commands were run using default settings.

Data integration and clustering

Data integration was performed on counts matrices from the following samples: 4 Round 1 teratomas perturbed with the full library, 2 Round 2 teratomas perturbed with the full library, 3 Round 1 teratomas perturbed with a library without *c-MYC* or *myr-AKT1*, and 2 Round 2 teratomas perturbed with a library without *c-MYC* or *myr-AKT1*. Integration was done using the Seurat v3 pipeline (Butler et al., 2018). Expression matrices were filtered to remove any cells expressing less than 200 genes or expressing greater than 20% mitochondrial genes, as well as to remove any genes that are expressed in less than 0.1% of cells. DoubletFinder (McGinnis et al., 2019) was used to detect predicted doublets, and these were removed for downstream analysis. The expression matrix was then normalized for total counts, log transformed and scaled by a factor of 10,000 for each sample, and the top 4000 most variable genes were identified. We then used Seurat to find anchor cells and integrated all data sets, obtaining a batch-corrected expression matrix for subsequent processing. This expression matrix was then scaled, and nUMI as well as mitochondrial gene fraction was regressed out. Principal component analysis (PCA) was performed on this matrix and 22–30 PCs were identified as significant using an elbow plot. The significant PCs were then used to generate a k Nearest Neighbors (kNN) graph with k = 10. The kNN graph was then used to generate a shared Nearest Neighbors (sNN) graph followed by modularity optimization to find clusters with a resolution parameter of 0.8.

To calculate change in the cell type abundance, the number of cells of each type was summarized for each wild type and each driver library sample. This table was input to edgeR (Ritchie et al., 2014; Robinson et al., 2010), which was then used to determine $\log_2$ fold change, p value and false discovery rate (FDR).

Barcode assignment

To assign one or more barcodes to each cell, we used a previously described method (Parekh et al., 2018). Briefly, we extracted each barcode by identifying its flanking sequences, resulting in reads that contain cell,

UMI, and barcode tags. To remove potential chimeric reads, we used a two-step filtering process. First, we only kept barcodes that made up at least 0.5% of the total amount of reads for each cell. We then counted the number of UMIs and reads for each plasmid barcode within each cell, and only assigned that cell any barcode that contained at least 10% of the cell's read and UMI counts.

Bulk RNA-Seq analysis

We mapped the bulk RNA-Seq fastq files to GRCh38 and quantified read counts mapping to each gene's exon using Ensembl v99 and STAR aligner (Dobin et al., 2013). The genes by counts matrix was then processed via DESeq2 (Love et al., 2014) to normalize counts and estimate differential expression. Log fold change in gene expression of *c-MYC* and *c-MYC* + *myr-AKT1* driven tumors compared to wild type teratomas was used as input for geneset enrichment analysis.

Data from the TARGET initiative was obtained from the UCSC Xena browser. TARGET data as well as experimental tumor samples were processed using edgeR (Robinson et al., 2010) and limma (Ritchie et al., 2015) to filter and normalize the data and to remove heteroscedascity. The 2000 most highly variable genes across the TARGET tumors were then selected for performing principal component analysis.

Barcode counting

Barcodes amplified and sequenced from genomic DNA were aligned using Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009), and then counted using MAGeCK (Li et al., 2014a).

Figure generation

All figures were generated using Rstudio, Graphpad PRISM, InkScape, GIMP and ImageJ.