8-2020

LONGITUDINAL CLONAL LINEAGE DYNAMICS AND FUNCTIONAL CHARACTERIZATION OF PANCREATIC CANCER CHEMO-RESISTANCE AND METASTASIZATION

Chieh-Yuan Li

LONGITUDINAL CLONAL LINEAGE DYNAMICS AND FUNCTIONAL

CHARACTERIZATION OF PANCREATIC CANCER CHEMO-RESISTANCE AND

METASTASIZATION

by

Chieh-Yuan (Alex) Li, B.S.

APPROVED:

_____
Giulio Draetta, M.D., Ph.D.
Advisory Professor

_____
Andrea Viale, M.D.
Advisory Professor

_____
Nicholas Navin, Ph.D.

_____
Anirban Maitra, M.B.B.S.

_____
Richard Behringer, Ph.D.

_____
Haoqiang Ying, M.D., Ph.D.

APPROVED:

_____
Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

LONGITUTINAL CLONAL LINEAGE DYNAMICS AND FUNCTIONAL

CHARACTERIZATION OF PANCREATIC CANCER CHEMO-RESISTANCE AND

METASTASIZATION

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Chieh-Yuan (Alex) Li, B.S.
Houston, Texas

Date of Graduation, *August, 2020*

Copyright

Chapter one of this thesis is based on the article titled *Pre-existing Functional Heterogeneity of Tumorigenic Compartment as the Origin of Chemoresistance in Pancreatic Tumors* (https://doi.org/10.1016/j.celrep.2019.01.048) and was published on the 5th of February in 2019 in *Cell Report*. This is an open-access article distributed under the terms of the Creative Commons Attribution - NonCommercial - NoDerivs 4.0 (CC BY - NC - ND). This license allow users to copy and distribute the Article, provided that this is not done for commercial purposes, and further does not permit distribution of the Article if it is changed or edited in any way, and provided that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. Under Creative Commons, authors retain copyright in their articles (in a thesis or dissertation, provided it is not published commercially) and I am the primary author of this article. All author contributions could be found here: https://doi.org/10.1016/j.celrep.2019.01.048. The rest of the thesis contents are original work and copyright by author Chieh-Yuan Li.

To my parents, Chung-Meng Li and Tzu San Yen

To my sister, Jill Li

To my dear wife, Susan Aung, who has always been there and supporting me.

I love you.

Acknowledgments

I would like to thank my mentors Dr. Giulio Draetta and Dr. Andrea Viale for the support and guidance throughout my training and accepting me into your laboratory. I am especially thankful for the open minded and exploratory style of your mentoring, and your focus on the discovery aspect of science. I especially remember the times when where we would discuss new topics and how to conduct experiments, and in the end, you would frequently end the discussion with "do whatever you want, we are still learning." It is that sense of discovery spirit that drives us. On the personal side, when I encounter hardship, you are both very understanding and really enforce the values of family by assisting me through hurdles. I cannot say thank you enough. Most importantly, thank you for always keeping the doors open for me, meeting on the weekends on top of your busy schedules. I would also like to thank my committee members, Dr. Navin, Dr. Maitra, Dr. Behringer, and Dr. Ying, for all your scientific inputs and guidance. Especially Dr.Nick Navin, who I had the honor to rotate with during my first year. Even after rotation, he continues to support me throughout my Ph.D. training both in personal development guidance and scientifically. You are one of the persons I look up to. In addition, I would like to thank all the members of the Draetta lab, Viale lab, and Dr. Kim Do's lab – Dr. Do, Dr. Lin and Delia Wang. Especially I-Lin Ho, one of the most dedicated and responsible person I know with uncompromising good moral value. I am grateful for my graduate school, GSBS and the program, which really provided a protective nursery for students. Especially Dr. Bill Mattox, Brenda Gaughan and Genes and Development Program Manager Elisabeth Lindheim, who always go above and beyond with genuine care to make sure that we are okay.

LONGITUDINAL CLONAL LINEAGE DYNAMICS AND FUNCTIONAL CHARACTERIZATION OF PANCREATIC CANCER CHEMO-RESISTANCE AND METASTASIZATION

Chieh-Yuan (Alex) Li, B.S.

Advisory Professor:

Giulio Draetta, Ph.D.   Andrea Viale, M.D.

In recent years, technological advancements, such as next-generation sequencing and single-cell interrogation techniques, have enriched our understanding in tumor heterogeneity. By dissecting tumors and characterizing clonal lineages, we are better understanding the intricacies of tumor evolution. Tumors are represented by the presence of and dynamic interactions amongst clonal lineages. Each lineage and each cell contributes to tumor dynamics through intrinsic and extrinsic mechanisms, and the variable responses of clones to perturbations in the environment, especially therapeutics, underlie disease progression and relapse. Thus, there exists a pressing need to understand the molecular mechanisms that determine the functional heterogeneity of tumor sub-clones to improve clinical outcomes.

Clonal replica tumors (CRTs) is an *in vivo* platform created specifically to enable robust tracing and functional study of clones within a tumor. The establishment of CRTs is built upon our current concept of tumor heterogeneity, intrinsic cancer cell hierarchy and clonal self-renewal properties. The model allows researchers to create large cohorts of tumors in different animals that are identical in their clonal lineage composition (clonal correlation amongst tumors >0.99). CRTs allow simultaneously tracking of tens of thousands of clonal lineages in different animals to provide a high level of resolution and biological reproducibility. CRTs are comprised of barcoded cells that can be identified and quantified. A critical feature is that we have developed a systematic method to isolate and expand essentially any of the clonal lineages present within

a CRT in their naïve state; that is, we can characterize each sub-clonal lineage at the molecular and functional levels and correlate these findings with the behavior of the same lineage *in vivo* and in response to drugs.

Here, based on the CRT model and its concept, we studied differential chemo-resistance among clones, where we identified pre-existing upregulation in DNA repair as a mechanism for chemo-resistance. Furthermore, through stringent statistical testing, we demonstrated orthotopic CRTs to be a powerful and robust model to quantitatively track clonal evolution. Specifically, we longitudinally tracked clones in models of pancreatic ductal adenocarcinoma (PDAC) from primary tumor expansion through metastasization, where we captured unexpected clonal dynamics and "alternating clonal dominance" naturally occurring in unperturbed tumors. Moreover, by characterizing pro- and none-metastasizing clones, we were able to identified key clonal intrinsic factors that determined the nature of tumor metastases. Finally, I will discuss distinct clonal evolution patterns that emerged under different environmental pressures, leading to the hypothesis of "tumor clonal fingerprint", where the characteristic of a tumor could be defined by actively maintained ratio of different tumor lineages, which could provide measurable insights to how we approach treatments.

Within this thesis, I provide detailed technical and analytical descriptions to establish the relevance of CRTs as a model, as well as experimental data generated to test hypotheses regarding tumor functional heterogeneity and evolution. This thesis includes three chapters, which all interconnect to form one large picture on how tumor cells behave. Each chapter contains key backgrounds to facilitate the reading experience and so each chapter may be read independently.

Table of Contents

Table of contents

Chapter 3

Abundance independent preservation of clone-to-clone ratio in cancer and the prospective concept of "tumor clonal fingerprint"

15

List of Tables………………………………………………………………………………………xi

Introduction - observation of a piece of land

There is a piece of land in front of MD Anderson, south campus, where I would pass by whenever I go to the animal facility from my office. The lot of land was flourished with greens, flowers, little lizards, crickets and butterflies, it was a beautiful sight and full of life. One day, to my surprise, everything on that land turned yellow. It turns out that weeding treatment were applied to that lot and it killed everything. I was stunned by how fast the landscape changed. Instead of what it once was, it is now filled with lifeless and brittle plants. The land is now "silent", as if it has taken a break and halted in time.

As the sun rises and falls, as tides come and goes, life moves on. Gradually, resurgence of sprinkled hues of vibrant green started to appear. Notably, a single kind of plant was seen to expand and steadily spread throughout the land. The barren was resilient, and life has returned.

From my green point of view, I continued to monitor the landscape throughout the years. From the initial monoclonal type of grass dominating the entire field, other types of grass and eventually flowers started to emerge. Spatially apart, and heterogeneous patches composed of homogeneous plants start to sprout out of the ground and colonize everywhere. Evening primrose, white thistle, dandelion, yellow wood sorrows, black-eyed Susan, clover…life is in full bloom.

As the weather changes and "Houston level" heat rises, most of the flowers starts to diminish, and progressively, the land is dominated by fewer types of plants, thriving in their own

specific niche. "Epi-flora" wise, insects and birds have returned, crickets can be heard during dusk, the ecosystem has re-balanced itself.

I cannot help but think of the uncanny similarities between this land and cancer. Just like the land is comprised of various species of plants, cancer, as a whole, is made up of heterogeneous sub-clones with different functional phenotypes. The unbiased treatment of "weed killing chemicals" mirrors chemotherapies targeting proliferating cancer cells; the initial proliferating mono-species grass post treatment mirrors chemo-resistant clones relapsing; and the inter-species competition to specific niches supporting growth mirrors intra-tumor clonal diversifying and competition for resource and tumorigenic niches. Furthermore, in the end, as a whole system, clonal equilibrium is established by unique interactions between inter-clonal dynamics. Just like this piece of land, cancer, is an ever-evolving system based on intricate dynamics among clones establishing a balance under different environmental perturbations.

How to find a chemical to target only weed but not flowers?

Start with studying understand subtypes

| Effectiveness? | What grew back first? | Which are resistant? | Composition of heterogeneous populations? |

To understand how cancer evolve and adapt to external perturbations such as treatment response and metastasization outgrowth, it is logical to breakdown the tumor and study their fundamental components, which are sub-clones with functionally diverse phenotypes. However, at the same time, we should be cautious to study these sub-clones within the context of natural

tumor progression and environment. Here, in this thesis, I will discuss how we aim to create a high resolution and sensitive model for unbiased tracking of clonal evolution in the whole tumor longitudinally, from tumor progression to metastasis, as well as relapse post-chemotherapy with good biological reproducibility. I hope through this study, we could capture and recognize the natural rules and dynamics amongst tumor lineages.

Chapter 1

Pre-existing Functional Heterogeneity of Tumorigenic Compartment as origin of Chemoresistance in Pancreatic Tumors

Abstract

Adaptive drug-resistance mechanisms allow human tumors to evade treatment through selection and expansion of treatment-resistant clones. Here, studying clonal evolution of tumor cells derived from human pancreatic tumors, we demonstrate that *in vitro* cultures and *in vivo* tumors are maintained by a common set of tumorigenic cells that can be used to establish clonal replica tumors (CRTs) - large cohorts of animals bearing human tumors with identical clonal composition. Using CRTs to conduct quantitative assessments of adaptive responses to therapeutics, we uncovered a multitude of functionally heterogeneous subpopulations of cells with differential degrees of drug sensitivity. High-throughput isolation and deep characterization of unique clonal lineages showed genetic and transcriptomic diversity underlying functionally diverse subpopulations. Molecular annotation of gemcitabine-naïve clonal lineages with distinct responses to treatment in the context of CRTs generated signatures that can predict the response to chemotherapy, representing a potential biomarker to stratify patients with pancreatic cancer.

Introduction

Tumors are complex ecosystems composed of sub-clones with diverse mutational landscapes driving cancer evolution

Sequencing technologies has made huge strides over the last several years. Researchers are no longer restricted to Sanger sequencing, where fragments of DNA must be amplified and decoded individually by capillary electrophoresis[1]. In 2005, pyrosequencing came to market, where it adapted the concept of sequencing by synthesis and can sequence massive amounts of

different DNA fragments in parallel[2-3]. The biotechnology sector continues to push the limits of next-generation sequencing (NGS) by creating advanced platforms of NGS, each with its own unique merits for suitable applications[4-8]. Echoing Moore's law, the exponential improvement over cost and throughput has made NGS a key technological platform for molecular characterization in biology research, and it is one of the representative technologies in the 21th century[9-11].

While NGS propelled the scale and speed, on the sample preparation side, breakthroughs for sensitivity such as extreme high-multiplexed PCR assays and low-sample quantity requirements for biological materials were made concurrently[12-14]. To name a few, a milestone of single-tube 23,000 plex PCR chemistry for target amplification of whole human exome with 10ng DNA input[15-17], random degenerate oligos as unique molecular identifier for improved uniformity and accurate biological representation[19], and microfluidics droplet-based nanoscale reactions for high throughput sample processing[20-22]. Advances such as single-cell sequencing have allowed researchers to study cancer at an unprecedented resolution[23-27]. Both single-cell genomic and transcriptomic characterization of tumors have started to shed light on the complex, dynamic bionetwork of tumors, wherein populations of cells harboring a unique mutational landscape coexist and progressively drift from a founder clone accumulating functional diversity[28-30]. "Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression – Peter C. Nowell", in 1976, describes the clonal evolutionary process well with much foresights [28]. As the disease progresses, clones may acquire genetic and epigenetic mutations, which may lead to functional growth advantages for sub-clones within the tumor while under selective pressure that direct the evolution of the tumor as a whole[31]. Clonal selection and expansion are apparent in the context of adaptive resistance to chemotherapy. During chemotherapy, the tumor mass often reduces; however, post therapy, minimum residual disease composed of a fraction of resistant clones may survive and eventually re-establish the tumor[32]. The relapsed tumor becomes more resistant due to its new clonal composition. A striking example is resistance to vemurafenib, a BRAF inhibitor, in melanoma, where the patient shows significant

treatment response initially, but residual tumor clones with adapted or innate resistance mechanisms expand, causing relapse in merely months[33].

Clonal progression and selection are also observed during metastasization. As tumors progress to late stage, clones with advantageous features disseminate to distal organs through the metastasization cascade and resides in specific niches that supports propagation[34-37]. Studies focusing on clonal characterization have yielded informative models regarding the time of seeding (early vs. late dissemination), latency (seeding vs. outgrowth) and corporation among clones during metastasis (cluster migration, poly- or mono-clonal lesions), as well as reduction in clonal heterogeneity in metastasis[38-42]. These observed genetic clonal bottle necks and clonal diversity each contribute to the selection process in cancer evolution during different stages of the disease, and are shown to be impacted by the treatment regimen[43].

Limitations of experimental models for clonal characterization and validation

<u>Resolution and accurate representation of the entire tumor</u>

Clone-level resolution of molecular profiling for clinical samples is often hindered by the cost and reduced level of sample representation of the entire disease. To be cost effective, only a fraction of the tumor is often profiled. To maintain single-cell resolution, profiling is usually achieved through droplet-based PCR or through FACS cell sorting after the tumor is processed and dissociated. Practically, the number of cells one can characterize is in the hundreds to thousands, which is generally represents only a small fraction of the entire tumor. Thus, while this approach can provide a general picture of the molecular composition within a tumor, it cannot capture the tumor in its entirety without sacrificing resolution. To capture the molecular profile with reduced resolution, regional or compartmentalized sequencing is commonly used[44-45]. Instead of single cell resolution, a tumor is divided into small sections and each processed for molecular profiling. The molecular profile of individual sections of the tumor are then pieced together to

represent the entire tumor profile. However, again, limited sampling prevents scientist from obtaining the full scope of tumor subclonal interactions and evolution. This problem is well discussed by Alves and Co., where limited sampling obtained (multiregional) tumor evolution trees are not phylogenies due to biased and limited sampling representation[141]. Furthermore, notably, functional definition of clones is not fully predicted by genomic and/or transcriptomic profiles; epigenetic signatures and other factors contribute as well[39]. Unfortunately, with current methods for tumor profiling, choices must be made regarding which analyses should be performed, with many interesting questions left unanswered after the tissue is used up.

Single time-point snap shot of the disease

Performing epi/genomic or transcriptomic profile of a tumor is informative; however, it only characterizes the current state of the tumor, with no information regarding how the tumor once were or how it might progress in the future. This is a common challenge in the field, as continuous re-sampling of solid tumors in the patient is limited. In addition, even though fine needle biopsy allows for re-sampling, besides the technical inherited problem of lacking in sample representation, the biopsy procedure perturbs the natural progression of the tumor. Preventing us from profiling tumor in its entirety as well as monitoring its progression in an unbiased manner. Therefore, to study clonal evolution, it is imperative to find a good experimental model that captures the relevant context of the disease.

Animal models for tumor clonal evolution

Genetically engineered mouse models (GEMMs) are useful to deconvolute specific aspects of the disease under study; however, GEMMs do not fully recapitulate the degree of complexity in a human tumor, which often evolves and accumulates mutations over a long period of time. Conversely, patient-derived xenograft (PDX) models can retain the histological and genetic features of the primary tumor, and these models are extensively used to evaluate tumor

therapeutic response in a pre-clinical setting[46]. Combining PDXs and lentiviral barcoding technology has proven to be an effective method to trace clonal lineages and study tumor evolution, as well as delineate normal tissue stem cell hierarchy[46-53]. But, barcode characterization of PDXs is overwhelmed with problems such as the inability to establish biological replicates, therefore, lacking in reproducibility and power to draw confident conclusions. This reproducibility issue prevent researchers from designing meaningful, statistically sounds experiments. In addition, following lineage tracing experiment, there lack a method for researchers to functional study or validate of clones of interest.

Current state of pancreatic ductal adenocarcinoma

Pancreas is the organ that regulates our body's digestion and glucose homeostasis. Majority of the pancreatic mass is comprised of exocrine functional cells, such as acinar cells and ductal cells, which secrete enzymes for digestion. The remaining portion of the pancreas is comprised of cells related to the endocrine system, the islet cells, which secrete hormones through the bloodstream to regulate glucose usage in our body. Pancreas is an organ that is vital in maintaining and supporting our body's metabolism and energy homeostasis[54].

PDAC has relatively defined stages based on histology and mutational landscape in each pancreatic intraepithelial neoplasia (PanIN stages I to IV). Histologically, normal pancreatic ductal structure becomes less and less defined as disease progresses accompanied by KRAS mutation, CDKN2A, TP53 and SMAD4 loss of function[54-55]. It is estimated that the pancreatic carcinogenesis progression takes more than 12 years, and another 7-10 years to develop metastasis[56]. However, due to the typical asymptomatic characteristics of the disease, most patient are diagnosed at a late stage of the disease, where tumor cells have already disseminated. This leads to PDAC being one of the worst prognoses rates with only ~7% 5-years survival rate, 4th leading cause of cancer death in US[57].

There is currently no cure for PDAC. Patients diagnosed with pre-metastasis by clinical standard has the option for pancreaticoduodenectomy (~20%) in combination with neoadjuvant therapy; however, ~80% of the patients have recurrence after surgery with a 5-year survival rate of less than 20%. Patients diagnosed with advanced tumor (metastatic) are given treatment as a palliative measure. Treatments include chemotherapy (gemcitabine, abraxane), radiation, erlotinib (TKI), FOLFIRINOX; targeted therapy includes herceptine, erbitux. Nonetheless, tumor resistance occurs within 6-8 months. Drugs with 2-5weeks of prolonged survival are being approved by FDA for patients with advanced PDAC [58-59].

Replicating tumor heterogeneity – creating clonal replica tumors

To overcome the limitation of lacking biological replica tumors in experimental cohorts with the identical clonal composition, we developed a novel molecular barcoding approach to model tumors in transplantation.

Focusing our studies on pancreatic cancer, we used lentiviral barcoding to barcode patient derived xenograft cell lines with a high complexity of barcode library[60-61]. Then, we track the evolution of the same barcode lineages (clonal lineages) under different environmental conditions, *in vitro* passage in culture and *in vivo* serial transplantation in recipient mice, in parallel over a long period of time. As time goes by, we observed that the clonal populations eventually come to an equilibrium state, where subsequent passages would contain relatively similar barcode composition and their relative abundance. Furthermore, through this continuous passaging process, it is revealed that lineages *in vivo* and *in vitro* share a similar evolutionary trajectory, where a common set of lineages that are endowed with long-term self-renewal (LTSR) ability can maintain the tumor *in vivo*, as well as the culture *in vitro* over time.

These clonal lineages with LTSR properties are able expand *in vitro* almost indefinitely. Using these cells, we can create large cohorts of animals baring tumors with identical clonal composition. We call these tumors Clonal Replica Tumors (CRTs).

CRTs allow the exploration of clonal evolution at a very high level of resolution and provide the biological reproducibility and statistical power from experimental replicates that is needed in the study of tumor heterogeneity. CRTs enable the evaluation of clonal differential response to single agent or combination therapy, where we could track individual sub-clonal populations in independent tumors (independent animals) with high precision.

In addition to high-resolution tracking of clonal evolution, the CRT system enables the isolation and functional characterization of essentially any sub-clone of interest based on its behavioral phenotype in the *in vivo* CRT experiment. Upon identifying tumorigenic clonal lineages with extreme therapeutic response via CRT cohorts, we employ a high throughput clonal isolation approach to isolate and expand clones of interest at a treatment-naïve state for further functional characterization. From the deep characterization of these individual treatment naïve clones with different degrees of response to chemotherapy, we demonstrate that the tumorigenic clonal lineages that sustain tumor and their evolution are genomically and transcriptomically diverse and display differential response to different external perturbations. Finally, differentially expressed transcriptomic profiles from the isolated treatment naïve clones with various degrees of sensitivity to chemotherapy generated a molecular profile that can predict therapeutic response to standard-of-care in patients with pancreatic cancer.
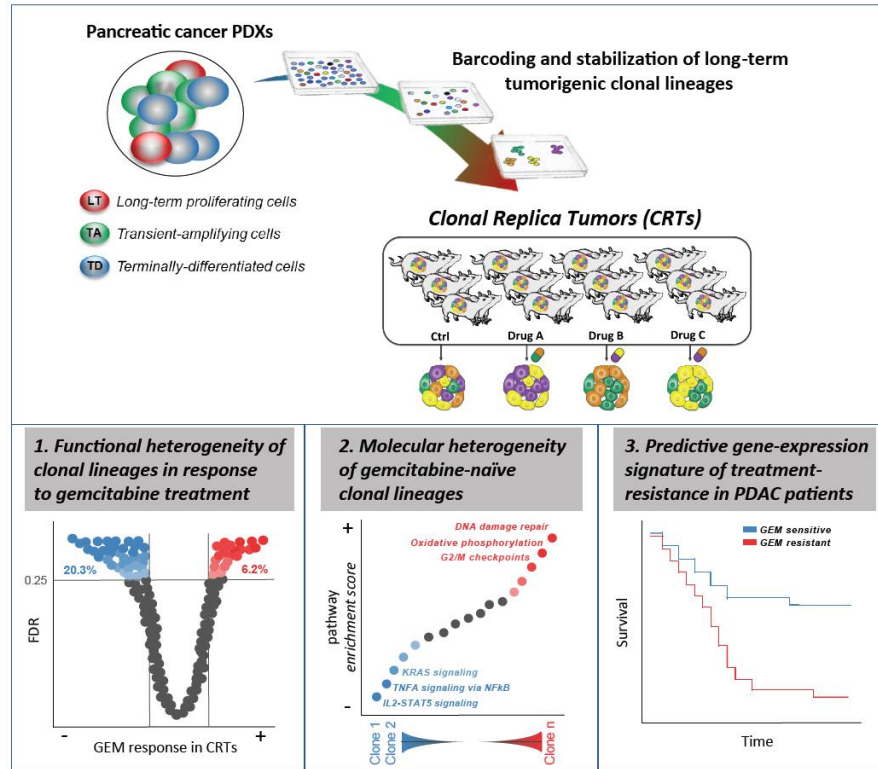
Figure 1. Schematic overview of clonal replica tumor establishment and functional characterization of lineage specific gemcitabine sensitivity. High-complexity lineage tracing enables the generation of large cohorts of Clonal Replica Tumors (CRTs). Applying the CRT technology to pancreatic cancer reveals functional clonal heterogeneity in response to therapeutics. Specifically, characterization of gemcitabine-naïve clones identifies a molecular signature able to predict chemotherapy resistance in patients.

Results

Barcoding PDX derived PDAC tumor cells

To establish clonal lineage tracing through an unbiased marker, we employed a highly complex barcoded lentiviral tracking library[62]. Each barcode in the library consists of two segments of 18bp and 18bp degenerate bases, and the combined sequence makes up for a unique ID. The variation of the barcode library is ~ 10 million. As for the tumor of study, we chose early passage of PDX cells from human PDAC primary tumor, which were directly embedded in mouse immediately after resection from patients. These well-characterized PDX derived cell

model after re-transplantation in mouse retains and mirror patient tumor's morphological features and molecular characteristics. Besides, the PDX derived cell are efficient to culture, expand and have high engraftment efficiency, which are essential characteristics to minimize variabilities in experiments and building a robust clonal tracking system[63, 64]. Here, in this thesis, we will mainly discuss two PDX derived cell lines, PATC124 and PATC53.

To retain as much of the original tumor heterogeneity, we used around 30 million cells and transduced them with the barcode library at a low multiplicity of infection (MOI ≤ 0.25). The logic is based on Poisson's distribution, where majority of the cells will have no barcode integrated, which increases the chance of one barcode integrant per cell (barcoded cells express RFP protein, the percentage of positive integration is checked by FACS analysis). With low MOI and a high complexity of barcode variations, we can ensure that most of each progenitor cells that establishes subsequent lineage carries a unique barcode ID, which is critical in clonal tracing experiments. Then, the infected cells were selected with puromycin and allowed to expand and passaged. During passaging, we split the culture with a minimum dilution of 1:2 to preserve as much heterogeneity as possible and to minimize the effects of passaging on clonal composition of the culture. The barcoded cells can then be processed (DNA extracted and barcode region amplified) for assessment of barcode identity through NGS readout.

Figure 2. Schematics of barcoding of PDX derived tumor cell lines. Early passage of PDAC patient PDX derived cell lines are infected with lentiviral barcode library (10M unique barcode types). Using a low MOI (Poisson distribution), vast majority of the cells will have one unique barcode integrant. Then, puromycin was used to select for cells with barcode (barcode vector also contains puromycin resistance and red florescent protein).

13Kx13K CB Library in pRSI16 vector: - size of Gex1-Bpi/NR2 amplicon is 267 bp.

ClaI-U6-FwdU6-1>FwdU6-2>EcoRI>FwdHTS3>Gex1-Bpi>BarCode18TTCG BarCode18<GexSeqS<XhoI<cPPT<Gex2-NR2<R2<RevUbiC1

```
                              FwdHTS3
                       TCGGATTCAAGCAAAAGACGGCATA Gex1-Bpi
          Bpi                 TCAAGCAGAAGACGGCATACGAAGACA
U6-ACCGGAGTCTTCTTTTTTGAATTCAGCAAAAGACGGCATACGAAGACAGTTCG-NNNNNNNNNNNNNNNNNNTTCGNNNNNNNNNNNNNNNNNN-TTCGGACTGTAGAACTCTGAACCTCGAGCAA
U6-TGGCCTCAGAAGAAAAAACTTAAGTTCGTTTTCTGCCGTATGCTTCTGTCAAGC-NNNNNNNNNNNNNNNNNNAAGCNNNNNNNNNNNNNNNNNN-AAGCCTGACATCTTGAGACTTGGAGCTCGTT
                                                                                                    AAGCCTGACATCTTGAGACTTGGAGA
          cPPT                                                                          GexSeqS
TTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAATT
AAATTTTCTTTTCCCCCCTAACCCCCCATGTCACGTCCCCTTTCTTATCATCTGTATTATCGTTGTCTGTATGTTTGATTTCTTAATGTTTTTGTTTAAGTTTTAA
                              EcoRI    AgeI   BstBI BspEI       UbiC promoter
TCTGCGTTGTTGTCGGTGCTCGTTCTCTGCTCTTCACGCTACTGAATTCATCACCGGTTCTTCGAAGGCCTCCGCGCCGGGTTTTGGCGCCTCCCGCGGGCGCCCCCCTCCTCACGGCG
AGACGCAACAACAGCCACGAGCAAGAGACGAAGTGCGATGACTTAAGTAGTGGCCAAGAAGCTTCCGGAGGCGCGGCCCAAAACCGCGGAGGGCGCCCGCGGGGGGAGGAGTGCCGC
AGACGCAACAACAGCCACGAGAGCCACCAGCGGCATAGTAA                                              AGGAGTGCCGCTCGCGACGGA
          Gex2-NR2           AAGAGACGAGAAGTGCGATGA                                             RevUbiC1
                              R2
```

1. Map of 13K×13K Lentiviral Barcode Library Vector

Check most cells are barcoded in PATC 53 1st infection

Merged RFP, BF at 100X

Figure 3. Clonal tracking vector and barcode region's design sequence. Barcodes consists of two segments of degenerate 18bp nucleotides, which provides up to ~10M unique combination of barcodes for high complexity clonal tracking. This figure is adapted from Cellecta™ (Top and bottom left); Barcoded cells also express RFP. Bottom right is a representation figure of tumor cell line (PATC53) post infection and puromycin selection, where all barcoded cells are red.

Enabling quantitative assessment of barcodes

Tracing barcodes alone can tell us the relative abundance between clones in the tumor, and thus, allow us to study dynamics amongst them. However, just by looking at "relative abundance" between lineages fail to capture the absolute quantity of the clones present in a tumor. The questions relating to "how many cells" cannot be addressed sufficiently by just looking at relevant ratios of lineages.

To be able to assess each barcode lineages quantitatively, and not just by relative abundance, we created a "NGS read to cell number conversion scale" which is spiked into the sample of interest while processing for barcode readout. The "scale" is composed of a set of

known number of cells (i.e. 5 cells, 50 cell, 500 cells, 5,000 cells, 50,000 cells) with each of their own unique barcodes that are not in the clonal tracking barcode library. By spiking the same scale alongside each of the samples during processing, the scale acts as an internal reference point and can be used to normalize clonal abundance across all samples. Furthermore, it allows us to convert individual barcode's NGS read into cells, which give us the quantitative aspect of clonal tracking. By doing so, we can now answer how many "cells" is in each clonal lineage.

In addition, the "scale" also serves as a function to gauge the detection sensitivity from the entire workflow, which is around 1 cell out of 7×10^7 cells from this study. It also helps us to define a "cell based "cut-off for accurate quantification across samples (i.e. 50 cells). Furthermore, the size of the "scale" in terms of cell number is adjustable, for a more sensitive application, the scales size could be adjusted downwards as needed.
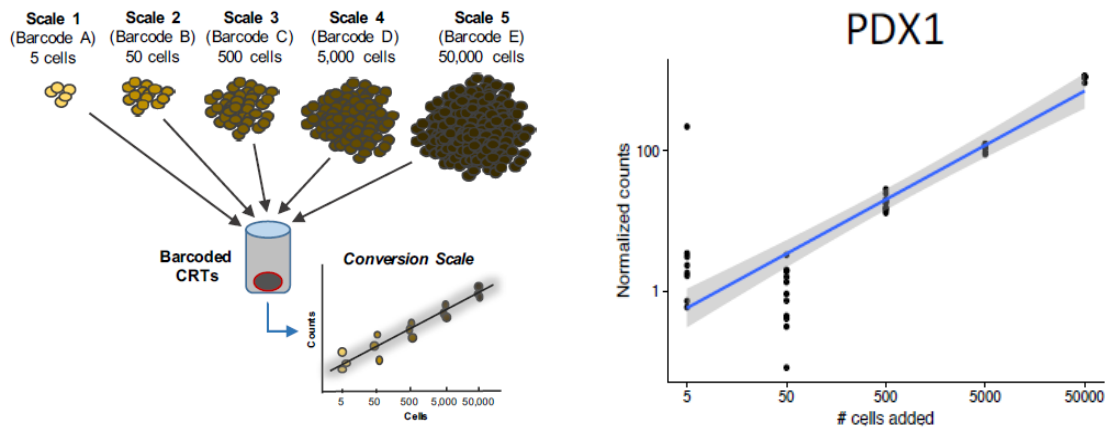
Figure 4. Schematic for the Conversion Scale spike-in to enable quantitative aspect of clonal tracking. Pools of 5 incremental number of cells (5, 50, 500, 5,000, and 50,000) are generated by mixing precise amounts of cells independently infected with 5 unique barcodes (A-E) not represented in the tracking library. Added to each sample before processing, the Conversion Scale can be leveraged to accurately convert barcode read counts into number of cells (Left); Fitting line for the sequencing counts (normalized) generated by the Conversion Scale mixed with the barcoded PDAC samples for PATC124 and PATC53 (right).

A common subset of lineages endowed with LTSR and tumorigenic potential maintain primary cell cultures *in vitro* and tumors *in vivo*

Since the barcode process is unbiased and universally applied to the entire population of tumor cells, clonal tracking is a suitable way to assess cell and lineage intrinsic differential self-renewal and tumor initiating potential. If a cell has self-renewal potential, we expect to see the barcode associated with the cell to increase in representation over time; conversely, if a cell that is differentiated, and not able to establish a lineage of its own, we expect to see the barcode associated with that cell to dropout over time. Based on this, we designed a functional study to survey the cell's LTSR potential by culturing the barcoded cells in parallel in both *in vitro* and *in vivo* conditions and study the clonal evolution.

Right after positive selection of barcoded cells, the cell is expanded (passage 2), collected and split into two pools with identical barcode composition. One half of the barcoded cells are subjected to continuous passaging *in vitro* (with 1:2 passage ratio from one passage to the other, ~50 million cells) for 8 weeks (P1 to P21), and eventually transplanted *in vivo* (at passage 26) to measure their clonal tumorgenicity; the other half of the barcoded cells are serially transplanted *in vivo* in NSG mice over 32 weeks (F1 to F4, 1:2 ratio from one passage to the next). Each passage *in vitro* and tumors *in vivo* are collected, and their clonal composition assessed via NGS barcode readout.
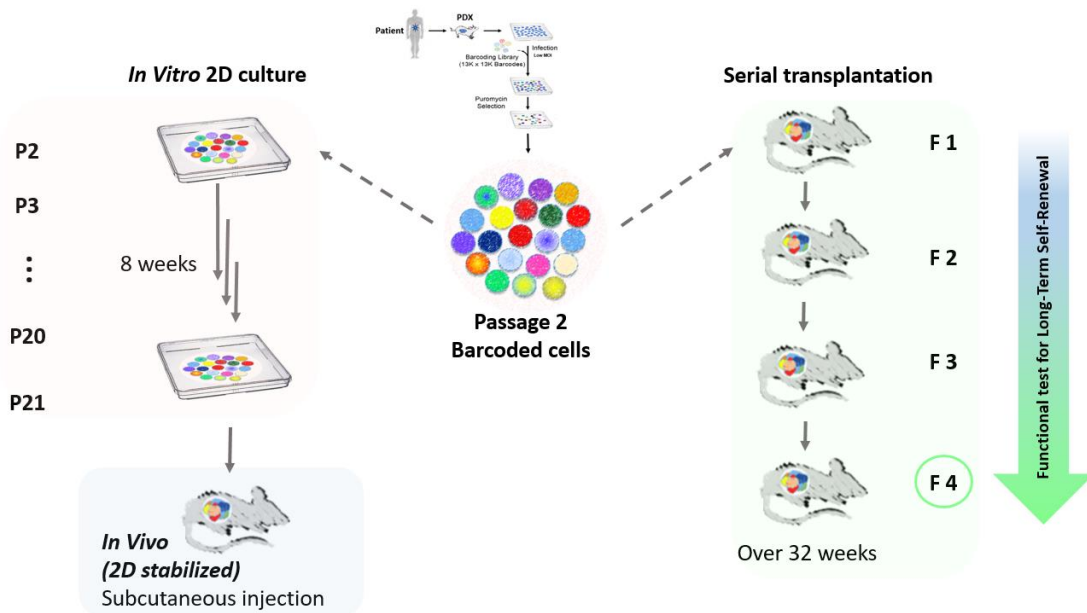
Figure 5. Schematics of *in vitro* and *in vivo* clonal tracking experiments. Barcoded cells were expanded and split into two experimental arms – *in vivo* and *in vitro*. For the *in vivo* serial transplantation, each passage consists of five mice. Between passages, all tumors were combined and mixed. Half of the mixed tumor is passaged to the next generation amongst five mice in equal amount. A total of four passages were done from F1 to F4 over a span of 32 weeks. As for the *in vitro* passaging, the cells were passaged at a 1:2 ratio to preserve as much clonal heterogeneity as possible. The *in vitro* passages were carried out for over 8 weeks and for 21+ passages before injecting into mouse.

PDAC are sustained by clones with LTSR potential and display cellular proliferation potential hierarchy

Following the clonal evolution from the *in vivo* serial transplantation experimental arm, initially, we can see a drastic reduction in barcode complexity (~ 2M initial barcodes) of ~80% drop in the first-generation tumor (F1). The reduction in barcode complexity continues to drop as the passages increase, eventually, with only around 0.67% of the initial barcode complexity left in F4. See Figure 6 below, each line represents a unique lineage. By following the clonal dynamics over the passages, we can see that most of the lineages fail to engraft and are exhausted over time (represented by blue); we can also observe lineages displaying transit amplification characteristics, where clonal expansion is observed initially, but followed by eventual exhaustion (represented by green); we also observe lineages that maintain the tumor over time and are

increasing in representation (represented by red). Finally, a subgroup of lineages that seems to be "dormant" in the initial passages F1 to F3 with low abundance and eventually starts to expand (represented by orange).

In vivo serial transplantation represents a universal recognized functional assay to determine cellular long-term self-renewal properties and stem-like cells in both normal tissue and tumors [65]. From the experiment, the clonal evolution pattern (from F1 to F4) where clonal lineages display differential sustainable expansion capability, ending with only a fraction of the clonal lineages carry the ability to sustain the tumor's long-term growth. These "cellular hierarchy" observation is consistent with the observations from studies in both solid tumors and leukemia[66-74]. This set of data supports these cellular hierarchy in terms of differential proliferation capability exists in PDAC, and with barcode tracking, we can identify these clones with LTSR capability.
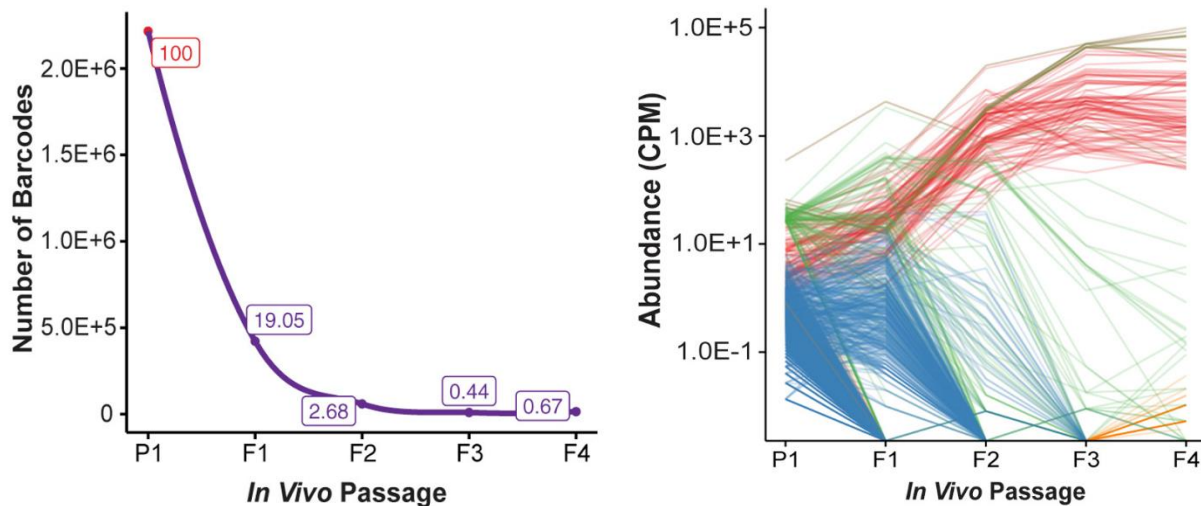
Figure 6. Tracing clonal complexity and individual clonal dynamics in *in vivo* serial transplantation. The left graph shows the percentage of unique barcodes detected in percentages normalized to P1 (P1 has ~2M unique barcodes). Only a portion of the clones have the ability to "engraft" and form tumor (19.05% in F1, and only a small fraction of the clones have the ability to sustain tumor (F3, F4 with 0.44% and 0.67%) ; The right graph shows the differential clonal dynamics over serial passages *in vivo* (F1, F2, F3, F4): functionally exhausted clones (blue), transiently engrafting clones (green), long term self-renewing clones (red), early passage low represented/undetected clones enriched upon *in vivo* transplantation at passage F4 (orange).

<u>PDAC culture (*in vitro*) are sustained by LTSR clones and follows similar cellular hierarchy structure *in vivo*</u>

On the other hand, *in vitro* clonal evolution is captured the same way by sequencing passages P1, P5, P11, P15, P21 *in vitro*, as well as the *in vivo* tumor transplanted at passage 26. The clonal dynamic behaves in a similar fashion where it displays clonal-specific differential proliferation potential. Notably, barcode complexity drops sharply in the initial passages and eventually comes to an equilibrium, where the culture is sustained by only a fraction of the initial clonal lineages (~3.6%). Throughout passages, there are clones that exhaust over time (blue); transit amplification lineages where initially expand and eventually exhaust (green); and fraction of the lineages that sustain the culture and can engraft to initiate tumor (red). As seen from the data, the engrafted clonal lineages represent only about 1.67% of the initial clonal complexity, however, those clonal lineages make up the majority (~94.5%) of the stabilized culture after ~passage 15 to 16. In addition, the common lineages present in the *in vivo* tumor (injected at P26) and the *in vitro* culture (P21) makes up the majority of the tumor and culture mass. This demonstrates that these clonal lineages, albeit cultured outside physiological conditions, are endowed with LTSR potential and are functionally important in representing the majority of the cell culture. Furthermore, these clones can engraft *in vivo* to form tumor. This data also demonstrates cell hierarchy in terms of differential sustained proliferation potential exists in the *in vitro* condition and reflects the intrinsic tumor cell properties.
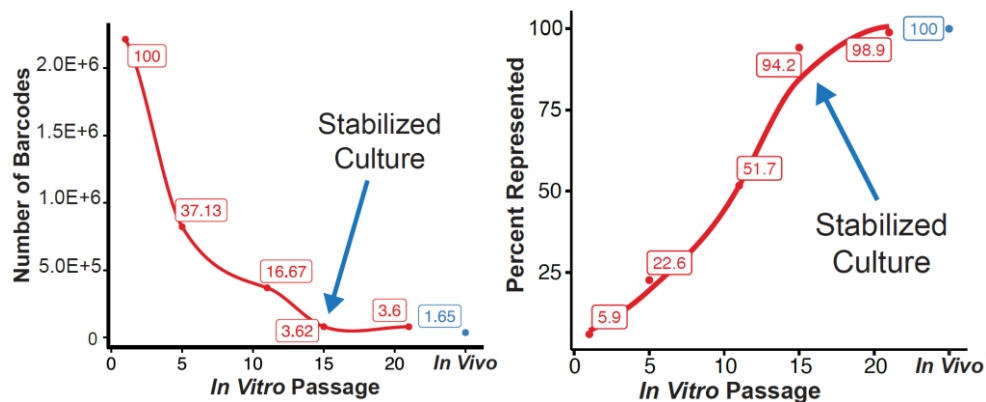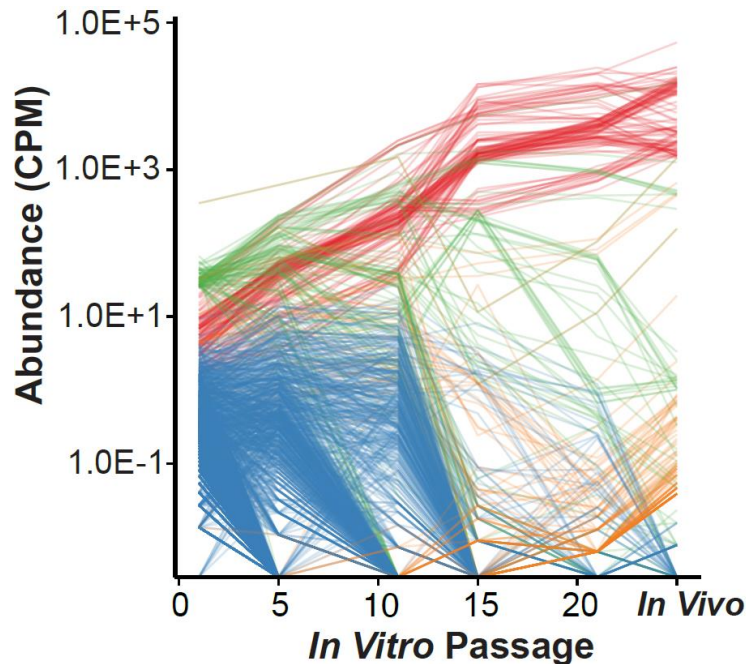
Figure 7. Tracing clonal complexity and individual clonal dynamics in *in vitro* passages. Percentage of initial unique barcodes (clonal lineages, normalized to P1, which has ~2M barcode types) detected at previous sequential passages *in vitro* (red: P1, P5, P11, P15, P21) and upon transplantation *in vivo* (cyan) (top left); Percentage of reads represented by all clonal lineages detected *in vivo* (cyan), at the previous sequential passages *in vitro* (red: P1, P5, P11, P15, P21) (top right); Differential clonal dynamics identified during sequential passages *in vitro* (P1, P5, P11, P15, P21): functionally exhausted clones (blue), transiently amplifying clones (green), long term proliferating clones sustaining tumor *in vivo* (red), and *in vitro* low represented clones enriched upon transplantation *in vivo* (orange) (bottom).

A common subset of lineages with LTSR properties with tumorigenic potential maintain primary

cell cultures *in vitro* and tumors *in vivo*

From clonal tracing in parallel in both *in vitro* and *in vivo* conditions, we see that over time,

the culture becomes stable and the clonal composition comes to an equilibrium. To compare

whether the clonal evolutionary dynamic in both the experimental conditions (*in vitro* and *in vivo*)

are similar, we represented the clonal behavior over time using t-distributed stochastic neighbor

embedding (t-SNE) method. Essentially, all the lineages are represented in the t-SNE as

individual "dots", and these lineages are "clustered" based on dynamic over time. For instance,

clones that display relative high dominance in the tumor/culture consistently, would be clustered close to each other. From the t-SNE plot, while each dot representing a clone, we also highlight their behavior with "red" representing the lineage is persistent and "grey" representing exhaustion of particular lineages. Interestingly, large set of clones that exhaust over passages are largely shared between the two experimental passaging conditions. Furthermore, the lineages identified to have LTSR property from the *in vivo* serial transplantation experiment (F4) are also found in the cluster of lineages maintaining the culture *in vitro*. From looking at these behavior trends over time in both the experimental arms, we identified common patterns of evolution. This suggests that despite the environment of passaging methods being different, overall, the tumor cells retain intrinsic behaviors that display differential hierarchy.
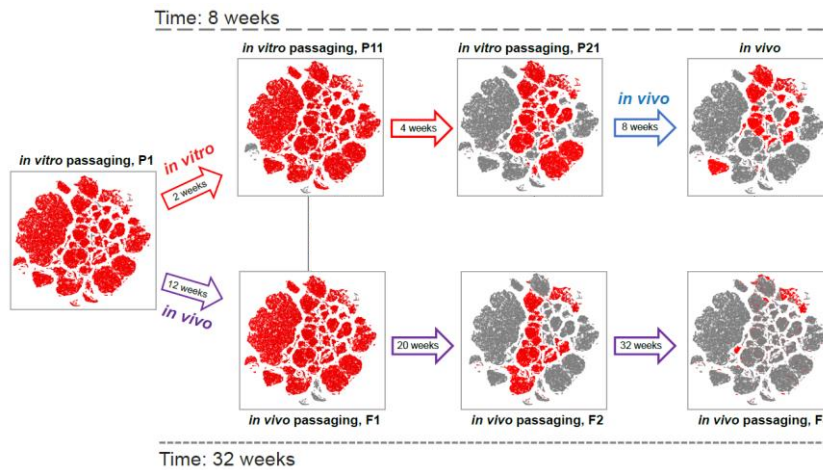
Figure 8. t-SNE representation of barcode dynamics for *in vitro* and *in vivo* clonal tracking showing convergence of long-term self-renewal clones over time. Each of individual small dots represents an unique barcode, while the colors represent whether or not a barcode is present in a particular time point and sample (red = present, gray = not present), a total of 88K lineages are represented in the graph out of 2.2M. Time points start from left to right, where all barcodes are present in "P1" (all dots are red). In two different conditions, *in vivo* (bottom lane) and *in vitro* (top lane), the majority of the same barcodes (clones) are exhausted over time (i.e. in "P21" vs. "F2"), and eventually, the long-term self-renewal clones (comparing "*in vivo*" and "F4") at the upper right corner of the t-SNE plot converge.

Furthermore, we assessed these cells' tumor engraftment efficiency by comparing *in vitro* culture at passage 21 to the *in vivo* tumor transplanted at passage 26. We see that the P21 and the *in vivo* tumor shares a common set of barcodes, and those barcodes makes up almost the entire culture of P21 and *in vivo* tumor's tumor mass (n=5863 barcodes in common; 96.8% of culture; 99.8% of the tumor). In addition, if we compare the clonal composition of serially transplanted tumors (F4) to the clonal composition of tumors derived from *in vitro* stabilized cultures (*in vivo*), we identified a set of common barcodes (801 barcodes). These set of common lineages represent a large proportion of F4 tumors (~67.9% of the tumor mass), as well as represent almost the entire mass of tumors derived from stabilized cultures (~96.5% of the *in vivo* tumors injected from *in vitro* passages cells). This data demonstrates that the *in vitro* primary cultures and *in vivo* tumors are sustained mostly by the same set of clones endowed with LTSR potential. In other word, this again shows that the intrinsic lineage self-renewal property is maintained despite the passaging condition in this experiment.
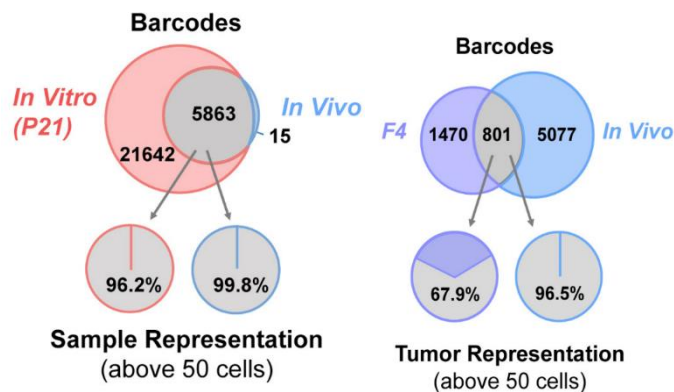
Figure 9. Venn diagram of the number of clonal lineages detected *in vitro* (red, P21) and *in vivo* (cyan); pie-charts of the percentage of sample/tumor mass represented by common clones above 50 cells (grey, 5863 clones) (left); Venn diagram representing the number of clonal lineages detected *in vivo* upon serial transplantations (F4, purple) and *in vivo* upon *in vitro* expansion and stabilization (*in vivo*, cyan); pie-charts of the percentage of tumor mass represented by common clones above 50 cells (grey, 801 clones) (right).

Generation of Clonal Replica Tumors through clonally stabilized culture

We have demonstrated that: 1. Majority of the lineages will exhaust over time; 2. The clonal composition of the culture (and tumor) will eventually come to an equilibrium and "stabilize", where the culture is maintained by a fraction of the clonal lineages; 3. These clonal lineages, by functional definition (from the *in vivo* serial transplantation) have tumorigenic and LTSR potential; 4. These LTSR clonal lineages represent the majority of the tumor mass (*in vivo*) and cell culture representation (*in vitro*) and are what is functionally maintaining the tumor / culture. Based on these concepts, once the *in vitro* "clonally-stabilized" culture is achieved (at around passage 16), we can expand these cell cultures, which is enriched in tumor maintaining lineages with tumorigenic potential and create a cohort of tumor baring mice via subcutaneous transplantation. This is exactly what we have done. Moreover, by monitoring these cohorts of animal with "stabilized" cells injected, we observed a similar growth rate in these tumors.

Once these tumors have reached significant size (~0.8cm X 0.8 cm), we collect them and process each tumor independently for barcode readout. By assessing the individual tumors from independent mouse of their clonal composition and their relative abundance (tumor representation in %), we find these tumors established through "stabilized" culture are essentially identical in their clonal composition both qualitatively and quantitatively.

As an example, we established two independent cohorts of "clonally stabilized" tumors in mice derived from two PDAC patients (PDX1 and PDX2). Within each cohort, we compare three individual tumors collected from independent mouse and assess their clonal composition. We find the tumors all share a common set of clonal lineages (PDX1 = 1499 barcodes; PDX2 = 3063 barcodes), and those common set of lineages makes up almost the entire tumor mass (PDX1 tumors 97%, 98%, and 98% respectively; PDX2 tumors 99%, 99% and 99% respectively). Additionally, we see high correlation of each clone's relative intra-tumor abundance amongst tumors (PDX1 Pearson's correlation >0.98; PDX2 Pearson's correlation close to 100% for all three

tumors). Furthermore, due to the spike in of the "NGS read to cell conversion scale" during the sample processing, we can quantitatively analyze each tumor's barcodes relative abundance in "cells" respectively. As seen in the Figure 10, amongst the three tumors from the same cohort, quantitatively, the clonal lineages with high abundance is high across all tumors, and vice versa, clonal lineages at low abundance are equally low across the tumors.

Therefore, with the process "stabilization", which naturally enrich in clones with LTSR property, we can create cohorts of animals baring tumors that are extremely similarly in their clonal composition, both qualitatively and quantitatively. Since these tumors are essentially clonal replicas with each tumor's heterogeneity replicated within a cohort of animals, we termed these tumor "clonal replica tumors" (CRTs).

Essentially, CRTs are biological research models which provides robust reproducibility within an experimental cohort. This opens the door for reliable study of tracking and identification of a common set of clones in different animals subjected to different experimental conditions, which was not possible before.
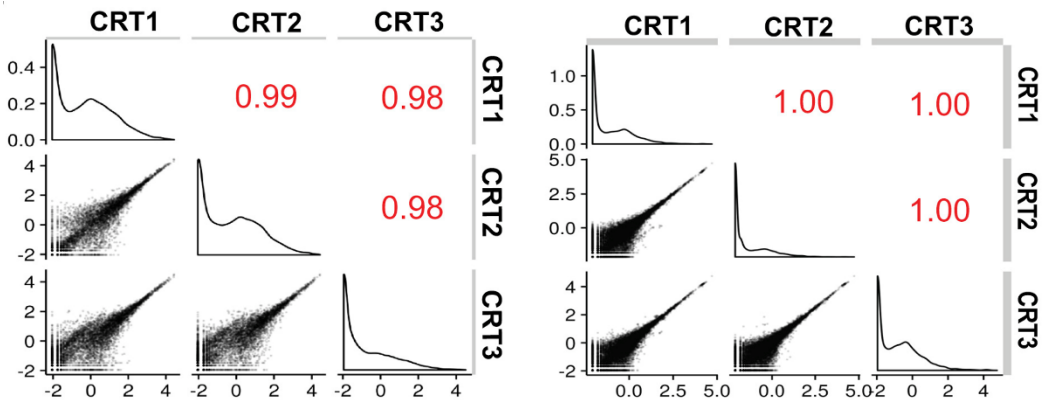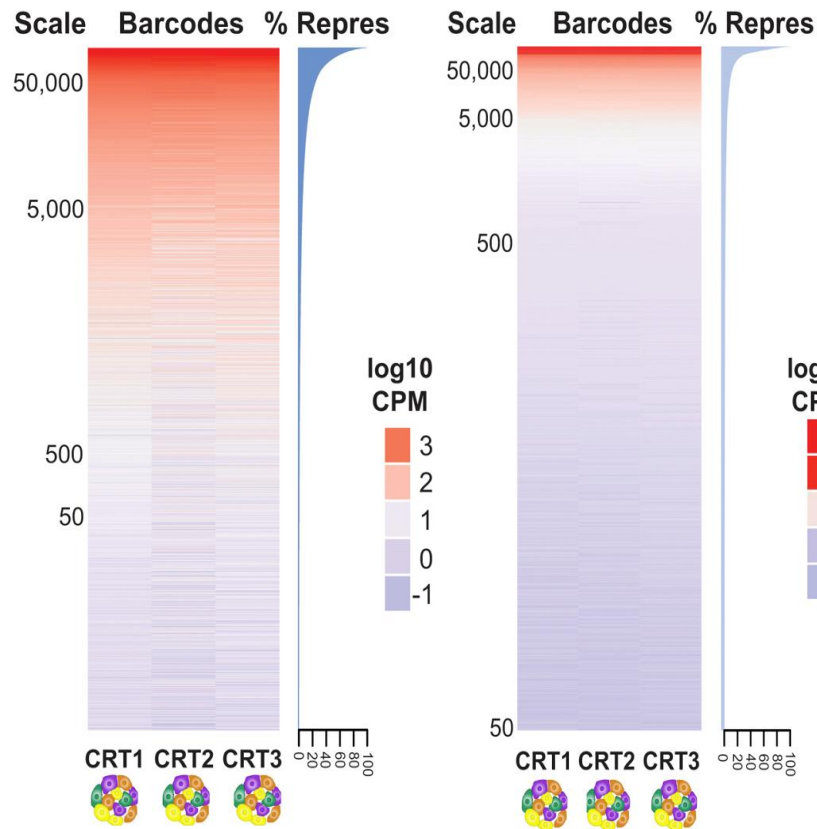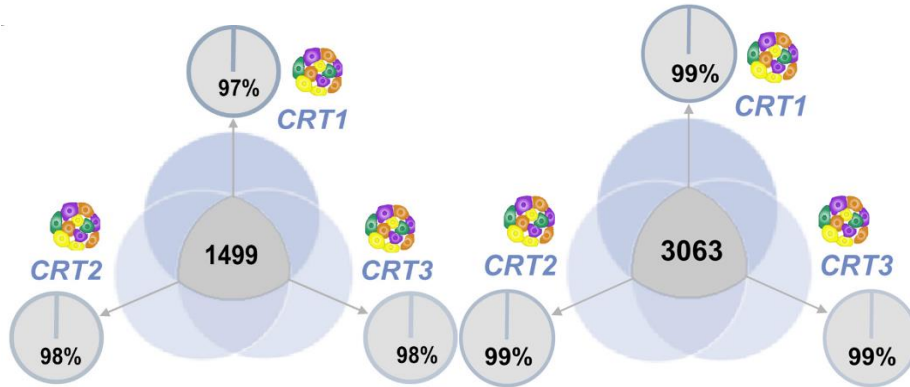
Figure 10. Barcode complexity and abundance correlations in clonal replica tumors. Venn diagram highlighting the common number of clonal lineages detected in three independent CRTs; pie-charts of the percentage of tumor mass represented by common clones above 50 cells (grey, 1499 clones) (top); Heat-map of the clonal lineage abundance of three CRTs (5683 common clones) (Log10CPM); Contribution of each lineage can be estimated as cell number using the reference scale (left: 5, 50, 500, 5,000, 50,000 cells) and as relative contribution to the totality of the tumor (right: cumulative percentage represented by clones below that lineage) (middle); Pearson's correlation amongst the three CRTs (bottom). The data shown are from two independent CRT cohorts based on left and right.

Systematic evaluations of clonal dynamics in CRTs inform of differential drug response exist in the tumorigenic compartment

The CRT platform, by having biological tumor replications, provides the unique chance to systematically and quantitatively model tumor behavior *in vivo,* study clonal response dynamics to extrinsic factors such as treatment at high confidence.

To demonstrate the effectiveness of the CRT platform, we evaluated PDACs clonal response with standard of care agent gemcitabine[75]. With the cohort of CRT tumors, we allow the transplanted tumors to grow and apply gemcitabine to the experimental mice. Consistent to clinical observations, the tumor size shrinks during the gemcitabine treatment, but relapse follows after drug suspension. This observation is essentially describing tumor resistance, where the treatment fails to eradicate the entire tumor, allowing residual tumor cells to expand and causing cancer relapse. Besides gemcitabine, we investigated the effects of targeted therapeutic agents such as AZD6244 (MEK1 inhibitor) and BEZ235 (PI3K/mTOR inhibitor) for distinct dynamics of clonal adaptation mechanisms. For the targeted therapeutic agents, we observed in general, a similar tumor resistance pattern in all the CRT tumors, where relapse occurred after treatment cessation.
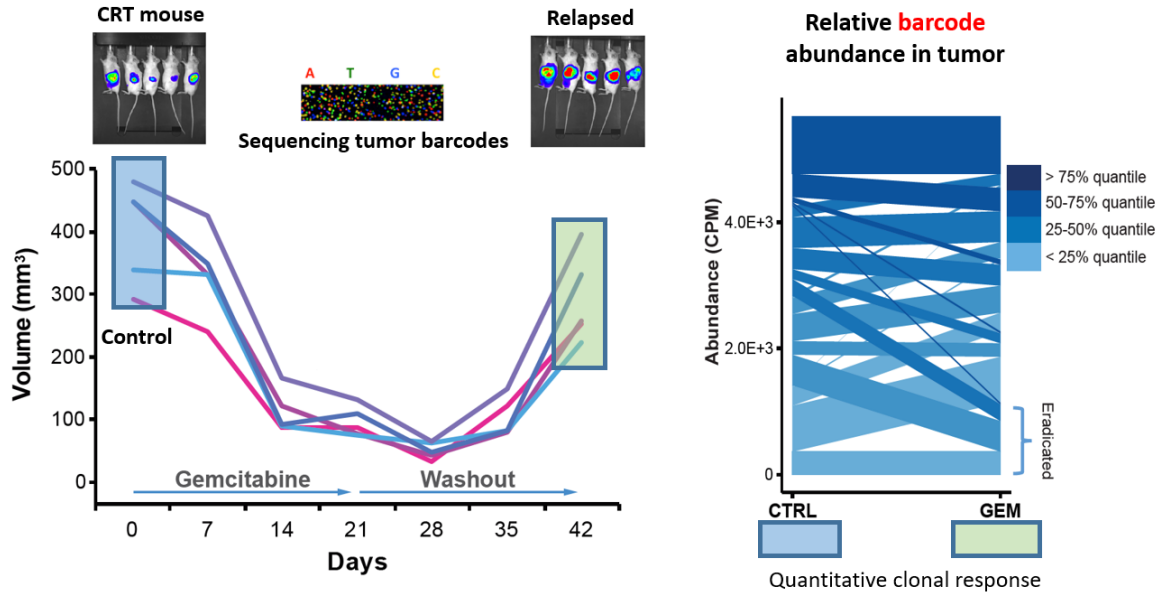
Figure 11. Treatment regimen of CRTs with gemcitabine and graphical representation of analysis approach. Schematic of tumor size (mm^3) and IVIS detection of CRTs treated with Gemcitabine for two weeks and washout to allow for relapse (left); Relative clonal treatment response captured by comparing the percentage of clonal representation post- and pre-treatment (right).
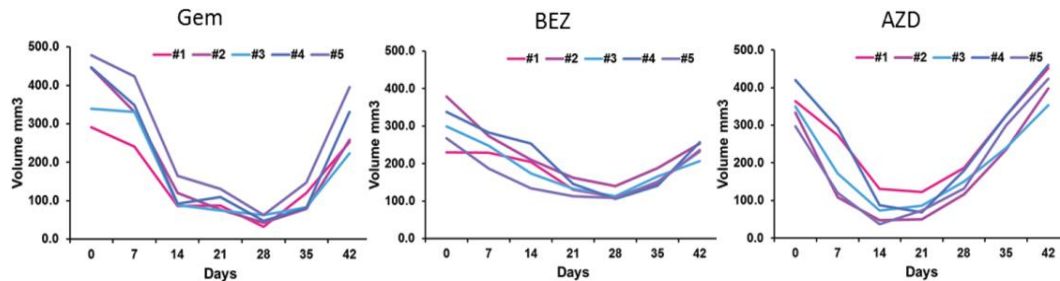
Figure 12. Tumor growth inhibition and relapse for CRTs (T1-T5) in response to Gemcitabine (n=5), BEZ235 (n=5), AZD6244 (n=5); CRTs are subjected to treatment 14 days with Gemcitabine, BEZ235 and AZD6244 in independent cohorts of animal followed by wash out to allow for relapse. Each line represents a unique mouse in the CRT cohorts. All tumors behave in similar dynamics, where tumors shrink in volume under treatment and relapse. On a side note, AZD6244 treated tumors seems to relapse faster compared to gemcitabine and BEZ235 treated tumor.

To identify the relative clonal sensitivity to specific therapeutic agents, we convert each clone's barcode read to percentage representation in a particular sample. By comparing the

difference of each clone's tumor representation in the relapsed tumor to the control tumor (untreated), we could calculate the relative fold change of each clone. For example, a positive fold change would indicate a particular clone has gained more tumor representation in the relapsed tumor, indicating a relatively more resistant phenotype. On the contrary, if the treatment is effective to a clone, the clonal tumor representation is expected to decrease or even disappear in the relapsed tumor. Furthermore, since the treatment group has biological replicas, we can assess the consistency of each clonal behavior among tumors, and therefore, assign a confidence value of how well we capture the specific clonal response to treatment. As an illustration, if a clone has similar "fold change of 2" across multiple animals in the cohort, that specific clone displays a more consistent "resistant" response to drugs, as well as technically, having a higher confidence that we are capturing such response more accurate.

Treatment effects on clonal complexity

We can represent the data in volcano plots, where each dot represents a specific barcoded clone. For the volcano plot, the X-axis represents the clonal fold-change when comparing the relapsed CRT tumor to the untreated CRT tumor, while Y-axis indicating the false-discovery rate calculated by the consistency on drug response across the CRT cohorts.

The CRT tumors in the control group has a high clonal abundance correlation amongst each other (Pearson's correlation >0.99); post treatment, the clones in the relapse group also has a high correlation (Pearson's correlation ~0.95). This demonstrates not only the tumor replica aspect the CRT platform brings, but also reflects the clonal lineage intrinsic factors shaping the response to treatment; hence the high correlation for post-treatment samples.

For gemcitabine specifically, the treatment effectively depleted clones that were responsible for ~20.3% of the populations (blue dots), but most of the clones (~73%) remained

relatively unchanged (grey dots). On the other hand, pre-existed clones (~6.0%) in the untreated tumor are enriched in the relapsed tumors, indicating clonal survival and selection advantage upon gemcitabine treatment.

For AZD6244 and BEZ235, the target therapies' effects on clonal complexity are similar to that of gemcitabine, where 17.3% and 29.6% sub-population is effectively depleted by AZD6244 and BEZ235, respectively. On the other hand, ~3.9% of the pre-existing clonal population is expanded in the relapsed tumor in both the cases of CRT cohorts treated with either AZD6244 or BEZ235.

Following the clonal relative tumor representation, we found that chemotherapy does perturb the relevant abundance of clones, but the overall effect in depleting the tumor's clonal complexity remained minimal. After chemotherapy regimen, most of the clones remained relatively unchanged in terms of tumor representation in relapsed vs. untreated. This observation could be specific to the treatment routine and duration, however, it provided us with a mean to observe and study drug effectiveness at a clonal level. Here, using CRTs, we demonstrated that mechanistically different drugs display similar effects on tumor size shrinkage and relapse, moreover, the degree of tumor eradication and the dynamics at a clonal level is surprising similar.

Furthermore, for all three drugs (gemcitabine, AZD6244 and BEZ235), the clonal lineages that display relative sensitivity is evenly spread across all quantiles of the tumor representation. This hinted that intrinsic clonal specific properties contributing to different susceptibility and fitness under treatment is a significant factor shaping tumor's evolution, and not necessarily determined by abundance.
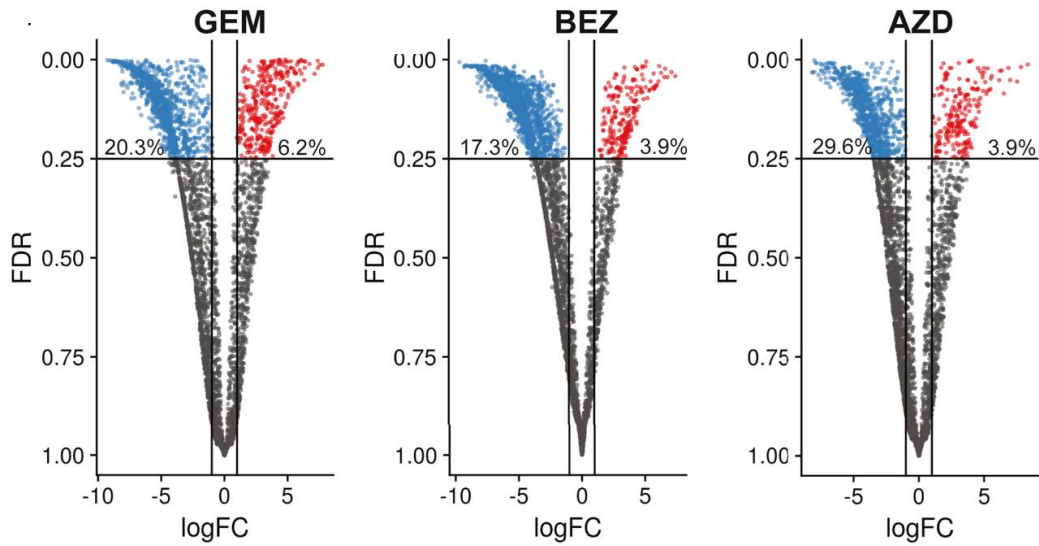
Figure 13. Volcano plot of clonal lineage response to gemcitabine, BEZ235 and AZD6244. Each dot in the plot represents a unique clone. X-axis represents the log2 fold change of comparing the % representation of each clones post- to pre-treatment. Y-axis represents the false discovery rate (FDR), where a smaller value indicates a higher confidence in the response.
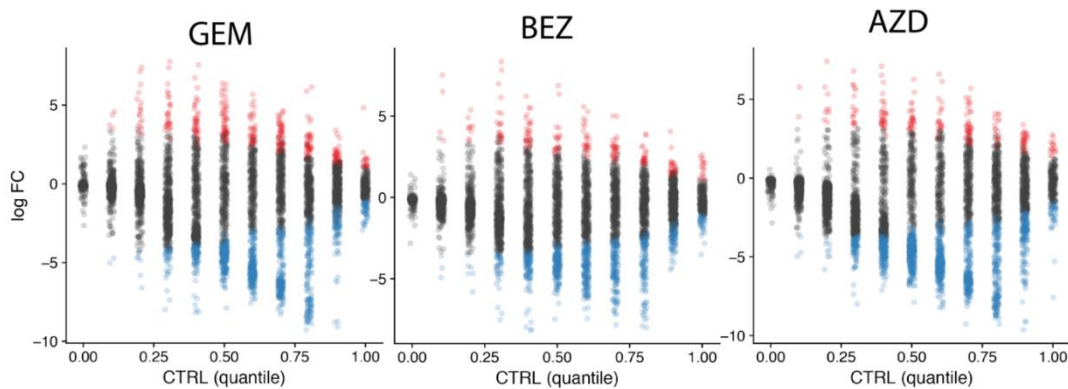
Figure 14. Plot of relative clonal fold change (log$_2$ FC, Y-axis) comparing post- to pre-treatment for GEM, BEZ235, and AZD6244 to individual clonal tumor representation in the untreated (control tumor, CTRL). The color red indicates clones that are relatively more resistant, while blue clones are relatively sensitive to treatment. The color scheme is in align with the volcano plot in Figure 13.

Table 1. Tumor clonal correlation of relapsed tumor treated with gemcitabine. Three CRTs treated with gemcitabine and allowed for relapse ("GEM1", "GEM2", "GEM3") vs. three control CRTs receiving vehicle ("V1", "V2", "V3") and their barcode relative abundance Pearson's correlation value.

PATC53 CRT clonal Pearson's correlation

	GEM1	GEM2	GEM3	V1	V2	V3
GEM1	1	0.856	0.8407	0.9035	0.8782	0.8954
GEM2	0.856	1	0.7732	0.8409	0.8267	0.8359
GEM3	0.8407	0.7732	1	0.855	0.8367	0.8545
V1	0.9035	0.8409	0.855	1	0.9731	0.9867
V2	0.8782	0.8267	0.8367	0.9731	1	0.9927
V3	0.8954	0.8359	0.8545	0.9867	0.9927	1

* V = non-treated control receiving vehicle.

Clonal intrinsic pre-existing differential response to treatments

By analyzing the CRT cohorts all together for all three treatments, we identified subsets of clonal lineages, within the same tumor, that exhibit distinct drug susceptibility profiles. For example, subsets of clones that are more resistant (or sensitive) to either only gemcitabine, or, AZD6244, or BEZ235.

For a more comprehensive view of all the clones in the tumor, we used constrained correspondence analysis (CCA) [76]. In short, the CCA plot reduces multivariate data into a two-dimensional graphical form while keeping the relative correlation amongst data points. In the CCA plot, tumors are represented by triangles and each dot represents an individual barcode lineage. We observe that the control untreated tumors are distant from either the treated tumors on axis1 (CC1), and that the three independent drug treated tumors are spread across the axis 2 (CC2). This tells us that treatment does indeed affect intra-tumoral clonality, and that there is only partial overlap of clones amongst different treatments. Furthermore, looking at the relative enrichment of tumor abundance by comparing relapsed tumor to untreated tumor (the color red representing positive enrichment, while blue representing negative enrichment), in other words, clone's relative resistance, we can see that clonal lineages displaying resistance (red) to a specific drug is far away from the lineages displaying resistance to other drugs.

This observation suggests that tumor harbors heterogeneous lineages functionally displaying a diverse spectrum of sensitivity to different external perturbagens. Once the equilibrium of these lineages is perturbed by pharmaceutical agents, sub-population of clones with fitness advantage may drive the relapsed tumor and form a new equilibrium amongst clones. Most importantly, since we demonstrated that the population of clones resistant to treatment are different depending on the agent used, this indicates that the clonal composition of relapsed tumor is associated with the type of pharmaceutical agents used.
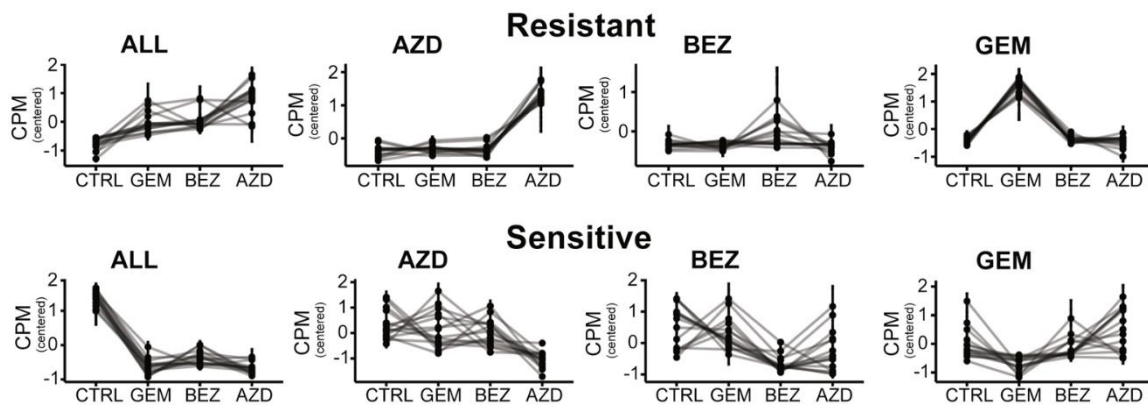
Figure 15. Representation of clones displaying a various degree of sensitivity to different drugs. Each line represents one unique clone and Y-axis indicates normalized counts per million of individual barcode. The top portion are representative clonal dynamics that are displaying selective resistance to one particular drug, or, to all drugs. The bottom portion are representative clonal dynamics that are displaying selective susceptibility to one particular drug, or, to all drugs.
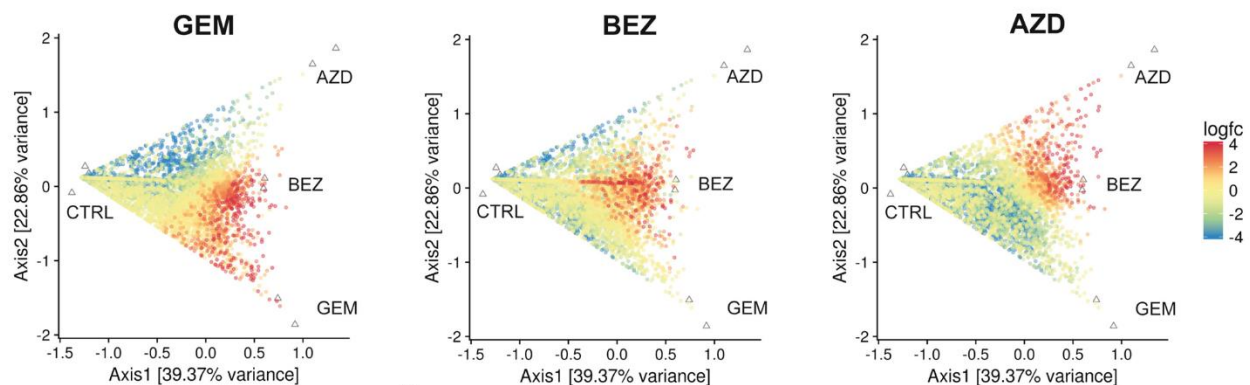
Figure 16. Constrained correspondence analysis (CCA) plot for clonal differential sensitivity to GEM, BEZ and AZD. Differential clonal dynamics as detected using constrained correspondence analysis (CCA), upon treatment (GEM, BEZ235, AZD6244) in comparison with untreated tumors (CTRL); Tumors are represented using triangles and points represent lineages, with control

tumors (left) further from treated (right) in each biplot. A higher logFC indicates a more resistant clonal phenotype (in red), while more sensitive clones to a particular treatment is bluer.

"Tissue culture-to-NGS library" workflow enables high-throughput isolation of clones of interest identified via CRT experiments

CRTs enabled us to see precisely individual clone's distinct phenotype to different treatments in an *in vivo* setting, providing a reliable mean for detailed clonal dynamic observation. However, being able to observe the clonal specific post-treatment phenotype is one thing, while being able to answer what is the underlying molecular features of treatment naïve clones that is driving such phenotype *in vivo*, and being able to functionally test and validate them is where we could really expand knowledge towards developing a better treatment strategy.
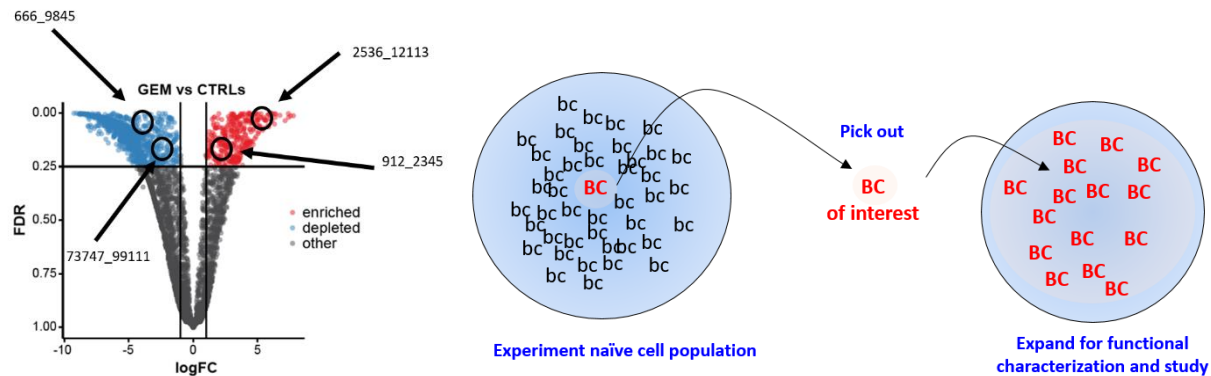
Figure 17. Schematics of "treatment response informed" clonal isolation. Through analysis of post- treatment clonal relative fold change (compared to untreated control), we identify clones displaying resistance or sensitive phenotype to gemcitabine (left); The goal is to isolate pure (treatment naïve) clones identified from the parental population of thousands of unique barcoded cells (right).

The technical challenge here is essentially how to pick out specific clones harboring a unique barcode sequence (~36bp) in its genome among tens of thousands of barcode variations differing in few nucleotides. To achieve a minimally purebred, high-throughput workflow to isolate clonal lineages of interest, we developed a workflow that is highly sensitive and efficient. In brief, the isolation steps involve "enrichment of clones of interest"; "establish conditions of single cell

cultures" with FACs; creating "single clone culture replica plate"; detection of clonal barcode enabled by a one-tube "cell culture-to-NGS-library" chemistry, where NGS library is directly built in the wells of trypsinized cell culture; and using multiplexed "NGS indices" for deconvolution and identifying single clones and their relative position in the culture plate ("positional sequencing"). Here I will describe in brief the workflow in a stepwise manner. For detailed work and rationale of the isolation method, please see the method section.

Enriching the clones of interest from a mixed population

The cell cultures used for the CRT *in vivo* experiments could be re-established at any time from liquid nitrogen. Since the barcode complexity of the cell culture is very high, we first must narrow down our scope for an efficient isolation of single clone. Thus, we first partitioned the cell cultures into multiple tissue culture wells (96-well plate) and allow them to expand. Once the expanded culture reaches near confluence, the culture is then split into two equal halves. Half of the culture will remain in the tissue culture while keeping its original 96-well position, while the other half of the culture will go through DNA extraction and build NGS libraries with unique indexes for each well to identify the barcode composition present in each well using a "cell-culture-to-NGS-library chemistry" (discussion in later section). Once the barcode composition in each of the 96-wells is identified, we then look for the wells that are enriched with the clone of interest. The wells that are enriched with the clones of interest is then used to create isogenic clonal culture through FACs. This procedure will increase the chance of creating clonal cultures of interest.
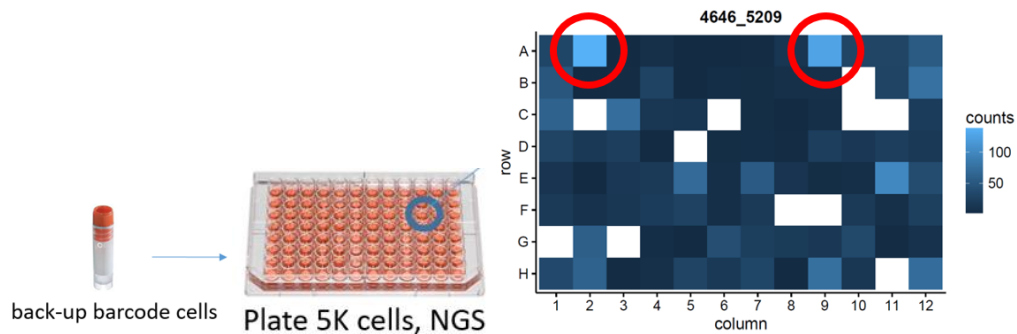
Figure 18. Schematics for enrichment of identified clones of interest. Barcoded cells were cultured at 5K cells per well and sequenced. The right heat-map represents the clone of interest (ID 4646_5209) and its relative abundance across the plate (white is absent, light blue has a higher abundance). The most abundant wells with clones of interest (circled in red) are further FACs single cell sorted in new tissue culture plate to establish pure isogenic culture.

Establishing clonal culture from single cells: only a fraction of single cells are able to establish new cultures while others exhaust or fail to proliferate

Cells from the identified wells are sorted into multiple of 96-well plates and allowed to expand. These plates with single cell in each well are then placed in incubators coupled with a high-content imaging system. Interestingly, we observed multiple phenotypes of morphology and different degree of proliferation rate from these clonal cultures. Moreover, only a minor fraction of the cells are capable of establishing a new culture, while the others either stop proliferating after initial expansion, or completely fail to proliferate at all. To dissect this observation further, we found out that most of the hyperploid cells (based on Hoechst dye) fail to proliferate, and the cells that are able to grow and establish new colonies are relatively independent of cell cycle (~10% of the cells can establish a new colony, while ~75% of the cells fail to proliferate). This observation is in line with the cell hierarchy data we saw from the *in vitro* serial transplantation of barcoded cells from F1 to F4, where only a fraction of the clonal lineages are endowed with long-term self-renewal potential. Furthermore, we then tested whether or not the culturing media condition would affect the ratio of cells that are able to establish a new colony. By comparing the ratio of cells that proliferate in conditioned media vs. fresh media and monitor via live imaging (Incucyte TM), we see a very similar ratio of cells that are able to proliferate (~15%). This suggests that tumor cells are

32

mainly organized through hierarchy, and that certain cells are intrinsically endowed with the ability to self-renew and establish a new colony.
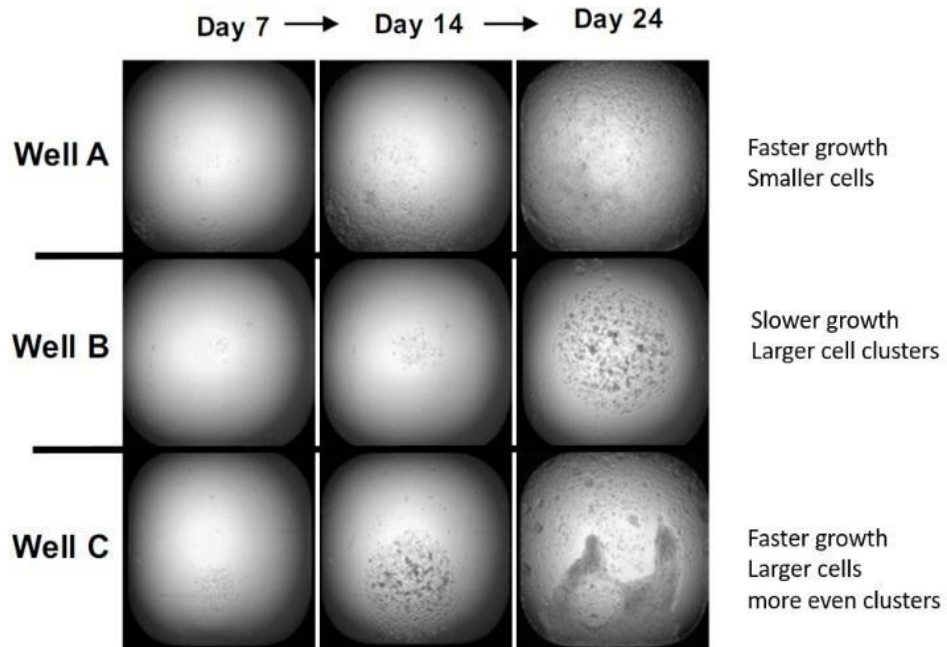
Figure 19. Isolated clonal population growth dynamic monitoring using high-content imaging system. Continuous monitoring of single clone growth using high-content imaging system (Operetta™). Growth dynamics and pattern of single clones can be monitored for all isolated colonies. In well "A", clone exhibit a smaller cell size phenotype and grew faster compared to clone in well "B" (over 24 days of growth). Clones in well "C" also exhibit a different growth rate (faster) compared to clones in well "B" (slower). Both clones in well "B" and "C's" growth pattern are more clustered than clone in well "A".
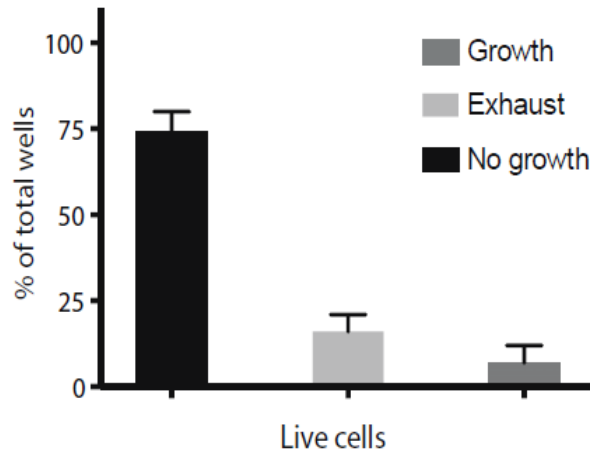
Figure 20. Percentage of total flow-sorted live single cells based on growth dynamics. Tumor cell line was single cell sorted for "live cells" (using Sytox™) and monitored for proliferation over time in tissue culture plates. In this particular tumor, only a fraction of single cell grew and formed new colonies (~20-30%), while the majority of the sorted cells did not grow (in black). Amid the cells that grew, around half of the colonies would eventually exhaust (light grey bar) and never grew to confluence (over 8 weeks period, n=384).
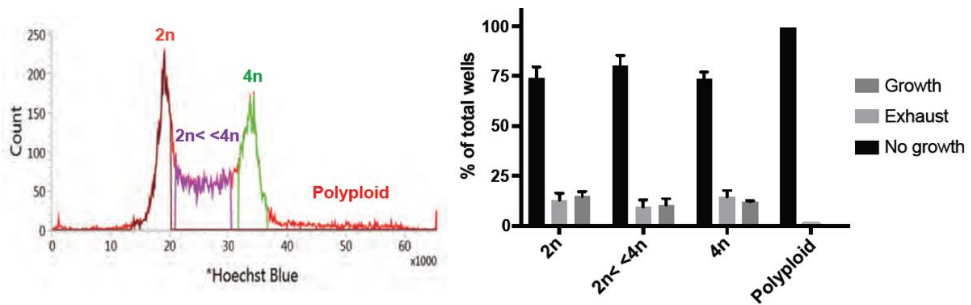
	No growth (StDev*) %	Exhaust (StDev*) %	Growth (StDev*) %	Number of wells (plates)
2n	74.0% (5.6)	12.8% (3.6)	14.6% (2.6)	576 (6)
2n< < 4n	80.2% (5.3)	9.4% (3.7)	10.2% (3.7)	576 (6)
4n	73.7% (3.3)	14.3% (3.4)	12.0% (0.6)	384 (4)
Polyploid	99.0% (n/a)	1.0% (n/a)	0% (n/a)	96 (1)

*StDev calculation based on number of plates

Figure 21. Single cell proliferation dynamics by cell cycle. Single cell sorted based on cell cycle G0/G1, S, G2/M, and polyploid population (top left graph n=864). In this tumor, majority of the sorted cells could not re-establish new colonies. Around half of the population that grew into colonies will eventually stop proliferating and exhaust. In particular, the polypoid single cells fail to establish new colonies (1 out of 96 grew, but eventually exhaust).

34
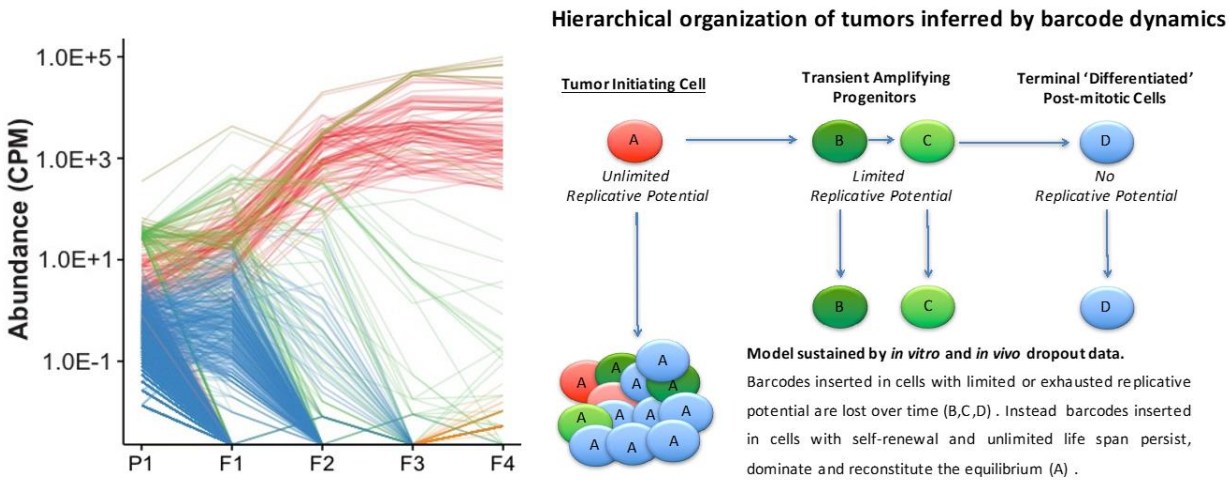
Figure 22. Hypothetical model of observed clonal hierarchy organization based on barcode lineages tracing. Left graph shows representative lineage dynamics across 32 weeks of *in vivo* serial transplantation, a functional assay to determine cells with self-renewal capability. Right graph is a suggested (based on lineage tracing data) graphical representation of cell hierarchy model in tumor. Barcode inserted in cells with limited or exhausted replicative potential are lost over time (B, C, and D). Instead, barcodes inserted in cells with self-renewal capability can persist and eventually re-establish self-renewal equilibrium.

Different tumors have different ratio of cells with ability to establish a new culture

Intrigued by the observation that only about ~15% of the PDX derived tumor cells were able to give rise to a new colony from single cell sorting, we wondered if this holds true with other PDXs. Thus, using the same setup, we sorted single cells (live cells) in 960 wells from another tumor (PATC124) and monitored its proliferation latency for 5 weeks. Every 4-5 days, we recorded the population in terms of proliferation and its relative confluence. We observed a wide range of clonal dynamics in terms of the time it took for the cells to divide after sorting (initiation) and their relative expansion rate and exhaustion rate (proliferation rate). Similar to the other tumor's observation, 87.81% (843/960 wells) of the cells failed to initiate expansion and remained single cell; 12.19% (117/960 wells) of the cells went through initial expansion. Out of those cells that could initiate expansion, they exhibited different expansion dynamic (captured below) with only 5.94% of the cells were able to establish a full colony without exhaustion. This experiment was repeated 2 more times and the result is consistent.

Comparing PATC124 vs. PATC53, while both tumors display similar cell hierarchy, where only a fraction of the cells can establish a colony, the ratio of those cells with self-renewal properties are different. PATC124 tumor displays a lower percentage of self-renewal cells (~6%) versus PATC53 (~10%). The ratio of the self-renewal cells could very well reflect the aggressiveness of the disease, as PATC53 was derived from a liver metastasis from the patient, whereas PATC124 was derived from a primary tumor. However, whether or not this is true requires further characterization.
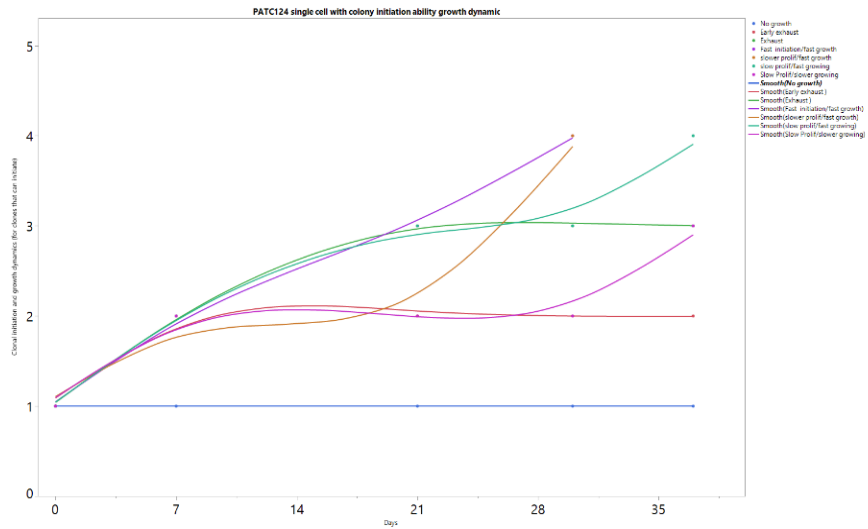
Figure 23. Isogenic clonal proliferation and expansion dynamics *in vitro*. Each line represents an isogenic lineage established from one cell. X-axis represents the days after incubation and the Y-axis represents the relative confluence over time. Data obtained from PATC124 cells.

Single Cell proliferation capability (PATC53 vs. PATC124)
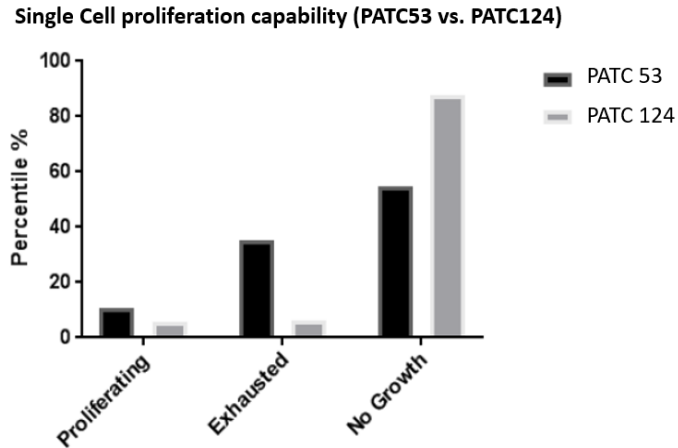
Figure 24. Bar graph comparing the percentages of cells with ability to re-establish culture between two patient tumors derived cell lines. The bar graph shows the % of cells after single cell FACs that could establish a new isogenic line, as well as % of cells that eventually exhaust or never divided. Comparing the two tumors, albeit following a similar trend where most cells exhaust, the percentage of cells with capability to re-establish colonies are different (PATC53 ~10%, PATC124~5%).

Self-renewal hierarchy exists in isogenic clones and is reflective of its parental population's

Next, we were curious whether or not, intra-isogenic clonal culture (established from a single cell) also display self-renewal hierarchy. Using the same colony monitoring assay by culturing single cells in individual wells, we sorted plates from four individual isogenic clones. Interestingly, individual clones displayed a higher percentage of cells that are able to initiate cell division and expand compared to the parental population (~2 fold), but many of them eventually exhaust and fail to sustain the culture indefinitely. Interestingly, each of the four isogenic clonal culture displayed similar percentages (10.7%, 11.7%, 13.3%, and 7.7%) of cells that were able to re-establish culture. The percentages cells that were able to re-establish culture were comparable to the parental heterogeneous population where these four clones were derived from (11.6%).

Clones*	Number proliferated**	Percent Proliferation	Number Exhausted***	Percent Exhausted	Number of no growth****	Percent of no growth
CL 09	21	10.7%	60	30.6%	115	58.7%
CL 11	23	11.7%	65	33.2%	110	56.1%
CL 01	26	13.3%	80	40.8%	90	45.9%
CL 12	15	7.7%	67	34.2%	110	56.1%
Parental	189	11.6%	184	11.3%	1266	77.6%

* 196 wells of single cells were sorted for each clone; 1632 wells were single sorted for parental culture (mixed clones); Cells were observed over 1 mo period for proliferation.

** Wells (cells) exhibit signs of proliferation and are able to establish confluence.

*** Wells (cells) exhibit signs of proliferation initially, however, stopped growing.

**** Wells (cells) stayed 1 cell throughout.

Figure 25. Intra-clonal cell's capability to re-establish a culture is heterogeneous. Isogenic clone's cell % that could re-establish an entire culture in bar-graph format (top) and table with detailed data (bottom). This suggests the "self-renewal" (ability to re-establish culture) hierarchy exists.

Colonies established by single cells with self-renewal capability can initiate tumors *in vivo*

To test the tumor initiating capability of these clonal cultures expanded from single cells, we transplant these clones in immuno-deficient mouse (NSG) via subcutaneous injection. Indeed, these isolated clones are able to initiate and establish tumors *in vivo*. These clones display a wide range of tumor growth rate and morphological phenotype characterized by tumor size and cellular activity (Ki67) and histology.
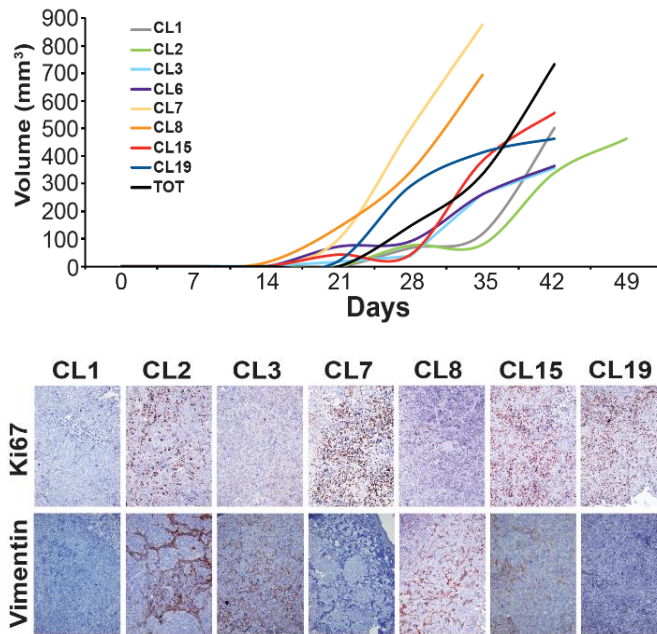
Figure 26. *In vivo* tumor growth rate and histology of seven isogenic clones. Tumor volume over time for isolated clones and histology stained with Ki67 and vimentin are diverse amongst the seven individual isogenic clones. Experiments performed by Denise Corti.

A highly sensitive "culture-to-NGS-library" chemistry tailored for low cell numbers to identify cellular barcode identity

From the above characterization of clonal culture established via a high-throughput workflow, we established means to isolate single clones from a mixture of cell populations. To identify the hundreds of isolated clones' barcode ID in a time and cost-efficient manner, we tailored a new chemistry and workflow to identify all the barcode ID across all the wells simultaneously with one NGS run.

In brief, the single cells were sorted into multiple 96-well plates and allowed to expand. After allowing the cells to grow for 2-3 weeks, some wells would approach 80-90% confluence, while other wells would have much fewer cells (10-100 cells). At this stage, all the wells were trypsinized regardless of individual well's cell number. The trypsinized cells from each well were split in half: one half re-plated in 96-well tissue culture plate and allowed to expand, while the other half into a 96-well PCR plate used to identify their barcode identity. These two plates are

essentially replica plates, with the same clones in their relative matched well positions. The rationale is to use one plate to identify the clone's ID in each well; once the barcode ID is known for each well, we can go back to the 96-well tissue culture plate and expand the clones of interest from the identical well position.

The technical challenge lies in the fact where in some wells, after trypsinization, there are only few cells present. Extracting DNA from wells with few cells are neither not practical nor efficient, where most likely there would be a loss of material due to the DNA extraction process. Therefore, to prevent material loss, the logical approach is to combine the cell lysis step together with the PCR amplification step. This way, after trypsinization, the whole process of cell lysis, barcode amplification via PCR, building NGS library would be in the same reaction tube from start to finish. Through optimization of each experimental steps, I was able to optimize a reaction condition that achieves building NGS library from tripsinized cell culture by addition of reagents only (no liquid replacement).
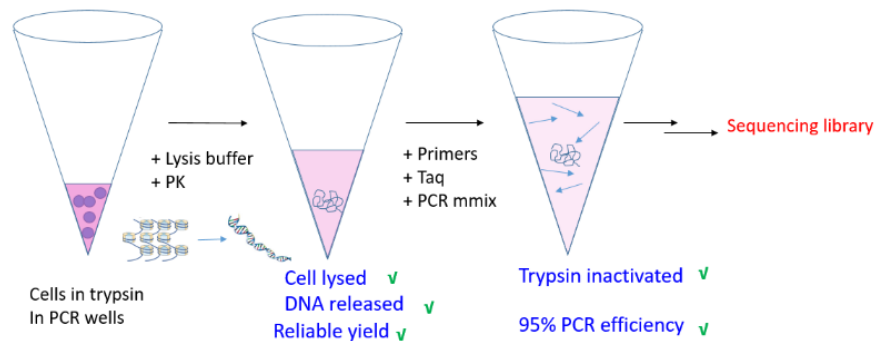
Figure 27. Step-wise rationale breakdown of direct single tube "culture-to-NGS" chemistry. "Tissue culture-to-NGS" lysis solution can effectively and consistently lyse cells. In addition, the post-lysis reaction is compatible with PCR reactions.

We need a cell lysis condition that can achieve complete lysis of cells while cells are in trypsin and media with FBS, as well as, a lysis condition strong enough to degrade histones to allow for linear DNA. At the same time, the lysis buffer needs to inactivate trypsin (a potent PCR inhibitor). In addition, the entire lysate would need to be PCR compatible. Finally, due to the

minute amount of starting material (~10 – 100 cells), we need to achieve the above without any liquid replacement; instead, by addition of components on top of the previous reactions.

We decided on an NP-40 based lysis solution with protein kinase (PK) as the final condition to achieve the above goals. For the lysis capability, we compared the PK lysis solution to several other conditions, including a commercially available lysis buffer that is advertised to be compatible with downstream PCR reactions. We used Hoechst dye to stain for nucleus and microscope to gauge the effectiveness of lysis. We found the PK lysis condition can completely lyse the cells and release the DNAs from its heterochromatin state while the other conditions cannot. Then, to test the reproducibility of the PK lysis buffer, we used a TaqMan™ assay targeting RNase P region of the genome (1 copy per genome) to amplify 8 replicates of cell lysate. We found the PK lysis condition gave the most efficiency and reproducibility (8 replicates, Cq = 24.22; STD 0.17).
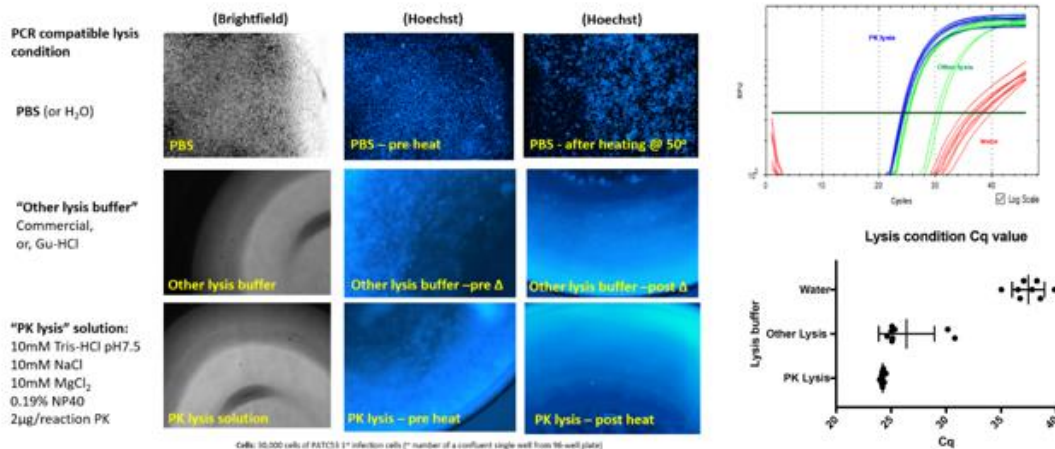
Figure 28. Microscope and PCR validation of direct single-tube "culture-to-NGS" lysis buffer efficiency. *PCR compatible lysis condition* (Upper left panel): PK lysis solution effectively lyses all cells as seen by the lack of Hoechst dye positive nuclei; *Cell lysis reproducibility* (Upper right panel): cells (~15,000 per well) are directly lysed (in same wells) with addition of PCR mastermix and DNA primers/probe targeting RNase P. PK lysis buffer has the highest lysis efficiency and reproducibility (8 replicates, Cq = 24.22; STD 0.17);

PK is potent PCR inhibitor. Therefore, PK must be inactivated before we could add the PCR components for downstream amplification. Indeed, without proper inactivation of PK, the PCR reaction does not work as seen from our result using the same RNase P TaqMan™ assay. However, after heat inactivation of PK, the entire cell lysate is now compatible with PCR with high amplification efficiency of 95.49%. Consequently, we now have a complete workflow that is compatible with low cell number, and a condition where we could amplify the barcodes and generate NGS libraries directly after trypsinization of the cells without DNA extraction.

PK lysis buffer PCR efficiency (n=2)

Lysis buffer	PK (inactive)	PK (inactive)	PK (active)	
Trypsin	No trypsin	Active	Active	
Before PCR	heat @95°C	heat @95°C	heat @95°C	DNA input/rxn
Average Cq	27.8	na	27.6	25ng
	30.2	na	29.9	5ng
	32.6	na	32.2	1ng
	34.9	na	34.8	0.2ng
slope	-3.38		-3.44	
PCR eff%*	97.78%	na	95.49%	
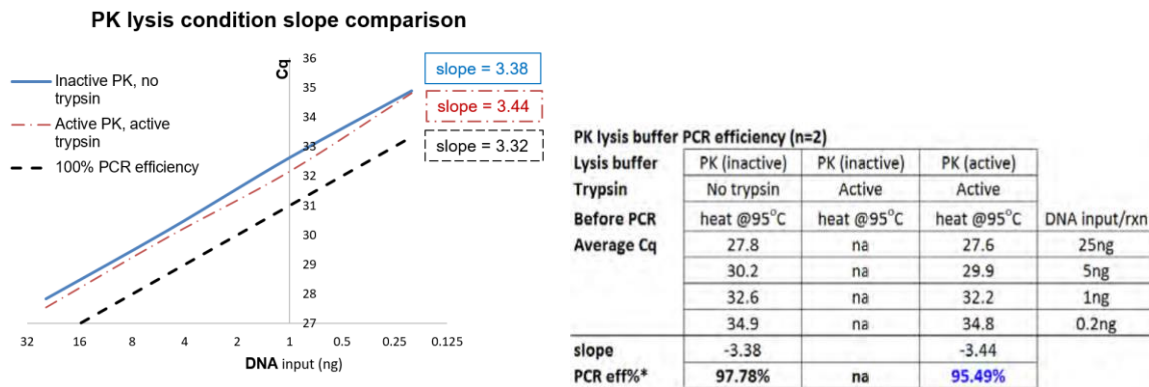
Figure 29. PCR efficiency of direct lysis protocol: Human gDNA reference and GAPDH assay is used to measure % PCR efficiency under various lysis conditions. The lysis buffer with inactive PK has minimum effect on PCR (PCR efficiency = 97.78%); Trypsin inhibits PCR reaction (no Cq observed); PK effectively inhibits trypsin and overall PCR efficiency remains high (PCR efficiency = 95.49%)

A schematic summary of clonal isolation workflow
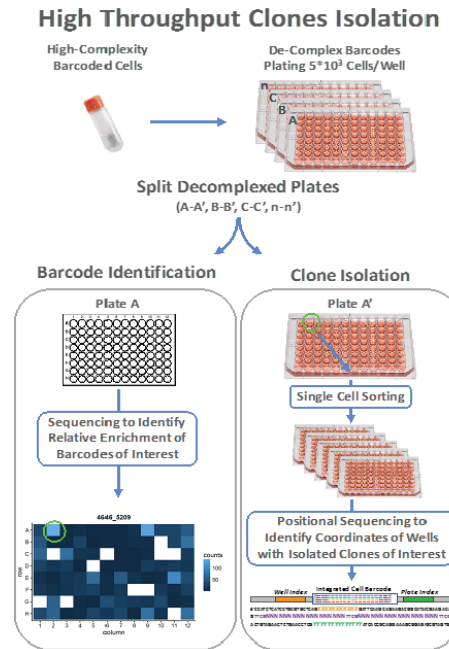
Figure 30. Two-step high-throughput individual clone isolation strategy (1st step – de-complex of mixed culture and identify wells enriched in clones of interest, 2nd step - identify clonal barcodes from single-cell sorted wells identified in the 1st step): Frozen high-complexity barcoded cells originally used to generate *CRTs* are thawed and expanded in 96-well plates at 5,000 cells per well in order to decrease complexity. After cell expansion, the de-complexed plates are split into two replica plates - one plate is kept in culture for clonal isolation and the other processed via "direct cell culture to NGS" workflow for barcode identification. Via NGS read analysis, heat-maps of clones (barcodes) of interest are generated and the well positions with relative higher abundance identified. From the corresponding wells in the replica plate, single cells are sorted into multiple 96-well plates and allowed to grow. Isolated clones, once expanded, are split to create two sister replica plates – one for storage/culture, and one processed for NGS using positional sequencing (built-in plate and well index) to identify coordinates of wells that correspond to specific clones of interest. This workflow enables cost effective, rapid identification of thousands of single clonal barcodes with extremely high sensitivity.

Isolation of gemcitabine resistant and sensitive clones identified through *in vivo* CRTs experiment

Using the high-throughput isolation strategy, we restarted the barcoded cell cultures used to generate *CRTs*. During the first de-complexing stage, (5,000 cells per well), we were able to capture all the differential clonal lineage behaviors in response to gemcitabine *in vivo* as seen

from the volcano plot, where enriched clones after relapse highlighted in red and the depleted clones highlighted in blue (LogFC = Log2 fold-change, Gem vs CTRL, FDR<0.25). The position of 19 clones of interest (7 enriched and 12 depleted) is highlighted in the volcano plot. Then, from the de-complexed plate, we generated heat maps of relative abundance of the clones of interest. As an example, one of the depleted clone 5125_5793 (LogFC = -1.26, FDR = 0.20 and one of the enriched clone 13767_14001 (LogFC = +2.00, FDR = 0.07) were picked from their most abundant wells as seen from the heat map of the de-complexed plate. Wells containing higher representation of the two clones were picked for single-cell sorting (circled wells). The, multiple single-cell sorted plates were processed for NGS library in parallel and combined for one NGS sequencing run. The well-coordinates of the isolated two clones of interest were traced back using 'positional' sequencing well and plate NGS indexes (bottom part); NGS barcode identification was extremely specific, where average read depth used for barcode identification were 122456 reads, the no template control (NTC) reaction had only an average of 4 read counts; chromatogram of Sanger sequencing for the isolated clones confirmed NGS determined barcode identities. Barcode identity of individually isolated clones were then confirmed by Sanger sequencing as seen from the chromatogram, matching the barcode sequence identified by NGS. Here, we isolated a total of 53 clones with one pass of the isolation workflow that spans across the treatment response, including the 19 clones of interest displaying strong relative gemcitabine sensitivity.

Figure 31. Schematics for informed isolation of clones of interest and validation of clonal barcode identity. Volcano plot from culture plates at 5,000 cells/well for differential clonal lineage behaviors upon Gemcitabine treatment; enriched clones highlighted in red; depleted clones highlighted in blue (LogFC = Log$_2$ fold-change, Gem vs. CTRLs, FDR<0.25); Clones of interest (depleted clone: 5125_5793, LogFC = -1.20; enriched clone: 2211_135, LogFC = 1.18) (top left); (right graph) Heat maps of relative abundance of two clones (blue = relatively sensitive, red = relatively resistant to treatment) from the de-complexed (5,000 cells/well) plate (upper two graphs). Wells with the most representation of the two clones were picked for single-cell sorting (circled wells); Well-coordinates of the isolated two clones traced back using positional sequencing of multiple single-cell sorted plates using well and plate NGS indices (bottom two graphs); Bottom left graph: Chromatogram of Sanger sequencing of isolated clones to validate barcode identified by NGS.

Functional gemcitabine sensitivity validation of isolated treatment naïve clones of interest

To illustrate whether or not the isolated clones would behave similar to the phenotype observed in the CRT *in vivo* cohorts, we took the two most representative clones at both end of the gemcitabine sensitivity spectrum (ID 5125_5793 being sensitive, and ID 13767_14001 being resistant) for functional characterization. Individually, the treatment naïve clones were cultured and treated with Gemcitabine for 48 hours and wash out. Mimicking the *in vivo* setting, we allowed the culture to grow back (relapse). As we can see from the figure and crystal violet staining and

viability quantification (absorbance, means of 4 measurements +/- SD), clone 13767_14001 displayed, intrinsically stronger resistance and ability to grow back than clone 5125_5793.
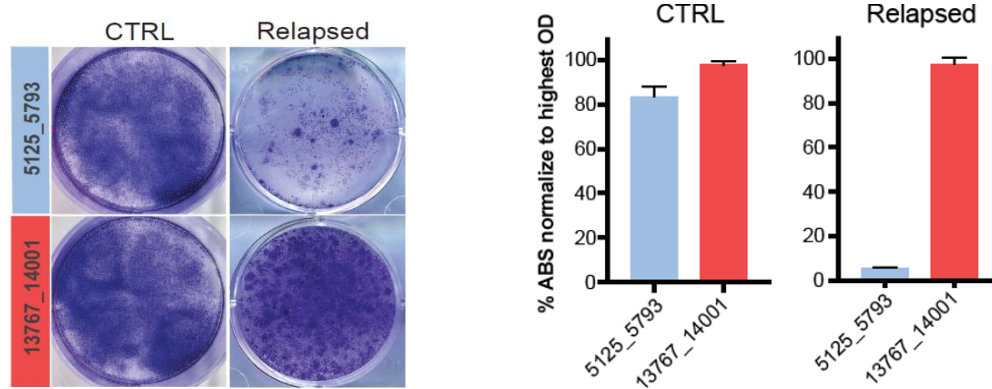
Figure 32. Isolated clones exhibit differences in tolerance to gemcitabine treatments. Crystal violet staining for the two isolated clones 5125_5793 and 2211_135 prior to treatment with 10nM gemcitabine (left "control") for 5 days and relapse (17 days) after gemcitabine washout (right "relapsed"); Crystal violet 570nm absorption readout (normalized to highest OD, n=4) for each condition (right). Prior to treatment, 2211_135 had 1.17X more OD measurement than 5125_5793; after 17 days of relapse, 2211_135 had 16.5X more OD measurement than 5125_5793.

Furthermore, we tested the clonal relevant resistance to gemcitabine while the two clones were mixed together. The sensitive clone was re-infected with vectors expressing 5125_57923 (GFP+) and mixed together with 13767_14001 (GFP-) equally (75,000 cells) and plated in 2D culture. The mixed culture was then treated with different concentrations of Gemcitabine (0nM, 100nM and 500nM) for 5 days and then washed-out. Then, the culture was left for relapse for 14 days. The relative abundance of the two clones were determined by FACs analysis. The result showed that without treatment, the two clones behaved and grew similarly after two weeks (43.8% GFP+, vs. 54.0% GFP-). For the group under treatment, FACs analysis showed clone 13767_14001 exhibiting more tolerance to gemcitabine and faster relapse rate than 5125_5793 (4.2% vs. 4.9% after relapse at 500nM) and displayed dosage dependent response.

This data set validates the lineages that maintain tumors, never exposed to gemcitabine treatment, intrinsically display differential gemcitabine resistance. Moreover, the clonal

gemcitabine sensitivity phenotype obtained via CRTs cohorts can be reproduced in an *in vitro* setting. The possibility of being able to isolate clones that is treatment naïve and characterize them, coupled with the treatment phenotype information (via *in vivo* CRT experiments) opens up a unique opportunity to determine clonal intrinsic molecular mechanism contributing to drug resistance.
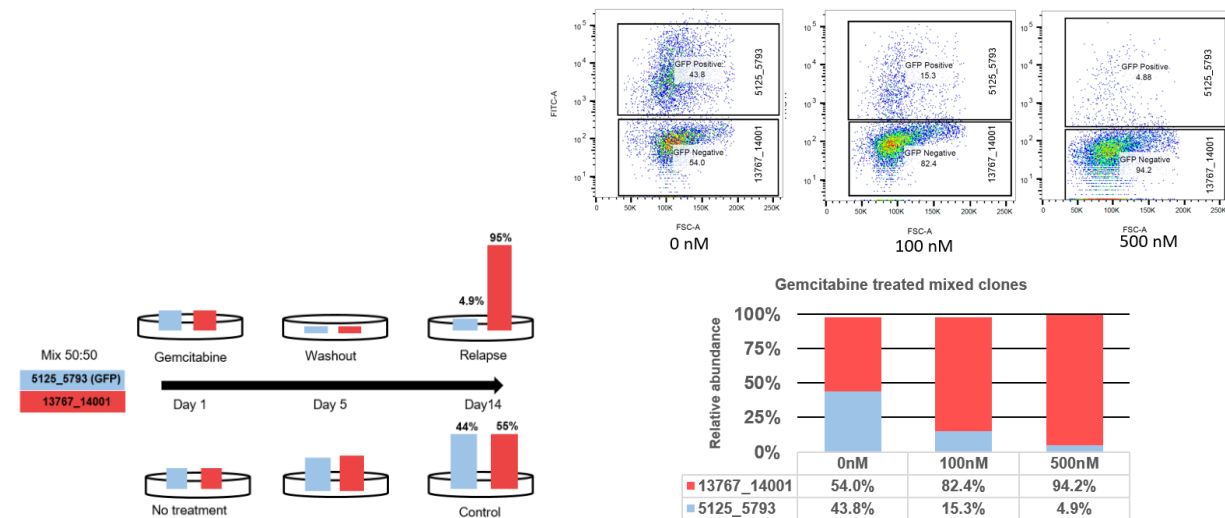
Figure 33. Co-culturing of resistant and sensitive clones under gemcitabine treatment. Mixed clones' treatment response; 5125_5793 (GFP) and 13767_14001 were mixed equally (75,000 cells each) and plated in 2D culture. Top group was treated with 500nM gemcitabine for 5 days and followed by wash-out and allowed for regrowth. Bottom group was not treated and allowed to grow to confluence. FACs analysis showed clone 13767_14001 exhibits more tolerance to gemcitabine and faster relapse than 5125_5793 (95% vs. 4.9% after relapse). The clonal relative representation of the culture was analyzed through FACs to determine GFP+ vs. GFP-populations (top right). For untreated culture, two clones are relatively equal in representation (~44% for "sensitive clone 5125_5793" vs ~55% for "resistant clone 13767_14001"). Under gemcitabine treatment, the "resistant clone 13767_14001" dominates the "relapsed" culture and displayed gemcitabine dosage dependent response from (0nM, 100nM and 500nM) (bottom right).

Genomic and transcriptomic analysis of isolated treatment-naïve clones displaying differential drug response determined via CRTs *in vivo* experiment

The isolated clonal cultures span across the entire gemcitabine sensitivity spectrum as determined from the CRTs *in vivo* experiment. In the figure below, we display 12 of the clones

across the gemcitabine resistance spectrum from sensitive to resistance and their relative position in the volcano plot. To study the strength of the relative chemotherapy resistance phenotype associates with the genomic or transcriptomic profile, we submitted these clones for molecular profiling (WES, RNAseq, RPPA).
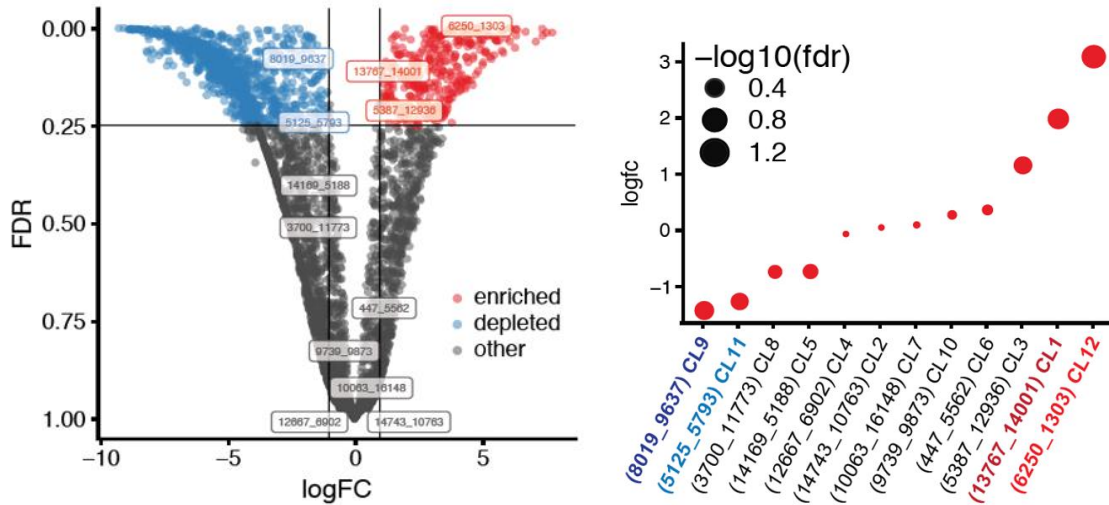
Figure 34. 12 isogenic gemcitabine resistance clones isolated. Position in the volcano plot of twelve isolated clones with differential gemcitabine sensitivity in PDX2-derived *CRTs* (red = enriched, blue = depleted, grey = not significantly modified). Each dot represents a clone (unique barcode). Clonal sensitivity to gemcitabine is calculated by comparing the relative fold-change of abundance in gemcitabine relapsed CRT tumors vs. controls (X-axis); false discovery rate (FDR) of treatment response change is calculated for each clone (Y-axis); number in the boxes indicate the unique clonal barcode ID of isolated clones; alignment of 12 isolated clones spanning across the continuum of response (logFC = log2 fold-change) to gemcitabine treatment; clonal lineages are displayed as red dots, with size representing the confidence in the treatment response (–log10 FDR) (right).

Clonal heterogeneity exhibited at the genomic level and does not have significant correlation with treatment response

Through copy number variance (CNV) analysis of the clones from WES data, we observed many of the chromosomal events shared amongst the clones. Common chromosomal events such as the deletion of SMAD4 on hr18 and KDM6A on chrX. Besides the common events, clonal specific events were also observed. As an example, on chromosome 4, a region containing 33 genes was detected to be amplified in the parental population as well as with 9 other clones. In

addition, we also used SNP6 array (Affymetric) to measure copy number variation alongside CNV

data obtained from WES. The CNV data from both technologies infers the parental cells in this

model are triploidy/pseudo-tetraploidy.
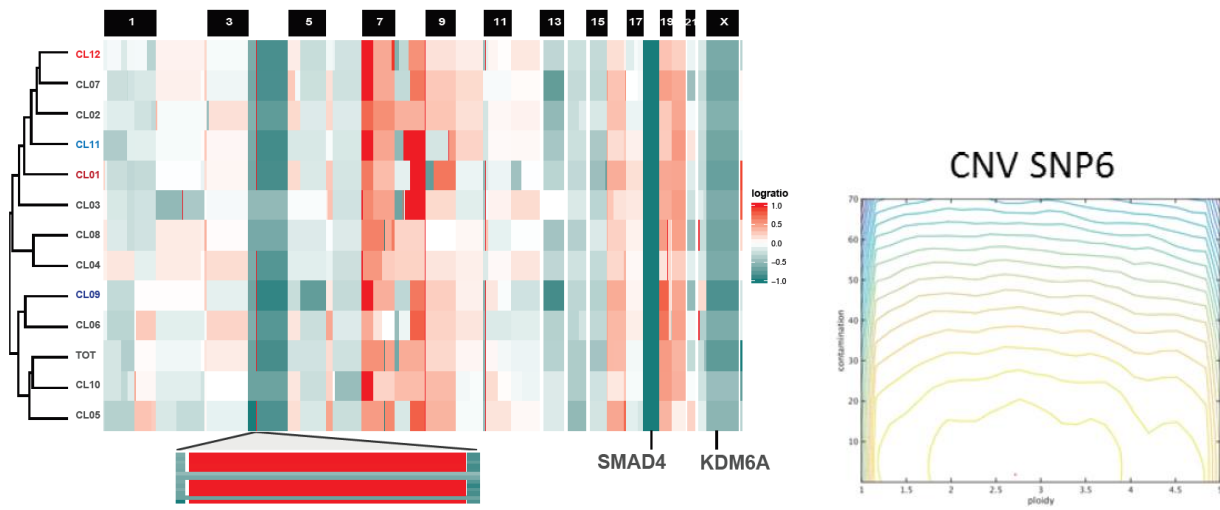
Figure 35. Heatmap of genetic alterations (WES) and clustering of copy number variation (CNV) of the isolated 12 clones. Genetic alterations are gene-level log2 ratios. Bottom panel shows a zoomed in version of a region in chr4 (Positions: 37245842- 41962589, number of genes: 33) which is amplified in parental population and 9/12 clones; The right panel shows PICNIC (predicting absolute allele copy number variation with microarray cancer data) for the clones and parental population, it specifically estimates normal contamination and ploidy and background hybridization levels, and arrives at a Bayesian prior distribution of ploidy (red point at 2.7); Genome level copy number (log2ratio, right panel) changes across all chromosomes as derived from SNP6 array on the parental PDX sample, aligns well with parental cell line data.

From SNP and allele frequency (AF) analysis, we observed many genetic alterations with

genes that were known drivers of PDAC, genes such as TP53 that were shared amongst all

clones, as well as the parental population. Taking a deeper look, we observed clonal

heterogeneity in AF variations in KRAS^{G12D} amongst clones (min= 0.41, max=1.0, the parental =

0.76). This show that the parental population harbors clones with different copies of KRAS

mutation, which is in line with PDAC patient genomic data. Furthermore, we observed that many

SNPs that has a low AF in the parental population, but high AF in the individual clones. This

suggest that these clones were genetically drifted from a common ancestor as seen from the

phylogenetic trees generates based on SNP and CNV. Interestingly, by cross referencing the genomic profiles and defined molecular state of individual clones to their *in vivo* treatment response, individual clonal growth kinetics, or relative tumor abundance, we did not find any genomic traits that could predict the drug treatment response. This suggests the different degree of drug resistance could be associated more with their clonal specific epigenetic and transcriptomic state.
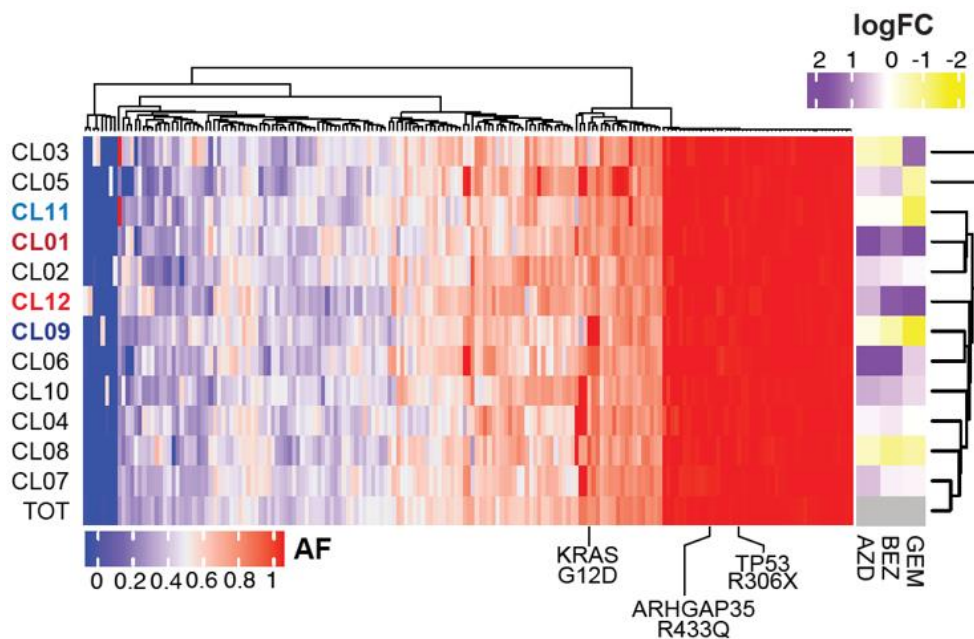
Figure 36. Heat-map of the single nucleotide variants (SNV) of 12 isolated clones. SNP included somatic mutations existing in at least one of the samples and previously detected in TCGA and ICGC cohorts, AF = allele frequency.
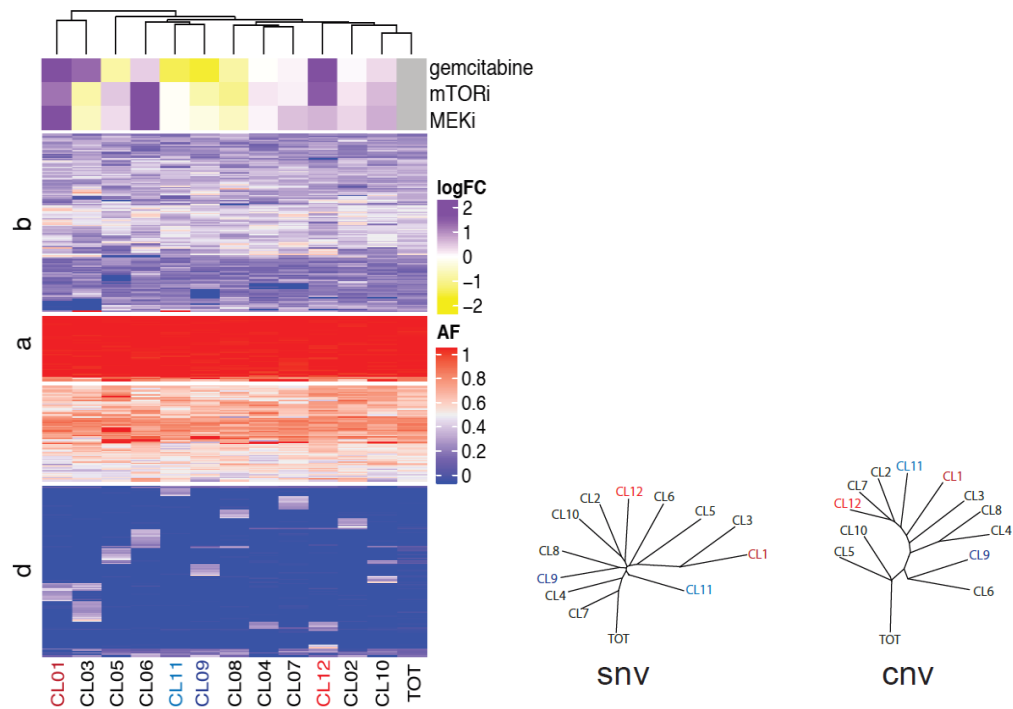
Figure 37. Allele frequency clustering of 12 isogenic clones and phylogenetic trees based on SNP and CNV data. Heat-map highlighting the k means clustering (k=4) of somatic alterations (AF = allele frequency) detected in PDX-derived parental population (TOT) and 12 independent isolated clones: cluster a. mostly clonal all three copies, cluster b. mostly 2 copies, cluster c. mutations with single copy in the majority of clones, cluster d. mutations detected in single copies in few of the clones. Right graph shows the unsupervised clustering substantiates a condition of triploidy/pseudo-tetraploidy for this PDAC model (data also corroborated by SNP6 array). Phylogenetic trees based on various –omics platforms (SNV, CNV).

Clonal differential transcriptomic profile associates with gemcitabine treatment response - DNA damage is a major mechanism of adaptive resistance to chemotherapy in pancreatic cancer

Through analyzing clonal transcriptomic profiles, we found strong correlations between their individual transcriptomic profiles and their relative sensitivity to gemcitabine *in vivo* through GSEA pathway analysis. Intrinsically, molecular pathways such as - DNA damage repair, oxidative phosphorylation, MYC/E2F targets, and folate biosynthesis are selectively enriched in gemcitabine-resistant clone.
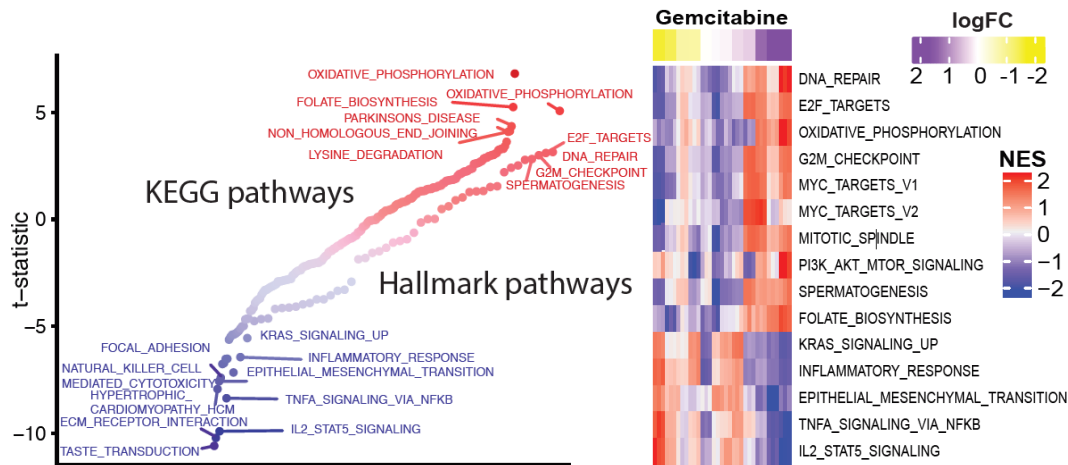
Figure 38. Pathway analysis (KEGG, Hallmark) of 12 isogenic clones and their gemcitabine resistance phenotype. Association of molecular pathway scores (ssGSEA) with gemcitabine response (logFC = log2 fold-change) for 12 isolated clonal lineages (Left graph). Y-axis represents t-statistic, with gemcitabine logFC as response variable and pathway scores as an independent variable (left, KEGG; right, Hallmark); (Right graph) Heat-map highlighting molecular pathways significantly enriched (NES) in gemcitabine resistant or sensitive clones (aligned by logFC to gemcitabine).

In line with these findings, upon *in vitro* validation of chemo-sensitivity, we took the top four clones with the most differential gemcitabine sensitivity *in vivo* for further characterization (two resistant and two sensitive). Consistent with the findings where the resistant clones have intrinsic upregulation in DNA damage repair, we demonstrate that the two resistant clones have a lower basal level of phosphorylated histone H2AX (γH2AX). γH2AX, phosphorylated by DNA damage check point proteins (ATM or ATR), is responsible for recruiting DNA damage repair proteins to DNA double strand break foci. Therefore, a lower level of phosphorylated γH2AX indicates the basal level of DNA damage in the two resistant clones is lower than the two sensitive clones. We further hypothesize that the gemcitabine-resistant clones, upon exposure to gemcitabine, due to having a higher DNA repair capacity, would maintain the DNA damage at a lower level by comparison to the gemcitabine-sensitive clones. The experimental result shows exactly that. Upon 24 hours of gemcitabine treatment, both the resistant and sensitive clones' DNA damage level increases (measured by γH2AX) compared to their untreated counterparts; furthermore, the resistant clones have definitively lower levels of DNA damage than those

detected in the sensitive clones. In addition, staining for RAD51, a protein that is actively involved in DNA damage repair, we found the nuclear localization of Rad51 at γH2AX foci in the resistant clones with lower levels of DNA damage post treatment, which further supports that DNA damage repair plays a key role in cell recovery and survival. Treating the resistant clones with AZD-6738, a potent ataxia telangiectasia and rad3-related (ATR) inhibitor, a critical master regulator of DNA damage response, sensitized the gemcitabine resistant cells to gemcitabine[77]. This suggests that cell intrinsic elevation in DNA damage repair is a key mechanism in cells adapting to gemcitabine treatment in pancreatic cancer.
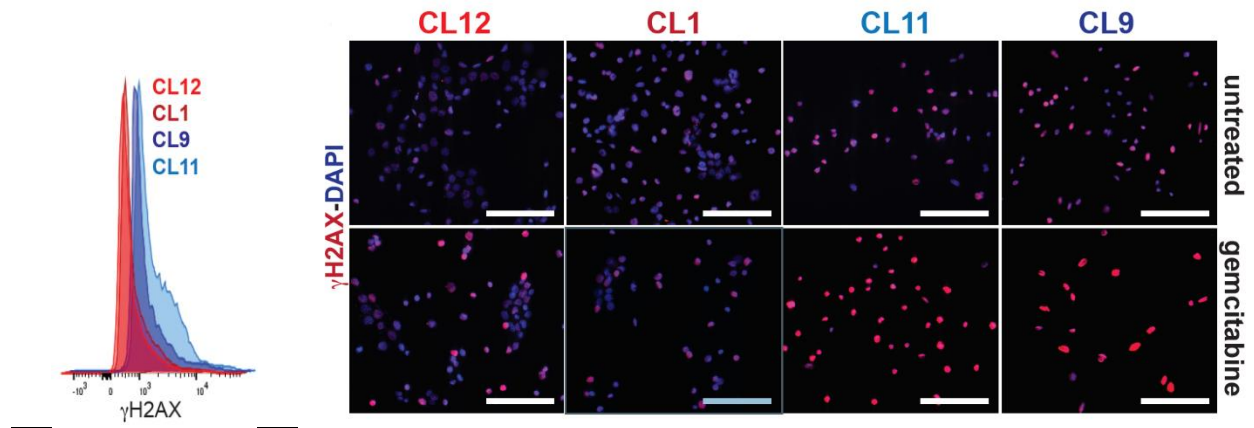
Figure 39. Differential γH2AX level of gemcitabine resistant and sensitive clones by FACs and IF. Flow cytometry analysis of basal level γH2AX signal of isolated clones with differential gemcitabine sensitivity (CL12 and CL1 are relatively resistant (red); and CL9 and CL11 are relatively sensitive (blue) (Left graph); Resistant (CL1, CL12) and sensitive (CL9, CL11) clonal lineages were treated with or without 1 µM gemcitabine for 24 hours (bottom and upper panels respectively) followed by immunostaining of γH2AX in red and DAPI in blue (Right graph). Scale at 100 µm.
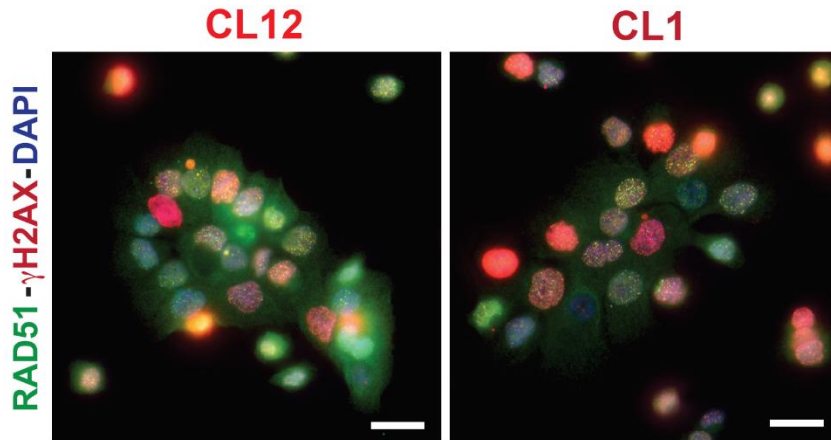
Figure 40. RAD51 and γH2AX staining (IF image) of gemcitabine resistant clones' post exposure to gemcitabine. Co-immunostaining of RAD51 (Green), γH2AX (Red) and DAPI (Blue) of resistant clones (CL12 and CL1) after 24 hours treatment with 1 µM gemcitabine followed by 3 hours recovering.
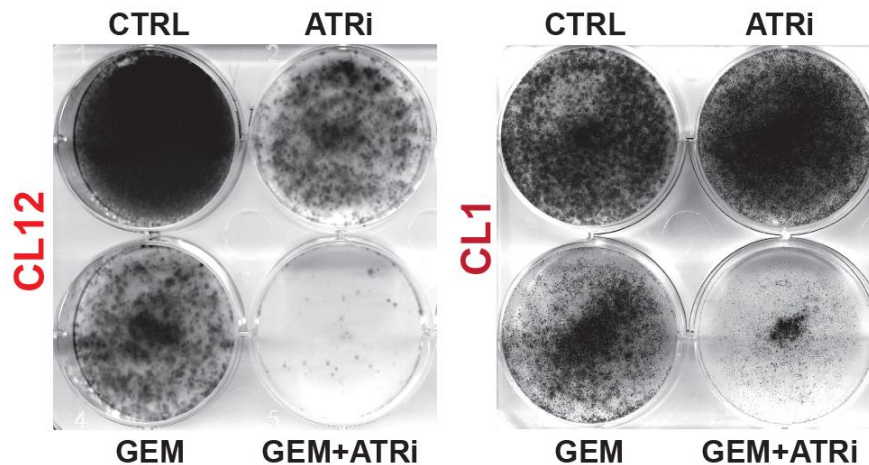
Figure 41. Colony formation assay of resistant clones under treatment of GEM, ATRi and in combination. Resistant clones (CL12 and CL1) were pre-treated with 1 µM ATRi AZD6738 for 3 hours followed by co-incubation with 1 µM gemcitabine for another 24 hours. Scale bar at 20 µm.

Gene signature identified through isolated clones predicts patient response to chemotherapy

Based on the 12 isolated clones' differential sensitivity to gemcitabine *in vivo* and their differential gene expression profile, we identified a molecular signature based on 200 most differential expressed genes. Using this gene set, matching with TCGA PDAC data-set, we were

able to detect pre-existing resistant populations of cells in pre-treated tumors; furthermore, it

allows for prognostic stratification of PDAC patients in terms of disease-free and overall survival.
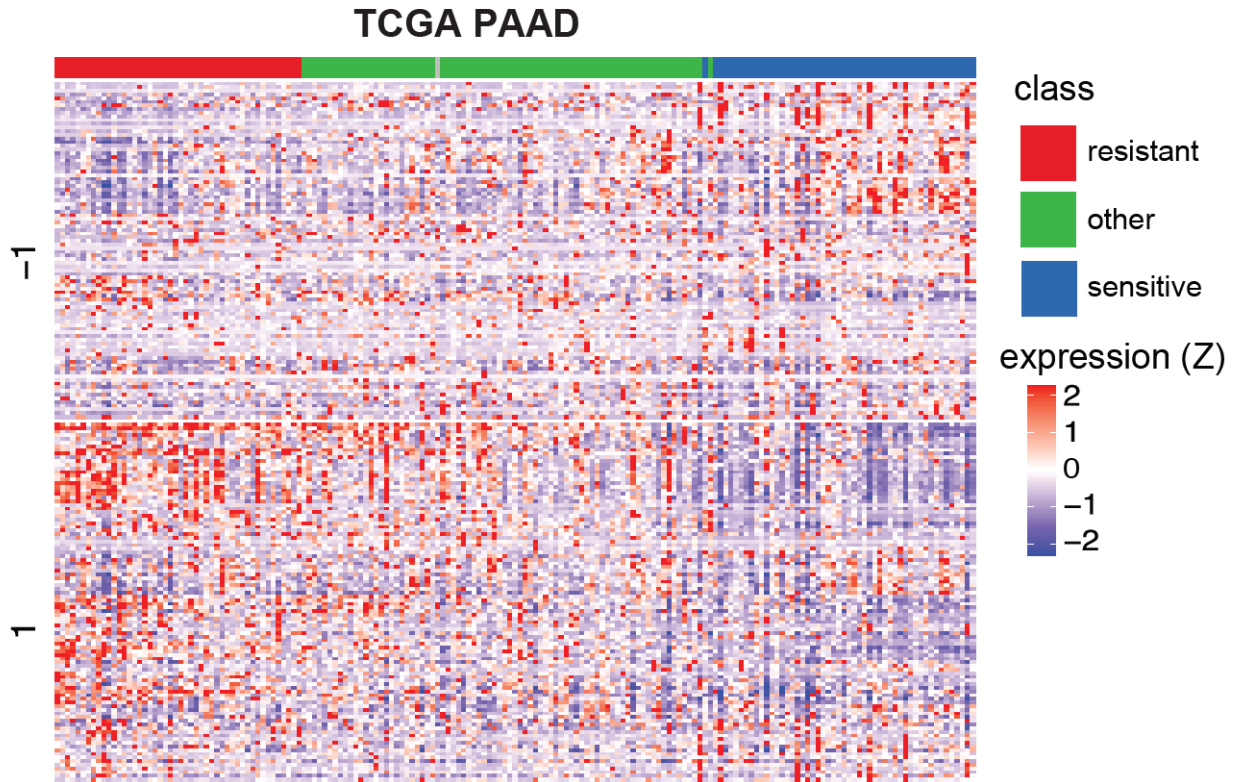
Figure 42. Top 100 most enriched and depleted genes (total 200) were prioritized to create a gene-signature. TCGA pancreatic cancer tumors (n = 179) were subtyped using this gene signature, with genes in rows and samples in columns. Top annotation bar represents the final class of each tumor, with each cell representing normalized expression of a specific gene in a tumor, clustered into two sets (enriched/resistance/1 or depleted/sensitive/-1).
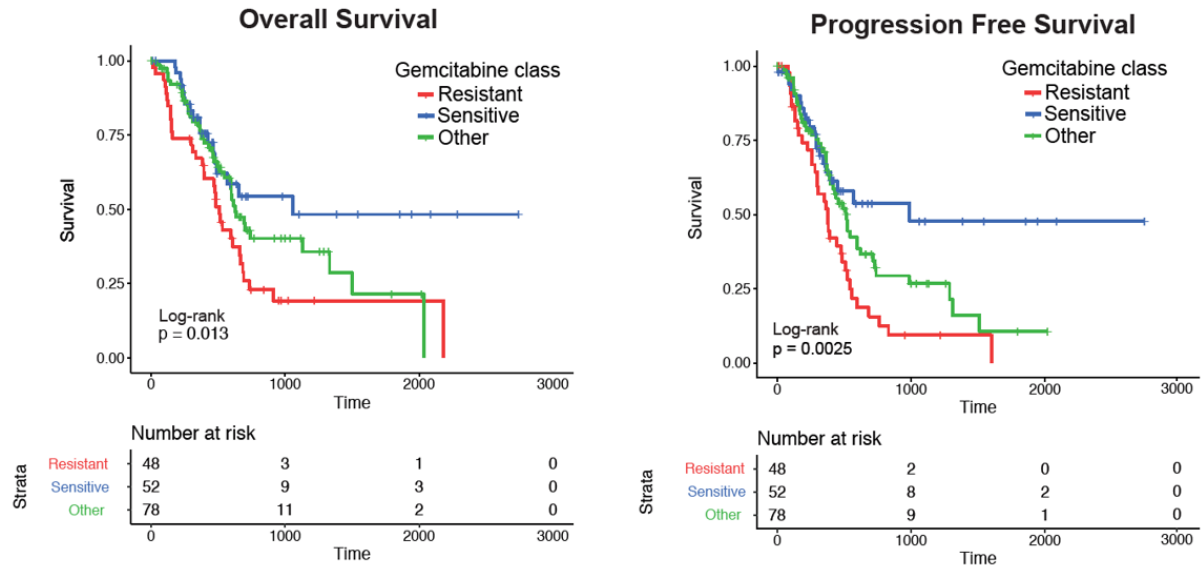
Figure 43. Survival analysis of TCGA pancreatic cancer patients, classified according to the gene-expression signature informed by gemcitabine-treated CRTs (red: gem resistant, blue: gem sensitive, green: others). Two panels show Kaplan-Meier curves using overall survival (Log-rank p=0.013) and progression-free survival (Log-rank p=0.0025), in left and right respectively.

Chapter Summary and Discussion

Chemo-resistance has been one of the major challenges in achieving complete remission in cancer treatment, during which a fraction of the tumors has the capability to escape treatment or adapt to external perturbagens. Tumors, which shrink and decrease in mass under treatment regimens, often grow back upon cessation of treatment. Furthermore, relapsed tumors usually become more aggressive and more resistant to drugs. The molecular mechanism underlying such resistance has long been a major and necessary focus in the field of cancer research. With advancements in molecular biology techniques in cellular genomic and transcriptomic profiling, the field has gained insights to another level of tumor complexity. Tumors can be viewed as complex ecosystems, where heterogeneous cells co-evolve and acclimate to environment. Thus, tumor heterogeneity, in which a wide array of genetic or epigenetic events have been acquired during tumor development, contributes to the relative functional resistance of these tumor cell sub-populations that drive relapse [78-80]. Currently, in concordance with clinical studies[81], it has

been widely accepted that tumor evolution and adaptation are inherited properties of cancer and plays a pivotal role in treatment resistance and tumor relapse. However, merely measuring tumor volume to determine the long-term efficacy of treatments is grossly insufficient [82]. To fully understand how a tumor responds to treatment and to pinpoint the mechanism of resistance, it is essential for researchers to study tumor's functional heterogeneity at the clonal level. To achieve this, we need a robust experimental model that can accurately measure tumor clonal complexity upon treatment.

To facilitate the understanding of chemo-resistance tumor heterogeneity, cellular barcoding lineage tracing experiments are particularly useful, as they are a relatively unbiased empirical method for interrogating tumor sub-clonal architecture and dynamics upon exposure to pharmacological agents. In addition, they can precisely identify small sub-groups of tumor cells that exhibit a resistant phenotype under treatment. Nonetheless, current methods for lineage tracing have severe drawbacks in terms of resolution, complexity, and lack of functional characterization beyond -omics profiling, as well as, the inability to trace the same clonal lineages in parallel animals due to low reproducibility of tumor clonal compositions amongst animals. These limitations, therefore, confound the questions that we can empirically answer through experimental design. In order to improve applicability, reproducibility of pre-clinical research results, and overcome these limitations, we demonstrate a patient-derived xenograft (PDX) transplantation model that is suitable for the robust tracing of complex tumor lineage dynamics when exposed to treatment. We call this model "clonal replica tumors" (CRTs), where, in a cohort of animals, each animal bears tumors that are essentially identical in their clonal compositions and, thereby, provide a high level of biological reproducibility.

By barcoding early passage, patient derived, pancreatic tumor cell lines and following their individual clonal dynamics throughout *in vivo* serial transplantation (F1 – F4, over 8 months period), we unveiled the hierarchal organization amongst clones. In terms of clonal ability to

sustain tumor growth *in vivo*, our data suggest that solid pancreatic tumors are organized in a manner similar to hematopoietic stem cell hierarchy, where most lineages exhaust over time with only a fraction of clones having the self-renewal capability to sustain tumors over long periods of time (~2% of cells). This observation is further confirmed with *in vitro* culturing of single cells and by monitoring the fraction of cells with the ability to proliferate and form new colonies. Echoing the *in vivo* serial transplantation data, we observed that only a fraction of the single cells could give rise to new culture and engraft *in vivo* (~7-12%), while other cells displayed dynamics of transient amplification or exhaustion or were unable to initiate proliferation. Furthermore, the percentage of such long-term potentiating lineage compartments are tumor dependent; specifically, the PDAC tumor cell lines PATC124 and PATC53 have ~5% and ~12% self-renewal cells, respectively. These data, which were measured by a TIC assay, agree with the higher percentage of tumor initiating cells found in relatively more malignant tumors.

During the *in vitro* passaging of barcoded cells, we observed a clonal dynamic similar to that observed from the *in vivo* serial transplantation. Over time, the culture reached an equilibrium state and was sustained by a fraction of the initial lineages, whereas the majority of the lineages would exhaust. Remarkably, we observed overlap when comparing the lineages that were enriched at the equilibrium state *in vitro* to the lineages endowed with long-term self-renewal ability obtained functionally through *in vivo* serial passaging. The initial 2M unique variations of barcoded lineages, after 8 months of passaging *in vivo* and 24 weeks of passaging *in vitro*, converged to a common set of barcoded lineages (~801 lineages). Furthermore, the overlapping set of common barcoded lineages represented the majority (96.5%) of the tumor mass after the mouse received a subcutaneous injection of *in vitro* "passage 26" cells, as well as represented the majority (67.9%) of the tumor mass of the serially transplanted *in vivo* tumors at 8 months. Contrary to the common belief that *in vitro* 2D cultures are not usually seen as optimal conditions for enriching tumor "long-term self-renewal" cells (or "cancer stem cells" in terms of long-term

potentiation ability) through the *in vitro* "stabilization" process, we were able to obtain tumor lineages that naturally achieved an equilibrium state and were enriched in cell lineages that could sustain tumors long-term. Moreover, through repeated experiments and in multiple models, the tumor consistently followed similar evolutionary configurations, where the barcoded clonal population eventually resulted in a hierarchical equilibrium organization state, irrespective of different culturing conditions and external stress. This suggests a model of tumor evolution is guided fundamentally by pre-defined sub-clonal hierarchical compositions, where external pressure plays a role in influencing tumor evolution routes through the kinetics of clonal exhaustion. By comparing *in vitro* serial passaging vs. *in vivo* serial transplantation, we observed much faster kinetics of exhaustion *in vitro*, where the same lineages that are functionally exhausted drop out at a much faster rate compared to their counterparts *in vivo*. This suggests that external conditions can influence a sub-clone's rate of exhaustion, but have limited influence on its core clonal hierarchical equilibrium amongst lineages. Obviously, our model has limitations in terms of representation, such as barcoding PDX derived cell lines that have undergone a selection of cells and can grow outside of their tissue of origin. However, our data may be hinting at the possibility that pancreatic cancer, and perhaps other cancers in general, are governed by the common rules of hierarchical equilibrium amongst tumorigenic lineages [83].

Utilizing the self-renewal, hierarchical nature of these tumor cells, we expanded these post-serial passaging barcoded cells (stabilized in clonal equilibrium) and created a large cohort of models bearing tumors that were enriched for long-term self-renewal lineages. Indeed, when comparing the lineage composition of tumors in a cohort, we found that these tumors were essentially identical in both clonal composition and their relative abundance (common barcodes made up >98% of individual tumor, Pearson's correlation value between tumors were >98%). In another words, the tumor's clonal heterogeneity was replicated in multiple independent mice— hence the name, clonal replica tumors (CRTs). In essence, we can use CRTs to study the

dynamic responses of tumors with precision and at the clonal level of resolution because they allow the ability to interrogate the same tumor (or sub-clones) in parallel animals potentially under different external perturbations.

Studies on the treatment resistant subpopulations of cells that have common features of quiescence, stemness, higher level of tolerance to DNA damage and detoxification capacity have resulted in a persuasive model in how tumors escape and survival pharmacological agents [84]. Here, we subjected CRTs, which were enriched in self-renewal lineages, to chemotherapy treatment to study their clonal dynamic response and evolution after relapse. Under gemcitabine treatment, we saw a minimum disturbance in clonality in the relapsed tumor despite observing a drastic impact of treatment on tumor mass. In other words, despite inducing a temporary regression, the transient effect of gemcitabine failed to penetrate the tumor's long-term self-renewal compartment ("stem cell like compartment"), and thus, most of the tumor-sustaining lineages were able to re-enter the cell cycle upon cessation of treatment. Interestingly, when applying targeted therapy (AZD and BEZ) in parallel animals within the same CRT cohort, we observed the fitness of different sub-clonal lineages being affected in a drug dependent manner. Certain lineages were susceptible to a specific drug more than the others, and overall, we observed a wide spectrum of lineage-dependent differential drug sensitivity. The result suggests that the tumorigenic, self-renewal compartment of tumors are functionally diverse under the context of various therapeutic regimens. Overall, upon considering the tumor clonal evolutionary dynamic from barcoding to stabilization, and the differential clonal response to drugs of different mechanism of action, our data depict a model where clonal equilibrium can be achieved over time. Further, our data indicate that it is under different external perturbagens that clonal evolution and their newly established equilibrium (relapsed tumor) can be fully appreciated. It also depicts the differential functional response of intrinsic tumorigenic cells to therapy, and the competition in fitness amongst pre-existing clones as a major determinant in shaping tumor resistance and its

evolution. These lineage-dependent differential drug responses are especially important for combinatorial therapeutic applications, a rationale that was introduced by Frei in the year 1965 [86]. While adaptive therapeutic resistance—during which new genetic mutational events lead to an increase in fitness and eventually contribute to the newly established clonal equilibrium—can certainly occur, it was not a part of the original experimental design; however, it could be addressed and explored in future experiments by surveying clonal specific –omics profiles pre- and post-therapy.

Through isogenic clonal isolation and detailed –omics characterization, we demonstrated that an incredibly high degree of molecular heterogeneity exists within the tumorigenic populations of cells. This genomic and transcriptomic diversity may account for the wide spectrum of differential functional phenotype and responses to pharmacological agents, and "long-term self-renewal" clonal lineages may drive distinct tumor evolution in response to treatments. This concept is also corroborated by recent findings in glioblastoma multiforme (GBM), in which there exists sub-populations of tumorigenic cells that may be responsible for shaping tumor evolution when exposed to temazolomide [85]. The orthogonal approach of being able to isolate clones that are treatment naïve and obtain molecular profiles that could be corroborated with an unbiased lineage dynamic response to treatment through barcoding can lead to the identification of clonal intrinsic properties contributing to drug resistance. We discovered that clonal intrinsic enhanced tolerance to DNA damage by DNA damage repair is one of the top features in pancreatic cancer resistant cells when exposed to gemcitabine. Interestingly, while we found that the genetic profiles of isolated clones did not correlate with resistance phenotype, we generated a 200-gene transcriptomic signature that is predictive of therapeutic response in PDAC patients. Although the predictive application of this gene signature is debatable and is limited by the number of clonal samples in the analysis that it is based on, our method of starting from an unbiased barcoding of early passage PDX derived cell lines, creating a cohort of CRTs, treating tumors *in vivo*, and

ultimately isolating clones for validation and profiling to obtain gene signatures that correspond well to clinical therapeutic responses is extremely encouraging and provide evidence that the CRTs are a biologically relevant model. Furthermore, while our findings are limited in scope, they encourage the further study of combinatory treatments of DNA damage response inhibitors, such as ATRi with gemcitabine, which may further sensitize chemotherapy resistant cells in PDAC patients and lead to better standard-of-care response.

Chapter 2

Looking through a Phenakistiscope: Modeling Lineage Crescendos & Decrescendos during Tumor Expansion and Dissemination through Unperturbed Longitudinal Clonal Tracing

Abstract

Tumors can be viewed as a complex ecosystem where multiple sub-clones harboring specific functional phenotype compete and co-evolve. To quantitatively capture the interplay of lineages during tumor expansion and dissemination in an unbiased and unperturbed fashion, we created an orthotopic clonal lineage tracing model by establishing clonal replica tumors. Using PDX derived cell lines from pancreatic cancer patients, the model revealed the complex and dynamic alternating dominance nature of cancer expansion. Regarding dissemination, the model captured distinctive, organ specific clonal dynamic profiles amongst distal sites and revealed the stochastic and explosive nature of local clonal expansion. Further lineage tracing experiment reveals that the cells with intrinsic self-renewal properties are likely to be responsible for distal tumors dissemination. Through molecular characterization of isolated clonal lineages that have differential potential for expansion and dissemination, we identified constitutive N-methyl-D-aspartate (NMDA) receptor upregulation in clones that are more aggressive and invasive. Treating reconstituted tumors with a non-competitive NMDA antagonist (memantine) effectively reduced tumor expansion and dissemination. Furthermore, we identified a set of molecular signatures enriched in more aggressive and invasive lineages that could potentially inform on clinical outcome.

Introduction

Cancer is not a static disease, rather can be regarded as a complex ecosystem wherein either diverse sub-lineages (or clones) and their environment co-adapt and evolve. Clonal evolution is involved in multiple stages of cancer development from the beginning to end [43, 44]. Examples of cancer clonal evolution include, pre-malignant clonal expansion [87], gain of fitness associated mutations driving clonal dominance, selection of sub-population of clones with ability to disseminate and metastasize to distal organs [88, 92, 93], as well as treatment induced selection of resistant clones able to sustain tumor relapse [81, 89]. Competition amongst lineages and adaptation to tumor intrinsic or extrinsic environmental perturbations, selecting for clones with survival advantages, is one of the fundamental mechanisms driving tumor progression and tumor heterogeneity [91, 94].

Clonal lineage tracing is a powerful experimental approach to gain insights into how tumor sub-clones naturally evolve and respond to perturbations [95]. However, currently there are no *in vivo* model to robustly trace clonal lineages over time, nor ways to functionally characterize specific lineages beyond static –omics profiling. The challenges arise from multiple technical aspects of clonal detection such as perturbation of system (tumor) during re-sampling, inadequate sample size for uniform tumor representation, and the lack of clonal detection resolution [96, 97]. For instance, the serial biopsy or serial sampling process itself from the same tumor disrupts the tumor ecosystem; furthermore, the regions sampled may not necessarily represent the clonal makeup of the entire tumor. To the same point, surveying few metastasis lesions fails to inform on the degree of dissemination and clonal dynamics in the distal organ as a whole. More importantly, most models provide only a descriptive and qualitative depiction of complexity, lacking the ability to capture quantitative information, such as clonal cell number. For these reasons, most experiments extrapolate models of clonal evolution from endpoint analyses of cohorts of independent experimental subjects taken over time.

To bridge this technological gap, we created a barcoding-based longitudinal clonal tracing model to quantitatively capture in a high-resolution and robust manner the clonal progression of the entire tumor *in vivo* without perturbing the system. Using human pancreatic ductal adenocarcinoma (PDAC) patient-derived xenograft (PDX)-derived cells we generated *orthotopic clonal replica tumors (CRTs)*, which are cohorts of animals bearing human tumors that are essentially identical in their clonal composition. Orthotopic CRTs enables us to treat individual clones across different animals as interchangeable single variables, and therefore, providing the means to track the entire population's clonal dynamics over time through different animals without perturbing the tumor's ecosystem during expansion. Through this model, by surveying the entire tumor clonal abundance present in the pancreas and other organs over time, we observed multiple distinct clonal dynamics, as well as a high degree of Alternating Clonal Dominance (ACD) during tumor expansion. Furthermore, we observed distinct populations of clones displaying unique organotropism in the context of seeding and outgrowth, as well as the oligoclonal with monoclonal dominance nature of metastasis lesions. Through analyzing the relationship of metastasis clonal lineages and their counterparts in the primary tumor, our data strongly suggests cell intrinsic ability for long-term self-renewal is a critical property of metastatic clones. Furthermore, a unique capability of the CRT approach is our ability to expand clones of interest that are of relevance for this current work and are metastasis naïve– for functional characterization. Synergistic with the longitudinal clonal tracing model, to functionally study clones displaying various degrees of fitness and their pro-metastatic potentials observed, through a high-throughput clonal isolation workflow, we generated isogenic cell lines from the same population of barcoded cells used in the CRT model. Upon molecular characterization and functional validation of pre-metastasis isolated lineages, we identified key pathways, such as N-methyl-D-aspartate (NMDA) receptor, constitutively upregulated in clones that are more aggressive and invasive. Targeting cancer cells with a non-competitive antagonist of NMDAR (memantine), effectively reduced the tumor expansion and degree of invasion. Finally, through differential gene

expression analysis between clones with various degree of dissemination potential, we generated

a set of pro-metastatic gene signatures that can potentially predict patient clinical outcome and

inform choice of therapeutic regimen.
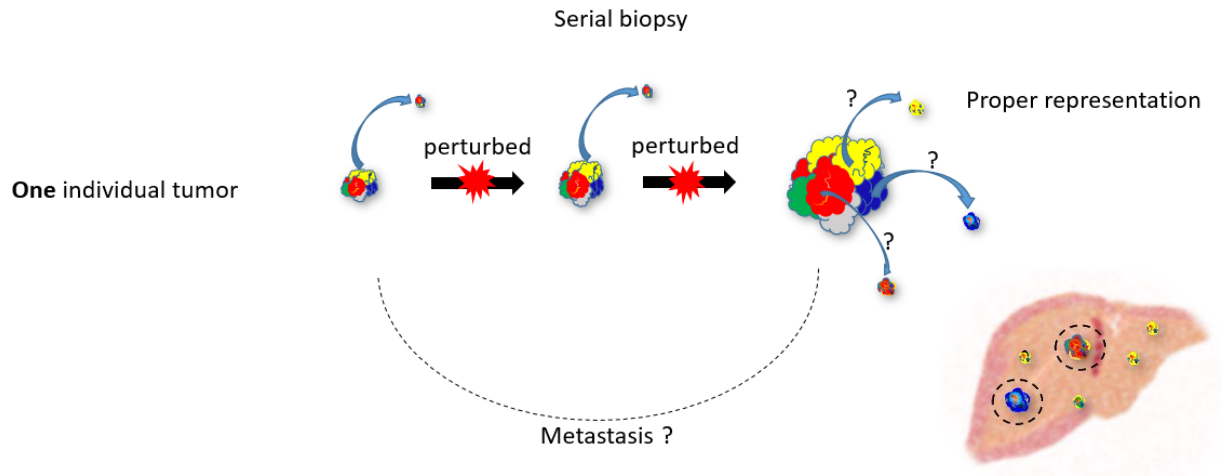
Figure 44. Limitations in current clonal evolution models. Currently, there are no models to continuously track clonal evolution without perturbing or acquiring enough sample size for adequate tumor representation. Schematic shows (assuming different color representing functionally diverse subclones), when performing serial sampling, each sampling of the tumor creates an artificial perturbation to the tumor, as well as lacking the ability to capture the clonal composition/ dynamics of the entire tumor. It is due to these practical limitations, capturing good representation of tumor expansion and dissemination clonal distribution or dynamics through serial biopsy is impossible.
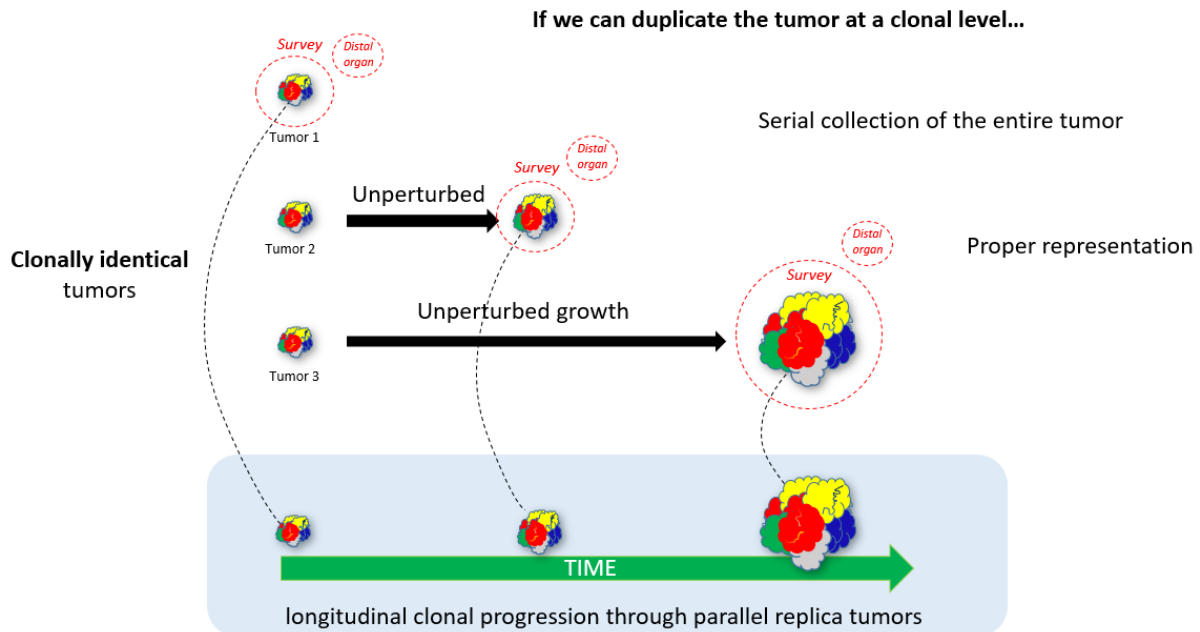
Figure 45. Orthotopic Clonal Replica Tumor longitudinal clonal evolution models. If we could establish clonally identical tumors, we could bypass the limitations of serial sampling and trace longitudinal clonal progression through parallel replica tumors. This method would give us an inclusive view of the entire clonal dynamic during tumor expansion and dissemination with good sample representation.

Pancreatic cancer and metastasis

PDAC has one of the worst prognoses among all cancer types with a 5-year survival rate of only ~5% [98]. Several factors contribute to this, including the fact that PDAC typically cannot be detected until it is at an advanced stage, frequently after invasion of nearby and distant organs has occurred. As a result, close to 90% of patients die due to obstruction or organ failure related to the metastatic disease[99]. Currently, it is generally believed that metastases occur late in tumor progression as tumor cells acquire new genetic alterations that confer invasiveness and survival advantages. This model seems to be corroborated by genomic studies where secondary lesions showed higher mutational load with respect to primary tumors, although no specific genes driving the metastatic process have been identified so far[39, 100]. Conversely, studies identified gene expression signatures of primary tumors that correlate with and are predictive of metastasis and

poor patient survival[101, 102]. These studies suggest that metastatic potential is established early during tumor progression and that metastases are probably driven by the same genetic alterations that drive tumorigenesis[100]. In other word, what is driving metastasis could be very much correlated to tumor fitness[101]. To gain clarity around these findings and gain perceptions in pro-metastasis factors to inform therapeutic strategies, a deep functional characterization of molecular mechanisms leading to the development of metastatic pancreatic cancer is urgently needed.

Focusing on the characterization of primary tumor cells with the potential to metastasize prior to dissemination

In order for metastasis to occur from a primary tumor, cells must undergo a multi-layered selection pressure despite the route of dissemination (i.e. vasculature, perivascular, nerve, or lymphatic system). The selection pressure could be both mechanical and molecular based, where cells must survival the external pressure during dissemination (i.e. shear pressure in blood), as well as adapting to tissue specific micro-environment (i.e. chemokines and growth factors), eventually establish or find a niche that supports metastasis outgrowth [56, 102, 103, 121]. Indeed, via studies of molecular characterization of paired metastasis samples and their primary tumors, a bottle-neck effect in terms of clonal heterogeneity was observed at metastatic sites [106, 107]. Studies also found metastatic site-specific mutations that were not present in the primary tumor [56], which suggests the nature of continuous evolution of tumor cells that are able to metastasize (parallel evolution). This posted a practical challenge for researchers, where characterizing and comparing established metastasis tumors with primary tumors might not be enough to determine key cell intrinsic factors driving metastasization. As illustrated below in Figure 46, where comparing distal tumor (mix of blue and red) with primary tumor (blue), with the color red representing potential post-metastasis acquired phenotype, may not necessarily reveal factors driving initial dissemination.

In line with the same concept, functional characterization of cells derived from metastatic sites is limited in informing us on aspects of the tumors post metastasis, where the tumor cells potentially undergone transformation due to the external selective pressure from the metastatic cascade and distal organ specific microenvironments. By only focusing on the study of these metastasis-site derived cells may confound our view in determining the driving factors that are inherited in the primary tumor toward a metastasis phenotype.

In our study, we focus our efforts on identifying clonal lineages that are pro-metastatic, as well as provide functional characterization on such clonal lineages that has the potential to metastasize, but are "metastasis naïve". In our experimental model, we choose early passages of cell lines derived from primary tumors as an effort to preserve tumor's natural diversity. These tumors are from patients that went through pancreaticoduodenectomy at stage III of the disease, where minimum to zero dissemination were clinically observed at the time of surgery, but eventually develop relapse months after surgery at distal sites. This indicates the cell lines derived from the primary tumor partly consist of cell lineage responsible for the recurrence in patients, and hence, having the potential to metastasize (pro-metastatic cell lineages, but are metastasis naïve).
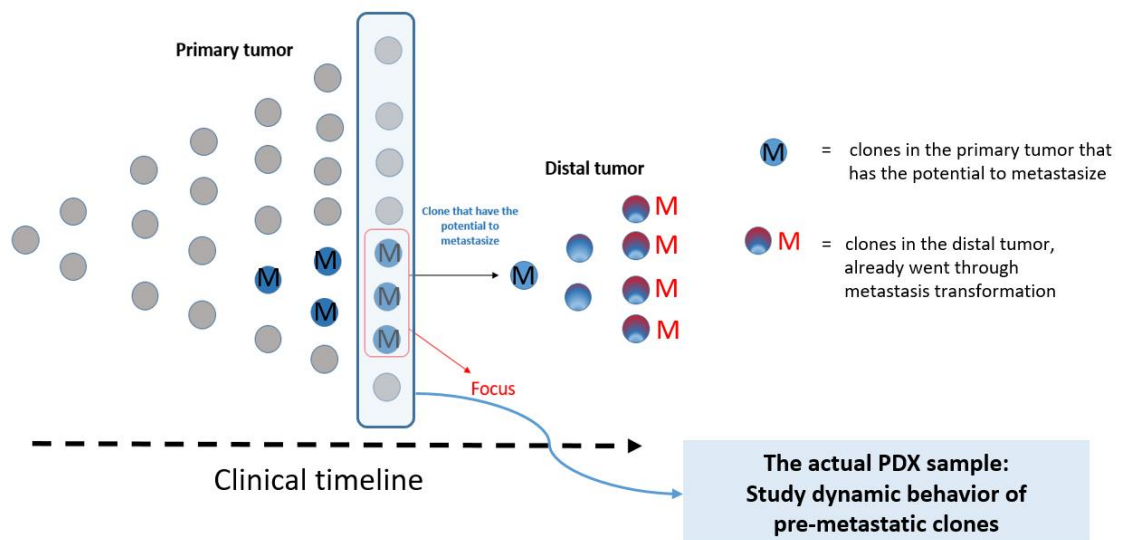
Figure 46. Schematic of tumor cells of study and the relative time frame of disease progression. The tumor cell lines in our study are derived from patient primary tumors with no clinical metastasis, however, eventually experienced distal metastasis relapse. Therefore, part of the tumor cells we focus our study on should contain clones (cells depicted in blue) that have the potential to metastasize, while the grey cells/clone representing the cells/clone that are "not able" (or significantly less able) to metastasize. Furthermore, such lineages (blue) derived from primary tumor has not gone through the metastatic cascade and acquire potential metastasis specific alterations (cells depicted in bluish red in the metastasis site).

Results

Orthotopic injection model to study metastasis progression

To establish a good representative model to study PDAC metastasis, I used an orthotopic (pancreas) transplantation approach due to its relevant organ microenvironment and blood circulation anatomy in the body. Mouse pancreas anatomy is different from that of a human, as the structure of the organ is less defined and dispersed throughout the peritoneal[111]. Therefore, the orthotopic injection is technically challenging and must be performed with high precision to avoid breaking any major vessels. When performing the injection, it is best to spread the pancreas apart on a sterilized cotton tip and inject near the lobes of the pancreas in a direction away from the vessels. This is to avoid injection associated vessel breakage and to insure the metastasis phenotype is not confounded by cells artificially leaking into the blood stream.
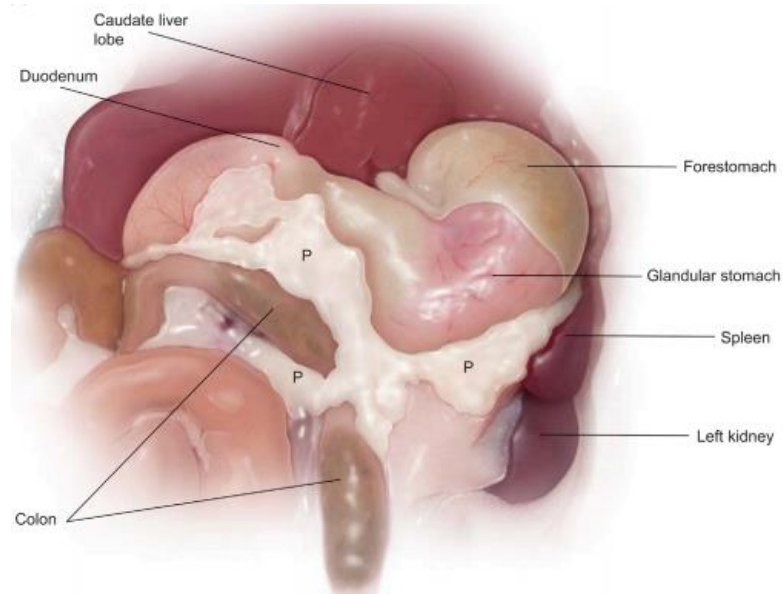
Figure 47. Illustration of mouse pancreas. Mouse pancreas structure is less defined and dispersed throughout compared to human pancreas, which has defined structures of head, body and tail. Illustration adapted from Denny Liggitt, Suzanne M. Dintzis, in Comparative Anatomy and Histology (Second Edition), 2018.

Experimentally, to gauge the potential cell leakage into the blood stream during injection, I used a cell line ASPC1 (from ATCC) derived from a metastatic site (ascites) and performed orthotopic injection (cells in FBS-free medium with 50% matrigel) while avoiding breaking the vessels in NSG mouse (n=3) [108]. After surgery, I waited 15 minutes for the matrigel to solidify in the pancreas. Afterward, I cauterized out and removed the injected cells alongside with tissues surrounding the injection site. Comparing to the animal that did not receive cauterization after injection (n=2), at the same time point post-transplantation, day 36, the cauterized animal has no visible metastasis, while the un-cauterized mouse has a massive degree of metastasis in the liver, lung and ascites confirmed by microscopy. This result serves as a validation on the injection technique, which is often overlooked, as well as provide confidence that the orthotopic transplantation model mimic physiological tumor dissemination originating from primary tumor.
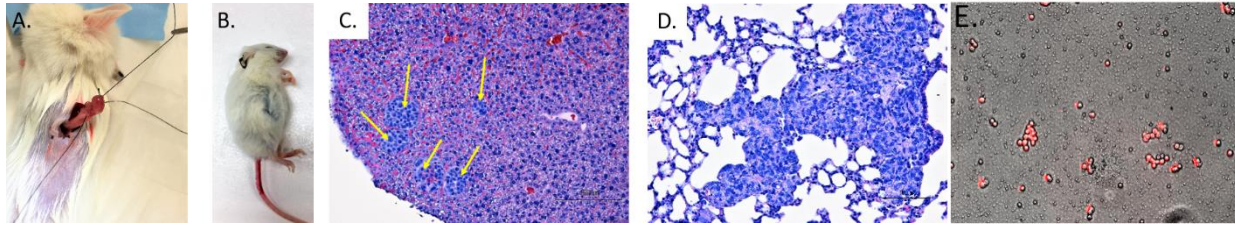
Figure 48. Orthotopic pancreas injection model experimental procedure positive and false-positive controls. A) Cells derived from human ascites were injected in the pancreas of mouse followed by removal of injected cells and part of the pancreas *via* ligation and cauterization. B) Mouse without removal post injection develop ascites, which have swollen abdomen. C) H&E staining of mouse liver for the groups without removal. Mouse without removal post injection developed liver metastasis indicated by yellow arrow. D) H&E staining of mouse lung. Mouse without removal post injection develop lung metastasis. E) Ascites collected from mouse without removal post injection (injected cells express RFP). Groups with post injection cauterization has no visible metastasis. Pictures are not shown.

Moreover, on the topic of choosing the right transplantation model to study metastasis, splenic injection (or tail vein injection, seeding) model is inadequate in recapitulating the natural process of liver metastasis and phenotype. Splenic injection, which is a common method used among researchers to enrich for cells with "metastasis" potential in the liver. In fact, cells enriched in the liver through splenic injection bypass many biological steps in the metastatic cascade. To name a few, the biological aspects of metastasization that are bypassed include stromal selection pressure and their fitness and vasculature distribution in primary tumor during expansion, selective pressure of extravasation, mechanic stress during the blood/lymphatic system, extravasation into distal organs and localization at metastatic friendly niche, and so on. An example below shows how liver metastasis lesions of a splenic seeding injection vs. an orthotopic injection model look like. The orthotopic liver metastasis matches clinical observations while the splenic injection does not. For all the aforementioned reasons, orthotopic transplantation model is best to study clonal dynamics during tumor expansion and dissemination.
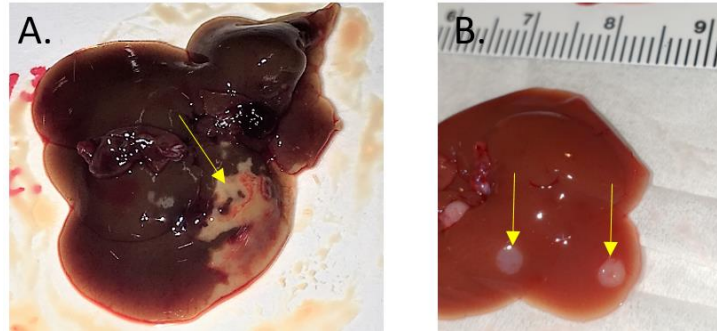
Figure 49. Splenic vs. pancreatic injection of PDAC cells and their liver dissemination phenotype comparison. A) Liver "metastasis" using splenic injection. Splenic injection bypass many natural mechanisms of dissemination. Cells injected is directly drained to the liver. As seen from the picture, many cells are mechanically trapped in the liver and does not recapitulate what is clinically observed. B) Liver metastasis through pancreatic injection. Metastasis indicated by yellow arrows. The metastasis phenotype obtained from orthotopic injection mimics the metastasis pattern observed clinically, where cells must originate from the primary tumor in the pancreas and undergo the metastasization cascade to disseminate.

Subcutaneous injection does not reproduce clinical metastatic phenotype

To have an idea of whether or not subcutaneous injection model is good for studying tumor dissemination, we did a comparison between subcutaneous vs. orthotopic transplantation of tumor cells in mouse and monitored their metastasis phenotype. The same amount of cells derived from PDAC patient liver metastasis tumor (PATC53) was injected subcutaneously, or orthotopically in the pancreas in individual NSG mouse. After allowing the tumors to grow for one month, there were no visible metastasis lesions by eye in the liver of subcutaneous injected animal; on the other hand, there were significant liver metastasis in the orthotopically transplanted mouse. Thus, orthotopic injection model is more suitable for us to study clonal dissemination pattern originated from pancreas, while subcutaneous transplantation model is not.
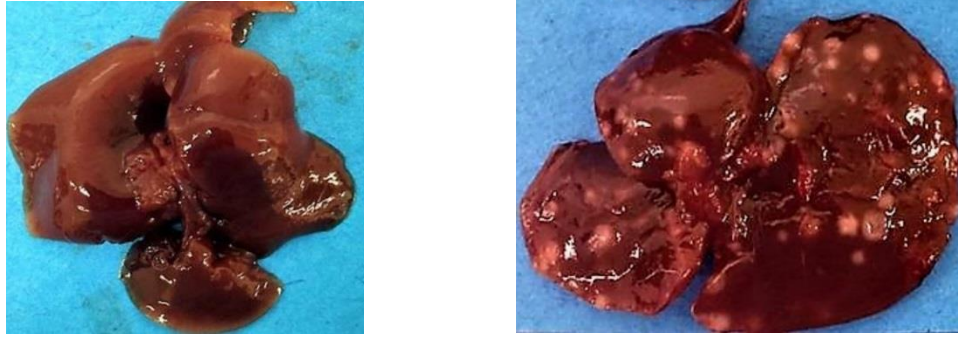
Figure 50. Picture of subcutaneous injection vs. orthotopic injection of human tumor cells in mouse and their metastasis phenotype in liver. Subcutaneous injected animal's liver after one month (left) versus orthotopic injection (pancreas) animal's liver after one month (right). Both "primary tumors" are comparable in size when the picture is taken. As seen in the picture, subcutaneous injection is not suited for studying tumor dissemination for pancreatic cancer.

Orthotopic PDAC transplantation model display inter-tumor heterogeneity in terms of metastatic latency and organotropism

To establish a representative model to study PDAC metastasis, I am using an orthotopic transplantation approach due to its relevant organ microenvironment and blood circulation anatomy in the body. ATCC cell lines and PDX cell lines derived from pancreatic lesions known for their metastatic potential were chosen. Specifically, for ATCC cell lines, PANC-1 and CAPAN-2 (cell lines derived from primary tumors with literature reports of metastatic potential) and AsPC1 (derived from a local invasion site) were chosen, with KP-1N (derived from a liver metastasis) included as a positive control[108]. PDX derived cell line models PATC124 and PATC69 were chosen based on patients relapse sites, where patients developed liver and lung metastasis after surgery respectively[61, 109]. To evaluate metastatic potential and latency, I transduced each cell line with a lentiviral vector encoding luciferase, and then injected orthotopically in immunocompromised animals (NGS). Subsequently, tumor progression is monitored by *in vivo* spectrum imaging (IVIS) to detect luciferase-positive tumor cells.
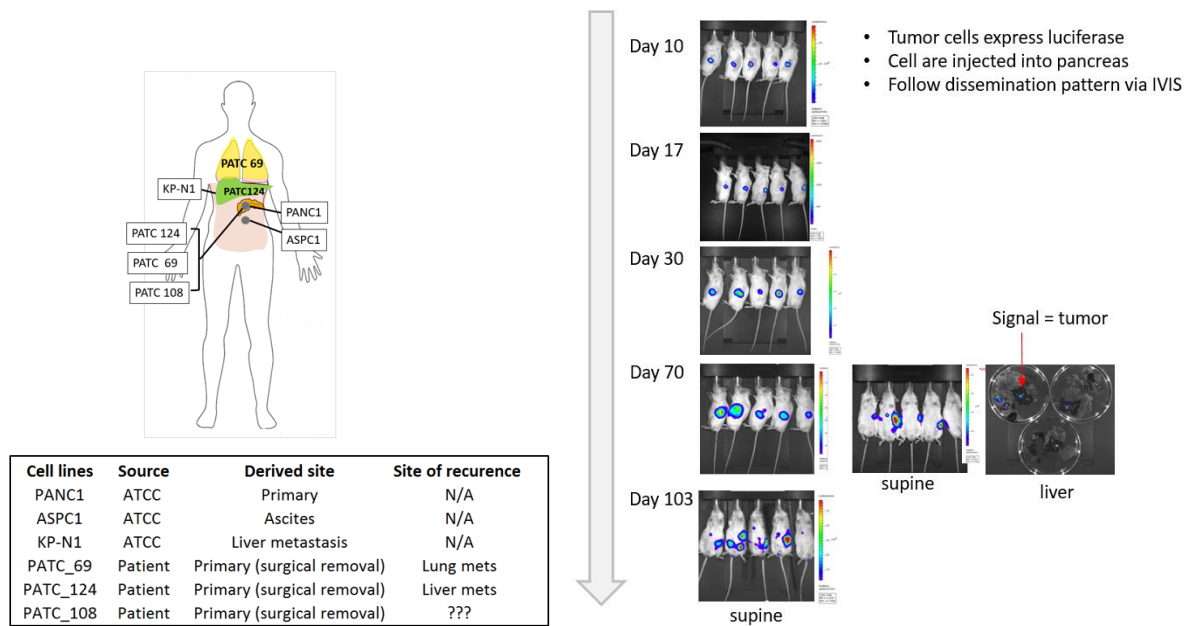
Figure 51. Summary of the PDAC cell lines and their respective sites of derivation and metastasis latency measured by IVIS. Information of PDAC cell lines chosen for validation of orthotopic transplantation model to study metastasis (left); An example of continuous IVIS monitoring of tumor cells (PATC124) expressing luciferase on the same cohort of orthotopically transplanted (pancreas) immune deficient mouse (right). In this experiment, metastasis in the liver is observed (detectable by IVIS) around day 70 post injection. The liver of the mouse was removed and put on petri dish to verify the signal was originated from mouse liver disseminated from primary tumor.

Concurrently, primary tumor, liver, and lung were collected at different time points and samples analyzed by H&E, immunohistochemistry (IHC) and immunofluorescence (IF) to detect for human cells, as well as to assess the extent and morphology of metastatic lesions. Through serial sectioning of FFPE blocks, we were able to determine the precise time points of micro-metastasis in the cohort for each specific cell line. Interestingly, through IF imaging staining for HLA, we observed clones from the same cell line displaying distinct levels of HLA expression co-localized adjacently; moreover, these clones display various degrees of morphology in terms of "invasive pattern" from defined to irregular "intrusive" boarder phenotype surrounded by hepatocytes (Figure 52).
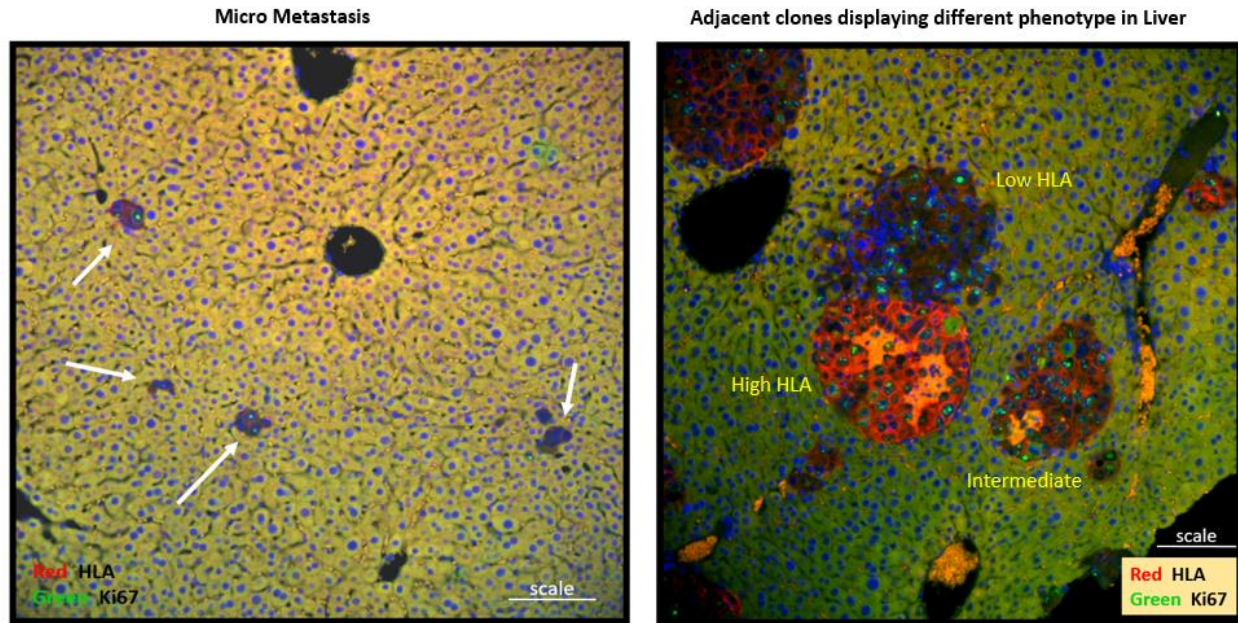
Figure 52. Representative immune-fluorescent image for micro-liver metastasis and intra-tumor HLA expression heterogeneity in the liver metastasis. Histology of PANC1 cells in liver stained with HLA (red), Ki67 (bright green) and DAPI (blue). The fluorescent background has been artificially increased to appreciate tissue structure. In the left graph, the white arrows indicate micro-metastasis in the liver. The right graph shows adjacent clones exhibit differential HLA expression and morphology. In this image, the top "clone" has a low HLA expression level and exhibit a more "irregular, invasive" pattern with peripheral cells more integrated into the liver. On the other hand, the adjacent clone expression high HLA (more red) has a more defined border. Scale = 100 micron.

Based on the combined analytical methods mentioned above, in our model, metastatic latency is highly variable among cell lines, ranging from 30 days (PANC1 and ASPC1) to 90 days (PATC69).
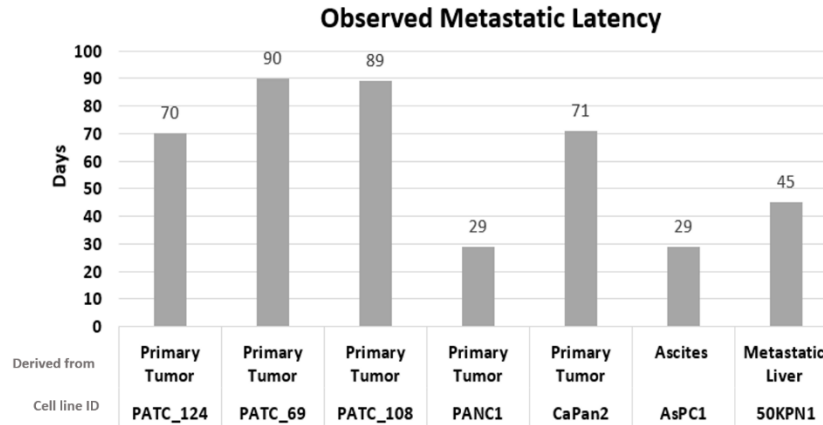
Figure 53. Summary of average metastasis latency amongst different tumor cell lines.

Last but not least, across tumors, we saw various degree of metastasis towards the liver or lung. We observed PATC124 metastasized primarily to the liver, while PATC69 metastasized primarily to the lung. Notably, in the PDX orthotopic transplantation model, the pattern of organotropism was consistent with the site of disease recurrence observed form the patient. In other word, the model could recapitulate clinical organotropic phenotype.

These inter- and intra- cell line heterogeneous phenotypes underscore the relevance of the research, wherein barcode tracking of distinctive clonal behaviors during tumor progression and metastasization may reveal unique molecular profiles relevant to their respective phenotypes. Most importantly, from the above results, I conclude that an orthotopic transplantation model in the pancreas is a valid model to study primary tumor expansion and capture distinct tumor dissemination patterns.
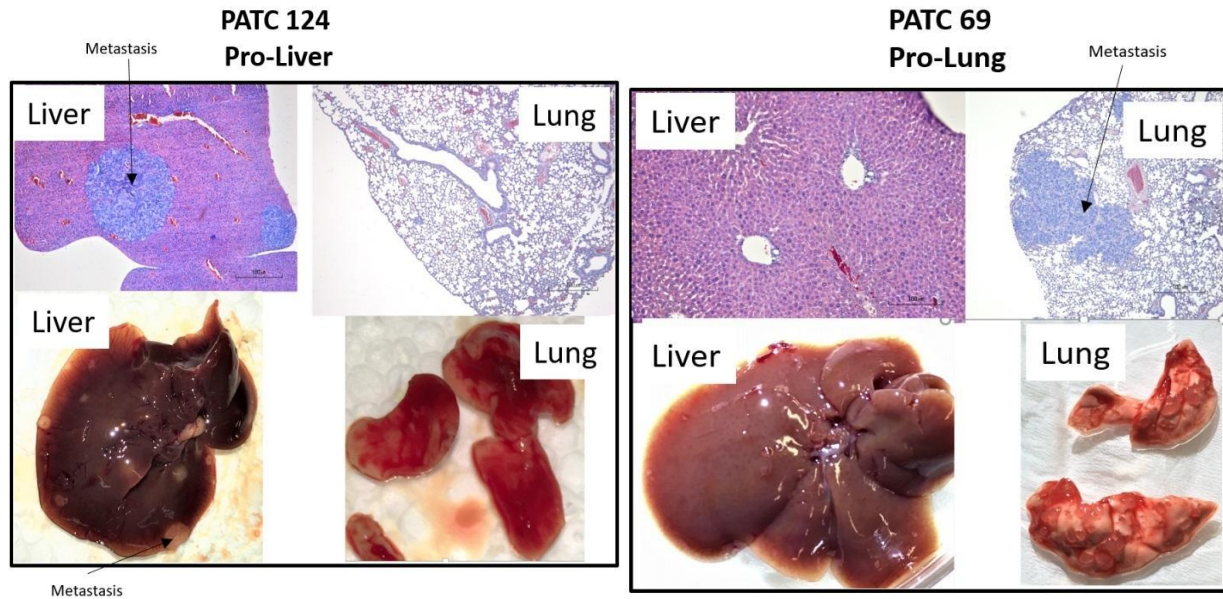
Figure 54. CRT tumor cell lines recapitulate patient metastasis site during relapse. Liver and lung after 2 months post orthotopic transplantation of PDAC PDX derived tumor cells. H&E staining of liver and lung (top) and pictures of the entire organ (bottom). PATC124 (left) developed mainly liver metastasis, while PATC69 (right) mainly developed lung metastasis. The organotropic phenotype is consistent with individual patient site of relapse.

Establishing long-term self-renewal barcoded cell lines to create orthotopic CRTs

To create a cohort of mouse bearing clonally identical human tumors in the pancreas, we followed a pre-established lentiviral barcoding workflow discussed in previous chapter. In brief, we took an early passage PDX derived cell line from patient's primary tumor (PDAC) and infected the cells with a pool of highly complex barcode library (~10M unique variations) at a very low molarity of infection (0.1-0.2 M.O.I) to ensure each cell ended up with one unique barcode integrant. Once the barcode has integrated in a cell, the barcode would be continuously passed on to its progenies, and hence allowing us to track cell lineages. The barcoded cell culture was then treated with antibiotics to select for positive barcode integrated cells (the barcoded cells are puromycin resistant and express RFP). The barcoded cells were then passaged at a 1:2 ratio *in vitro* for ~21 passages. During the *in vitro* passages, the cell culture would naturally enrich in cell lineages that are endowed with long-term self-renewal properties while most of the other lineages exhaust (~99%), the culture would eventually achieve equilibrium and stabilization in clonal

composition. Once the culture achieves stabilization, it could be expanded and transplanted orthotopically in a cohort of animals.
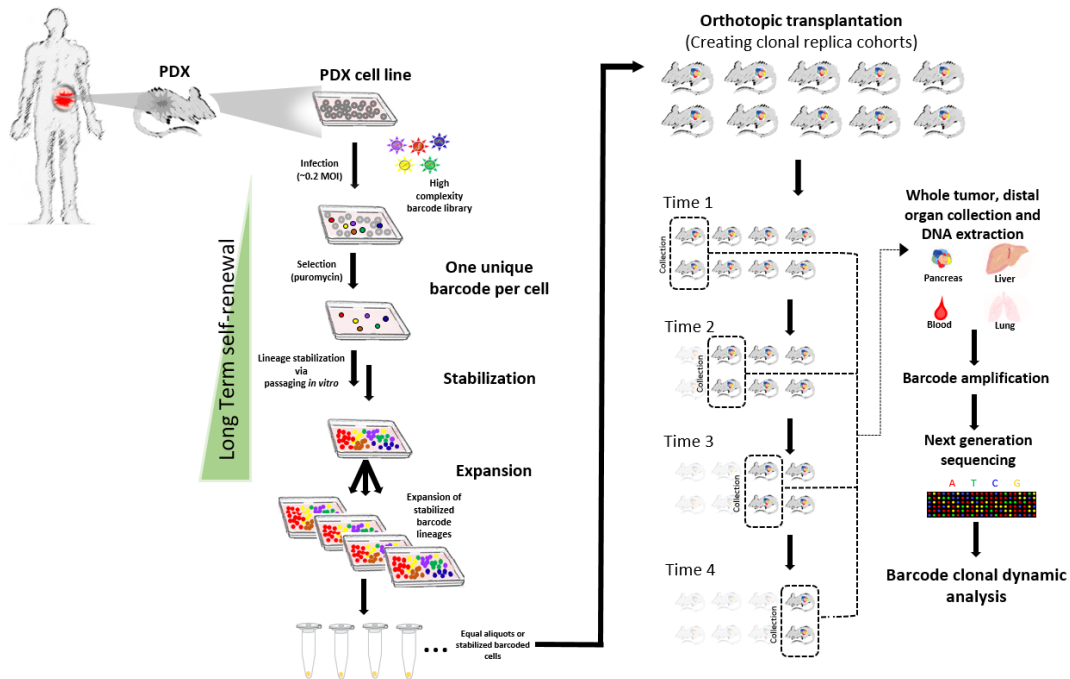
Figure 55. Schematics of the orthotopic clonal replica tumors workflow. Primary tumor cell lines were established from patient's primary tumors. Early passage of each cell line was infected with high complexity of barcodes at a low MOI to ensure one unique barcode per cell. Barcoded cells were stabilized through serial passages *in vitro* at a 1:2 passaging ratio for natural enrichment of long-term self-renewal clonal lineages. Stabilized cells were expanded and injected orthotopically into cohorts of mouse. A pair of mice with tumor were sacrificed at different time points, where barcode composition and abundance where measured via NGS for lineage tracing of clones over time.
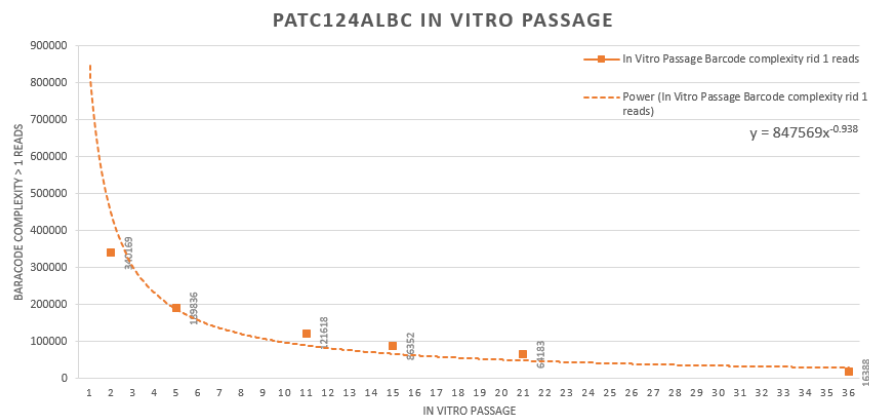
Figure 56. Barcode complexity during "stabilization". An example of clonal lineage complexity drop during *in vitro* culturing and enrichment of clonal lineages that are able to sustain culture.

Barcode complexity drops over time, which is contributed by lineage exhaustion and natural enrichment of clones that are able to sustain culture / tumor long term. This observation was consistent with what was observed and discussed in chapter one, where barcoded culture will come to a natural equilibrium, enriching in tumor-maintaining lineages.

Lineages in post stabilized culture continue to maintain tumor growth long-term *in vivo*

From the previous chapter, we have concluded empirically that post stabilization, lineages that remain in the culture has the ability for self-renewal and maintain the tumors long-term[109]. As a short summary of the experiment to determine the self-renewal nature of stabilized cell lineages, starting from passage 1 after barcoding, the cells were passaged in parallel by serial passage *in vitro* and serial transplantation *in vivo*. Over time, the lineages in both experimental arms converge with identical set of barcode lineages representing majority of the tumor mass (*in vivo*, after 8 months) and cell culture (*in vitro*, after 21 passages).

Since I am using the stabilized culture to generate a longitudinal CRT model to study clonal dynamics, it would be important to know whether the lineages in the clonal replica tumors post transplantation are indeed relevant in terms of sustaining tumor growth. In order to evaluate whether or not the cell lineages use in the model would indeed engraft and continue to maintain tumor long-term, we performed *in vivo* serial transplantation of the tumor (post stabilization) for three generations (F1 to F3 for 6 months, subcutaneous transplantation) and evaluated the barcoded lineages present in F3. By analyzing the lineages in F3, we could validate whether or not the lineages post stabilization has the capacity to continue to sustain tumor.

First, comparing the subcutaneous tumor F1 to F3, the tumors share ~1,606 common lineages, where those common lineages makes up almost the entire tumor mass in both F1 and F3 (97.48% and 99.99% respectively). This informs us the lineages used in the model can continuously engraft and form tumor *in vivo*, and that they are indeed long-terms self-renewal lineages.

Comparing the serial transplanted tumor at F3 to orthotopically transplanted tumors at two independent time points (day 50 and day 114), we observed ~1,300 common barcode lineages shared amongst F3 and orthotopic tumors (1327 and 1390 common barcodes respectively). The common barcodes make up almost the entire tumor mass of F3 (>99.8%), day 50 orthotopic tumor (87.1%), and day 114 orthotopic tumor (96.77%). This data indicates the orthotopic tumors created through the barcode stabilization process are made up of almost entirely tumor sustaining cell lineages, and that they are endowed with long-term self-renewal properties.
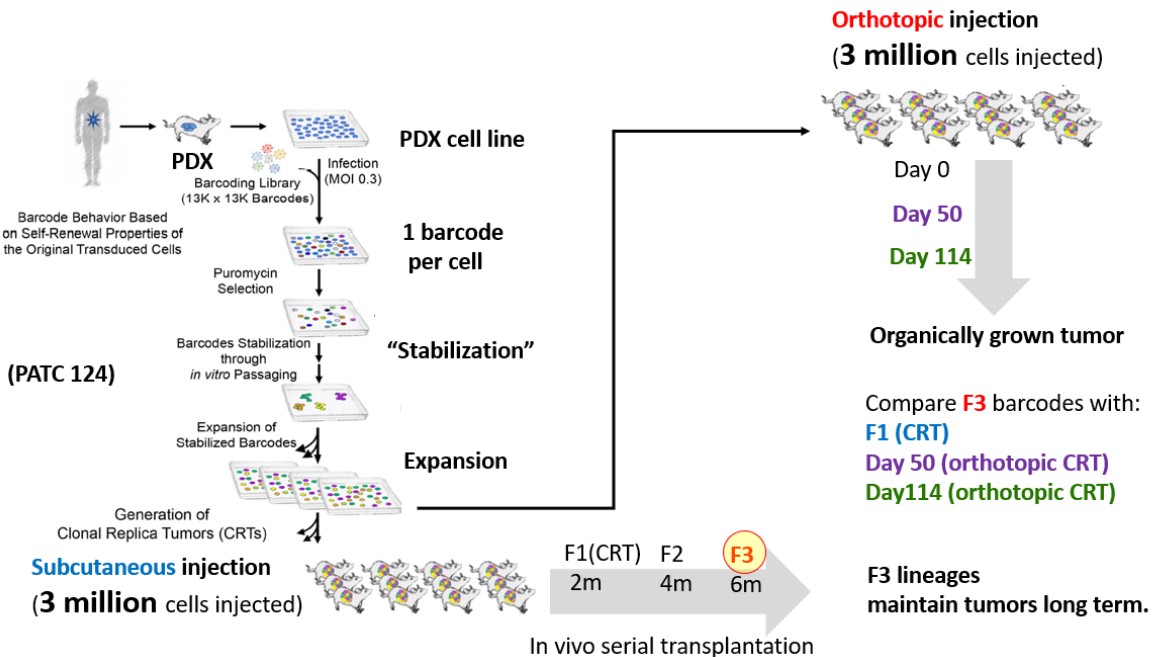
Figure 57. Experimental schematics to measure post-stabilization culture used to create CRTs for studying clonal dynamics are composed of lineages with long-term self-renewal (LTSR) properties. Post-stabilized barcoded cell culture was expanded and subjected to *in vivo* serial transplantation for six months (F1 to F3) to functionally measure lineages with LTSR; the same population of stabilized cell were injected in cohorts of mouse orthotopically in parallel. Orthotopic tumors were collected at day 50 and day 114. By comparing the barcode composition of the orthotopic tumor to the serially transplanted tumor F3 informs us of the LTSR nature of lineages in orthotopic tumors.
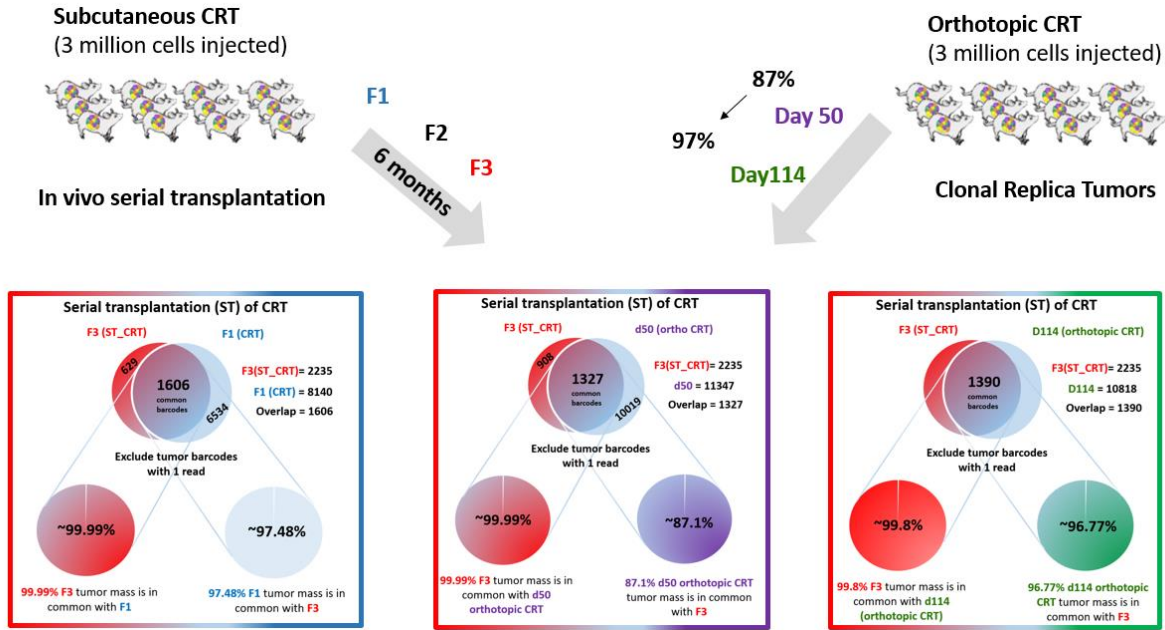
Figure 58. Clonal makeup of orthotopic CRTs vs. serial transplanted tumors to gauge the LTSR nature of lineages in orthotopic CRTs. Number of common barcode lineages and their relative tumor representation shared amongst CRTs (F1, orthotopic tumor, which are tumors generated right after *in vitro* stabilization) and serially transplanted tumor, F3. The barcode cutoff criteria used in this analysis is more "inclusive", where barcodes with reads above 2 are included. In the most left graph, it showed the "F1" tumor and "F3" tumor shared ~1,606 common barcodes, and those common barcodes makes up essentially the entire tumor mass of both "F1" and "F3" tumor. In the middle and right graph, "F3" tumor (informs of LTSR lineages) shared ~1,300 to 1,400 clones with orthotopically generated tumors. Furthermore, amongst those common clones, they make up essentially the entire tumor mass in "F3" and orthotopic tumors (day 50 = 97.1% and day 114 = 96.77%). This indicates that orthotopic tumors (CRTs) are made up of essentially all lineages endowed with LTSR which can sustain tumors long-term.

Barcoded early passage PDX derived primary tumor cell line is heterogeneously diverse

To assess the biological heterogeneity present in the primary tumor cells lines used in the longitudinal clonal tracing model, we submitted single cell sequencing of post-barcoded cells (~5,000 cells) and looked at their genomic copy number variance profile. This is to ensure that the barcoded culture used for clonal tracking is not a homogeneous culture. As shown in Figure 59 below, the barcoded culture (at passage 4) indeed consists of multiple CNV diverse clones. Furthermore, to assess whether the "stabilization" (*in vitro* passaging of barcoded cells at a 1:2

ratio) process is relatively neutral in preserving the biological heterogeneity, concurringly, I sequenced ~5,000 single-cells from a late passage (at passage18) and look at the overall copy number change by comparing to passage 4. From the analysis, the overall CNV profile from early and late passage are very similar, aside from small regions on chromosome 4 and 12, which tells us there are some degree of clonal enrichment, however, the "stabilization" process over all is relatively neutral in disturbing the CNV profiles of the tumor culture.
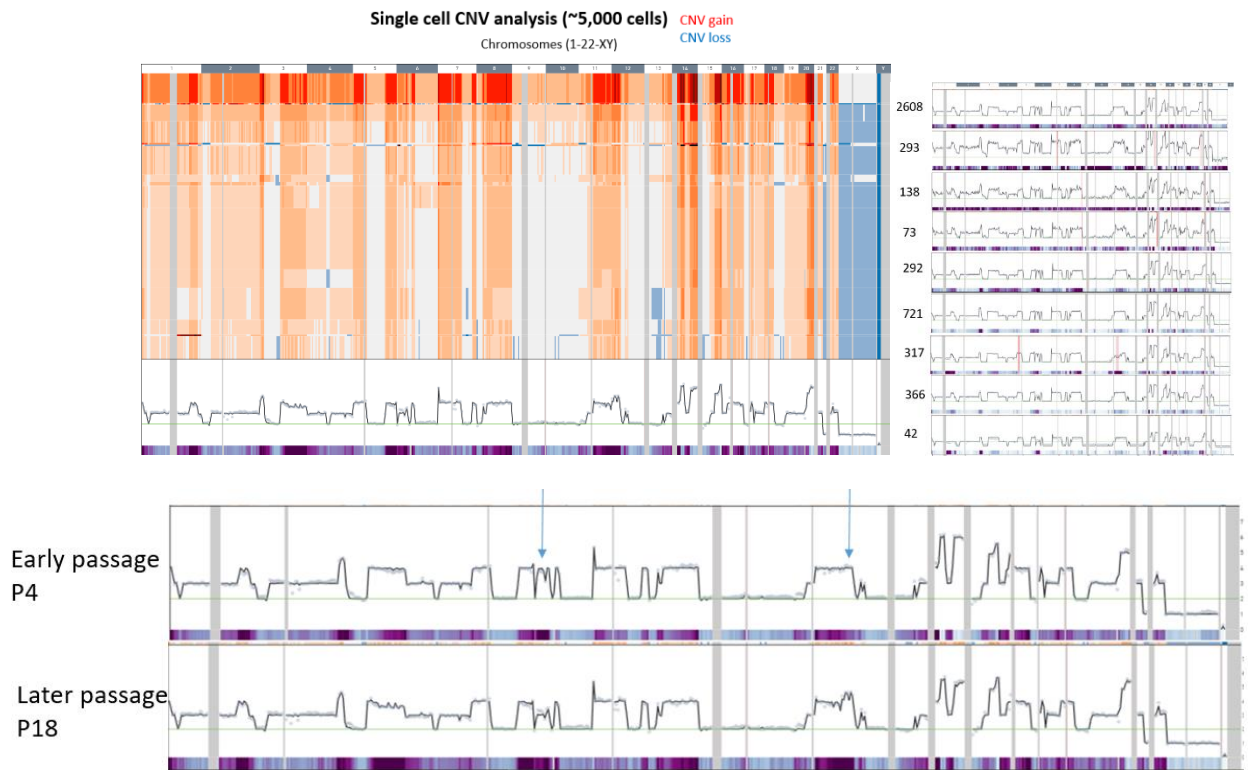
Figure 59. Early passage vs. late passage single cell CNV profile comparison during stabilizaton. Single-cell CNV data from early passage of barcode cells (passaged 1:2 each passage) derived from PDAC PDX. The top indicating chromosome number and the color red indicates gain in copy number while blue represents loss of copy number. The bottom profile represents the average CNV profile of all the cells (top left graph). A higher resolution view of individual clusters of CNV profile, the number on the left indicates cell number associated with specific profile (top right graph). Comparison of the average CNV profile between early passage and late passage barcoded cells (bottom graph). The data informs us that the stabilization process is relatively neutral and maintains tumor's biological heterogeneity.

The barcoding process is neutral to barcoded cells in terms of cell proliferation

To gauge the effects of the barcoding process on the barcode-inserted cells in terms of expansion fitness relative to the overall cell culture, we performed the barcoding process on an aggressive cell line (KPC cells) at 0.1 molarity of infection without antibiotic selection and passaged the cell culture 2:1 for eighteen (18) passages. At passage 2, 5, 11, and 18, we measured the relative percentage of RFP positive cells (barcoded cells express RFP) amongst overall cell culture both in terms of total cell population and live cell population. We found the percentage of relative passages to be 4%, 3.9%, 3.1%, 2.7% of total live cell population (see table below). Considering that the loss of florescent protein expression is a common phenomenon, and the fact the relative RFP+ population to be relatively stable (3-4%) over 18 *in vitro* passages, we could conclude that a gain of fitness due to barcoding was not observed. We could be relatively confident that the barcoding process, overall, is neutral to the fitness of the cells. Obviously, this experiment was only aimed at assessing whether or not, in general, barcoding have an effect on drastic fitness change relative to the entire population and not on a genomic/gene level. To obtain the precise location of insertion of barcodes, primers could be designed to extend outwards from the inserted vectors and the amplified regions mapped to genome, which is an ongoing effort in the lab.
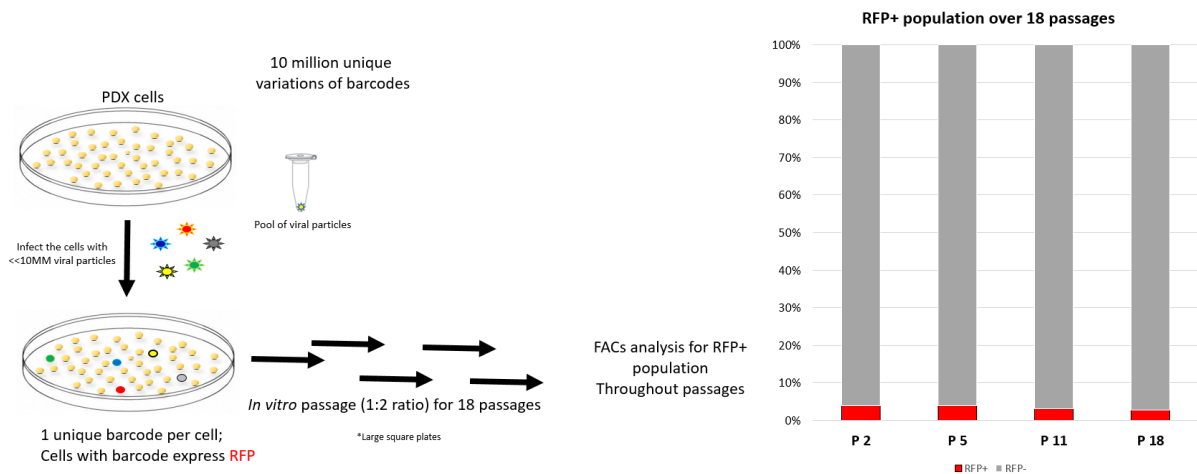
Figure 60. Experimental setup to gauge the effect of barcoding on fitness. The barcoded cells were co-cultured with the parental population over 18 passages (1:2 ratio, left); FACS analysis of barcoded cell population over 18 passages (RFP+ population, right). The barcoded population stayed relatively stable at 3-4% over 18 passages. Indicating no drastic gain of fitness.

	Negative control % RFP+ Live cells	Barcoded cells %RFP+ total cells	Normalize to passage 2	Barcoded cells %RFP+ Live cells	Normalize to passage 2
Passage 2	0	2.8%	100%	4.0%	100%
Passage 5	0	2.5%	89%	3.9%	98%
Passage 11	0	1.7%	61%	3.1%	78%
Passage 18	0	1.5%	54%	2.7%	68%

* total observed events from FACS >20,000.
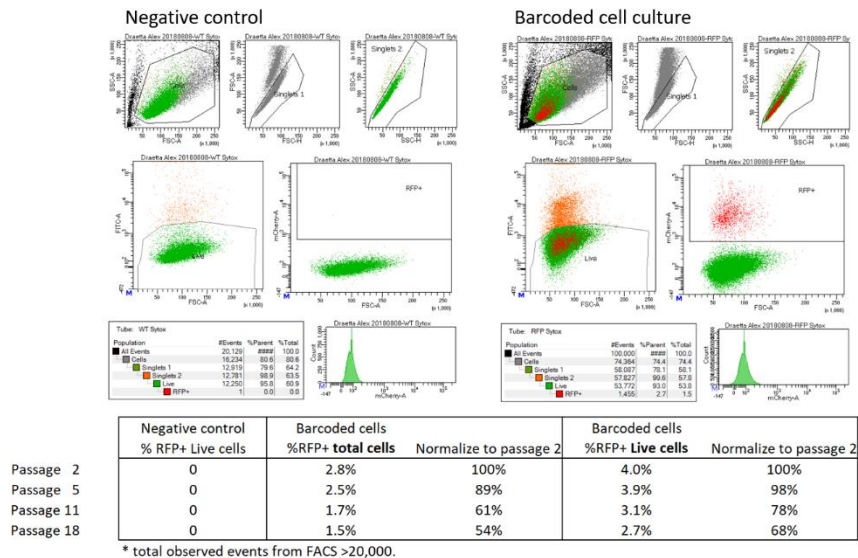
Figure 61. Barcoded cells' relative fitness in cell culture by FACS. Examples of FACS of infection-null cell line (top left) and barcoded cells (top right). The barcoded cells express RFP, where non-barcoded cells do not express RFP. Bottom table is the % RFP positive percentages in the context of total population and live cells throughout passages.

Clonal growth in subcutaneous injection vs. orthotopic injection are similar

Subcutaneous injection of tumor cell lines is commonly used in the research community as *in vivo* studie[110]. A subcutaneous transplantation model fulfils criteria of cell engraftment and growth in an *in vivo* setting with relatively simple procedures compared to orthotopic transplantation. It is also commonly used as a model to study for drug response due to the ease of monitoring tumor size in response to therapy.

In our model, we ask the question, "for the PDAC cell lines consists of long-term self-renewal lineages, how different, or, similar would these lineages behave in an orthotopic setting vs. a subcutaneous setting?" This information would help us understand the nature of these lineages under study and have a deeper understanding of our model.

The experiments consist of the stabilized CRT cell lines from PATC124 after barcoding. Cells were split into two groups, one group for orthotopic pancreas injection (~3M cells, n=2) and the other for subcutaneous injection (~20M cells, n=3). The tumors were collected after 1 month and processed for barcode readout. In brief, DNA from tumors were extracted with spiked-in scales of known cell count with unique barcodes ranging from 50,000 cells to 5 cells; then, barcodes were amplified via PCR with a common set of primers and sequenced via NGS. Each barcode was then normalized to "percent representation per sample" by dividing individual barcode counts by total read count of the particular sample. The barcode representation at this point can be interpreted as "clonal representation in tumor"; furthermore, the barcode representation was averaged between samples in each injection condition cohort before comparison.

By looking at the barcodes composing of the top 99.5% of the tumors, the barcode complexity is 1,402 barcodes in the orthotopic setting and 9,233 in the subcutaneous setting. The entire set of barcodes in the orthotopic tumors could be found in the subcutaneous tumor as

shown in the Vann diagram. This was likely due to a smaller sub-set of cells were transplanted orthotopically (3M vs. 20M). Comparing the relative tumor mass representation of the common sets of barcodes (1,402 barcodes) they make up 100% of the orthotopic and 98.7% of the subcutaneous tumor mass respectively. This indicates that the majority of the tumor mass from different sites were made up by the same set of clones.

By plotting the XY scatter plot of each barcode's representation of orthotopic vs. subcutaneous tumors, we could see each clone's relative representation in the tumor were highly correlated (Pearson's =98.34%). The data suggests that in the CRT model, with the stabilized cells containing long-term self-renewal lineages, the clonal engraftment properties from orthotopic injection and subcutaneous injection are highly similarity.

Since the goal is to create "clonal identical" tumors in different animals, this experimental result suggests that in our model, the CRT clonal lineages are indeed robust tumor maintaining clonal lineages, and that the cell intrinsic engraftment and proliferation capability are strongly preserved despite different transplantation sites *in vivo.* However, this does not indicate that the orthotopic and the subcutaneous environments are the same and would not influence clonal behavior over time (i.e. designing serial transplantation of orthotopic tumor vs. subcutaneous tumors would answer this question). All in all, we can conclude that in terms of lineage intrinsic engraftment and proliferation capability, these stabilized lineages have similar behavior under subcutaneous and orthotopic setting.
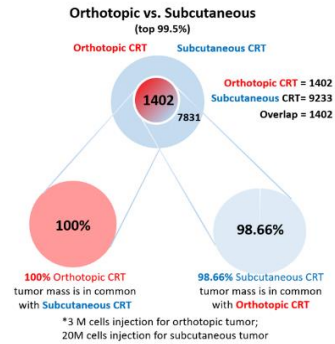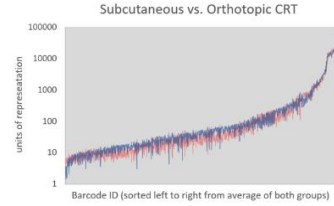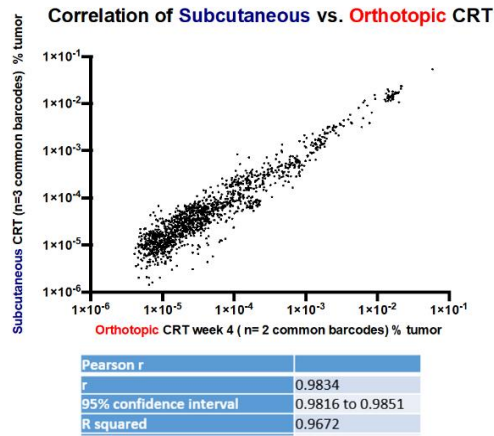
Figure 62. Subcutaneous vs. orthotopic tumor clonal composition and correlation. Comparing the common barcode lineages and their relative abundance correlation between orthotopic and subcutaneous transplanted tumors using stabilized clonal tracking cell lines. X-Y scatter plot of common barcodes and their relative abundance (%) in orthotopic and subcutaneously injected tumor. Each dot representing a unique barcode lineage (left); Converting each lineage into cell number through a spiked in scale, and then overlaying of each barcode lineage's representation ranked from low to high from left to right. Blue represents subcutaneous tumor lineages, while red represents orthotopic tumor lineages (top right); Common barcode lineage and their relative tumor representation between subcutaneous vs. orthotopic tumors (bottom right).

Orthotopic CRTs are maintained and progress similarly over time

From previous sections, we now have a better understanding that the tumor cells transplanted orthotopically are derived from long-term self-renewal cells and are biologically diverse; in addition, these lineages will continue to sustain tumor growth *in vivo* long term.

To briefly recap on the experimental design, a cohort of animals were injected with ~3M stabilized barcoded cells orthotopically. Two tumors were collected at week4, week6, week10, and week 12 respectively and their lineage composition determined by barcode read out through NGS.
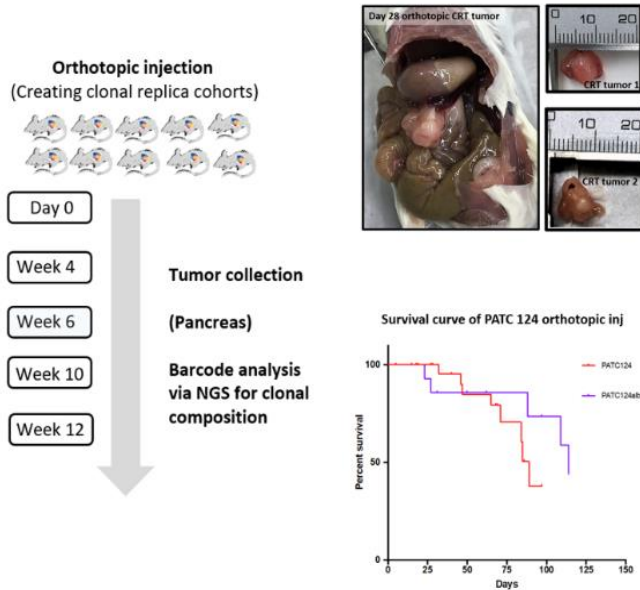
Figure 63. Experimental design of longitudinal clonal tracking model. Cohorts of animals orthotopically transplanted with stabilized barcoded cell lines were established. Tumors were collected at various time points post injection in pairs (left); Pictures of the tumors in parallel mouse at 4 weeks were similar in size (top right); Two independently barcoded (PATC124) orthotopic experimental cohorts and their survival curve (bottom right).
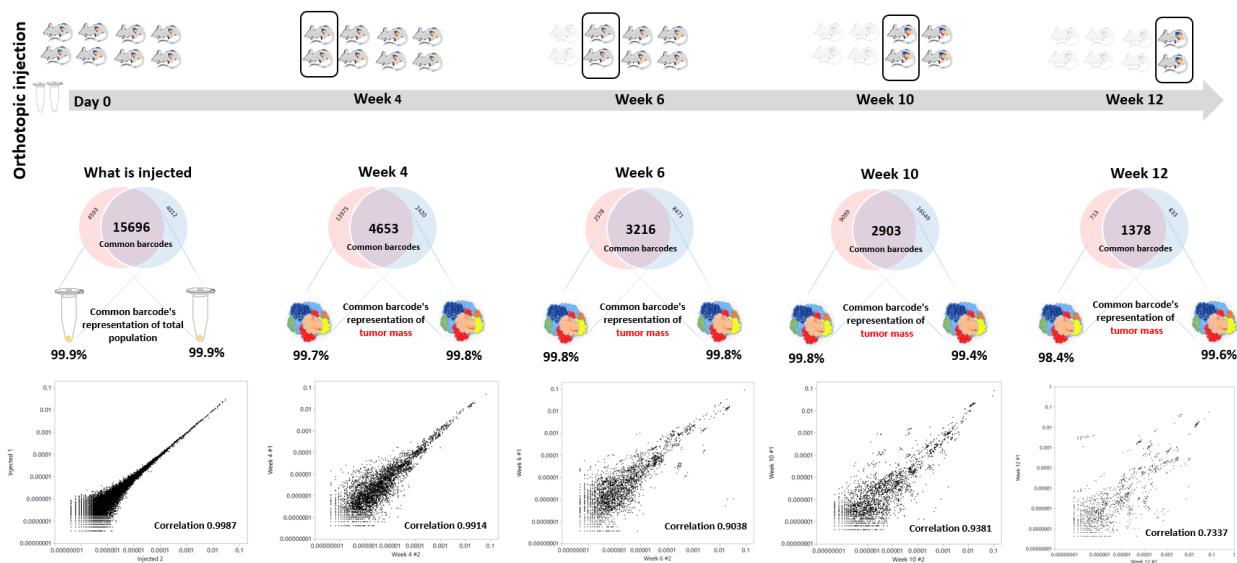
Figure 64. Orthotopic CRT clonal correlations over time. Numbers of animals collected per time point (Top); Barcodes in common and shared between parallel animals at a particular time point and their relative culture/tumor representation (%) (middle); X-Y scatter plot (log scale) of individual barcodes and their (%) correlation value (bottom) in parallel mouse at the same time

point. The CRTs stayed relatively similar and maintaining high correlation until week 12, where we observe a natural divergence amongst CRT animals. The degree of spread and divergence reflects the unpredictability of tumor expansion, considering we injected the same cells in the beginning.

Looking at the barcode complexity from two of the many replica vials of cell aliquots for injection (each animal received one vial of ~3M cells during injection), the two vials of "what is injected" shared around 15,696 barcodes. Those common barcodes represent almost the entire content (>99.9%) of cells in each vial. Furthermore, each of the barcode's abundance were highly correlated (Pearson correlation 0.9987). This indicates the tumor cells transplanted into each animal were basically clonally identical in both composition as well as relative abundance.

Following established orthotopic tumors and comparing two tumors at week 4, we see there are ~4,653 barcodes in common. A reduction of barcodes comparing to "what is injected" indicating not all lineages could engraft and expand under this specific experimental setting. The 4,653 common barcodes represent majority of the tumor mass in both tumors (99.7% and 99.8%) and have high correlation (0.9914). This indicates that these two tumors are essentially identical in their clonal composition and individual lineage's relative abundance is essentially the same. In other words, when a barcode lineage is highly represented in one tumor, it is also highly represented in the other tumor. Therefore, this satisfy the criteria of establishing clonal replica tumors.

Furthermore, analyzing lineage composition and relative abundance of tumors collected at week 6 (3,216 common clones, common clones makes up 99.8% of tumor mass, correlation of barcodes in two tumors 0.9038), week 10 (2,903 common clones, common clones makes up 99.4%-99.8% of tumor mass, correlation of barcodes in two tumors 0.9381) and week 12 (1,378 common clones, common clones makes up 98.4%-99.6% of tumor mass, correlation of barcodes in two tumors 0.7337), the data indicates that the CRTs were maintain throughout their progression. However, interestingly, throughout the tumor progression, we see natural clonal

reduction from initial ~4,653 barcodes at week4 down to ~1,378 barcodes at week 12, indicating clonal competition and clonal dropout during tumor expansion. Furthermore, we observed the degree of nature divergence of tumor lineage expansion. This was indicated by the drop of correlation from >90% to ~70% from week 10 to week 12 and by observing the increase in spread of clones in the X-Y scatter plot over time. In other words, in our experimental model, the CRT tumors progress very similarity over time, however, each clonal lineage also progressively diverge over time. This observation showcases the unproductiveness of cancer growth and the possible extrinsic factors shaping tumor progression.

Longitudinal clonal progression between CRTs are similar over time through statistical robust test

To further validate CRTs for longitudinal lineage tracing, we applied a series of statistical analysis to test for aspects of the CRTs in terms of their similarities over time: (This section was performed by Delia Wang).

Statistics 1 (*T1*) measures the sum of common barcodes shared amongst the two subject (tumors) at the same time point. For example, samples A and B. Therefore, the larger the value of *T1* the more A and B are similar. This informs us of the lineage composition aspect of CRTs.

$$T_1 = \frac{n_{A \cap B}}{n_{A \cup B}}$$

Statistics 2 (*T2*) measures the Pearson correlation coefficient between barcode populations from sample A and B (P_A and P_B). The closer *T2* is close to 1, the more similar the barcode populations are between A and B. This informs us about the relative abundance similarity between CRT animals.

Statistics 3 (*T3*) measures the Euclidean distance between P_A and P_B. The closer the value of *T3* is to zero (0), the more similar A and B are. This informs us of how closely related CRTs' clonal differences are as a whole at each given time point.

$$T_3 = ||p_X - p_Y||_2 \; (d_{A,B} = \sqrt{\sum_{i=1}^{n=A\cup B}(p_A - p_B)^2})$$

For the above statistical tests, we used Bootstrap resampling approach to perform similarity test (1,000X). The criteria we used to guide the simulation's parameter was based on the empirical obtained differences from the two technical replicas vials of cells of "what is injected", which is the most stringent criteria we could empirically achieve.

In Figure 65 below, the curves represent the simulated expected true distribution value based on the stringent criteria of similarity and the two CRT samples/tumors, and the dotted line representing the observed value at day 0 ("what in injected"), week 4, week 6, week 10 and week12 from top to bottom. We see that before week 12, for all three statistics, the observed value either falls within or close to the value that indicates similarities between the two samples. This means we cannot say these CRTs, before week 12, are not similar under these simulation criteria.
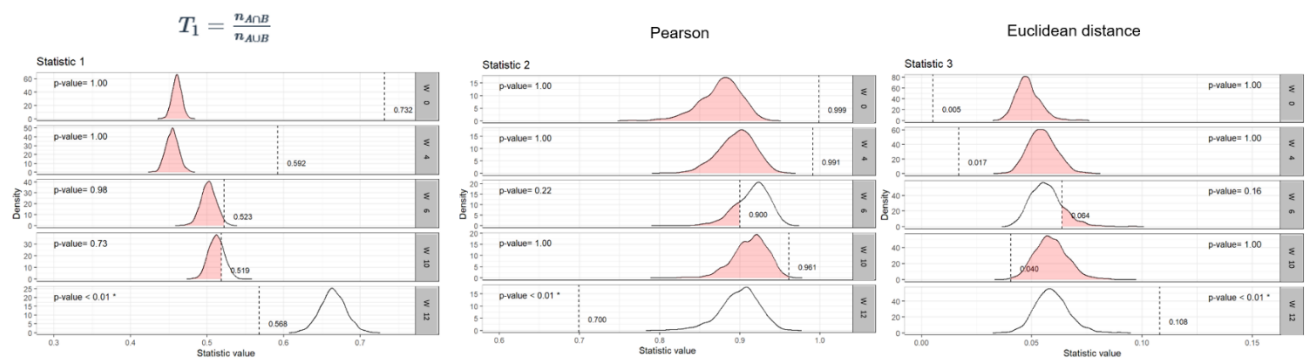
Figure 65. Statistical robust test for paired orthotopic tumors over time. From left to right, the statistical tests on simulated data: common barcodes shared between tumors (left), the Pearson's

correlation of each of the barcode abundance (middle), and finally, the Euclidean distance between the individual tumor's barcode analyzed as a whole (right). The distribution curve in each panel at each time points represents the simulated "true" value (1,000 times) by bootstrap method. The dotted line in each panel represents the observed value. If the experimental (observed) value falls within the "true" distribution or above (for "common" and "Pearson", the higher the value, the more similar) or below (for "Euclidean distance", the smaller the distance, the more similar), we do not have enough evidence to say these samples are different.

Interestingly, as previously described, although CRT tumors as a whole behaves similarly over time, we observe a natural divergence of clonal expansion over time as well. Such as, under this stringent criterion of the similarity test, week 12's CRTs, as a whole, are not "similar". Taken into consideration that we injected essentially clonally identical tumor cells in every animal, this observation captured the degree of the unpredictability and randomness nature of tumor progression.

Last but not least, keep in mind that the simulation is conducted on criteria of similarity guided by a very stringent set of data (two technical replica of aliquot vials of cells post stabilization). Therefore, even though week12's tumors as a whole reject the null hypothesis and are shown to be "not similar" in this set of robust test, they are still considered to be CRTs in nature given the common barcodes shared (n=1,378) representing 98.4%-99.6% of the entire tumor mass. On this note, from the analysis below, where the difference of the relative representation (in %) of each unique barcode between the two CRT samples/tumors are calculated and graphed out. We can see the vast majority of the barcode lineages are similar (that makes up >99.9 of the tumor mass), while only part of the barcoded lineages diverges gradually over time.
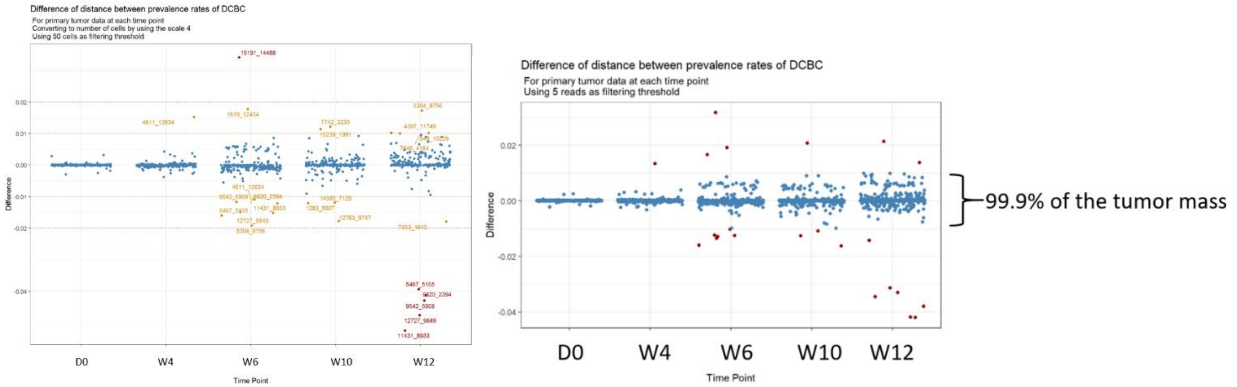
Figure 66. The difference of distance of barcodes in paired CRT mouse by time. Each dot representing a barcode and the Y-axis indicates the % difference of the same barcode in two different CRT tumors collected at the same time point. The data are analyzed based on quantitative analysis of the barcodes (cell units). The difference between each clone amongst CRTs at the same time point from day 0 to week 12 was calculated. Left graph was analyzed with 50 cells as cutoff, whereas the right graph used 5 cells as cutoff. We could see the majority of the barcodes remain very similar in abundance in CRT animals over time. Analysis by Delia Wang.
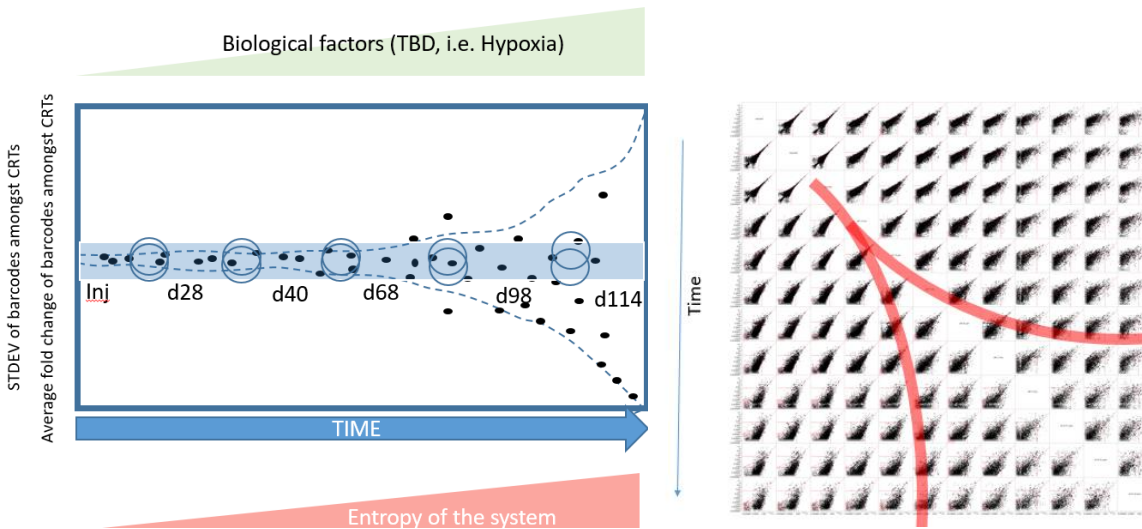
Figure 67. Representation of natural clonal divergence in unperturbed tumor expansion. Graphical representation of CRTs are maintained over time, with vast majority of the clonal lineages behaving similarly in replica tumors, while a subset of clones becomes more divergent and behaves differently starting around week10 post injection (left); X-Y scatter plot of the second cohorts of all the orthotopic CRT's primary tumor (each tiny dot is a unique barcode, value plotted in % representation, in log scale) from day 0 to day 114, from top to bottom and left to right orientation, where a general trend of decrease in clonal abundance correlation (right).

Orthotopic CRTs are maintained over time and the model is reproducible

To verify the orthotopic CRT model is reproducible, using the same PDAC cell lines (PATC124), we independently performed the entire process from barcoding, stabilization, to independent transplantation in another cohort of animals and collected tumors at various time points. This time, we extended the time points to week 16, which was the limit before animal subjects die of tumor burden.

For each paired tumors collected at each time point, we performed the same statistic robust test described in the previous sections. The conclusions were the same: orthotopic CRTs were well maintained over time as a whole. Natural tumor divergence over time was also observed with the vast majority of the barcode lineages that makes up the essentially the entire tumor mass behaves similarly over time. All in all, we are confident of the robustness and reproducibility of the orthotopic CRT process and outcome. Since these CRTs are maintained over time, we could exploit this property to track longitudinal clonal dynamics in parallel CRT animals.
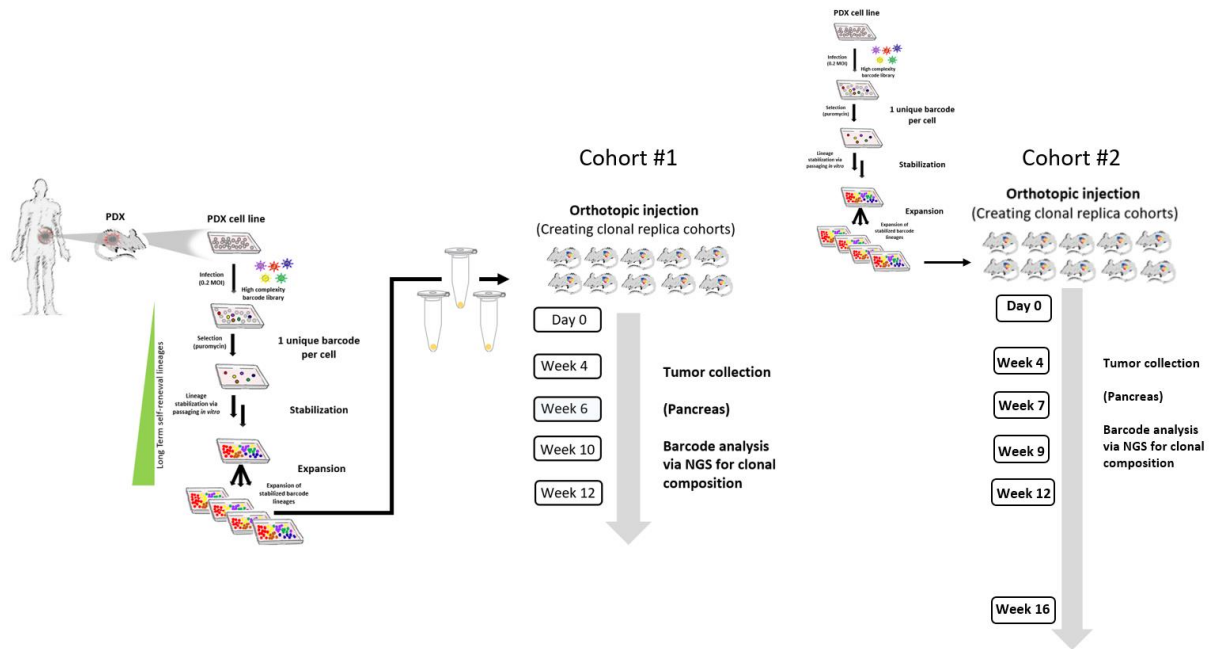
Figure 68. Experimental workflow schematics of two independent cohorts of animals carrying orthotopic CRTs. Using the same cell line derived from the same patient, through independent barcoding and stabilization, we created two independent cohorts of orthotopic CRTs. In the first cohort, the longitudinal study spans 12 weeks, while in the second cohort, the study lasts 16 weeks. The mice subjects bearing CRTs die around 12 - 16 weeks of metastasis throughout the body in general.
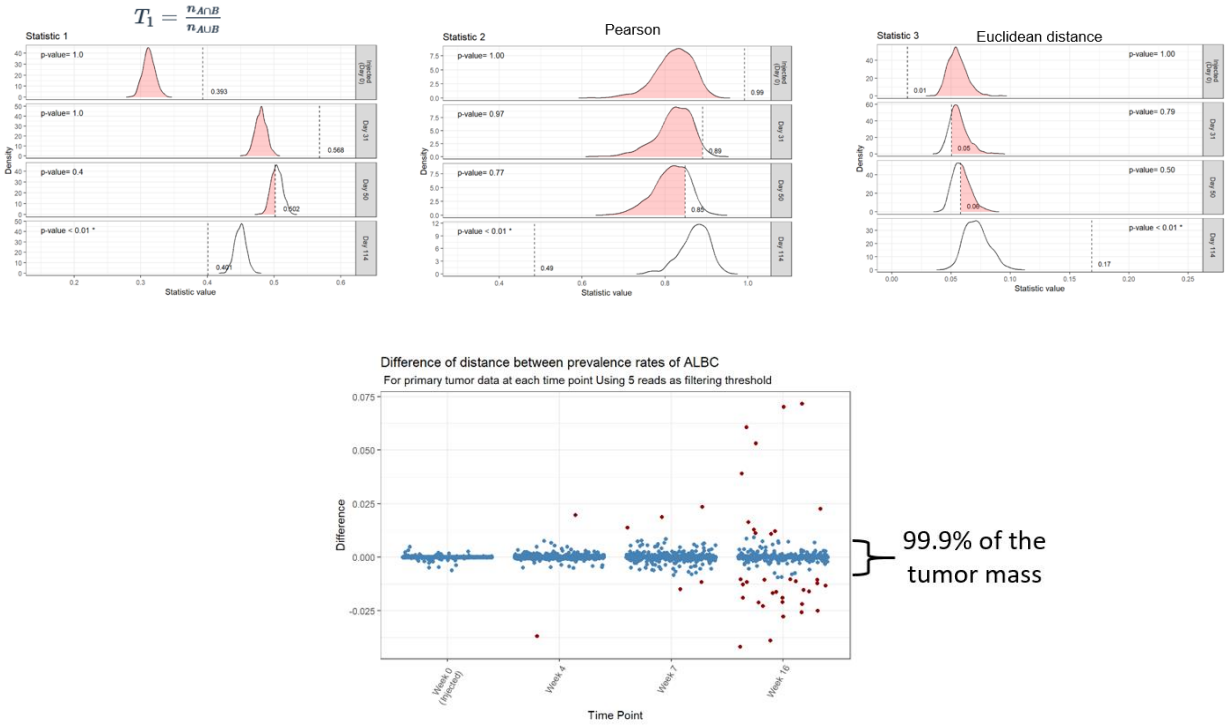
Figure 69. CRT robust test on the second independently generated orthotopic animal cohort. The same simulation of "common barcode types", "Pearson's correlation", and "Euclidean distance" (top panels from left to right) was applied to the second cohort of orthotopic CRT mice using a stringent criteria of similarity. The observation was in concordance with the CRT cohort 1, where orthotopic CRT tumors progress similarly till around week 9. In this cohort, while most of the lineage progress similarly, we also observe the natural divergence of tumor expansion between replica tumors over time (bottom panels, the red dot clones).

Majority of the clonal trends are similar amongst CRTs in terms of gain or loss in fitness

To further gauge the similarities of how CRT behaves in the context of individual clones gain or loss of fitness over time, we look at early vs. late time point CRTs (week 4 and week12) and compare the change in relative abundance change per lineage. For example, if a barcode abundance difference between week 12 and week 4 has a positive value, it indicates that the particular barcode has a gain in fitness over the 8 weeks span; vice versa, if the difference in value is negative, it indicates that particular barcoded clonal lineage lost fitness. Using this criterion, we can obtain four values for each barcode from the difference between "week 12 mouse #1 and week4 mouse #1", "week 12 mouse #2 and week4 mouse #1", "week 12 mouse #1 and

week4 mouse #2" and "week 12 mouse #1 and week4 mouse #2". Then, we further transform these four values into a simple "positive" or "negative" categories, where a positive value (gain in fitness) will be assigned a new value of "1", and a negative value will be assigned a new value of "0". Afterwards, we take the summation of these four values to gain five possible values of "0", "1", "2", "3", "4", where each value gives an indication of how consistent each barcode's behavior are. For example, if the summation is "4", it means consistently, for that barcode, comparing week 12's relative tumor representation (in either CRT animal) to that particular barcode's relative tumor representation at week 4 (either week4's CRT animal), it all indicates a positive gain in fitness. Therefore, a category of "0" and "4" indicates a most consistent outcome (all loss in fitness, or all gain in fitness), where "2" indicates there are discrepancies amongst individual CRTs.

Using the analytic methods described above, looking at the top 1,000 barcodes from week 4, around 82% of the barcode lineages' relative fitness trends are consistent amongst the four CRT animals (category "0" and "4", where 79% loss of fitness, 3% gain of fitness); 5% of the barcode lineages (category "1", "3") has partial in consistency; lastly, ~13% of the barcodes have inconsistency in gain or loss of fitness (category "2"). While this analytic method is somewhat crude, where we don't consider the absolute value of the differences, or, whether or not if the difference is significant, the set of data indicates that majority of the barcode trends amongst CRTs are somewhat consistent. As for the 13% of lineages fitness trends that are "inconstant" (category 2), if we consider the observation where clonal divergence happens naturally over time during tumor expansion, we might be measuring the same randomness of clonal behaviors in tumor. In other words, it is possible that "13%" of the clonal lineages will behave randomly, or, "13%" of the tumor in this particular setting are unpredictable due to unidentified factors.
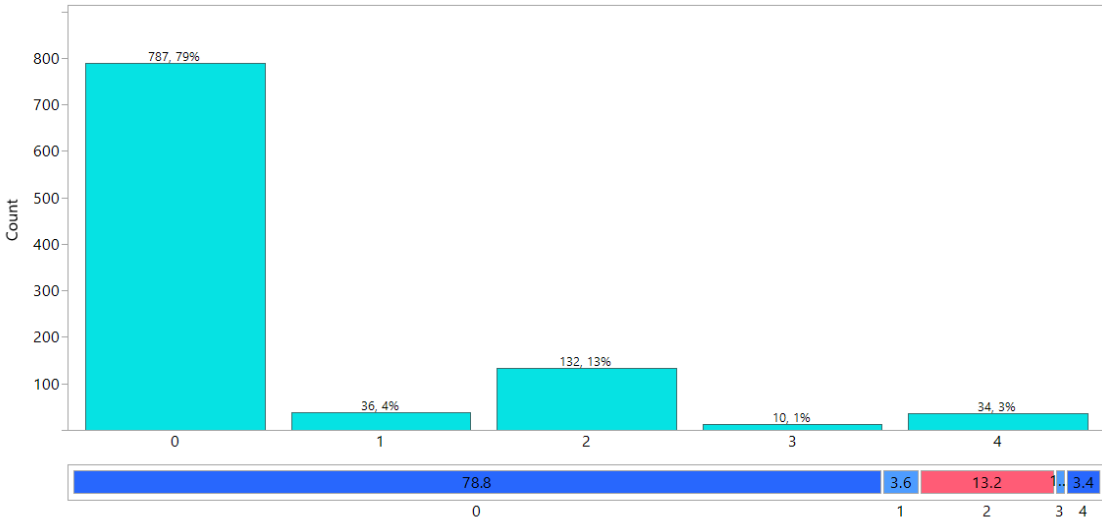
Figure 70. Category of clones that share consistent fitness dynamics between CRT tumors. Category to measure "consistency" of clonal loss or gain of fitness from week 4 to week 12, where the left number on each bar-graph indicates the count of barcodes in that category, while the right number indicates the % (n = 1,000). The X-axis, category "0" (loss of fitness, total n= 787, 79% of the clones) and "4" (gain of fitness, n= 34, 3 % of the clones) indicates the fitness trends are consistent amongst individual CRT animals. Category "1", "2" and "3" indicates clones that are partially consistent amongst replica tumors. The mosaic colored side way bar-graph at the bottom (blue and red) conveys the same information, where the numbers represent % of the total clones that falls into each bin (78.8% + 3.4% = 82.2% of the clones share consistent fitness trends in replica animals).

Orthotopic CRTs robustly replicates clonal tumor heterogeneity *in vivo* amongst individual animals from the same cohort and enables the orthogonal tracking of complex lineage dynamics and evolution

So far, we have validated that the orthotopic injection is a valid model to study tumor expansion and dissemination. Through barcoding and stabilization of the cell culture (consists of long-term self-renewal lineages), we are able to create cohorts of clonally identical tumors in terms of composition and relative abundance. Moreover, through statistical simulation robust test (using a stringent criterion as similarity baseline) and clonal trend (gain or loss of fitness) test, we validated that CRTs are maintained over time during unperturbed progression. On a side note, the CRTs clonal behavior diverges over time, where it is especially prominent during the late stage

of the tumor transplantation. This observation indicates yet to be identified factors (i.e. spatial vasculature distribution, hypoxia, nutrient deprivation) could drive and influence clonal evolution. Furthermore, considering clonally identical cell population were transplanted into each animal in the CRT cohort, this set of data also captures the natural unproductiveness of tumor expansion.

Due to the clonal biological replica nature of CRTs, it provides a way to accurately measure clonal evolution behavior while monitoring lineage abundance in independent animals longitudinally. Much like a cartoon flip book, without perturbing the tumor's growth, we can analyze individual clonal abundance in the CRTs at different time point and string them together to survey their dynamics over time. With this method, we are able to empirically obtain, and not by modeling nor extrapolation, an accurate representation of clonal lineage evolution pattern during tumor expansion.

Furthermore, by sampling individual barcode population in the blood (~800ul to 1mL), liver and lung (entire organ) of the CRT cohorts at different time points, we could gain insights to the clonal evolution dynamic of tumor dissemination.

Longitudinal model for tracing tumor's clonal progression
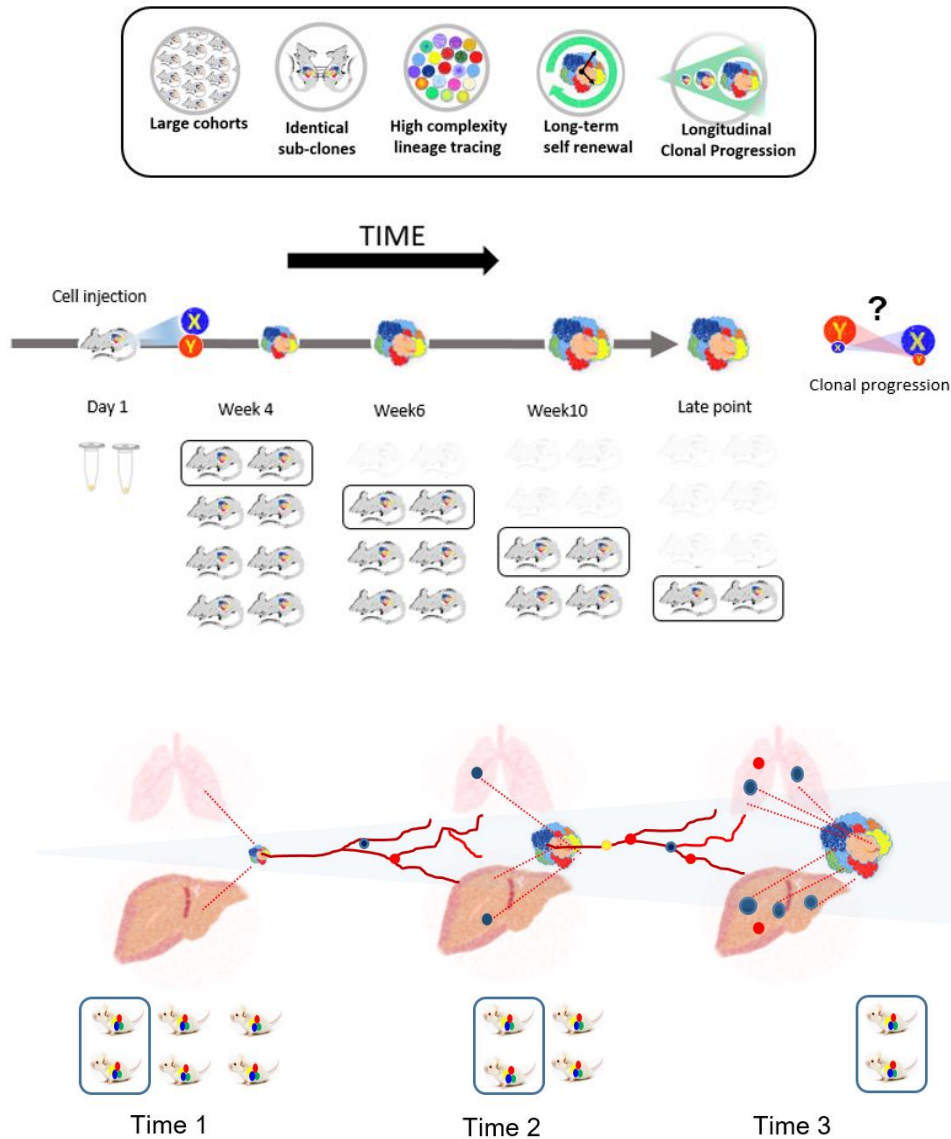
Figure 71. Graphical representation of longitudinal clonal tracking application by establishing clonal replica tumors. Top graph shows the key characteristics of CRTS, where it consists of long-term self-renewal lineages; and have the ability to simultaneously track a high complexity of lineages; can create a large cohort of CRT animals for experiments; also, by surveying CRTs over time, we could longitudinal trace clonal evolution in an *in vivo* setting without perturbing the tumor's system. Bottom graph shows the study of clonal dynamics over time is not limited to primary tumor, but clonal dissemination dynamics in the entire organism as well.
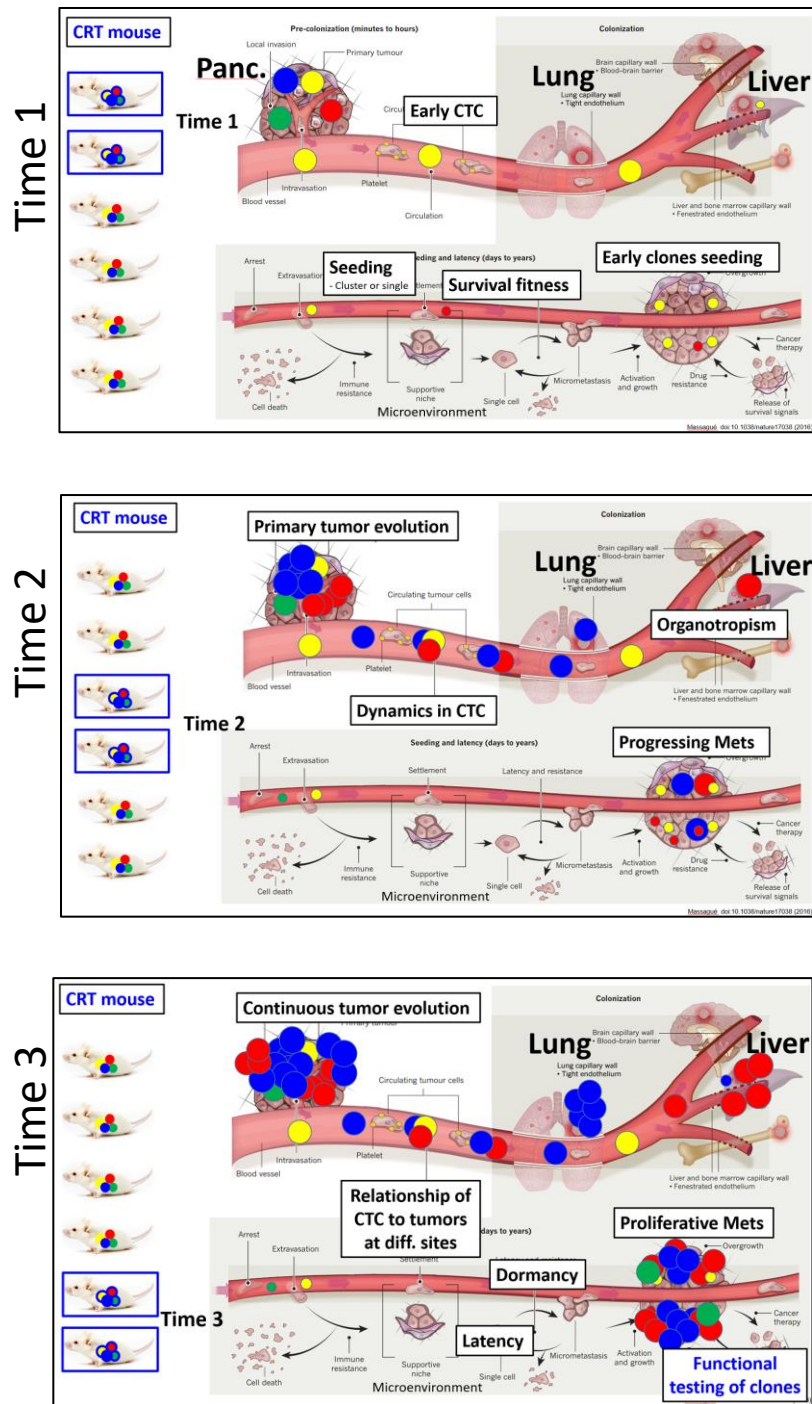
Figure 72. Schematic representation of experimental design for the longitudinal tracking of clonal dynamics using orthotopic CRTs. At each time point, at each time point relating to tumor progression, a pair of CRTs' (reproducibility) clonal composition (via barcode analysis) in primary tumor (pancreas), blood (for CTC) and liver and lung could be obtained to gain insights to the unperturbed clonal dynamics during tumor expansion and dissemination.
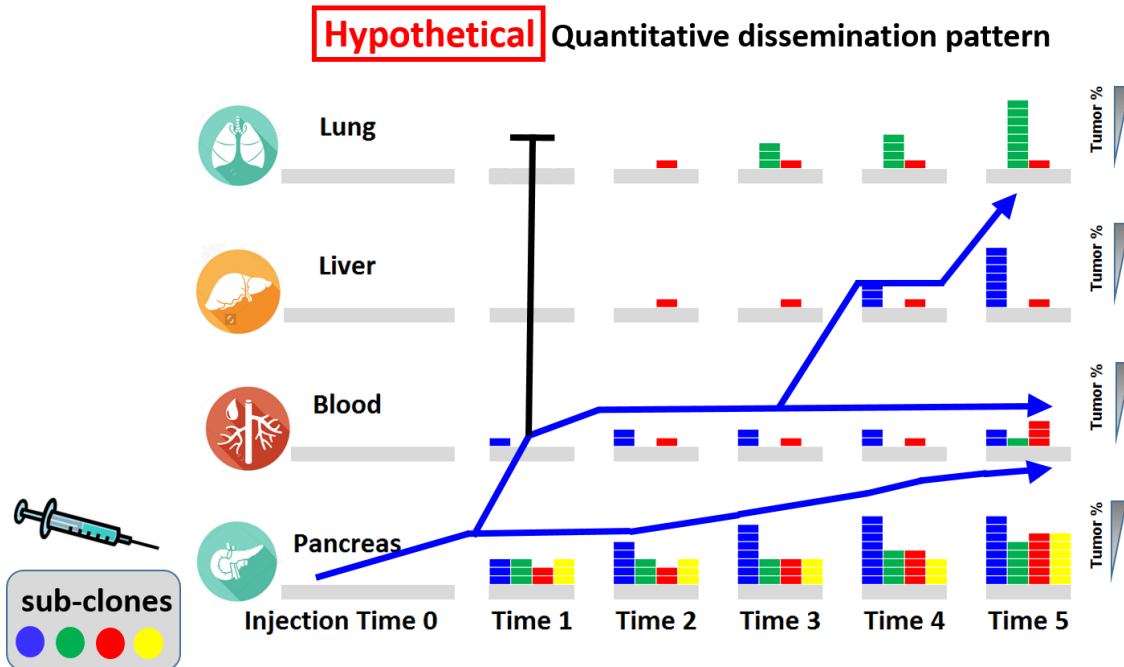
Figure 73. Graphical representation of hypothetical analysis of clonal evolution during tumor expansion and dissemination to distal organs. Simplified representation of four barcoded sub-clones in blue, green, red and yellow. The mixed clones are injected orthotopically into the pancreas of mouse to generate CRT cohorts. By analyzing the barcode abundance over time at different sites (primary tumor/pancreas, blood, liver and lung) we could trace each clone's unique expansion and dissemination pattern. For example, in the pancreas, the blue clone gradually expand and dominate the primary tumor over time (from "Time 1" to "Time 5"); enters the blood at an early time point ("Time 1"); Does not disseminate to the lung, however, is detected and dominate the liver metastasis from "Time3" to "Time 4". On the contrary, we may also identify clones such the yellow clone, while significantly present in the primary tumor, however, does not metastasize. Data in this figure are hypothetical examples.

Qualitative longitudinal analysis of top representing lineages revealed an unexpected degree of clonal expansion

By looking at the top few most abundant clones from day 0 to week 10 in the primary tumor, we see a high degree of clonal expansion. To look at the relative clonal abundance in a tumor, each barcode has been converted to % tumor representation at each time point. Briefly, the qualitative tumor representation is calculated by dividing each clone's read by the sum of total barcode reads detected in a sample. I.e. if barcode "A" has 100 reads detected, and the total

tumor's barcode read sum is 10,000, then barcode "A" represents 1% of the tumor mass. From

the graph below, we can see the top 12 clones in the tumor each represents around 1.4% to 3%

of the population at day 0 (what is injected); however, after 4 weeks post injection, one of the

clones went through an explosive expansion from 3% (day0) initially to ~6% (week4/day 28), to

~8.8% (week6/day40), to ~8.2% (week10/day70) of the tumor representation, while the other

clones maintained at 1-2% of tumor representation at week4 and gradually decrease in tumor

representation over time. By connecting each clone's % representation over time, we observed

within the top 12 clones, the expansion patterns are diverse. Looking at the graph below, some

clones experience initial expansion followed by decrease in representation, while some clones

seems to be continuously dropping in tumor representation over time. Keep in mind these are

empirically obtained clonal growth dynamics where tumors are left unperturbed and not

extrapolated model of dynamics. For the first time, relying on the CRT's robust biological clonal

reproducibility across cohorts, we are able to gain insights to clonal dynamics during natural tumor

expansion.
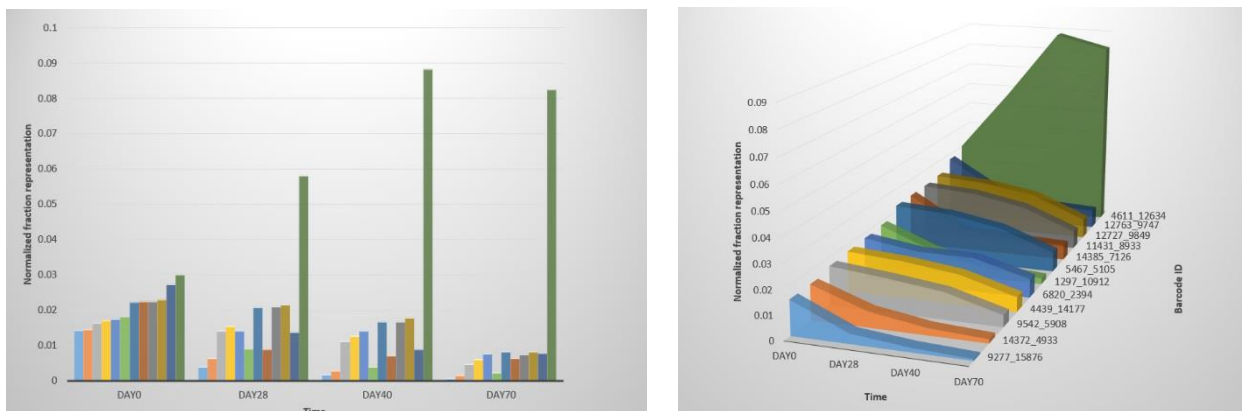
Figure 74. Qualitative analysis (% tumor representation) of top abundant clones in the primary tumor over 70 days of unperturbed expansion. Top 12 most abundant clones in the primary tumor and their relative tumor representation from day0 (injection) to day70 is displayed. X axis indicates time of collection (n=2 tumors for each time point) and Y-axis indicates % representation of the entire tumor. The graph on the right shows the continuum between clonal percent tumor representations across time points with barcode ID information. Each bar on the graph represents

a unique clone. Clone ID "4611_12634" (the green clone), compared to other clones, displayed a burst of clonal expansion once injected *in vivo* ("injected" ~3%, "28 days" ~ 5.8%, "day 40" ~8.8% and "day 70" ~8.2%).

Qualitative longitudinal analysis of top abundant clones reveals 6 major patterns of clonal dynamics during tumor expansion

T-SNE clustering was performed on the top 500 clones (represents majority of the tumor mass) in the primary tumor based on their tumor representation (%) over time[112]. From the analysis (Figure 75 below), we observed 6 major patterns of clonal dynamics over a 10-week period. Major patterns include clones having sharp decrease in fitness within 4 weeks post injection, leading to a decrease in tumor representation from day 0 to week4 and then maintained at a lower % till week 10 (green); clones that display a relatively slower loss of fitness, where the rate of loss of fitness gradual from day 0 to week 6, and then maintained relative fitness till week 10 (red); clones that maintain relative fitness through week4, and then experiencing gradual loss of fitness till week 10 (yellow); clones that continuously have gain in relative fitness throughout the 10 weeks (blue); clones with initial gain of fitness but eventually stabilize (teal); and clones that have sharp gain in relative fitness, plateaus and followed by sharp decrease in relative fitness (pink).
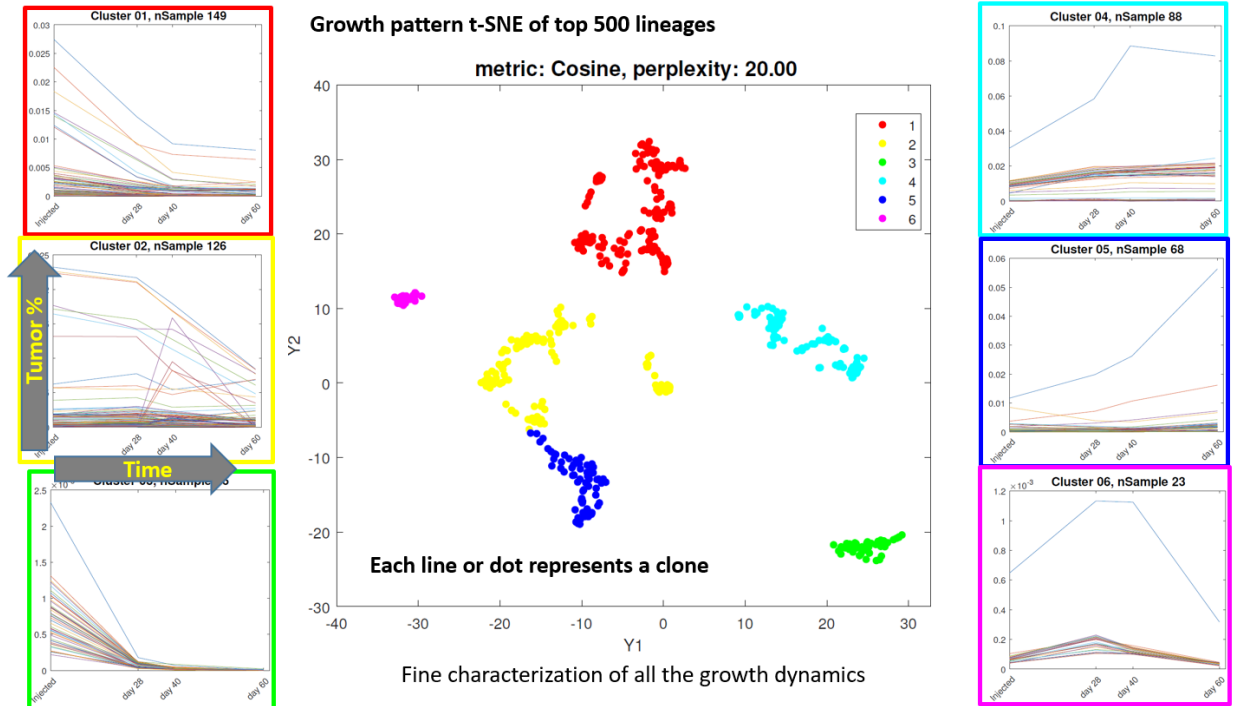
Figure 75. t-SNE clustering of distinct clonal dynamics in primary tumor (qualitative analysis in clonal % representation of tumor). t-SNE plot of the top 500 clones in the primary tumor based on relative tumor representation (%) over 60 days. Each box represents the parallel plot of individual clones from a cluster (each line is a unique clone, with y-axis in %, x-axis in time). For the top 500 clones, there are roughly six distinct growth patterns within the first 2 months of primary tumor expansion. Graph by Yinsen Miao.

Qualitative analysis - a look into early time point vs. late time point lineage that gain or loss fitness distribution stratified by quartiles of primary tumor mass at week12

Let us analyze the overall net gain or net loss of fitness of each sub-clones over the entire longitudinal study (this data set is analyzed with >1 read, primary tumor must have value at all times). Stratifying the primary tumor into four quartiles, we see most of the clones at the higher quartiles have an overall net gain in fitness, whereas the relatively lower representation clones have a harder time gaining fitness overall.
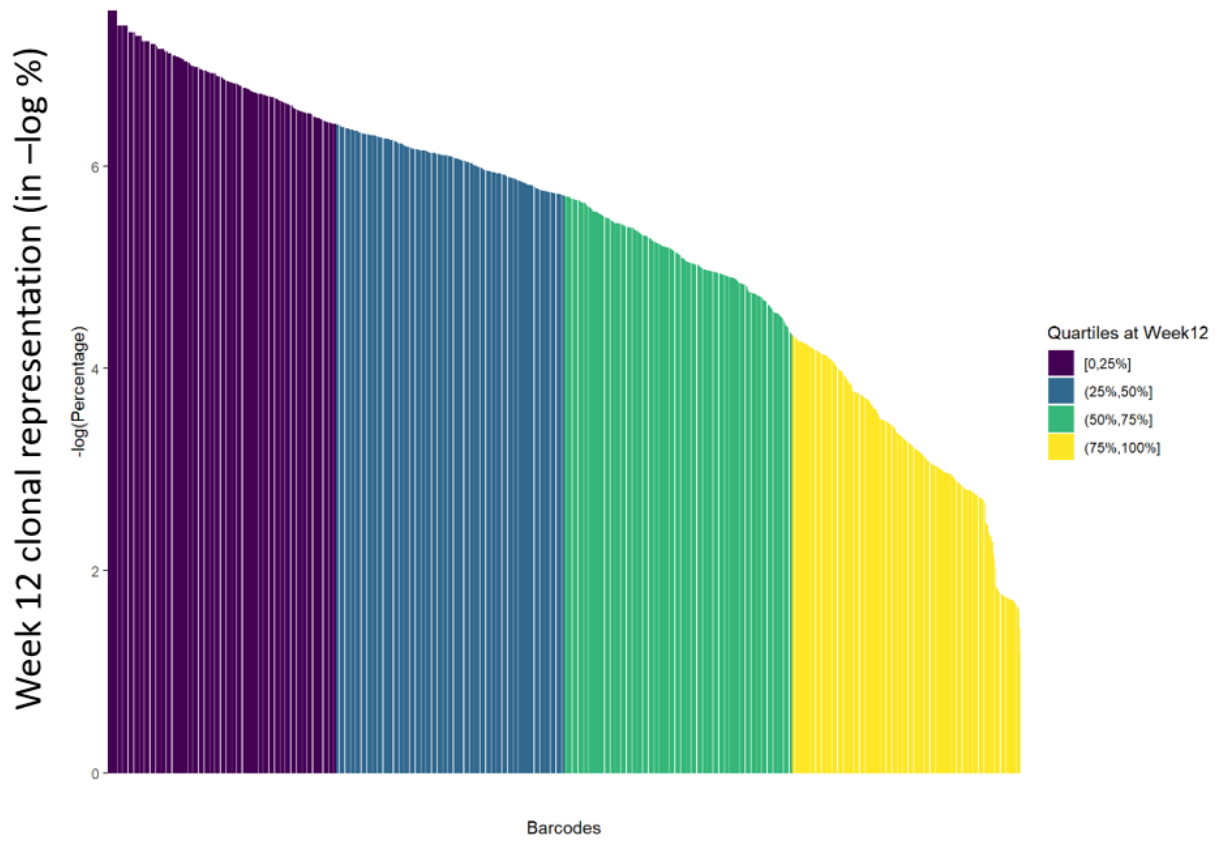
Figure 76. Primary tumor stratification into four quartiles based on value. Total 1336bc, blue=315, yellow = 701, red = 320.
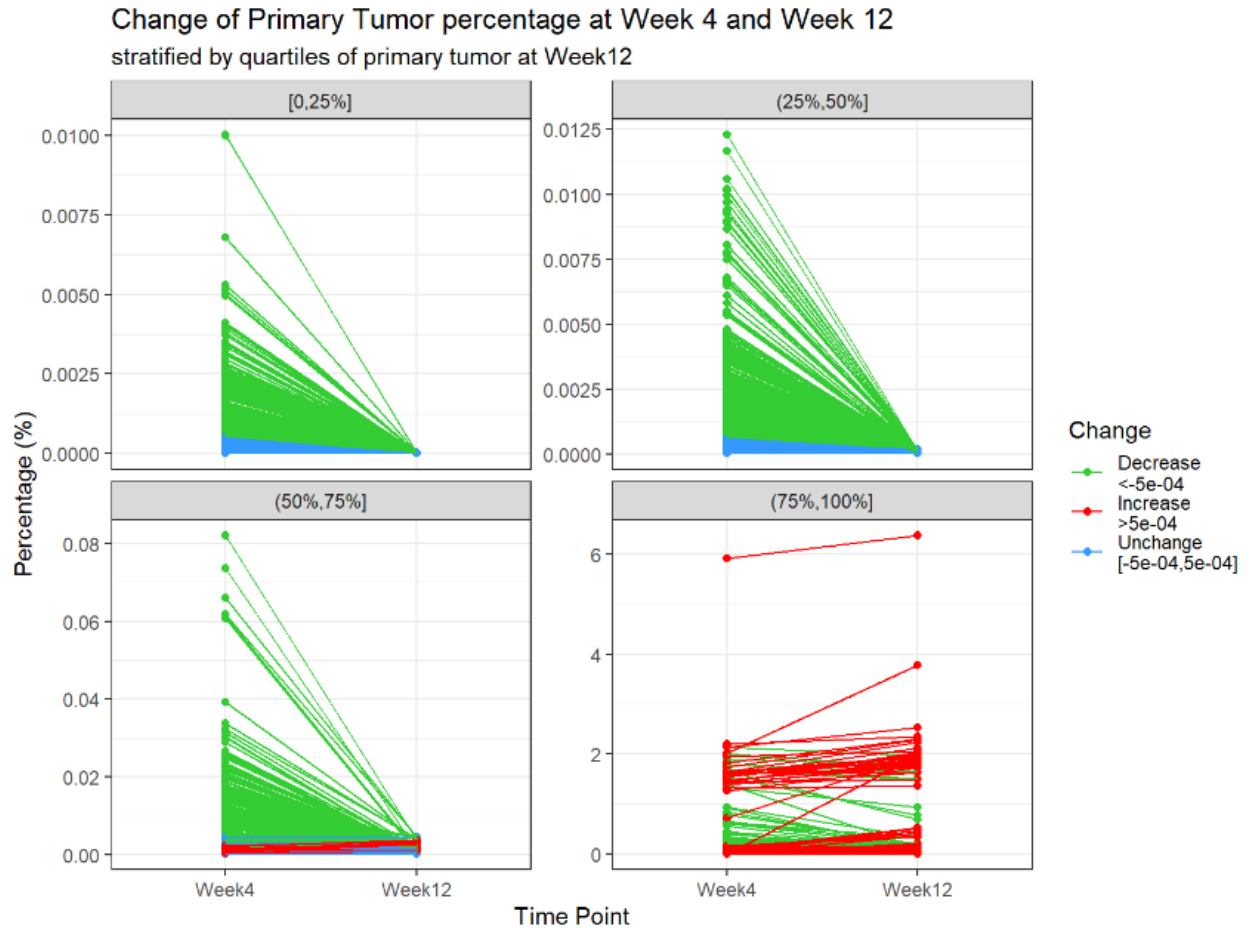
Figure 77. Qualitative analysis of overall fitness gain or loss stratified by quartiles of primary tumor at week 12.

Comprehensive view of primary tumor clonal dynamics during unperturbed tumor expansion

Clustering of similar clonal relative fitness dynamic (% tumor representation) in the primary tumor revealed a plethora of unique competitive interactions amongst sub-clones. Revealing that intra-tumor clonal competition is a dynamic process that occurs naturally during unperturbed tumor expansion. We will discuss each specific dynamics patterns in detail in the later sections.
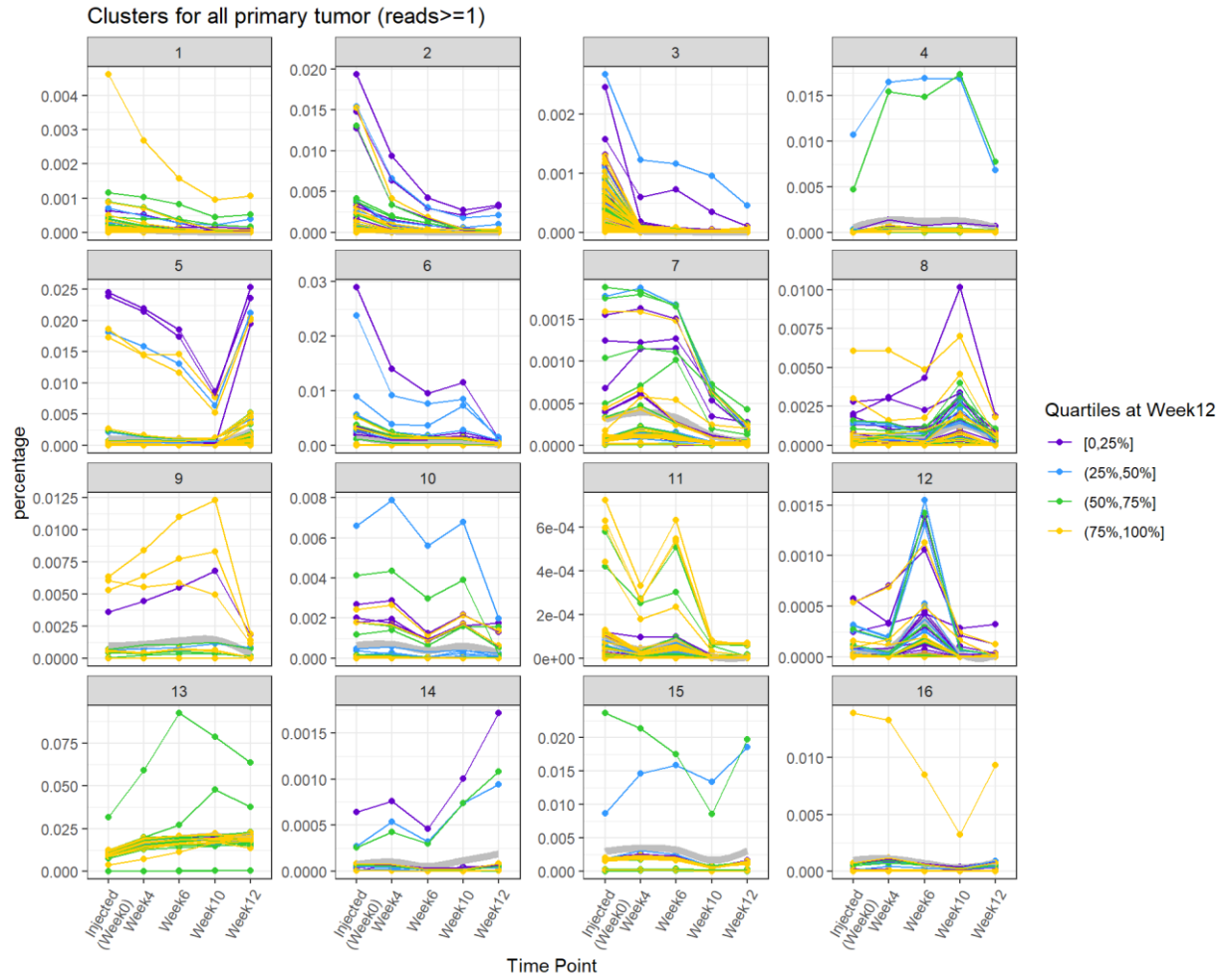
Figure 78. Qualitative analysis of clonal fitness evolution over time in the primary tumor. Y axis indicates the % representation of individual sub-clones, while each line represents a unique sub-clone. The color of the lines indicates the relative quartiles at week 12's primary tumor.
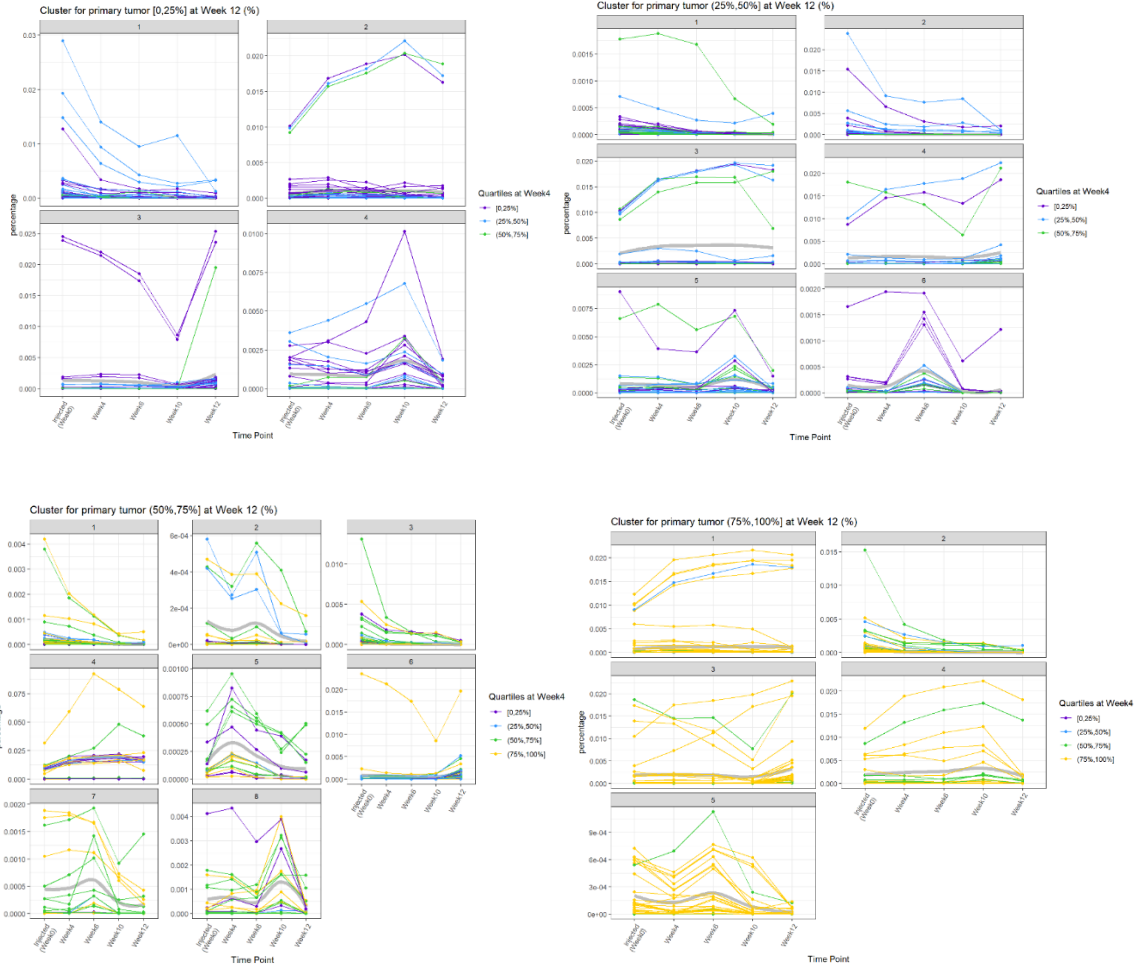
Figure 79. Qualitative analysis of clonal fitness evolution over time in the primary tumor. Y axis indicates the % representation of individual sub-clones, while each line represents a unique sub-clone. The color of the lines indicates the relative quartiles at week 4's primary tumor. Each panel is separated by quartiles at week 12's tumor.

Quantitative analysis reveals alternating clonal dominance and the ever-expanding nature of tumor

Converting each unique barcoded to "cell" unit using the spiked in "NGS read-to-cell" conversion scale during sample processing, it enables us to trace clonal dynamics in a quantitative manner. Here, we used a tiered-clustering (three rounds of clustering) approach to

analyze the data. First, all the barcodes that pass the noise filter is retained and clonal lineage dynamic clustered via KLM clustering[113]. In the first round of clustering, we see mainly three dynamics. The cluster "A" (in red) are the lineages with lower abundance. Cluster "B" (in green) representing the clones that are expanding over time since the beginning, and continue to be well adapted as reflected by the steady increase in their "cell" number. And then, there is the interesting clonal behavior in cluster "C" (in blue), where these clones are relatively low in abundance (almost dormant) for long periods of time (weeks), and then starting from week 9, all of a sudden went through an explosive degree of expansion. Most interestingly, these clones in cluster "C", after displaying a high degree of expansion, the quantity of these clones even surpassed those clones that were dominating the tumor during the initial few weeks (cluster "B"). We term this observed dynamic "alternating clonal dominance (ACD)", where lineages switch in tumor dominance naturally during natural, unperturbed expansion.

Since the clones in cluster "A" are low in abundance, it is hard to appreciate their evolution dynamics. To overcome this, we applied a second round of clustering on the lineages in cluster "A". The re-clustering of cluster "A" lineages (now named "AA") has four major clonal dynamics, as seen in the middle of the figure: cluster "AC" (in baby blue) represents clones that expand rapidly from the beginning, where near week 7, it starts to lose fitness and gradually decrease in number; cluster "AB" (in green) represents the clones that steadily expand and gain abundance over the entire timespan; cluster "AD" (in purple) represents clones that are relatively dormant in the first few weeks and expand exponentially at the later time points. In addition, looking at individual lineages in panel "AA" and "AB", we can appreciate that at any given time points sampled, there are clones with relatively lower abundance starting to proliferate and expand; however, it seems like most of those clones fail to continue to expand and eventually decrease in abundance at the following time point. This observation suggests the ever competing nature between clones, where lowly abundant clones are persist throughout and ever trying to initiate

clonal expansion. At the same time, for the clones that dominates throughout the longitudinal study, at a certain time point, may experiences a dramatic shift and lose its competitiveness and be taken over by other clones. This quantitative analysis data set also supports the alternating clonal dominance (ACD) observation mentioned in the previous section. The factors inducing the ACD phenomenon is under study.
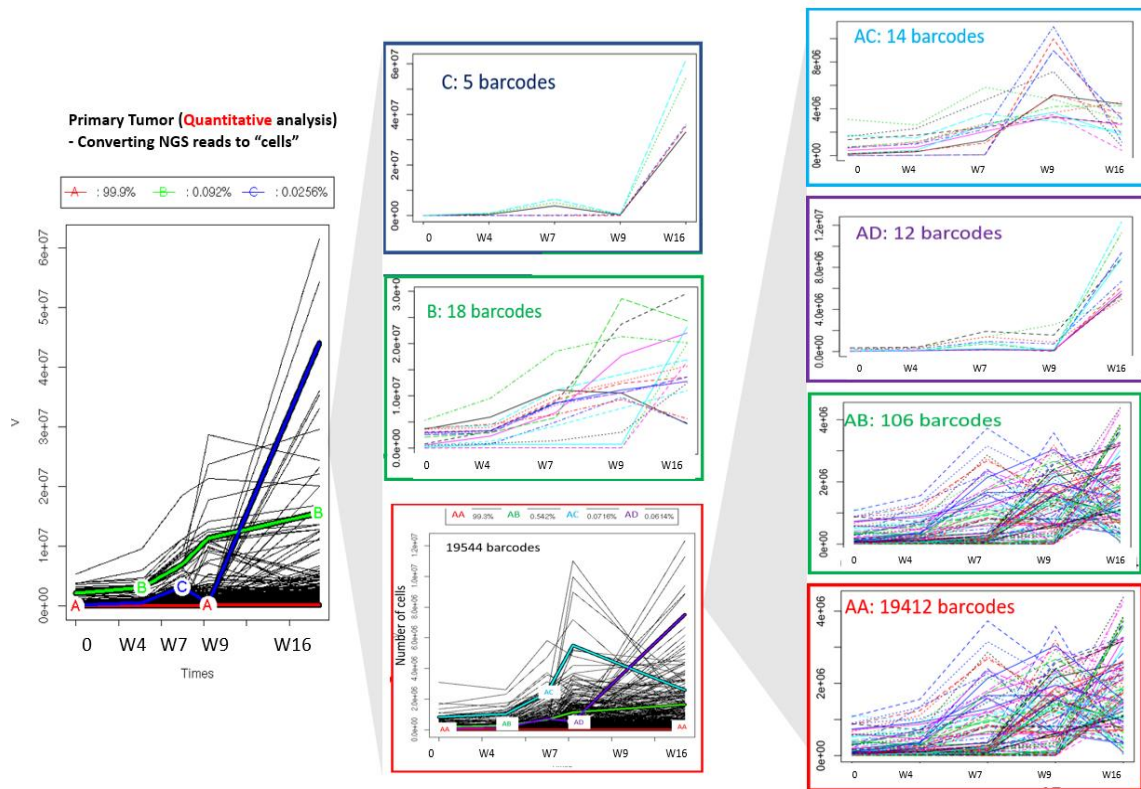
Figure 80. Tiered KLM clustering of clonal lineages in "cell number" over time. Bold line representing means of clusters. All lineages were clustered into three clusters (most left panel), where the mean of the clusters are highlighted. Behaviors such as clones with higher fitness over time (in bright green "B") and clones that have lower fitness throughout (red "A"), as well as clones exhibit lower fitness followed by outburst of gain of proliferation (in blue "C") were observed. In addition, alternating clonal dominance were observed, where the blue clones "C" over take the dominant position of the initially dominant clones "B" during later time points; re-clustering the initial cluster for higher resolution (the right panels) revealed alternating clonal dominance is also exhibited in lineages with lower abundance in the tumor (i.e. mid panel, bottom figure "AA"), as well as perpetual rising of low abundance lineages throughout tumor expansion. This data suggests the highly competitive nature amongst intra-tumor lineages, as well as a rich repertoire

of lineages, potentially with different intrinsic functional features consistantly trying to pop up. Representative data set from cohort 2. Graph analysis by Faezeh Darbaniyan.

Quantitative analysis of cumulative clonal abundance in primary tumor vs. distal sites reveals the rapid expansion nature of tumor cells outside its primary site

Looking at the summation of total "cells" of all lineages in the primary tumor, as well as in the blood, liver, lung and ascites at each time point, we can see that the total number of "tumor cells" are accumulating over time in different organs (below figure). In addition, the rate of tumor cells accruing at distal sites are exponentially rising over time while the primary tumor remained relatively similar in cell number. This observation is in line with the ever-expanding nature of tumor cells outside of the primary tumor site, where at the late stage of the tumor progression, the abundance of tumor cells at distal sites increases exponentially. Multiple factors could contribute to this, such as different spatial constrains, passive shedding of tumor cells from the primary tumor, clonal expansion at later time point.
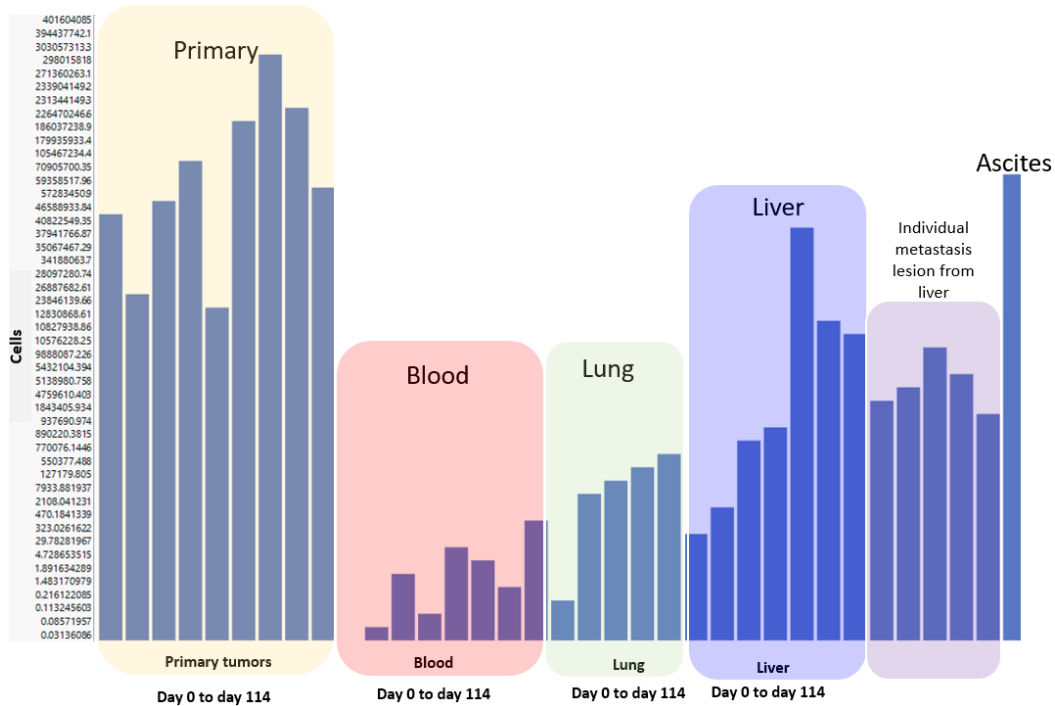
Figure 81. Cumulative cell number at each site (primary, blood, lung, liver) over time. Y-axis is " cells" in order of ranks (top); X-axis represents samples surveyed at different time points. Quantitative aspect of orthotopic CRTs allow us to survey sub-clonal accumulated mass in the tumor, as well as those that disseminate. Overall, the disseminated lineages accumulate in mass over time. At the late time point, the tumor mass accumulated in the liver metastasis, as well as the population in the peritoneal (ascites) could be comparable to that of the primary tumor's mass.

Statistical robust test of lineages that are able to disseminate outside of primary tumor

For the lineages that are able to disseminate, to reaffirm and be confident that they behave similarly in CRT animals for longitudinal tracking, an independent analysis of the statistical analysis was performed on this subset of clones. Using the same criteria described in previous sections, these clones in individual CRT animals indeed maintain and progress similarly. The natural clonal divergence over time is also observed in this subset of clonal lineages that has the ability to disseminate outside of the primary tumor.
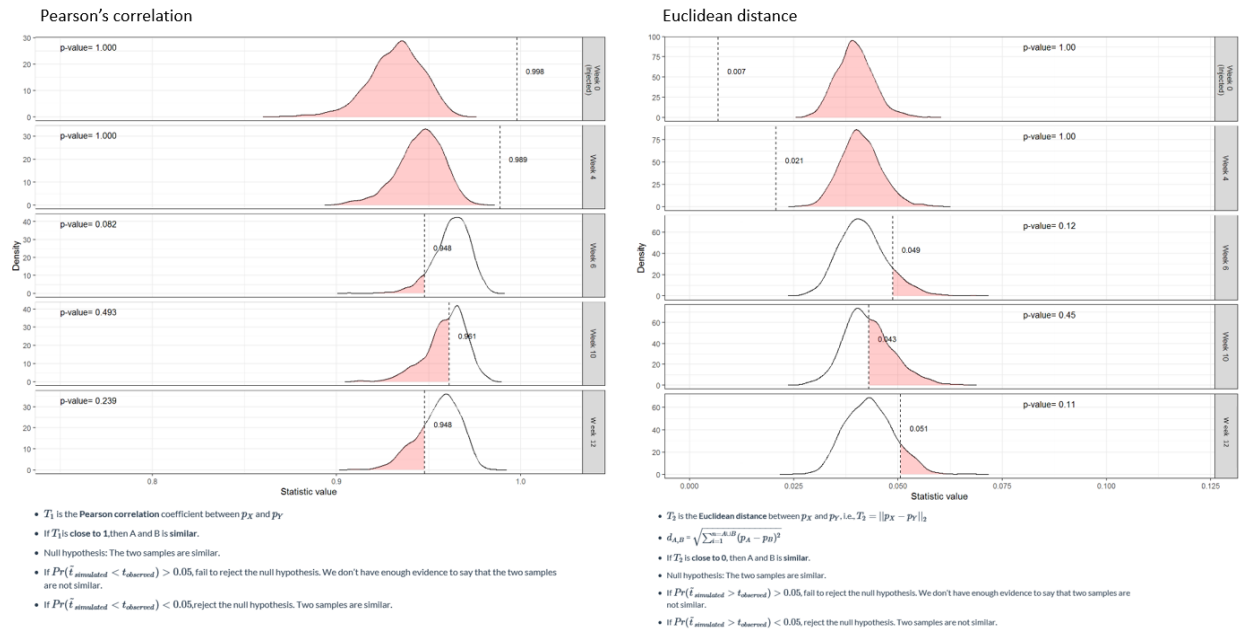
Figure 82. Statistical robust test on clones that are able to disseminate. Pearson correlation (left) and Euclidean distance (right). All sub-clonal lineages' qualitative data pass the statistical robust test, therefore, we have high confidence in the dynamics of these clones reflecting true biological dynamics. Analysis by Delia Wang.
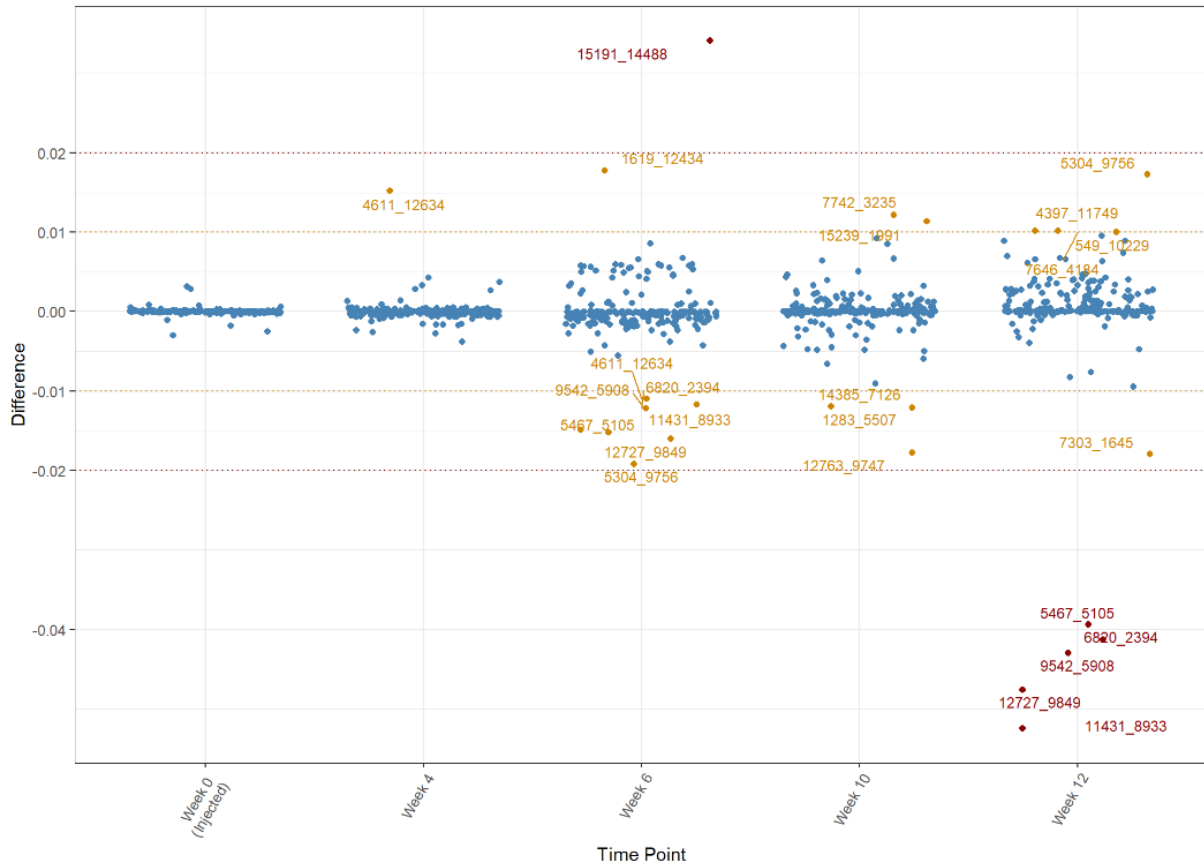
Figure 83. Difference (%) between sub-clones in replica CRTs primary tumor that are able to disseminate. Most lineages that are able to disseminate have a difference between replica animals within 1%. Throughout time, more clones diverge and have increased difference between replica animals of ~4% at week 12. Analysis by Delia Wang.

Tumor cells that disseminate into the blood increases up to 2,000-fold comparing early phase to late phase of tumor expansion

Analyzing the clonal abundance detected in the blood and compare the relative fold change from early stage (week4, where a solid mass of tumor is formed) to late time point tumor (week 14, where tumor burden was too much for the animal), we see an ever increasing dynamic in clonal abundance in the blood, as seen from the volcano-plot below comparing week 4 vs week

14. Also, the expansion of circulating tumor cells (work flow surveys CTC) in the blood seems to expand exponentially during the later stage of the time points. This could be seen from comparing CTC abundance increased from week 10 to week 14, where most of the increase in fold change has been observed.
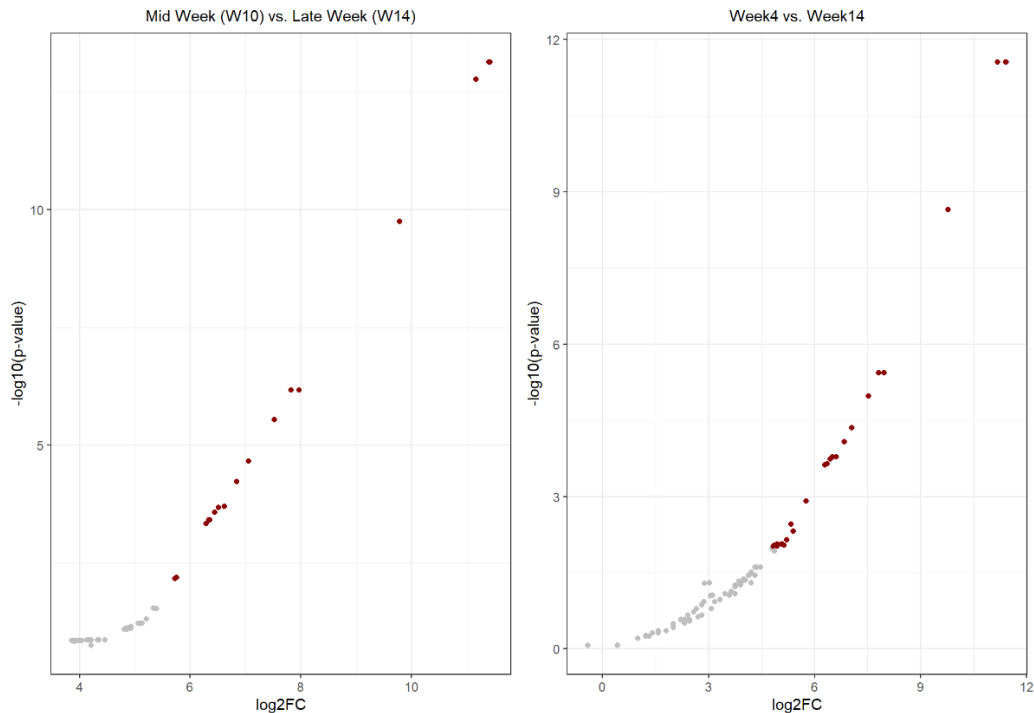
Figure 84. Volcano plot of individual clonal lineage fold change in the blood by comparing late stage animal (week 14) vs. early stage animal (week 4); as well aslate stage animal (week 14) vs. animal at week 10. The Y-axis indicates FDR and X axis indicates log_2 fold change comparing week 14 to week 4 (right) and week 14 to week 10 (left). This data indicates the total abundance of circulating tumor cells is always increasing over time in blood.

Longitudinal CRT model enables quantitative analysis of clonal complexity and clonal abundance in different organs

Our model allows us to assess the degree of dissemination by counting the number of unique clones and cell abundance observed in each organ. For example, there are ~200 unique clones

that has the ability to disseminate to the liver, and those ~200 clones makes up a total of ~10M

cells at the last time point (week 12). On the other hand, ~1,500 unique clones exists in the primary

tumor at week 12, and their total abundance is ~100M cells. This is a representative data to

showcase the quantitative analytical aspect of the orthotopic CRT model, which can obtain

accurate "cell count" of tumor mass with high resolution.
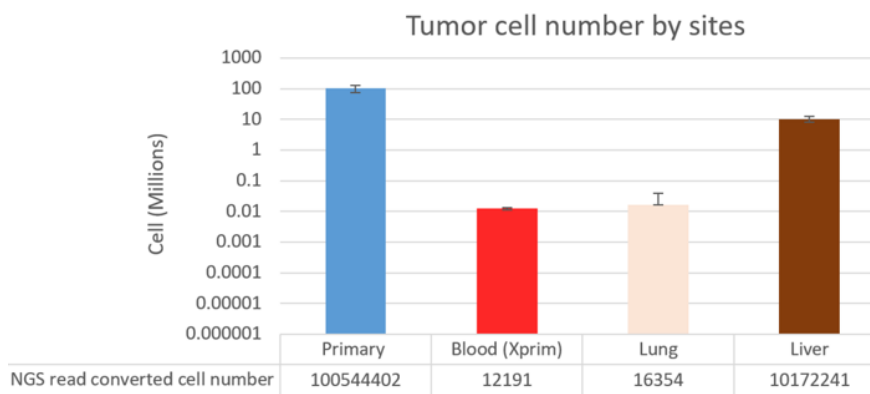
Figure 85. Cumulative cell number of tumors detected at different sites at the last time point. Liver metastasis at week 12 is ~10% of the primary tumor mass. Lung metastasis at week 12 is ~1.6% of the primary tumor mass. This is a representative data to showcase the quantitative analytical aspect of the orthotopic CRT model, which can obtain accurate "cell count" of tumor mass with high resolution.

By tracing clonal complexity in different organs over time, we see that in the primary tumor,

the clonal heterogeneity decreases over time. The disseminated clones detected at distal sites

experiences a "bottle-neck" effect where only a sub portion of clones having the ability to

disseminate (roughly 1% of the barcode complexity). The clonal complexity at distal sites

increases over time, suggesting continuous seeding and outgrowth of unique clones at distal

sites.
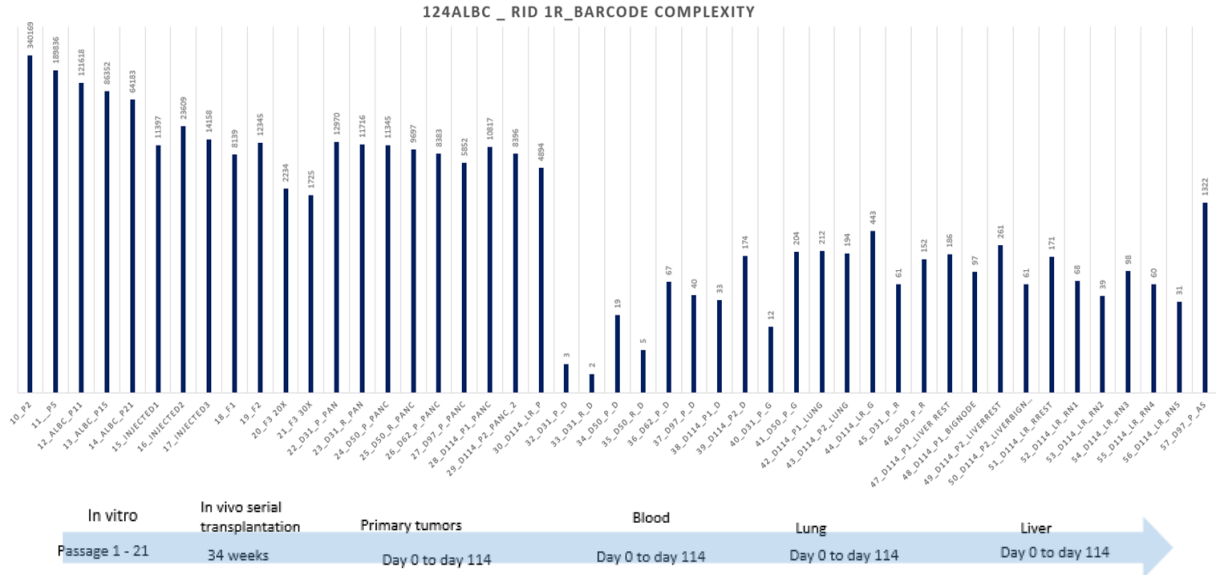
Figure 86. Clonal complexity present at different organs over time. Suffix ending with "Pan" or "panc" indicate pancreas primary tumor; suffix ending in "D" indicating "blood"; suffix ending in "G" indicates "lung"; suffix ending in "R" indicates "liver"; suffix ending in "As" indicates ascites. The number on top of each bar graph indicates unique barcode complexity.

Of those clones that can disseminate, most clones decrease in cell abundance in the primary tumor – quantitative analysis

Figure 87 below indicates the clonal cell abundance in week12 and week4's primary tumors. Of those clones that can disseminate, most clones decrease in cell number comparing week 12 vs. week 4 in the primary tumor. This could suggest that at week 12, while the tumor becomes more necrotic, the tumor mass might actually decrease despite the tumor size from the outside remain relatively stable (as measured by caliper, Figure 88). Further experiments are ongoing to address this point.

Change of Primary Tumor number of cells at Week 4 and Week 12
stratified by quartiles of primary tumor at Week 12
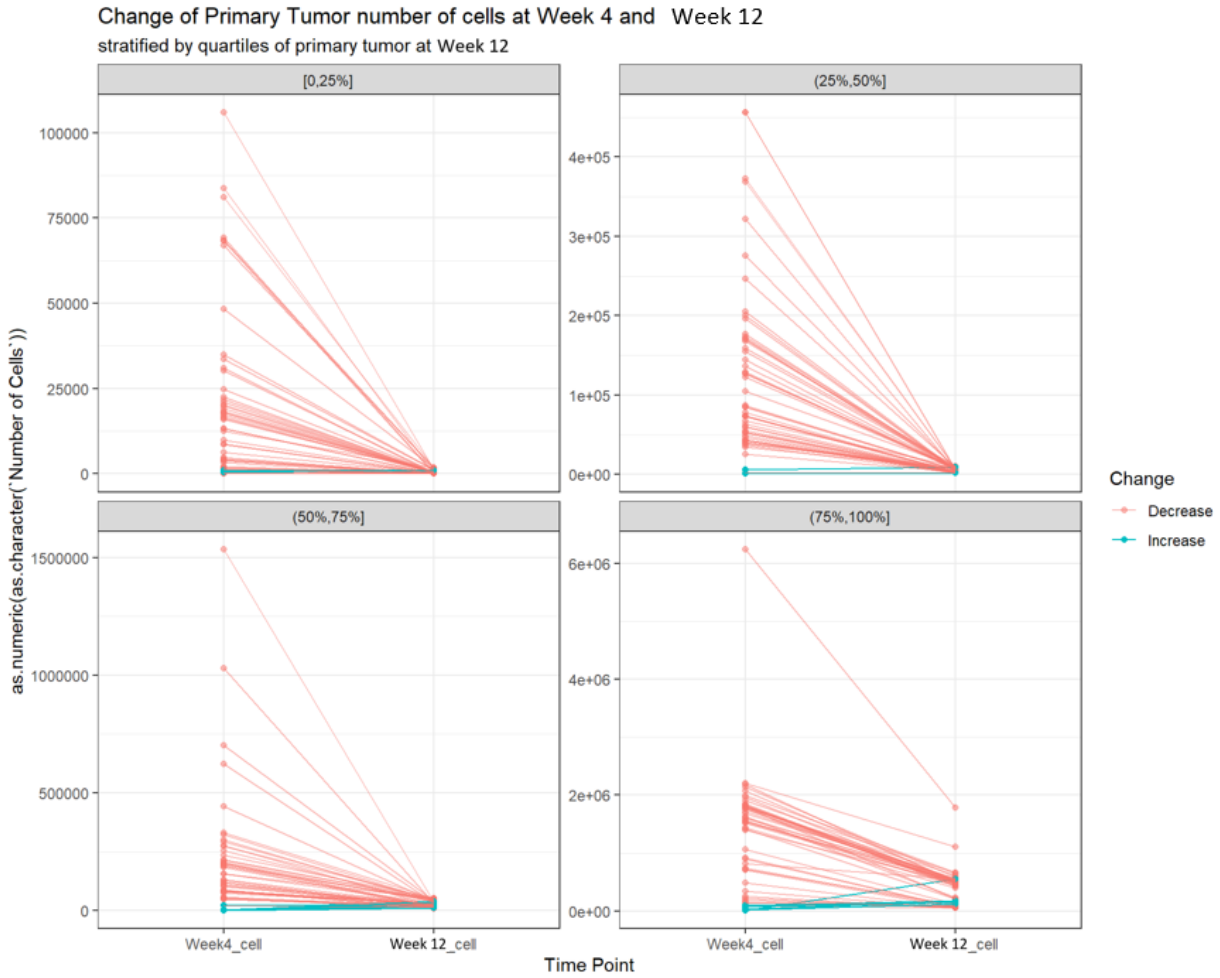
Figure 87. Quantitative - overall gain or loss of abundance per lineage comparing late stage (week12) vs. early stage primary tumor (week4). Breaking the primary tumor at week 12 in to quartiles (each panel represents a quartile) and plotting out all clones that have the ability to metastasize and whether or not the abundance decrease (in red) or increase (in blue) from week 4 to week 12.
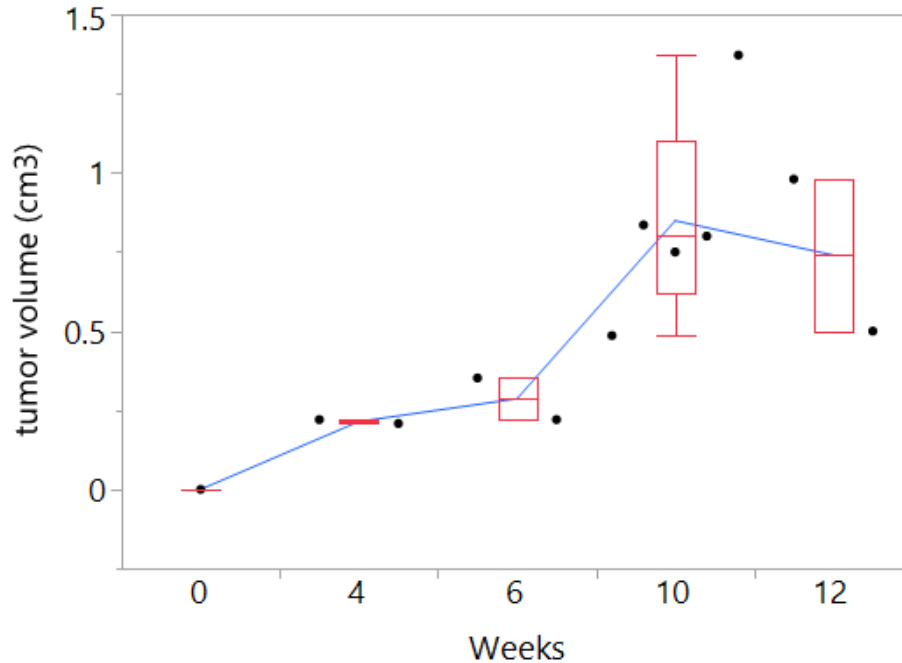
Figure 88. Orthotopic tumor volume measured by ruler. Orthotopic CRT cohorts's tumor volume measured (cm^3) over weeks. Week 0 assume volume = 0 (cell pellets).

The lineages that are able to disseminate have diverse evolutionary dynamics in the primary tumor

Out of the clonal lineages that have the capability to metastasize, we take a deeper look into their lineage dynamics in the primary tumor over time. Below sections will be focused on clustering lineages with similar behaviors and presenting it in a visually simplistic manner.

High volatility in clonal specific abundance during tumor expansion - quantitative analysis of clonal lineages

Figure 89 below shows the quantitative analysis of clonal lineages that were able to disseminate out of primary tumor. The clustering was done by dividing the primary tumor at the last time point (week 12) into quartiles (25% increments, the four big boxes) based on lineage accumulative tumor abundance. Each line represents a unique lineage dynamic over time, while the color of the line indicates the week 4's tumor representation quartile of that lineage. This way,

we could also be informed on the different clonal dynamics of lineages that are highly abundant vs. lowly abundant in the particular tumor. From the analysis, we can see clonal dynamic patterns that we logically think exists, such as a smooth gradual decrease (or increase) in clonal abundance over time. However, there are certain clonal dynamics we did not previously expect to see. The interesting patterns such as the alternating of intra-clonal fluctuating gain and loss of fitness, resulting in a general "W", "M", "sigmoidal", or "V" pattern. These patterns may seem perplexing or counter intuitive at first, however, considering that these data were obtained in an unbiased fashion and analyzed under the same stringent method and criteria, the pattern we are seeing here should reflect the true dynamics in the tumor during expansion. Especially given that if we accept certain generic clonal dynamic patterns, such as "a smooth gradual increase or decrease" over time, due to dogma, or simply that they are easier to believe.

From the plots below, for quartiles 0% - 25%, we can see that at week 12, there are no clones that exhibit a "dormancy and out growth during late time points" dynamics. However, most of the lineages in this quartile exhibit a "gradual decrease and a continuous loss of fitness" dynamic. Also, notice that there are no clones that are highly represented in early stage tumor (week4, 75%-100%). In addition, there are clones that seem to have a transient amplification dynamic (cluster 1, 2, 8), which suggests that most clones in this quartile failed to gain lasting fitness or adapt to the tumor environment albeit having the ability to expand initially.

As for quartile 25%-50%, we observe a more versatile lineage dynamics. There are lineages that experienced a longer phase of expansion continuously, however, due to unidentified factors, lost its competitiveness and started to gradually decrease (clusters 2, 3, and 6). There are also clones that display "cyclic gain and loss" of fitness throughout each time points (cluster 1 and 5); as well as, lineages that gradually lose fitness over time (cluster 4).

As for lineages in the quartile 50%-75% of the primary tumor at week 12, we see most of them exhibit the "alternating gain and loss of fitness" over time. This "alternating gain and loss" of

fitness could be reflecting the clones' higher intrinsic ability to change and constantly adapt to environmental changes (clusters 2, 4, 5, 6, and 7) compared to the lineages in the 0-25% quartile, where most clones exhibit "transient" expansion and are not able to regain fitness and exhibiting a continuous loss of abundance over time. Furthermore, interestingly, we see a subgroup of lineages that exhibit "dormancy" in the early time points followed by exponential growth during the last time point (cluster 3), suggesting that there are lineages "hiding" and maintaining low abundance and eventually break out.

As for lineages we see more of the same dynamics as the quartile 50% -75%. However, most of the clones in quartile 75%-100% have a higher abundance at week 4, suggesting clones that gain fitness early on have advantages in maintaining that fitness. Interestingly, there are also clones that exhibit "alternating loss and gain" of fitness initially but breaks out and continue to gain fitness (cluster 6), suggesting clones may have to go through a phase of adaptation before gaining fitness over competing clones; on the contrary, there are clones that exhibit "alternating gain and loss" of fitness but eventually continue to lose fitness (cluster 2, 3, 5). Lastly, there are clones that maintain a relative stable fitness and eventually breakout and grow exponentially (cluster 4).

All these dynamics were analyzed the same way across the samples and represents an unperturbed picture of how complex tumor lineages interact and compete with each other during tumor expansion.
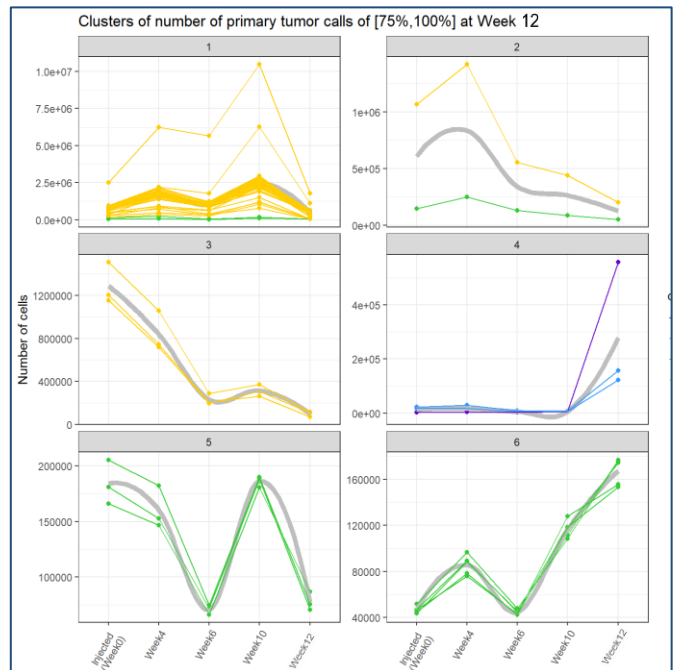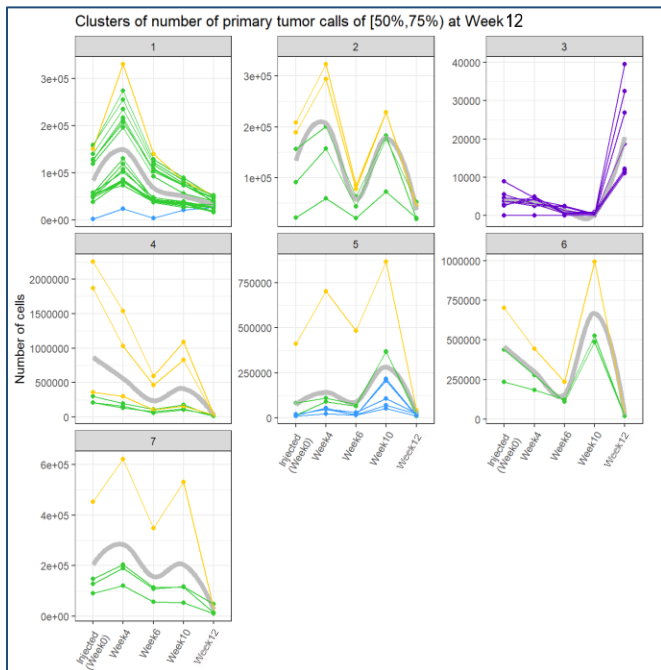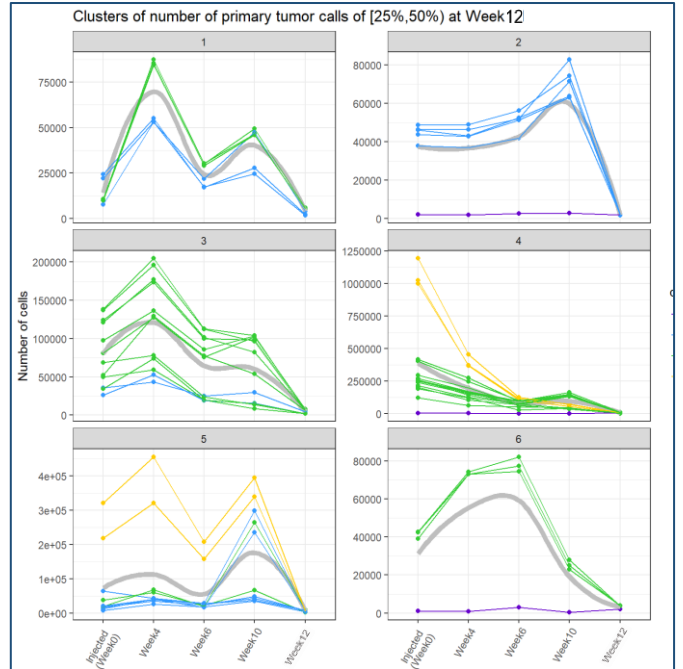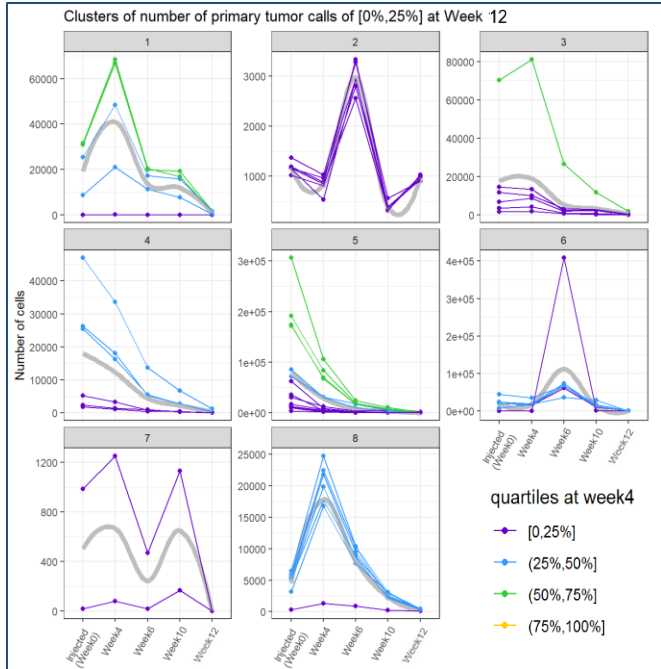
Figure 89. Longitudinal clonal dynamics in the primary tumor of clones that are able to metastasize. KLM clustering of individual clonal abundance based on quartiles of week 12's primary tumor. Individual line color reflects the abundance of clones at week 4, which is also stratified in quartiles.

Independent cohort of lineage tracing experiment reveals similar alternating clonal dominance dynamics (CRT cohort 2) - quantitative analysis

Quantitative analysis of orthotopic CRT cohort #2 reveals similar clonal dynamics seen in cohort #1. By analyzing the longitudinal clonal dynamics over 16 weeks, we see that in this cohort of tumors, they display all the major clonal dynamics as seen in the week 12 cohort. Notably, the "dormant followed by proliferation at later time point" clonal dynamic, the "transient amplified" dynamic, as well as the "alternating clonal dominance" dynamic around week 9.
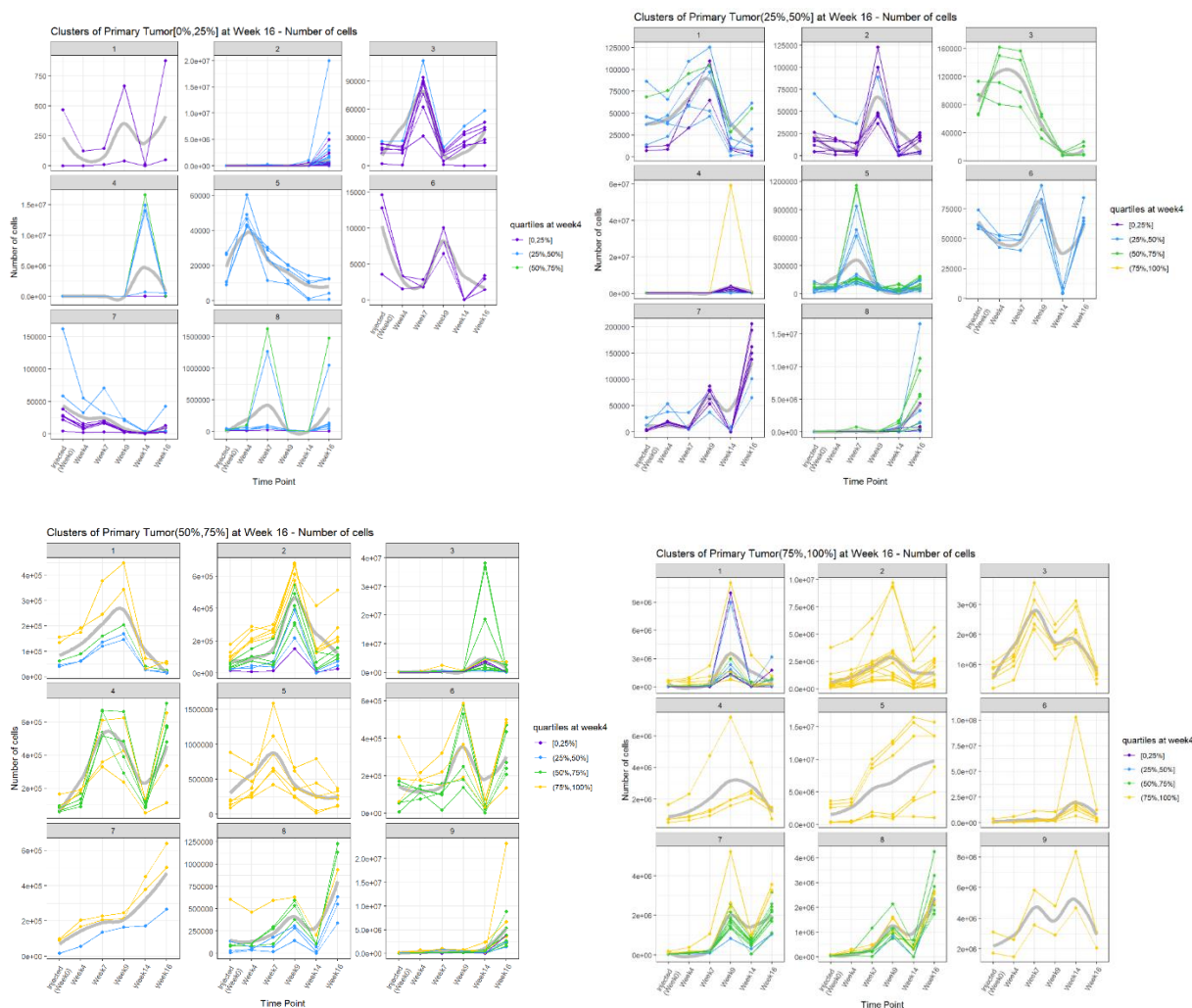
Figure 90. Clonal dynamics in the primary tumor of clones that are able to metastasize in an independent orthotopic CRT cohort. KLM clustering of individual clonal abundance based on quartiles of week 16's primary tumor. Individual line color reflects the abundance of clones at week 4, which is also stratified in quartiles.

<u>clonal evolution in the primary tumor – clones with ability to metastasize display high level of alternating clonal fitness during tumor expansion - qualitative analysis</u>

Converting the lineages into relative percent representation of the primary tumor at week 12 (end stage) allow us to look into the "relative fitness" amongst clones. Clustering the lineage relative fitness dynamics over time revealed the highly dynamic nature of clonal competition, as well as alternating clonal dominance naturally occurring during tumor expansion.

Here, we break down the clustering into quartile of the week 12's primary tumor representation (0-25%, 25%-50%, 50% - 75% and 75% to 100%); furthermore, at the same time, indicating the relative abundance of each lineage during week 4 tumor, also represented in quartiles (0-50%, 50% -75%, 75-100%) for clustering. For the lower quartile lineages (at week 12), clones display mainly four types of fitness dynamics: 1. Transient gain of fitness followed by loss of fitness after week 4; 2. Gain-loss-gain of fitness; 3. Dormancy followed by sudden gain of fitness towards the later time point; 4. Gradual loss of fitness over time (Figure 91).
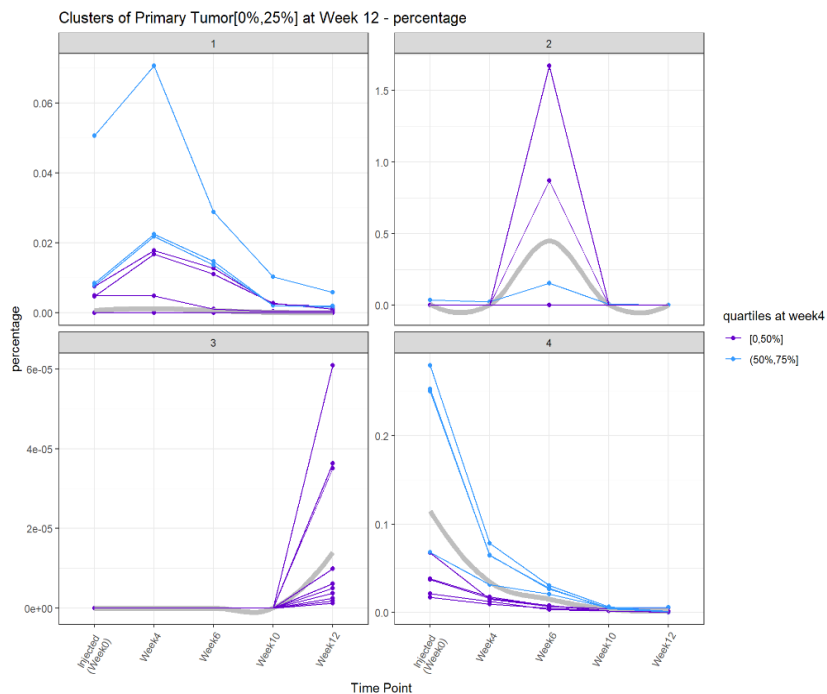
Figure 91. Clonal relative fitness dynamics in the lower quartile of the primary tumor in terms of abundance. Y-axis indicates % representation of each lineage; X-axis indicate time; each line indicates a unique sub-clone. The color of the line indicates the relative abundance of particular lineage at week 4. Four main cluster dynamics observed: 1. Transient gain of fitness followed by loss of fitness after week 4; 2. Gain-loss-gain of fitness; 3. Dormancy followed by sudden gain of fitness towards the later time point; 4. Gradual loss of fitness over time.

As for the 50%-70% tumor representing lineages, clones display mainly five types of fitness dynamics: 1. Transient gain of fitness followed by graduate loss of fitness; 2. Gradual loss of fitness over time; 3. Gain– loss –"big gain" –"big loss" of fitness dynamic; 4. Dormancy followed by sudden gain of fitness towards the later time point; 5. Gain – loss – loss – "big gain" of fitness. (Figure 92). Interestingly, from clusters 3 and 5, at the time of week 10, clones in this particular set of data display a large alternating fitness switch, resulting in a change of relative dominance of clones in the primary tumor.
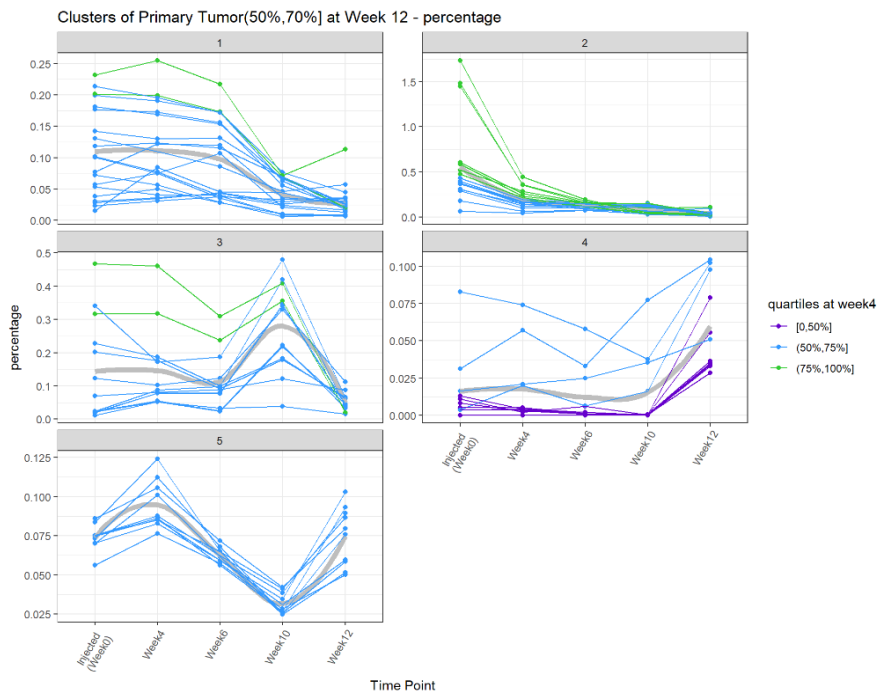
Figure 92. Clonal relative fitness dynamics relating to 50%-70% abundance of the primary tumor. Y-axis indicates % representation of each lineage. Each line indicates a unique sub-clone. The color of the line indicates the relative abundance of particular lineage at week 4. Five main cluster

dynamics observed: 1. Transient gain of fitness followed by graduate loss of fitness; 2. Gradual loss of fitness over time; 3. Gain-loss-GAIN-LOSS of fitness dynamic; 4. Dormancy followed by sudden gain of fitness towards the later time point; 5. Gain – loss – loss – GAIN of fitness.

As for the 75%-100% tumor representing lineages, clones display mainly eight types of fitness dynamics: 1. Gradual loss of fitness over time followed by a small gain of fitness towards the end at week 10, but continue to drop in fitness afterward; 2. A relatively stable fitness overall followed by slight gain of fitness at week 10; 3. Gradual increase in fitness followed by sudden loss of fitness at week 10; 4. A sigmoidal shaped dynamic where slight increase followed by loss of fitness from week4 to week 10, then, an aggressive gain in fitness after week 10 ; 5. Gain – loss – gain – loss of fitness; 6. Stable but slight increase of fitness overtime followed by slight drop in relative fitness; 7. Gradual drop of fitness followed by a sudden gain of fitness at week 10; 8. A persistent gain of fitness over time (Figure 93). From Figure 93, clusters 3 and 4, we also see lineages display an alternating fitness switch.
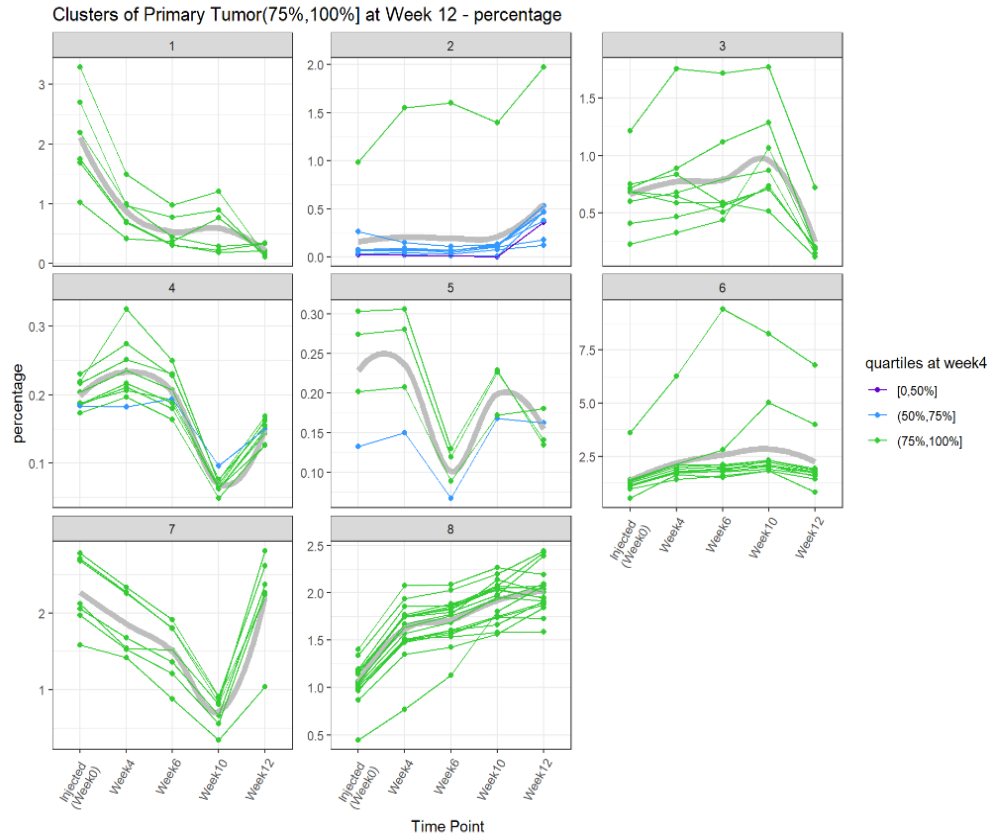
Figure 93. Clonal fitness dynamics displayed relating to 75%-100% of the primary tumor abundance. Y-axis indicates % representation of each lineage. Each line indicates a unique sub-clone. The color of the line indicates the relative abundance of particular lineage at week 4. Eight main cluster dynamics observed: 1. Gradual loss of fitness over time, followed by a small gain of fitness towards the end at week 10, but continue to drop in fitness afterward; 2. A relatively stable fitness overall followed by slight gain of fitness at week 10; 3. Gradual increase in fitness followed by sudden loss of fitness at week 10; 4. A sigmoidal shaped dynamic where slight increase followed by loss of fitness from week4 to week 10, then, an aggressive gain in fitness after week 10 ; 5. Gain – loss – gain – loss of fitness; 6. Stable slight increase of fitness overtime followed by slight drop; 7. Gradual drop of fitness followed by a sudden gain of fitness at week 10; 8. A persistent gain of fitness over time.

Another independent clustering (base on the quartiles of the injected population) of lineage fitness dynamics in the primary tumor reveals similar observation of diverse, unique dynamics of relative clonal fitness evolution and alternating clonal fitness naturally occurring in the tumor (Figure 94). For a comprehensive clustering, please see the methods section on clustering.
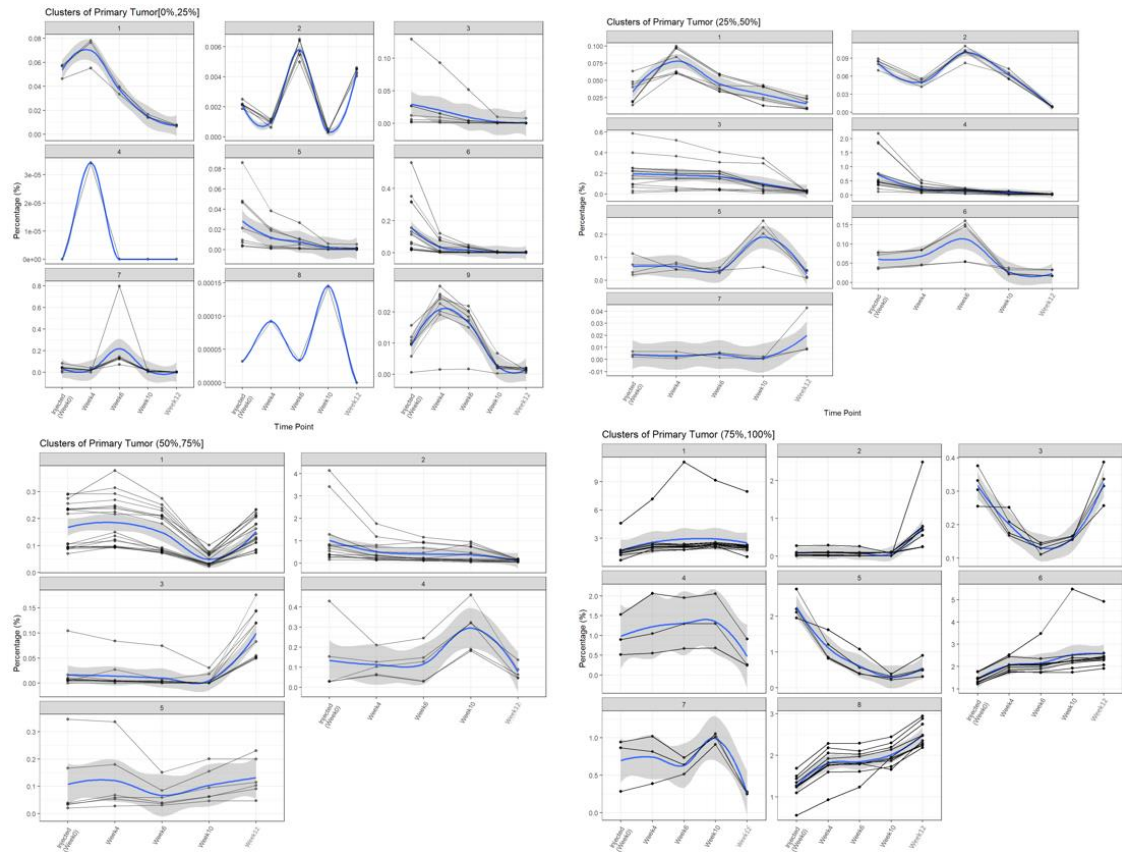
Figure 94. Clonal fitness dynamics of the primary tumor in quartiles of the injected population. Y-axis indicates % representation of each lineage. Each line indicates a unique sub-clone.

Longitudinal tracking of liver and lung clonal dynamics reveals the unpredictable explosive growth nature of metastasis and the existence of transient amplified sub-populations

We analyzed the clone's abundance fold change of late stage vs. early stage of tumor expansion (week 14 vs. week 4), and plotted a volcano plot (Figure 95 below), where each dot represents a unique clone found in the liver. On the plot, the x-axis represents $\log_2$fold change of each clone's abundance fold change between week 14 and week 4, and the y-axis represents

the false discovery rate (-log P value, the larger the number the more difference there is and consistent between the CRT animals).

From our model, there is a sub-set of lineages that are able to expand ~2^{15} times throughout the 10-week period in the liver and the lung. Furthermore, looking at the volcano plot, within those lineages that were able to expand ~2^{20} fold in 10 weeks, the ones that have a high – log(p value) means those clones behave similarly in both the CRT animal, while the low –log(p value) clones only appeared in one of the two CRT animals. Considering these lineages in the CRT primary tumors being similar, the "inconsistency" for the metastasis fold-change phenotype may be the degree of randomness introduced by the criteria required for a metastasis expansion.

Besides the massive metastasis outgrowth population, there are clones that display a relatively static in abundance dynamic without much expansion (clones with fold-change ~ 0), which infers to the phenotype of seeding and dormancy. Moreover, interestingly, we observe a sub-set of clones that have a higher abundance initially at week 4 then at week 14 in both the liver and the lung (the lineages that have a negative log FC in the volcano plot). This could be suggesting the scenario where clones that "seed" in the liver or lung early on, but could never establish a fitness towards growth, resulting in a transient amplification dynamic. This data suggests only those lineage that could adapt to the organ microenvironment will continue to dominate and form metastasis.
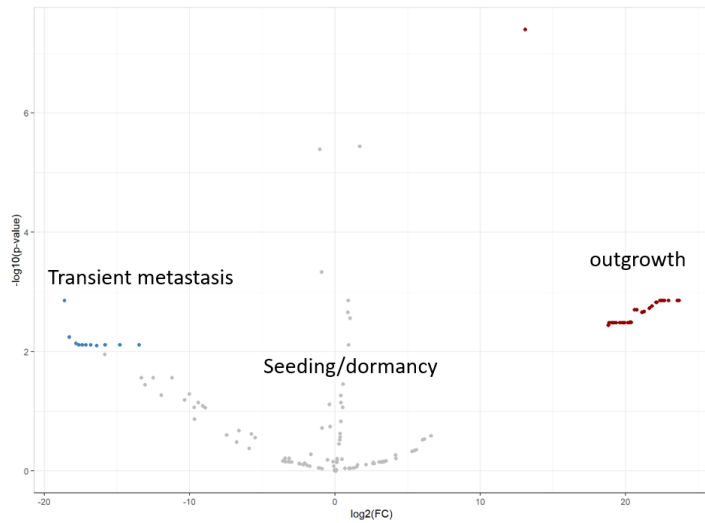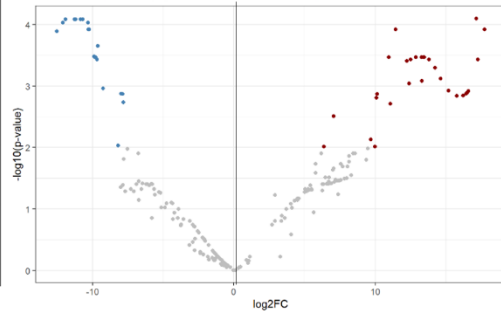
Figure 95. Volcano plot of the fold change of individual clones detected in the liver and the lung (Left= liver; Right = lung); Y-axis = FDR, X-axis = log₂FC. Fold Change is calculated by lineage abundance at week 12 divided by week 4.

A look into the primary tumor's lineage dynamics correlation with metastasis fitness

Metastasis in liver – many clones can "seed", but only certain clones would metastasize and grow at distal sites

Figure 96 below is the volcano plot from plotting the lineage abundance fold difference of liver metastasis week 14 vs. week 4 in terms of log_2 fold change (X-axis) and its P-value (Y-axis). From this graph, we can see there are clones that gained fitness over the 10 weeks in liver (positive logFC), and there are clones with decreased fitness (negative log FC); also, there are clones that maintained their relative abundance over time (|logFC| close to zero).

132
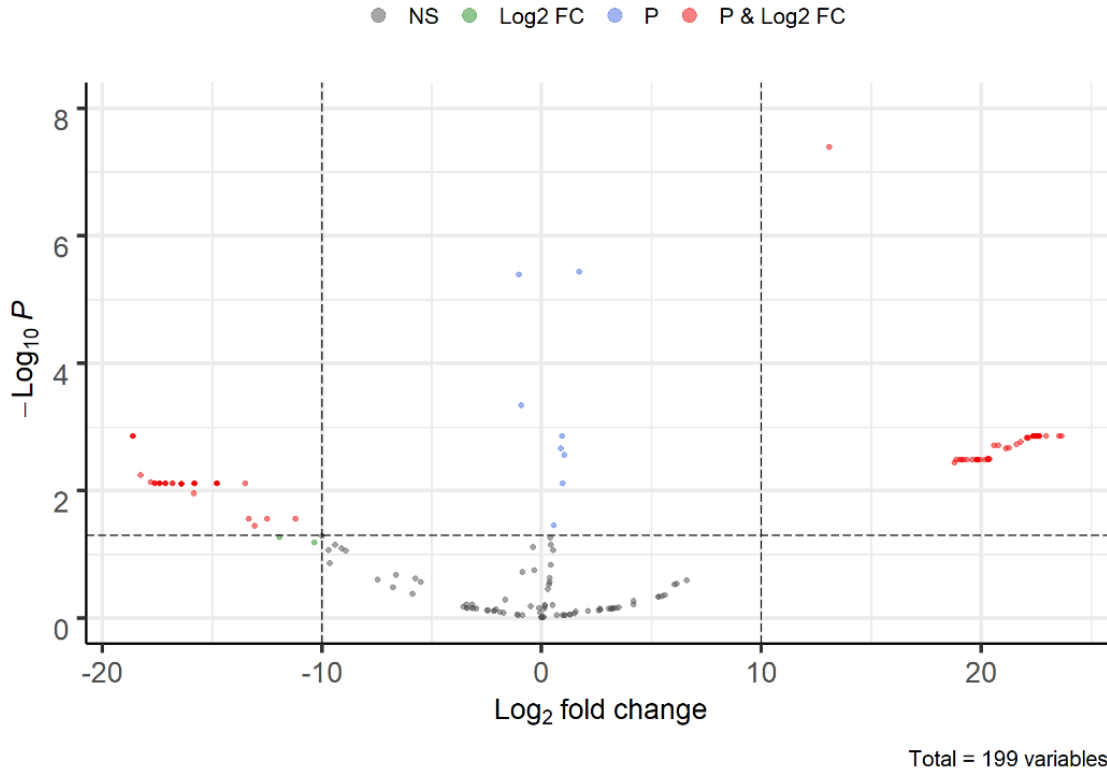
Figure 96. Volcano plot of the fold change of individual clones detected in the liver(week14 / week 4); Y-axis = FDR, X-axis = $\log_2$FC.

To probe into the relationship of specific lineage dynamic in the primary tumor contributing to liver metastasis, we plotted lineage dynamics in the primary tumor of all the lineages (passed the statistical robust test) that can be found in the liver. Furthermore, we bin the clones into four quartiles based on their relative abundance representation of week 12's primary tumor (0-25%, 25%-50%, 50%-75% and 75% - 100%); and then for each quartile, we performed clustering based on individual lineage dynamics. At the same time, each lineage's relative abundance at early stage (week 4) is indicated by color (0-25% in purple, 25% - 50% in blue, 50% - 75% in green, and 75%-100% in yellow).

From Figure 97 below, we can see that the lineage dynamics of "loss of fitness in the liver over 10 weeks period" mostly belongs to the 0%-50% primary tumor group (the blue color

lineages, with "FC < -10"); however, this dynamic is not exclusive to this quartile and could be found in all four quartiles of abundance in the primary tumor. This suggests that, regardless of their cell abundance in the primary tumor, many clones, at week 4, can "seed" and "anchor" in the liver but not necessarily thrive. Moreover, most clones that lose fitness in the liver (blue lineages) does seem to lose fitness in the paired primary tumor over time as well. This is especially apparent in the lineages that are in the 0-25% quartile, with exceptions of few (blue) lineages in the more abundant (50% - 75%, 75-100%) quartile. This suggests that a lineage's fitness in the primary tumor does have a positive correlation even after dissemination to distal organs.

On the other hand, the clones that display a "gained fitness" in the liver (depicted by the red lines) can be found spreading across all quartiles of the primary tumor. This indicates that although there certain clones are at low abundance in the primary tumor, can still seed and expand well in the liver. That being said, in general, we do see the majority of clones that gained thousands fold in abundance in liver over time mostly lies in the 50%-75% quartile in the primary tumor, suggesting that clonal primary tumor abundance is also a key factor contributing to the chance for lineage expansion in liver. As for the relationship of clonal fitness in the primary tumor vs. liver metastasis, whether or not it is an intrinsic property, we will delve into that in the next section.

As for the clones that did not show extreme fold change of abundance in the liver over time ($|logFC| > 10$, the yellow clones), we suspect that these clones maintain relatively high in abundance from seeding to late stage in the liver (week 4 to week 14). This is due to the volcano plot emphasize only on the fold difference, and the fact that most of the yellow colored lineages are from the 75%-100% quartile in the primary tumor.
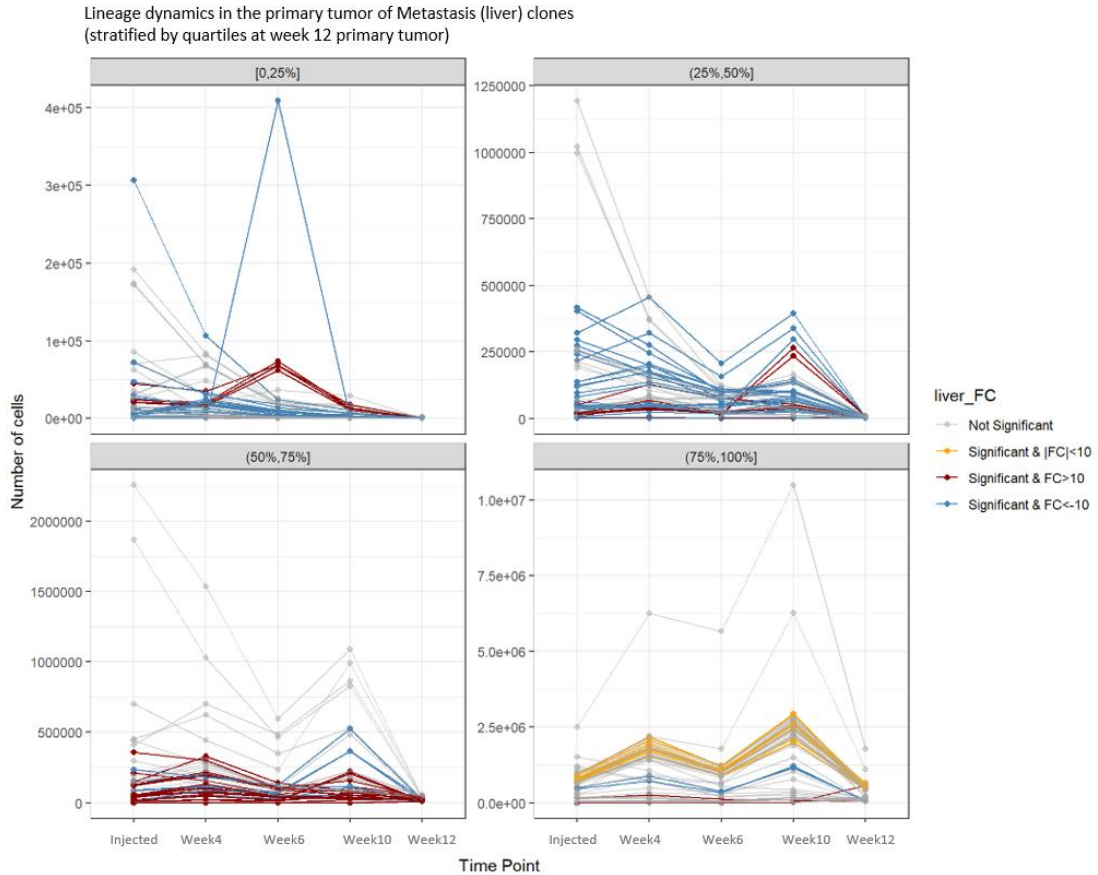
Figure 97. Lineage dynamics in the primary tumor (quartiles based on week 12 primary tumor) of the clones that are able to metastasize. Blue lineages indicate the clonal lineages that lose fitness in the liver from week 4 to week 14 ($log_2FC < -10$); red lineages indicate the lineages that gained fitness ($log_2FC > 10$); the yellow lineages indicate clones with $|log_2FC| < 10$.

Lineages that display loss of fitness in the primary tumor also exhibit a loss in fitness in the liver metastasis

Figure 98 below plots the dynamics of the lineages in the primary tumor that have the ability to metastasize based on the abundance of primary tumor at week 12 in quartiles, with each quartile further clustered to appreciate distinct lineage dynamics. The color of the lineage indicates whether the lineage has a loss (blue) or gain (red) a significant level of fitness. From Figure 98, we can see that most of the lineages that has a loss in fitness over time in the liver (blue) falls in the primary tumor abundance quartiles of 0-25%, and most lineage dynamics in that

135

quartile display a transient amplification, or, a continuous loss of fitness over time. On the other hand, most lineages that has significant gain of fitness and abundance over time in the liver metastasis falls in the quartiles of 50%-75% or 75% - 100% primary tumor abundance, where most dynamics display an alternative gain and loss of fitness, with some lineages gradually gaining abundance over time.
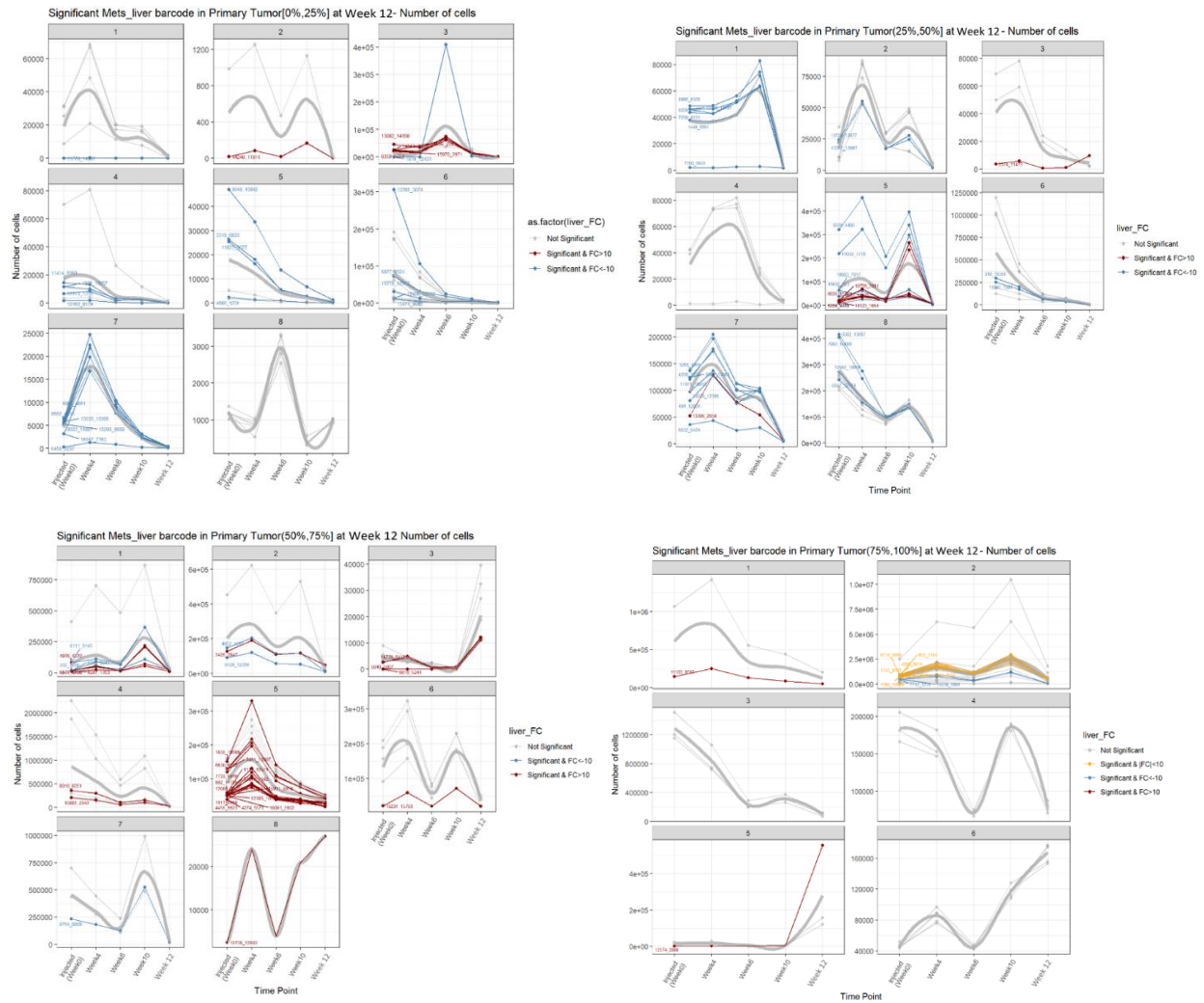
Figure 98. Lineage dynamics in the primary tumor that have the ability to metastasize to liver. Data stratified in four quartiles based on the abundance of primary tumor at week 12, and each quartile are further clustered to appreciate distinct lineage dynamics. The color of the lineage indicates whether the lineage has loss (blue) or gained (red) a significant level of fitness

Intrinsic lineage properties guiding the growth of metastasis that is independent of seeding abundance

By plotting out the cell number of clones that are in the liver at week 4 and week 12, and then stratify the data into 3 bins based on liver metastasis abundance (0-50%, 50%-75% and 75% to 100%), we can see that most clones that are able to disseminate to the liver can grow. Furthermore, this phenotype of growth seems independent of the initial "seeding" abundance at week 4, where lower abundant clones could also actually gained in cell number over time. On the other hand, there are clones with initially higher number "seeded" cells that actually decreased in abundance over time. This suggests there are intrinsic lineage properties guiding the growth of metastasis, independent of seeding abundance.
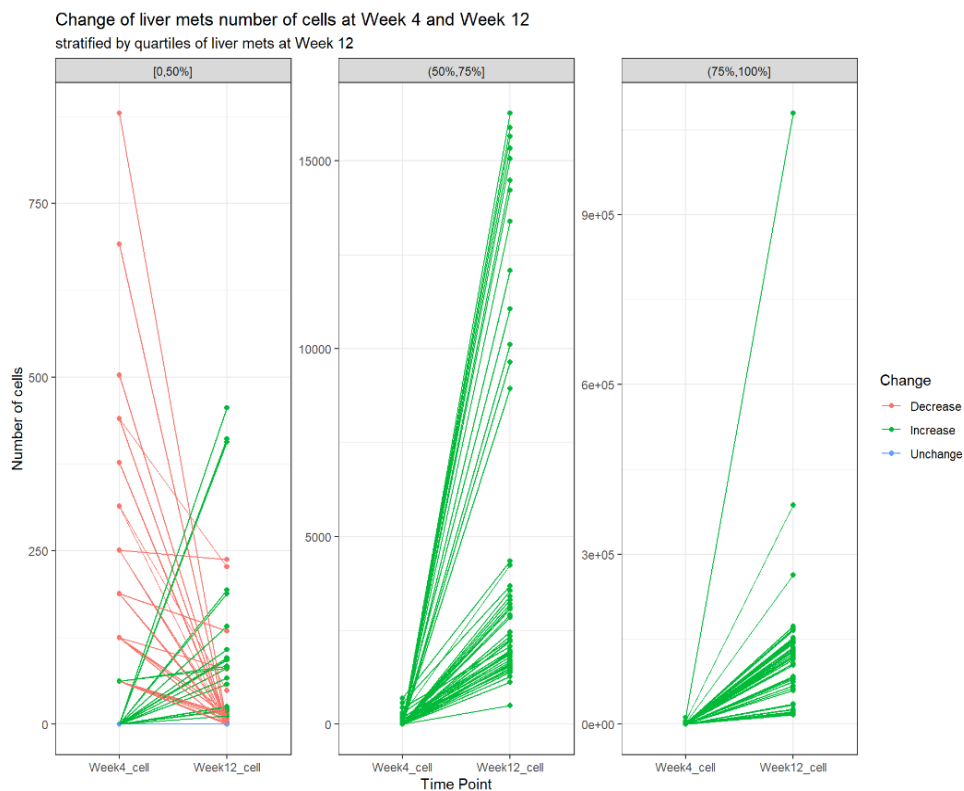
Figure 99. Cell number of clones that are in the liver from week 4 and week 12. The data is stratified into 3 bins based on liver metastasis relative abundance representation (0-50%, 50%-75% and 75% to 100%). Green represent lineages gained in cell abundance; red represents lineages that lost abundance comparing week 12 to week 4.

Liver metastasis lineages mostly exhibit a burst and exponential growth dynamic, and the burst in growth might be linked to primary tumor dynamics

By plotting the clones that are able to metastasize to the liver and their cell abundance over time, and then binning into three bins based on liver metastasis relative abundance representation in % (0-50%, 50%-75% and 75%-100%), we observe that most clones, once disseminated into the liver stayed relatively stable followed by a burst of exponential outgrowth at week 10.
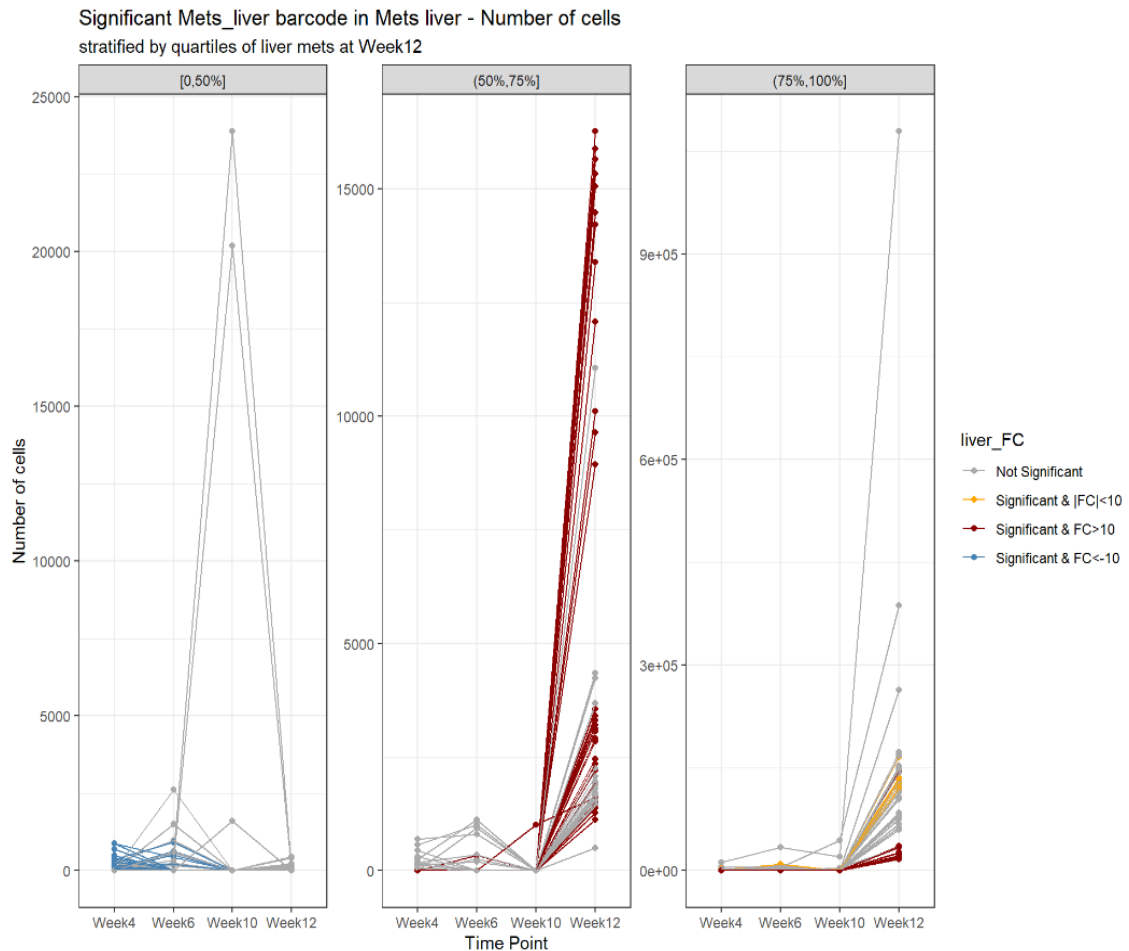
Figure 100. Lineage growth dynamics in the liver over time. This data is to be looked alongside the volcano plot in the previous section. Clones that have extreme change in abundance are

displayed in different colors (red = positive fold change of $\log_2FC > 10$, blue = $\log_2FC < -10$, yellow = $|\log_2FC| < 10$, grey are deemed not significant).

<u>Liver metastasis clonal growth dynamic seems to be more straight-forward, where no alternating dominance observed</u>

By clustering the lineage dynamics over time in the liver (

Figure 101) and overlay the clusters on top of the volcano plot ($\log_2FC$ between week 12 vs. week4 and their FDR, Figure 102), we can see in the clusters of lineages with $\log_2FC > 10$, which are the lineages responsible for metastasis outgrowth, mostly exhibit one type of growth pattern – initially dormant until week 10, and then followed by explosive exponential growth. As for the lineages in the cluster with $\log_2FC < -10$, which are the "transient metastasis" lineages, they display the dynamic of decrease in cell number over time. Considering most metastasis lesions in this model are comprised of monoclonal or oligoclonal lesions, this may suggest the competition amongst clones in the liver are not as complex as the clones in the primary tumor. The relatively long dormancy before outgrowth could be contributed by the lag in adapting to a new niche favoring growth, and also the fact that seeding in the liver is likely spatially sparse.
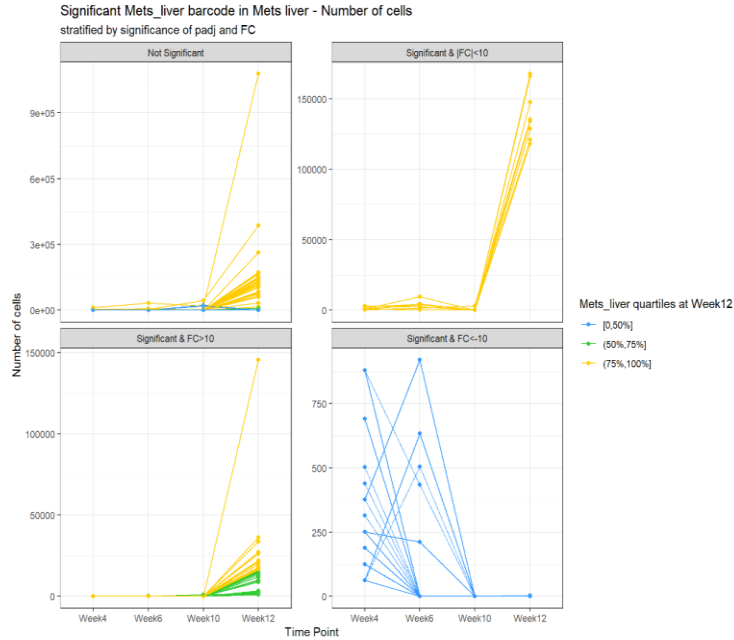
Figure 101. Longitudinal clonal dynamics in the liver (quantitative analysis). Lineage that metastasize to the liver in cell number are clustered based on the overall abundance fold change comparing week 12 vs. week 4. Color represents lineage representation in the liver (yellow =more abundant, blue=lowly abundant).
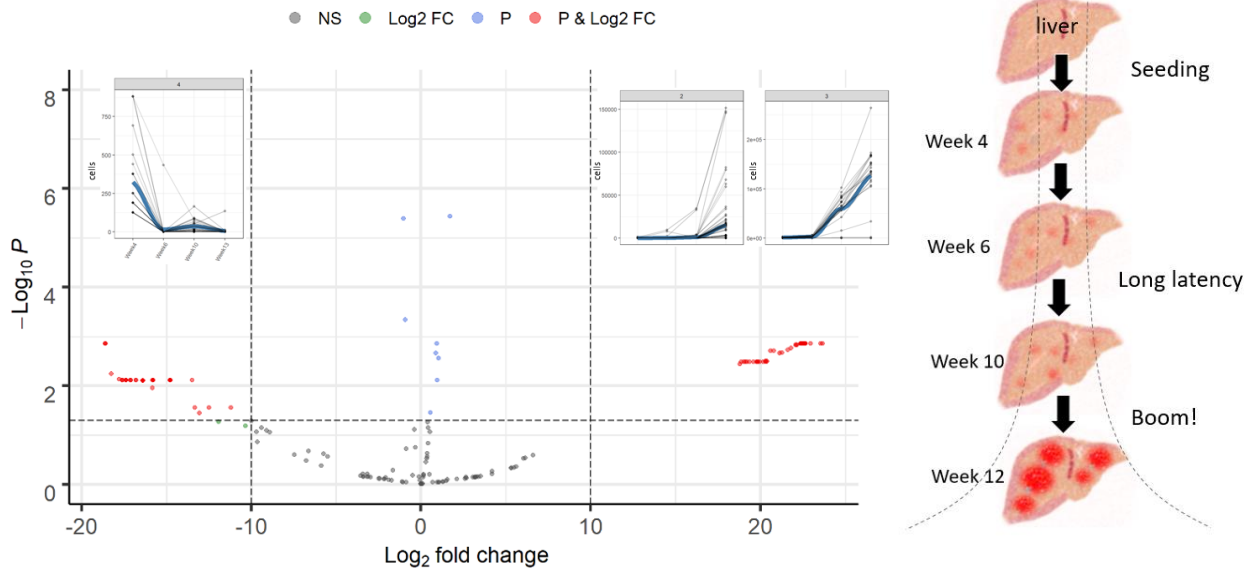
Figure 102. Longitudinal clonal dynamics in the liver (quantitative analysis) overlay onto the volcano plot displaying lineage fold change comparing week 12 vs. week 4 in liver. The volcano plot: Each dot represents a unique sub-clone, the fold change (FC) of lineage dynamics in the primary tumor according to the increase or decrease in abundance at the liver from week 4 to week 14. False discovery rate is calculated and displayed on y-axis ("Significant" – indicating high

confidence in a change, or "not significant" – indicating not confidant in the observed change). The lineage dynamics in the each black box are matched dynamics of those clones in the liver over time. The lineage dynamics of the out growing metastasis is straight forward, where it went through long dormancy followed by sudden outburst of proliferation (upper right box). The clones that displayed "transient metastasis" displayed also a straight forward dynamic of loss of abundance over time. Right schematic is a cartoon representation of the growth dynamic in the liver.

Liver and Lung metastasis longitudinal clonal fitness dynamics - qualitative analysis of clonal dynamics in terms of relative fitness

Instead of analyzing clonal abundance quantitatively (in "cell units") over time across the CRT cohorts. Here, we analyze each clone in terms of their relative fitness (percent representation). Looking at clonal abundance quantitatively (in "cell unit") gives us a more accurate representation of the tumor's "expansion dynamics in mass". On the other hand, by analyzing the clonal abundance qualitatively (in % representation of sample) allows us to appreciate the relative fitness amongst clones, and therefore, provide data in the context of "which clones are dominating", or in other words, "who is winning" in a particular sample regardless of the total tumor mass. Qualitative analysis of clonal dynamics across tumors would reflect a normalized clonal behavior in terms of "tumor composition" and provide insights more related to the clonal "competition dynamics" longitudinally.

Again, the normalization process of barcode is straight forward, when we sequence barcoded tumors using NGS, it will provide a "read number" for each unique barcode. To normalize each barcode to "percent representation of the total sample", we divide the "read number" of a particular barcode by the "total sum of all reads" of all the barcodes. For instance, if a tumor has only two barcodes "A" and "B" and the NGS read of clone "A" and "B" is 100 and 50 reads respectively, then, clone "A's" relative % tumor representation is 100 ÷ (100+50) = 66.6%, and clone "B" is 33.3%.

Metastatic clones may gain a unique dynamic of fitness once in a different environment – qualitative analysis

Analyzing data in terms of clonal relative fitness of primary tumor. Of those sub-clonal lineages with the ability to metastasize, we observe that there are clones, in the primary tumor that have a dynamic of loss in fitness, as well as clones that have a gain in fitness overall. This suggests that overall clonal fitness loss or gain alone may not be enough to be used as an indicator to determine sub-clone's ability to disseminate and grow. This also suggests clones may gain a unique dynamic of fitness once in a different environment.
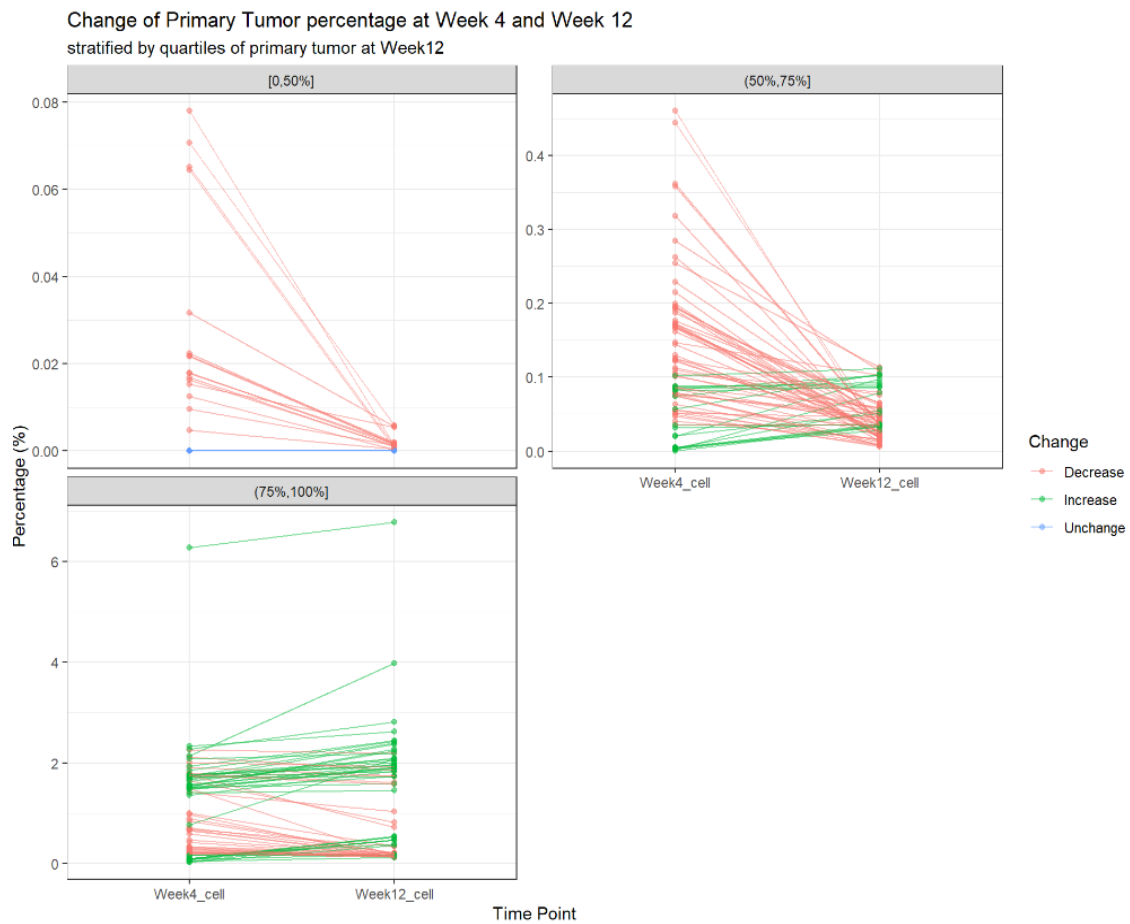
Figure 103. Liver metastasis clones' overall gain or loss of abundance in the primary site. Breaking the primary tumor in week 12 in to quartiles, plotting out all clones that have the ability to metastasize to the liver and whether or not the fitness decrease (in red) or increase (in green) from week 4 to week 12.

Liver metastasis – a comprehensive view of lineage dynamics in liver

Here we cluster the entire clonal relative fitness dynamics over time in the liver. A few major clonal fitness dynamics were observed (Figure 104): 1. Gradual loss of fitness; 2. Alternating fitness dynamic of gain-loss-gain; 3. Gradual loss and followed by significant gain in fitness; 4. A relative stable fitness overall.
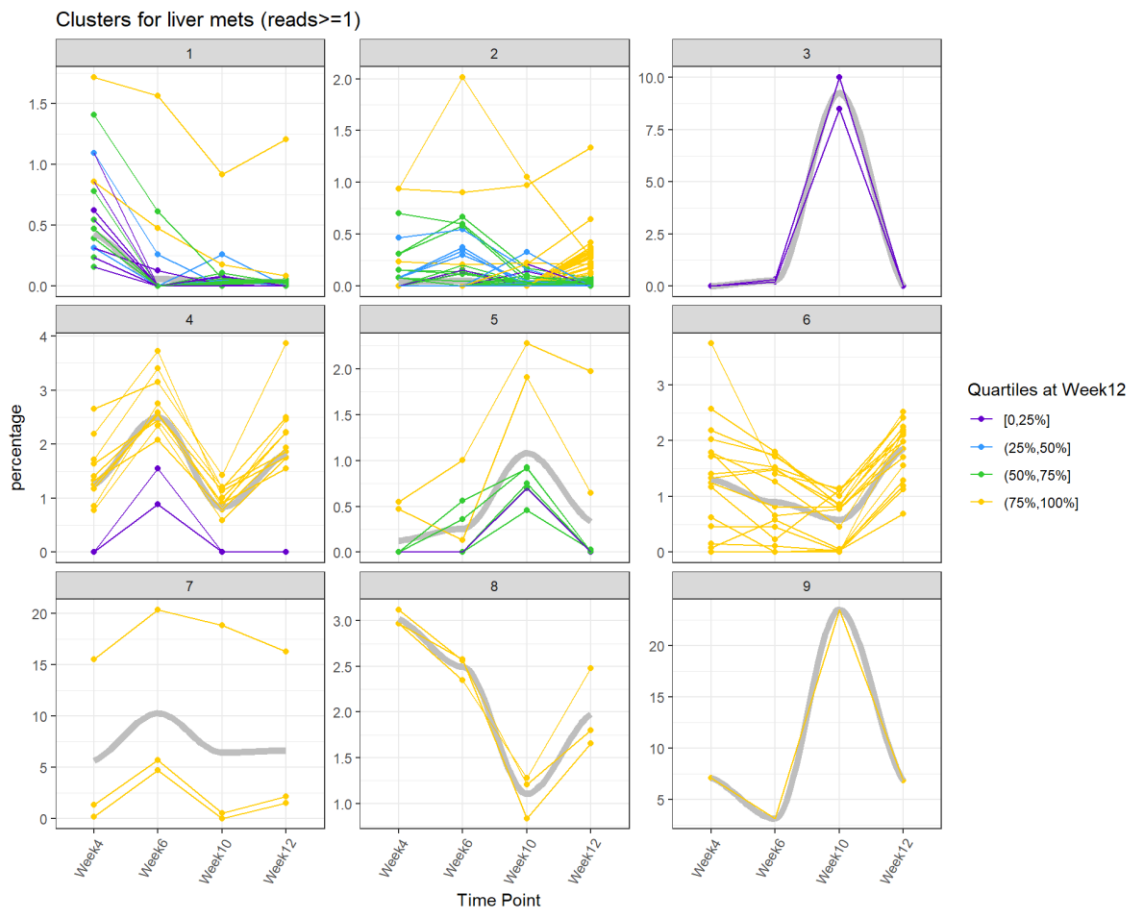
Figure 104. Relative clonal fitness dynamics in the liver over time. Each line represents a unique lineage, where the Y axis indicates relative % of liver metastasis representation. The color of the line indicates the relative abundance of a lineage in the primary tumor at week 12.

Lung metastasis – a comprehensive view of lineage dynamics in liver

Here we cluster the entire clonal relative fitness dynamics over time in the lung. A few major clonal fitness dynamics were observed (Figure 105): 1. Loss of fitness; 2. Loss of fitness followed by gain of fitness; 3. Sudden gain of fitness at a certain time point; 4. Gradual increase in fitness; 5. Dormancy followed by significant gain of fitness. This data also reflects the stochastic nature of the lung metastasis, where some of the fitness dynamic may not seem as smooth or gradual as those observed in the primary tumor.
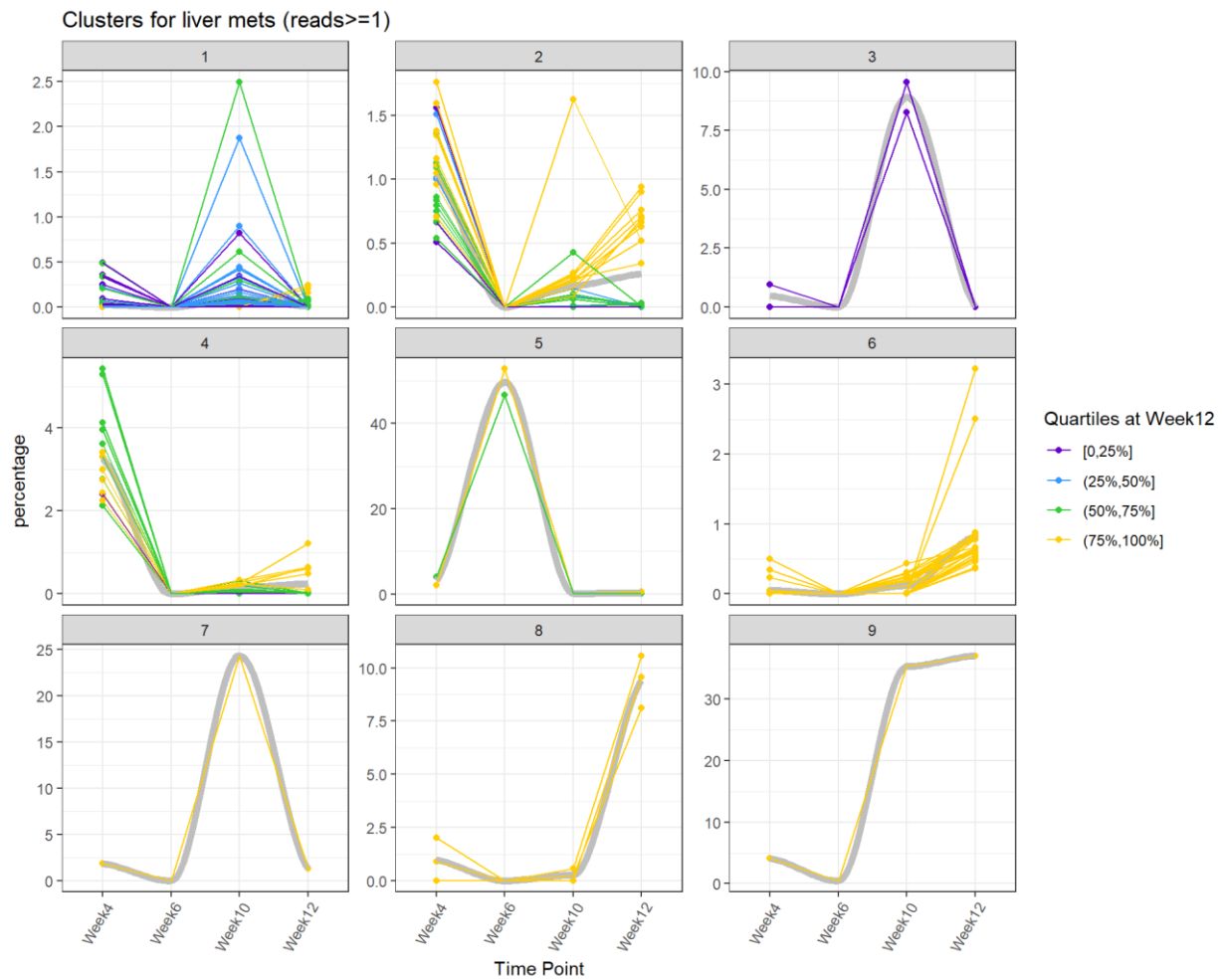
Figure 105. Relative clonal fitness dynamics in the lung over time. Each line represents a unique lineage, where the Y axis indicates relative % of lung metastasis representation. The color of the line indicates the relative abundance of a lineage in the primary tumor at week 12.

A comparison of the complexity of lineage dynamics across primary tumor, liver and lung metastasis – liver and lung display a lesser degree of clonal competition with simpler dynamic

Here we generate a clonal fitness matrix plot by looking at each lineage's gain or loss of fitness comparing to its previous time points or collection. From Figure 106 below, each dot represents a unique clone, and with the colors representing gain (red), loss (blue), or unchanged (green) in fitness. In the primary tumor, at each time point, the matrix plot has a relatively mixed pattern of lineages gaining (red) or losing fitness (blue), or relatively unchanged (green). On the other hand, liver and lung display a much simpler and uniform change in dynamic at each time point. For example, almost all clones experience a decrease or unchanged fitness (blue and green) before week 10, then almost every clone experience a gain in fitness (red). This suggests at each time point, overall, clonal fitness dynamics in the primary tumor are more complex than those in the liver or the lung. This also suggests clones in liver or lung experiencing a lesser degree of clonal competition; however, this also suggests the clones in liver or lung may have to overcome environmental pressure before expansion.
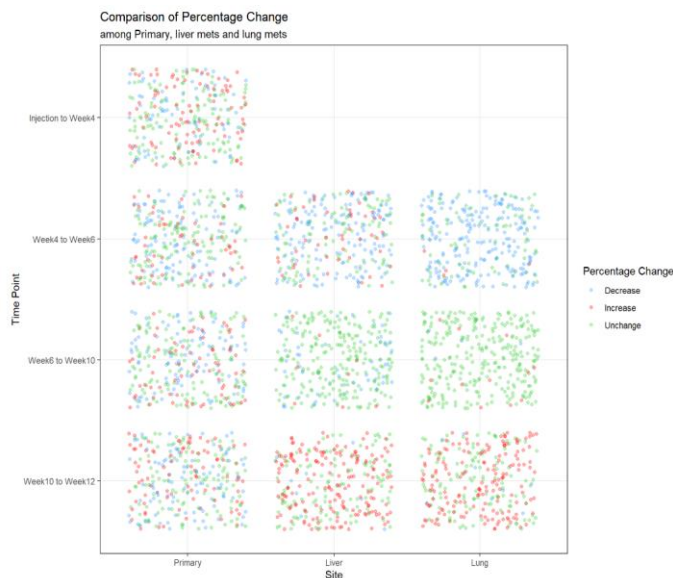
Figure 106. Lineage matrix plot showcasing the degree of overall clonal fitness dynamics in the primary, liver and lung. Each dot represents a unique clone, and each color represents gain (red), loss (blue) or unchanged (green) in fitness comparing to the previous time point.

For most lineages, abundance in the primary tumor has a significant positive correlation with dissemination

Overview of clonal abundance over time vs. liver metastasis represented by t-SNE

For a high level overview of how the clonal lineages evolve, we transformed the clonal data point into t-SNE plot. In the t-SNE plot, each dot represents a unique clone and the color indicating the abundance of the clones (red = higher, grey = exhaust). We can see that over time, there is a decrease in the level of clonal heterogeneity (grey clones exhausting) and that only a sub-cluster of the clones dominating the primary tumor until week 12. Interestingly, the clonal complexity seems to stabilize around week 10 and did not change much from week 10 to week 12. By looking at the clones that metastasized into the liver at week 12, we see that most clones of the clones, if not all, belong to a subset of the lineages that are more abundant in the primary tumor at week 10 and week 12. From this analysis, it seems to suggest that the relationship of clones that are able to expand and dominate the primary tumor vs. the clones that are able to establish distal metastasis in the liver are positively correlated.
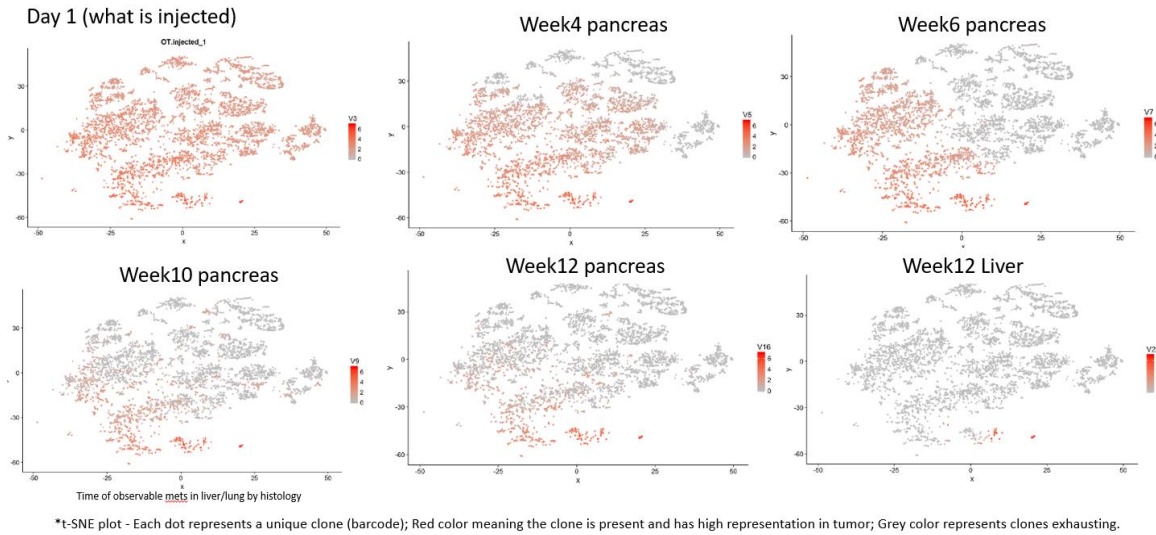
Figure 107. T-SNE plot of clones and their abundance over 12 weeks post injection. Red color represents the clones with higher representation in the tumor; the grey color represents the clones losing fitness and exhausting (plot by Nick Yen).

Tumor representation in the primary site (pancreas) of lineages that are able to disseminate – abundance has a positive correlation

To assess the relationship between clones that are able to disseminate (that can be found outside of the primary tumor) vs. clones that cannot be found outside of the primary tumor, we separate the two populations of clones and look at their relative accumulated abundance in the primary tumor. In addition, we plot each individual clone's relative tumor representation (%, last day's primary tumor) and look at their distribution in the primary tumor. From the analysis, we can see that, in general, clones that are able to disseminate have a higher accumulated primary tumor representation. However, despite that most of the clones that can disseminate having positive correlation with their primary tumor abundance, there also exists clones with low abundance that could disseminate, as well as, clones that are highly represented in the primary tumor that does not disseminate. This suggests clonal abundance could increase the chance of dissemination in general, however, there also exist lineage intrinsic factors guiding the process of dissemination and metastasis.

147
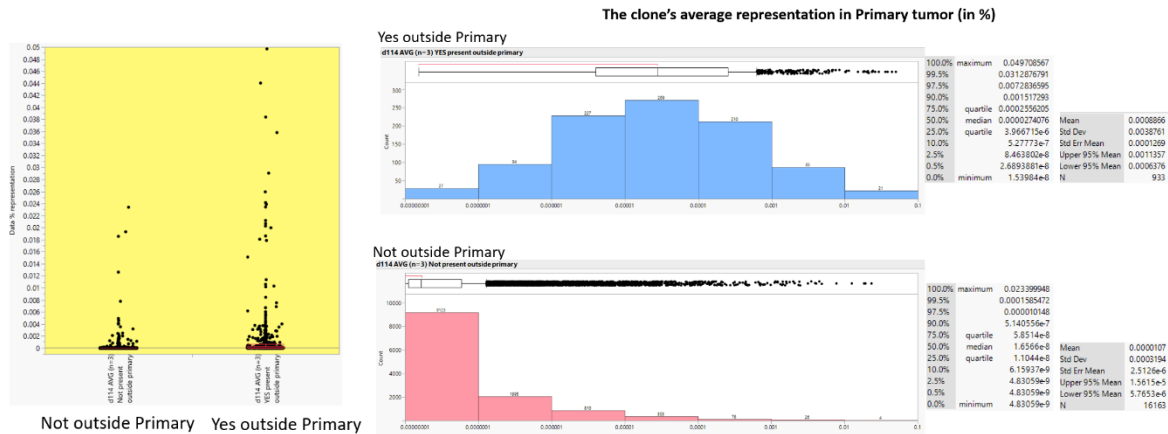
Figure 108. Relationship between lineage abundance vs. dissemination. Clones are separated into two categories: "able to disseminate" and "not able to disseminate". Left plot looks at individual lineage and their individual relative tumor representation (%) in the primary tumor (last time point); Right graph shows the distribution of clones and their relative tumor representation in the two groups. X-axis indicates % representation bins in log scale; Y-axis indicates the count of clones that falls into each specific bin.

Majority of the lineages that make up most of the tumor mass are found in the liver and/or the lung
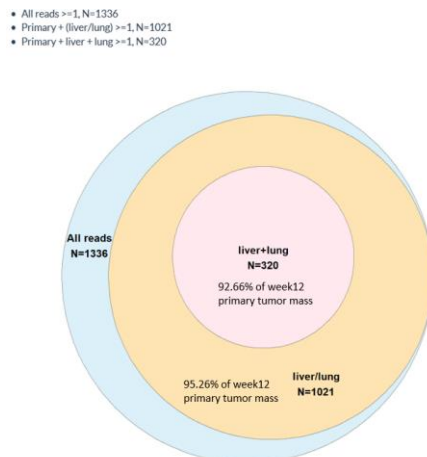
Figure 109. Diagram displaying primary tumor abundance representation of clones that are able to metastasize to the liver and/or the lung. Using another set of cutoff to analyze the data (NGS read>1, primary tumor must have value at all the time points, as well as appended an extra data set at week 10 for liver and lung metastasis); the conclusion remains the same, where the

accumulative % representation in the primary tumor of those lineages having the ability to metastasize to either liver and/or lung are high (92.66% for liver and lung lineages, 95.26% for liver or lung lineages).

Looking at individual lineage's clonal abundance in the primary tumor vs. in the liver

Below heat map shows the % abundance of each clone in the primary tumor from day0 to week 12, as well as the clonal abundance in the liver at the last time point (week 12). The heat map is further sorted based on the abundance of liver metastasis representation from high to low. From this analysis, we can gauge in the relative correlation of individual clone's abundance vs. their abundance in the liver. As we can see, the clonal abundance seems to have a positive correlation, consistent with the various analysis before; however, very importantly, not all of the clonal lineages abide by this rule. Once again, there are clones that are highly abundant in the primary tumor but are not able to expand in the liver; on the other hand, there are clones that are lowly represented in the primary tumor but are able to establish a relatively dominant presence in the liver. The X-Y correlation graph below show a similar story, where in general, there seems to be a positive abundance correlation between primary tumor and liver. The right most plot in Figure 110 shows the Venn diagram of overlapping clones and their cumulative tumor representation in the two CRT animals at the last time point. As seen from the analysis, the overlapping clones in the two CRTs makes up ~97% of the primary tumor mass, whereas the overlapping clones in the lung and liver represents 63% and 68% of the primary tumor mass. In other words, clones that makes up 63% of the primary tumor could be found in the lung, and clones that makes up 68% of the primary tumor could be found in the liver.
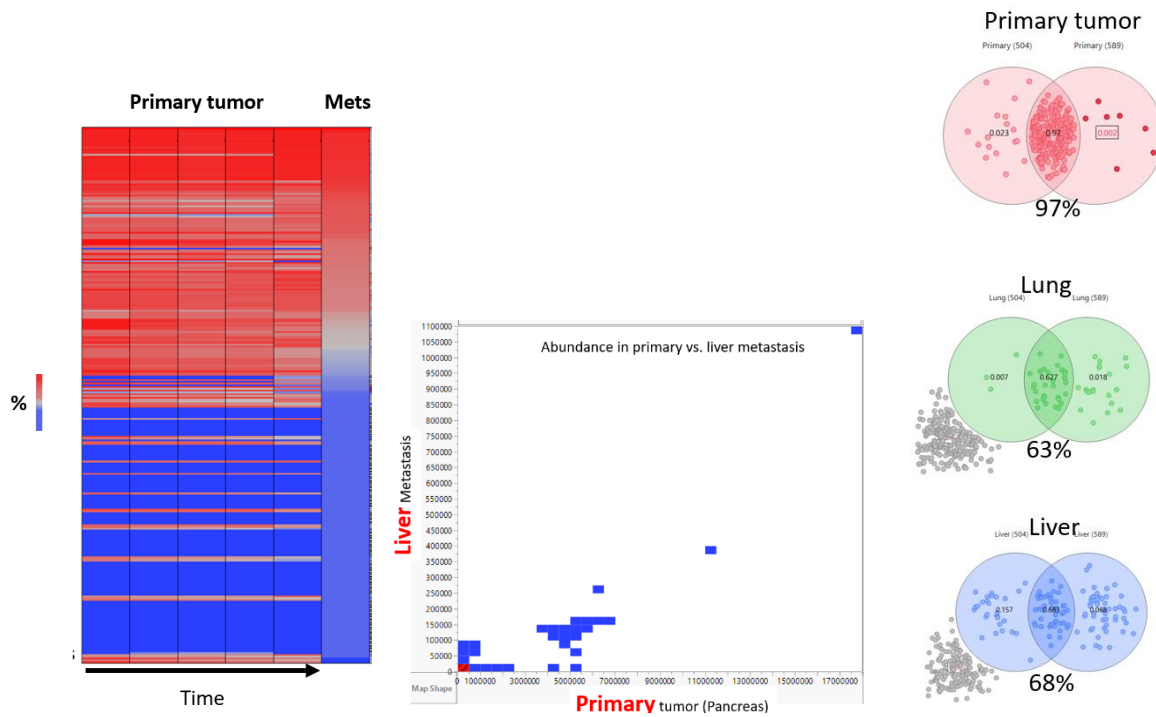
Figure 110. Graphs showing clonal abundance vs. dissemination outcome. In the left heatmap, from left to right on the X-axis, indicates primary tumor day 0, week 4, week 6, week 10 and week 12, as well as the week 12 liver on the right, where lineage abundance is sorted according to the liver sample. Each row represents a unique lineage. The red color indicates clones that are more abundant (left graph); XY - plot of clonal abundance (cells) found in liver vs. clonal abundance in the primary tumor. Color indicates the barcode count that falls into specific bin (red = high, blue = low). From the figure, we see a general positive correlation between the clone abundance in primary vs. liver metastasis (middle graph); the right most graph shows the two CRT mouse's common clones' cumulative tumor representation (averaged between n=2).

Additional data – cell abundance heaTCFP
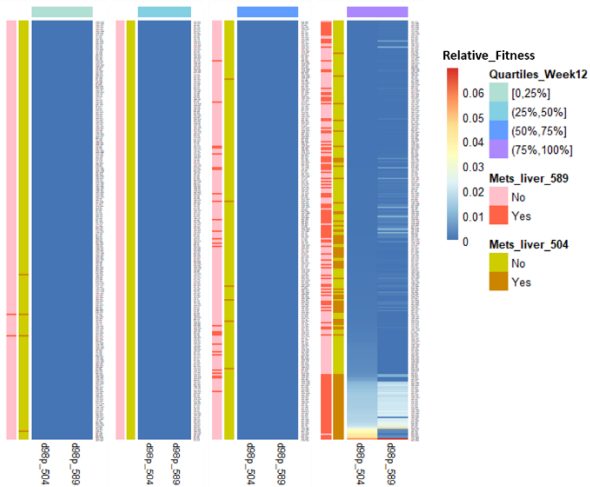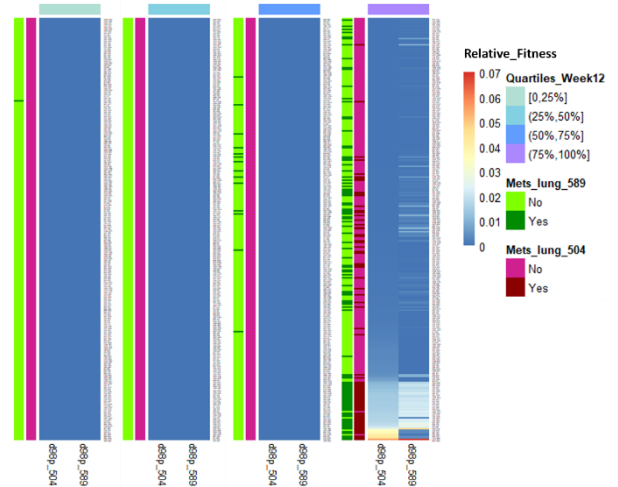
Figure 111. Heat map of clonal relative fitness in the primary tumor vs. liver (left)/lung(right) metastasis outcome. Heat map displaying the relative clonal fitness in the primary tumors at late time point (week13) for two CRT mice (mouse 504 and mouse 589). The metastasis outcome in the livers (left)/lung (right) in both mouse is indicated at the left side of the heat map. The metastasis outcome in both CRT mouse in the livers are consistent in general, however, there are also stochastic nature observed in metastatic outcome, indicating metastasis could be influenced both by clonal intrinsic and extrinsic factors.

Analysis of common clones shared amongst different dissemination organs – blood, liver, lung and primary

In order to have an idea of the clonal overlapping pattern between different metastasis sites, we took the clones that make up 99.5% of the primary tumor mass and performed clonal complexity overlapping analysis between blood, liver, lung, and primary tumor. We also displayed the number of clones shared amongst different sites and their cumulative relative primary tumor representation (%). In this analysis, the two CRT animals collected at each time points are analyzed independently. From the Venn diagram below, we can see that clones that make up the majority of the primary tumor mass (~85% to 88%) can be found in the liver, lung or blood. The

common clones between lung and liver make up a large portion of the primary tumor (61.6-62.8%). Liver has more unique clones that are in common with the primary tumor.
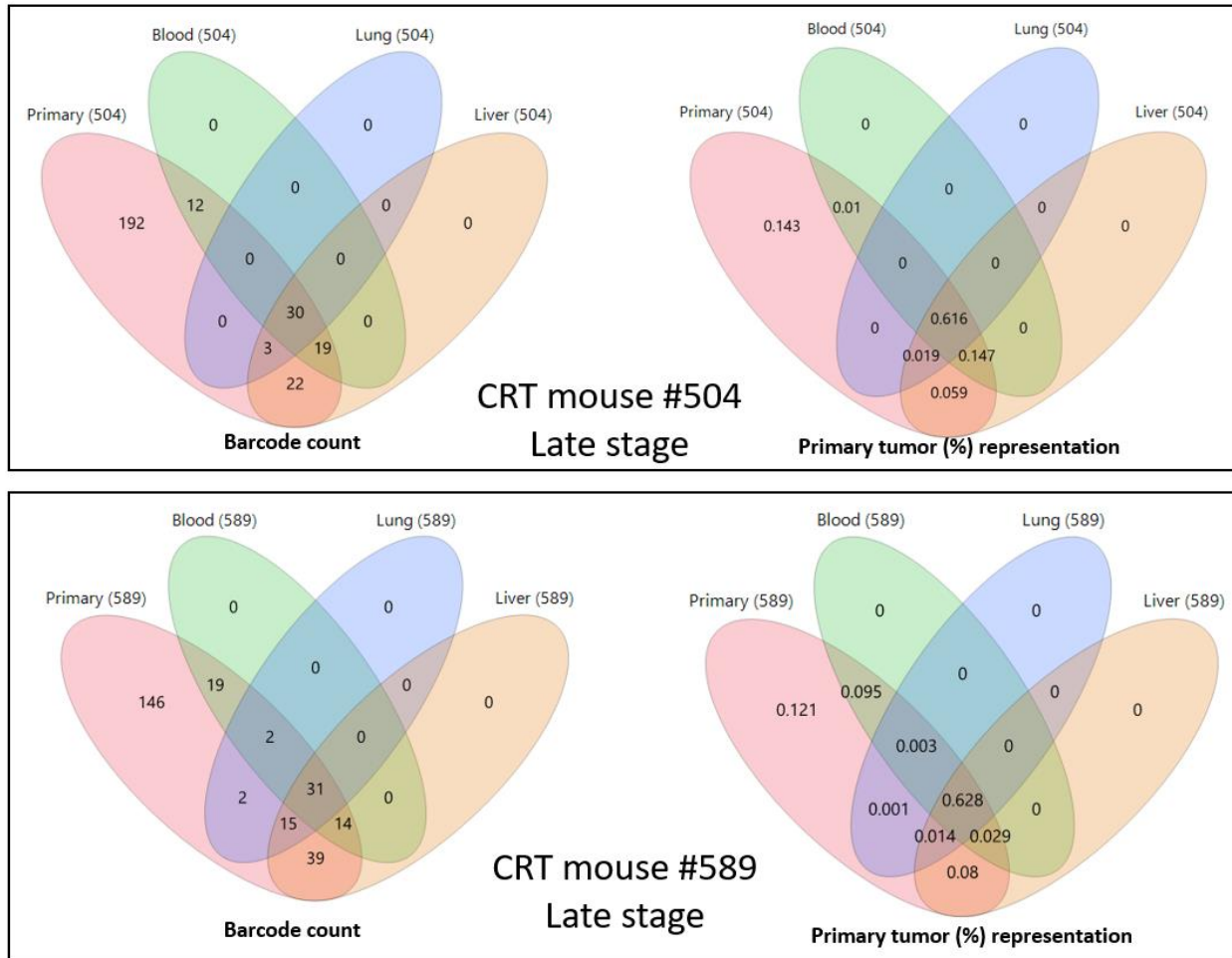
Figure 112. Venn diagram of the number of common lineages shared between primary tumors, blood, live and lung at last time point of clonal tracking experiment. The top graphs represent CRT mouse tag#504 while the bottom represents mouse tag#589. The left diagrams' number indicates the count of unique lineages shared; the right diagrams show the cumulative % primary tumor representation of those common clones.

Other analysis was performed that captured the unpredictability of metastasis growth in the liver and lung, however, it is not discussed here in detail due to the complex nature of the data. An example of interesting observation suggests that liver metastasis reflects the primary

tumor's clonal distribution more than the lungs. Also, the clonal correlations between liver vs. liver in the CRT animals at the same time point are higher than the clonal correlation between lung vs. lung, which suggests lung metastasis is more stochastic than that of the liver.

Figure 113. Pearson correlation of all samples in the CRT cohorts, including *in vitro* passaging and serial transplantation. Clonal abundance converted into "cells". Suffix ending with "Pan or "panc" indicates pancreas primary tumor; suffix ending in "D" indicat "blood"; suffix ending in "G" indicates "lung"; suffix ending in "R" indicates "liver".

Primary tumor, nearby vessels and duct metastasis tumor nodules are high in lineage complexity

In a few animals in the CRT cohorts, we observed "local" invading metastasis along vessels and ducts in proximity to the primary tumor. In order to survey and have an idea of the clonal composition of these locally invasive tumor nodules, we collected them and sequenced its barcode. Summarized in Table 2, looking at the Pearson's correlation of individual lineage abundance between each samples (primary, liver, lung, blood and local invasive vessel/duct nodules), we see the highest correlation value between the primary tumor and the liver (Peasron's ~0.93). This is consistent with the observations across different animals that liver recapitulates the clonal abundance representation of what is in the primary tumor better than lung (Peasron's ~0.77) or blood (Peasron's ~0.8). As for the locally invasive nodules along the vessels and duct, it also has a high correlation value with the primary tumor (Pearson's ~0.86), as well as a high correlation with of those lineages found in the blood (Peasron's ~0.9).

Moreover, looking at the barcode complexity and their distribution correlation in the X-Y scatter plot (Figure 115), we see there are many barcodes present in the locally invasion nodules. Albeit visually, from the picture during surgery (Figure 114), where we see only a few large nodules present, the barcoding data suggests that they are high in tumor lineage complexity. In addition, the lineages found in the local tumor nodules have high correlation with the lineages found to disseminate into the blood and in the primary tumor.
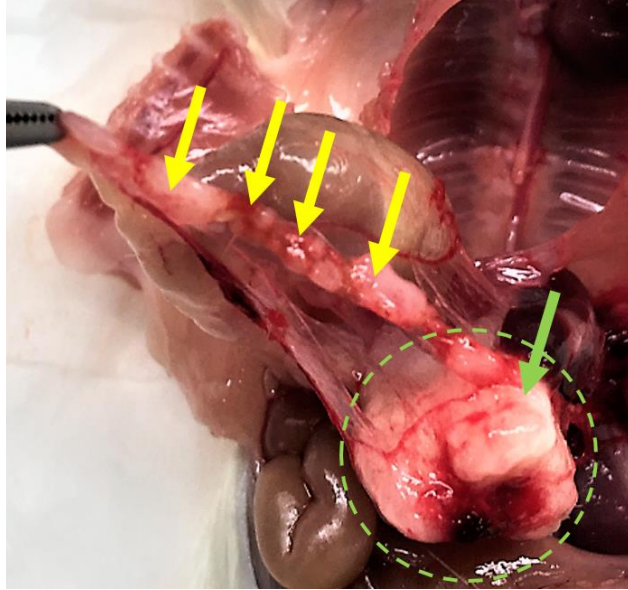
Figure 114. Image of the locally invasive tumor nodules along duct and vessels. The primary tumor is circled in green dotted line and with a green arrow pointed at it. The yellow arrows indicate the "locally" invasive tumor mass along vessels and duct.

Table 2. Paired analysis Pearson's correlation of clonal abundance between primary tumor, liver, lung, blood and locally invasive nodules. All samples are derived from one CRT in one mouse at week 10 post injection.

Pearson's (within one mouse)	W10 Blood	W10 Liver	W10 Lung	W10 Proxi ductal/vessel	W10 Primary
W10 Blood	1	0.7125	0.5657	0.898	0.7972
W10 Liver	0.7125	1	0.8823	0.7654	0.9314
W10 Lung	0.5657	0.8823	1	0.6464	0.7693
W10 Proxi ductal/vessel	0.898	0.7654	0.6464	1	0.8659
W10 Primary	0.7972	0.9314	0.7693	0.8659	1
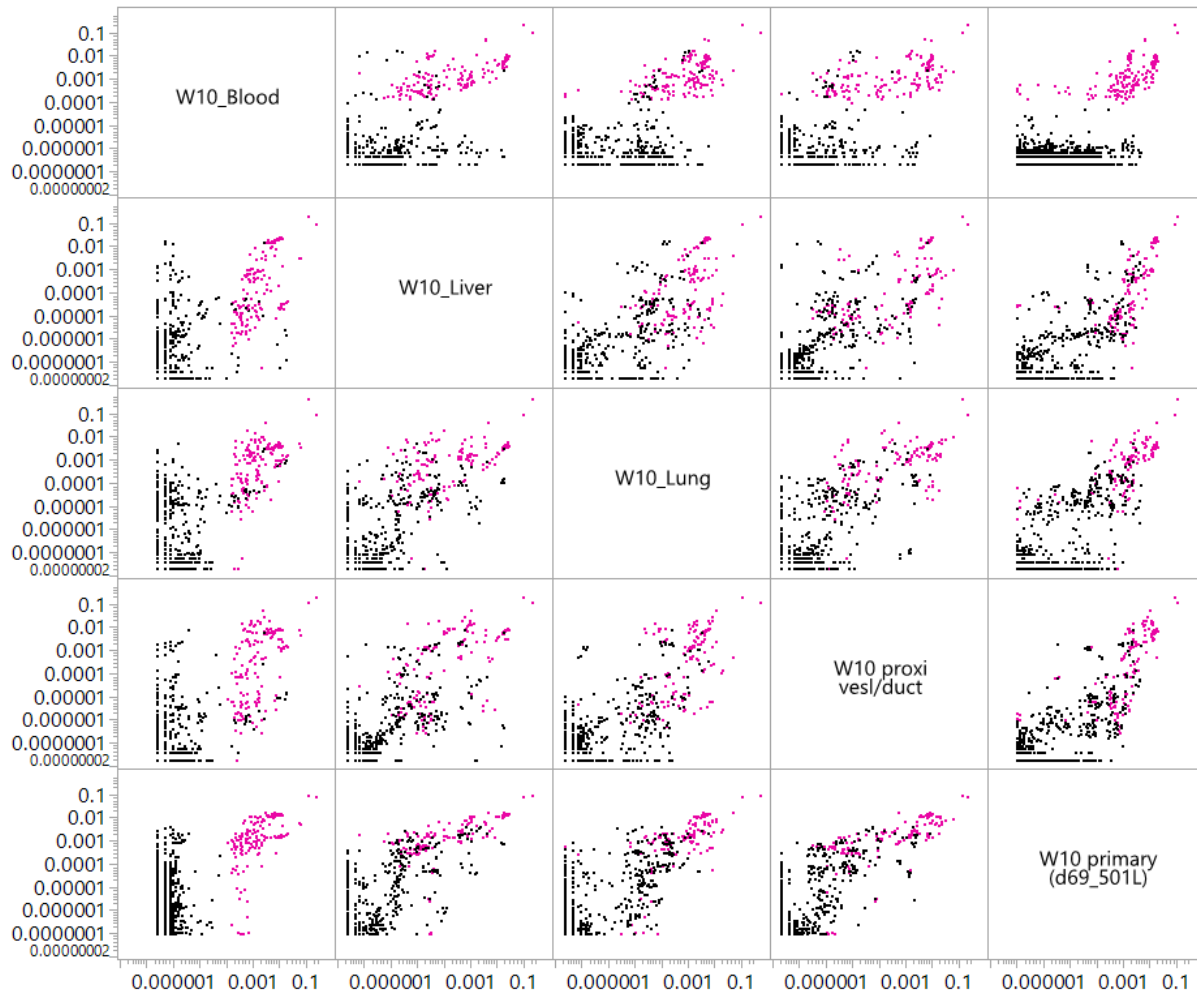
Figure 115. Pair-wise X-Y scatter plot of week 10 primary tumor, liver, lung, local invasive nodes and blood clonal abundance (in % representation). Within one single CRT (one of the week10 tumor CRT), we sequenced their primary tumor, blood, lung and liver, as well as the locally invading tumor nodules. The barcodes are converted to % representation within each sample. The axis is in log scale with the top abundant clones present in blood highlighted in "magenta" color.

Clonal fitness vs. gemcitabine sensitivity

<u>Lineages with less fitness are more susceptible to gemcitabine treatment</u>

Building upon the longitudinal clonal tracing aspect of the model, one question we could

asking is: "Are there a specific clonal dynamic that correlates to clonal gemcitabine response?"

To answer this, we treated, in parallel, another cohort of CRTs created with the same stabilized

barcoded cell lines that was used in the longitudinal metastasis study (PATC124) with gemcitabine (refer to Chapter 1 treatment condition). In this set of cohorts, barcoded cells were injected subcutaneously instead of orthotopically for ease of measurement of tumor volume vs. drug response. In addition, in the previous sections we demonstrated orthotopic tumor and and subcutaneous tumor have high correlation. The mouse baring tumors were treated with gemcitabine (Syd Labs) 100mg/kg IP every three days. After treatment for 2 weeks, the treatment was stopped and the tumor was allowed to relapse. By comparing individual clonal lineage's abundance change between relapsed and treatment null control, we can get a relative phenotype of the clonal sensitivity to gemcitabine. For example, if the clonal abundance is increased by at least 2-fold in the relapsed tumor compared to the none-treated tumor, we consider the clone is relatively more resistant to the drug, and vice versa. Therefore, by pairing the information together, for each uniquely barcoded lineage, we have a gemcitabine sensitivity phenotype that can associate with their longitudinal growth dynamics.

The histogram below captures the distribution of the treatment response in $\log_2$ fold change. The longitudinal tracing of relative clonal abundance in the primary tumor, from Day 0 to day 28, day 40, day 70 and day 98, are clustered and shown in the graph.

First of all, from the data, most of the clonal lineages have a negative value of fold change (the grey histogram), indicating that most of the lineages does have a decrease in representation and only a subgroup of lineages can dominate the relapsed tumor.

By looking at Figure 116 below, gemcitabine seems to be more effective on the clones that are being out competed throughout the three months growth period (the cluster of lineages with an overall downward trend has a gemcitabine sensitivity of negative value in fold change comparing relapsed tumor vs. untreated control tumor. I.e. the brown cluster). This could suggest that gemcitabine acts like an accelerator in helping the clones that have more fitness to dominate the tumor faster by inhibiting clonal growth of those that are less competitive.

158
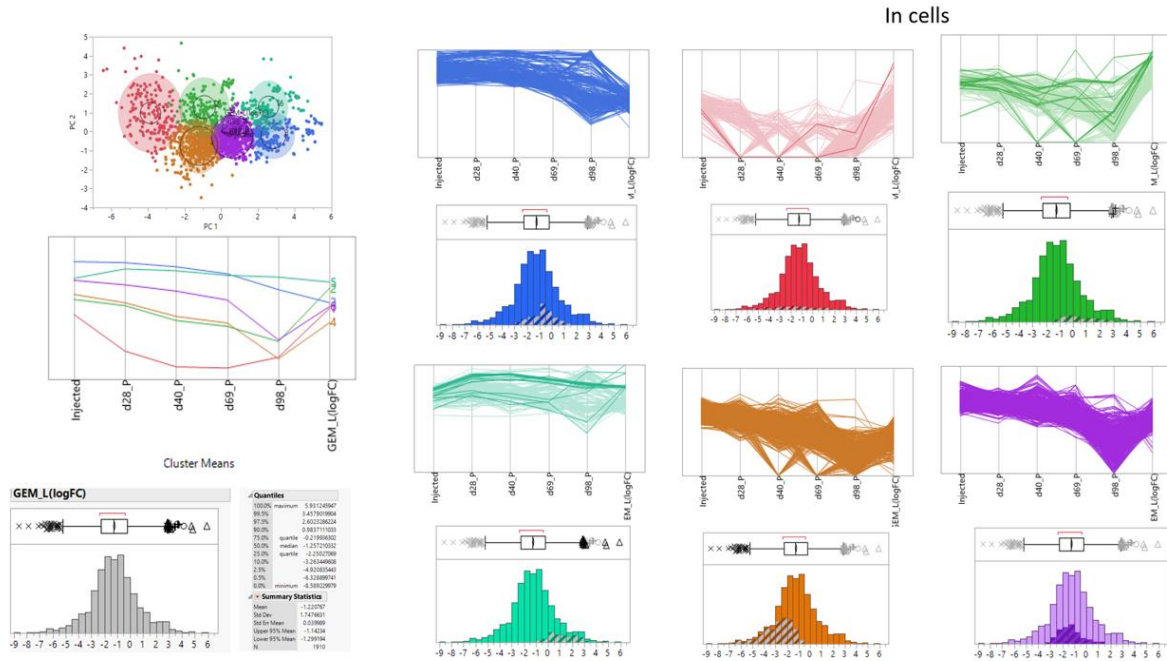
Figure 116. Clonal lineage dynamics and their intrinsic gemcitabine sensitivity. By clustering the clonal dynamics over 12 weeks (in "cell unit"), we obtain six major clusters (top left graph indicates the six clusters, and the middle left parallel plots are the average lineage dynamics of those six clusters). The cumulative number of lineage intrinsic sensitivity to gemcitabine distribution are shown in histograms (bottom left graph), where the X–axis indicates the $\log_2$fold change of individual clone's tumor representation comparing relapsed tumor vs. control tumor (positive value = resistant, negative value = sensitive). The height of each bar indicates relatively how many lineages falls into the $\log_2$fold change value. The right portion of the figure combines the lineage dynamic and the relative sensitivity to gemcitabine together. The shaded portion in the histogram indicates the lineage with gemcitabine sensitivity information in a particular cluster.

Not all lineages that have the ability to metastasize to the liver are more intrinsically resistant to gemcitabine

Furthermore, let us ask the question: "For the clones that have the ability to metastasize to the liver, are they more resistant to gemcitabine treatment?" To answer this question, let us plot the distribution of all the clones that are able to metastasize to the liver and look at the gemcitabine sensitivity distribution. From Figure 117 below, the grey colored histogram exhibits a normal distribution centered at $\log_2$ fold change at the value "0", indicating that not all the lineages that have the ability to metastasize to the liver are more intrinsically resistant to

159

gemcitabine. Furthermore, to look at the relationship of clonal dynamics and resistance, let us cluster the clonal dynamics in terms of "cell" unit and treat the gemcitabine sensitivity as the end point of the lineage's dynamics. From Figure 117 below, we see that there are mainly four clusters, and again, the cluster of lineages with the dynamic of "decreasing in cell abundance over time "being more sensitive to gemcitabine, which is indicated by the negative fold change in tumor representation comparing relapsed tumor to control tumor (the purple and blue clusters). As for the cluster of lineages that could maintain cell abundance over time, the lineages with more relative fitness, are more intrinsically resistant to gemcitabine (lineages in red and brown clusters). This suggests most gemcitabine resistance could come from clones already with higher fitness.
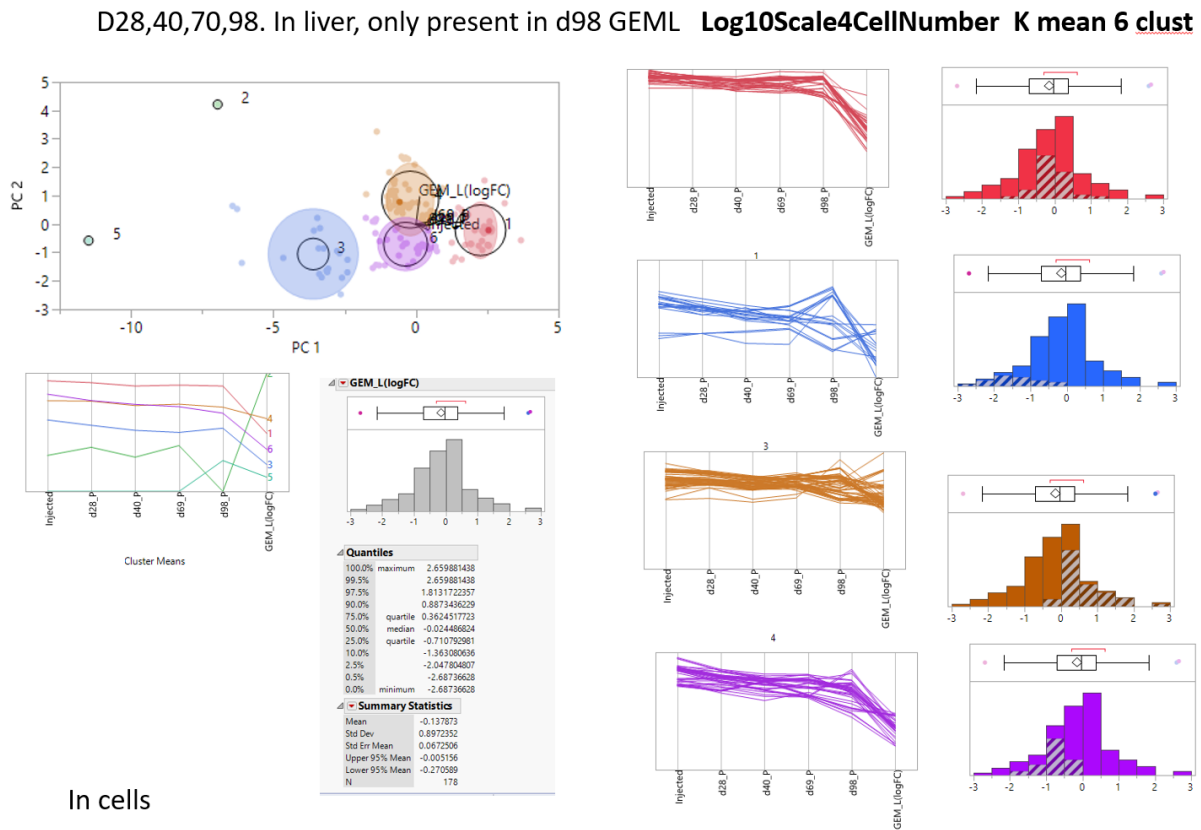
Figure 117. Liver metastasis clonal lineage dynamics and their intrinsic gemcitabine sensitivity. By clustering the liver metastasis clonal dynamics over 12 weeks (in "cell unit"), we obtain four major clusters (top left graph indicates the four clusters, and the middle left parallel plots are the

average lineage dynamics in those four clusters). The cumulative number of lineage intrinsic sensitivity to gemcitabine distribution are shown in the grey histograms, where the X–axis indicates the log_2fold change of individual clone's tumor representation comparing relapsed tumor vs. control tumor (positive value = resistant, negative value = sensitive). The height of each bar indicates relatively how many lineages falls into the log_2fold change value; the right portion of the figure combines the lineage dynamic and the relative sensitivity to gemcitabine together. The shaded portion in the histogram indicates the gemcitabine sensitivity in the particular cluster.

Lineage intrinsic long-term self-renewal property (from "cancer stem cells") is a key factor determining clonal fitness and metastasis

To get a better understanding of the observed positive correlation between clonal primary tumor abundance and their likelihood to metastasis, I designed an experiment to test the metastasis outcome when the lineages are relatively equal in abundance. One of the potential confounding factors surrounding the clonal abundance vs. metastasis correlation is the unbalanced distribution of lineage population used for injection. In the longitudinal experiment, the relative clonal abundance was inherited from the clonal equilibrium established post *in vitro* stabilization. In other words, although "stabilization" was critical in establishing CRTs, the lineages we transplanted in the animals were not equal in representation, with some higher than the others. Here, in the interest of studying lineage intrinsic factors leading to metastasis, we have to equal the playing field for all the clones in terms of relative abundance. To do this, we orthotopically transplanted cells in NSG mouse right after barcoding with minimum expansion, when the cells population were composed of more complexity of barcodes and were relatively equal in abundance. Then, we could survey the distal sites after a period of time to see which clones have more ability to metastasize.
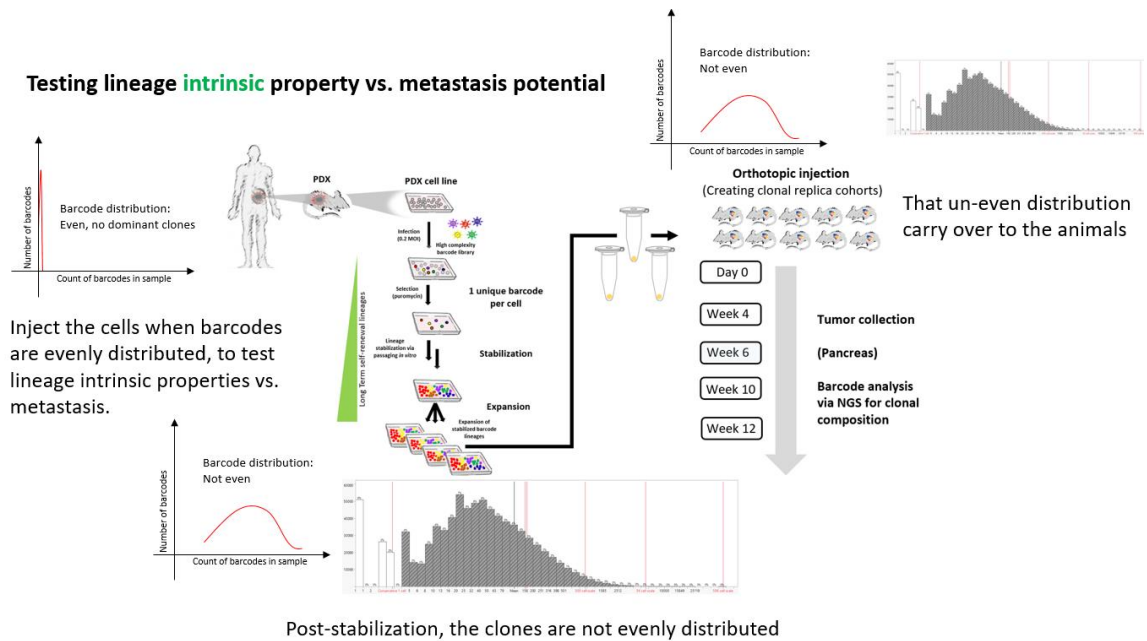
Figure 118. Schematics of inherited lineage abundance bias in previous experimental setting.

Additionally, we focus on the intrinsic long-term self-renewal property as a key factor for metastasis potential as a hypothesis. Therefore, in parallel to the orthotopic transplantation, we took the same pool of barcoded cells (equal clonal abundance) and performed *in vivo* serial transplantation to identify which barcoded cells in the beginning can give rise (to LTSR lineages) and sustain tumor long term. By piecing the two experiments, we could measure each barcoded cell's LTSR intrinsic properties (cancer stem cells) and their correlation to metastasis.
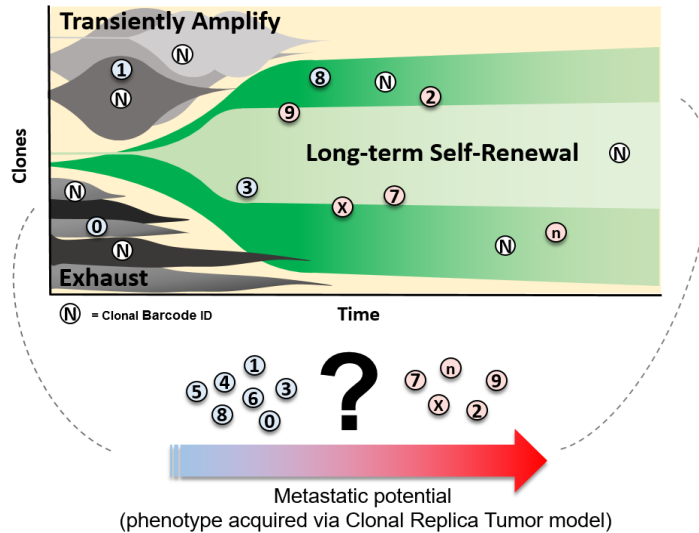
Figure 119. Schematics of cell intrinsic LTSR properties and their correlation with metastasis potential.

Serial transplantation *in vivo* reveals which cells were cancer stem cells

Currently, cancer stem cells are identified mostly through sphere formation assays in 3D culture and then the tumor cell's ability to engraft *in vivo*, as well as using identified "stem cell markers". As an example for the stem cell markers, a list of "CSC" markers are shown in the table below (for PDAC) [114]. However, these CSC markers are insufficient in capturing the cancer stem cell population in an unbiased fashion, as CSCs can (transiently) express lower amount of such individual surface marker and could be influenced by environmental factors. Therefore, the only way to identify cells that are endowed with CSC properties is through functional *in vivo* serial transplantation. If a cell and its lineage could re-establish a whole tumor and sustained it throughout multiple *in vivo* transplantations, by functional definition, that cell is endowed with LTSR ability [115-119]. Here, we use serial transplantation assay to identify cells (barcoded) that are CSCs in nature, and in parallel, assess their metastasization potential.

As discussed in the prior chapter, starting from 2M unique barcoded "single cells", only a few thousands of these cells will survive and sustain tumor after 8 months of serial transplantation.

If we think about it, the few hundred barcoded cells in the beginning must be "tumor initiating cells" or "cancer stem cells", while the other few millions of barcoded cells "exhaust" and could not be detected (above detection threshold) in the F4 generation of mouse after 8 months. This assay therefore not only measures which clonal lineages are endowed with LTSR properties, but also reveals during the initial barcoding (1 barcode per cell) which individual cell in the population are CSCs that give rise to clonal lineages that sustain tumor.
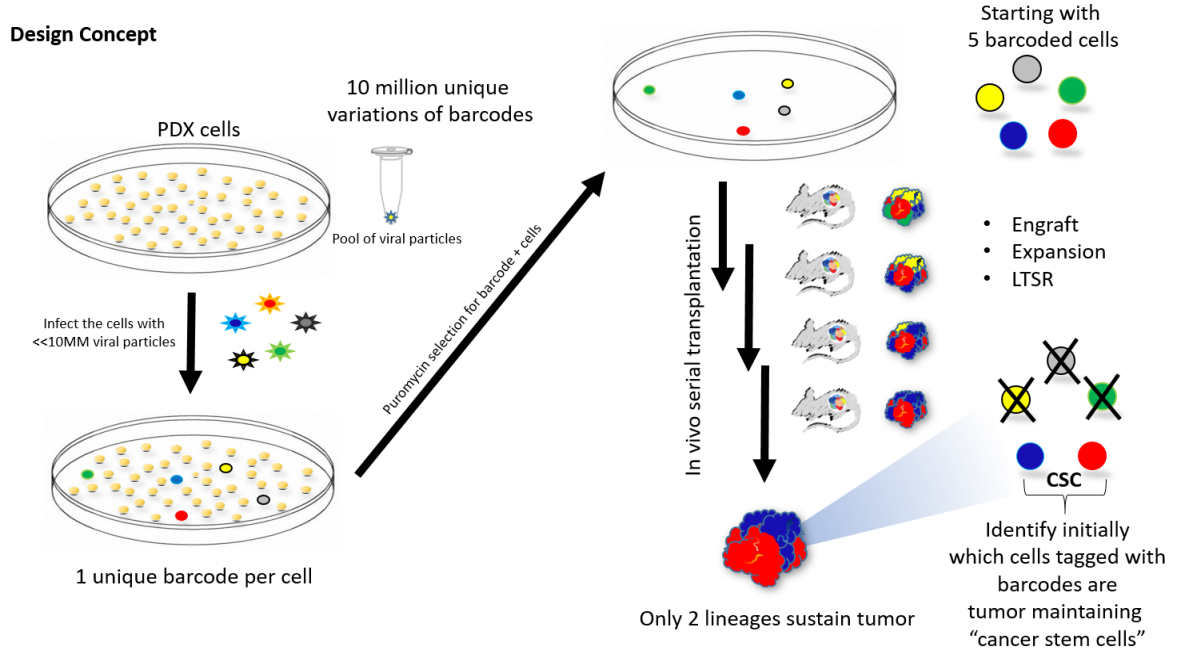
Figure 120. Schematics of the logic behind how to identify initial cancer stem cells that give rise to tumor sustaining lineages through *in vivo* serial transplantation.

Table 3. A list of published cancer stem cell markers. Adapted from Ishiwata, T., Matsuda, Y., Yoshimura, H. et al. Pathol. Oncol. Res. (2018) 24: 797.

CSC markers	IHC (%)	FCM (%)	Reference
ABCG2		0.4–7.33	46
ALDH		16.2	74
ALDH1	78.9		45
CD24	26	0.071–45.3	32
		3–28[*]	30
		57.8–70.1	51
CD24/CD44		2.1–3.5	51
CD44	15.1	46.1–100	32
		2–9[*]	30
		5.1–17.5	51
CD44/c-Met		0.5–5[*]	52
CD44/CD24/EpCAM		0.2–0.8[*]	30
CD133	14.5	0–1.61	32
		1.09–3.21	29
		1.98–69.85	46
CXCR4	35.6	0.274–38.2	32
EpCAM /ESA	60.6	1.36–93.7	32
		11–70[*]	30
Nestin	13.5	0.662–11.5	32

IHC: immunohistochemistry, FCM: flow cytometer, [*] Xenografted human pancreatic cancer cells

Transplanting minimally passaged post-barcoded cells allows us to track clonal dynamics during tumor expansion and metastasis when all barcodes are relatively equal in abundance

We have shown that CRTs' metastasis clones have a significant positive correlation with their abundance and fitness in the primary tumor. In other words, the more fitness and abundant a certain lineage had in the primary tumor, the more likely we can observe that lineage disseminating and grow in distal sites. To measure clonal metastasis potential when they are equal in abundance, we transplant minimally amplified post-barcoded cells orthotopically at passage 4. At this stage, all the barcode lineages are relatively equal in abundance, therefore, eliminating the clonal abundant bias in the previous longitudinal metastasis experiment. In other words, the clonal growth and metastasis phenotype here would be more related to "cell intrinsic properties". We focus our hypothesis on the LTSR as an intrinsic property for fitness and metastasis. Therefore, in parallel, using the same passage 4 barcode cells, we perform serial

transplantation for four passages *in vivo* (F1 to F4, 6 months) to identify which cells/lineages could sustain the tumor long term. Combining the lineage specific metastasis phenotype and the LTSR properties obtained empirically, we could answer whether or not LTSR is a key factor in determining fitness and metastasization.

The experimental schematic is shown below. The reason we choose passage 4 is due to only ~8-10% of the cells could give rise to another colony (discussed later in the single cells isolation section, empirically obtained number specifically for this cell line), and passage 4 is estimated to have gone through 10 doubling events. This would provide enough cells/CSCs to satisfy the experiment design in each group and replicates.
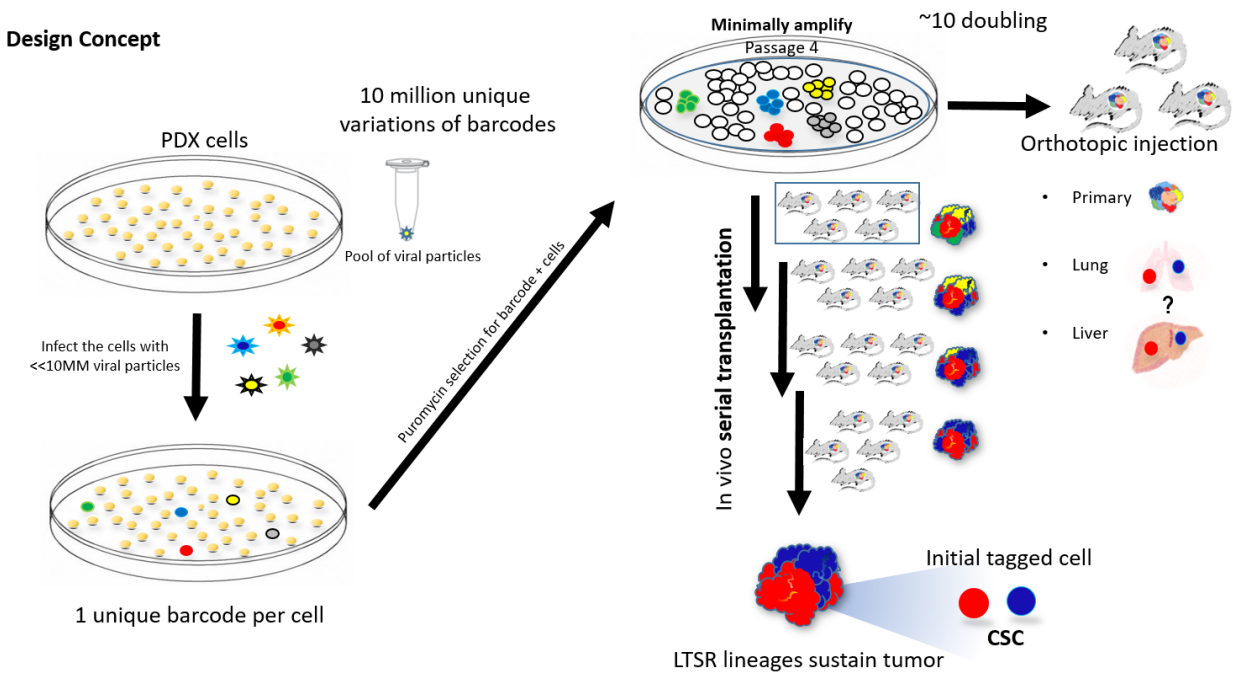
Figure 121. Schematics of the logic behind how to identify initial cancer stem cells through serial transplantation.
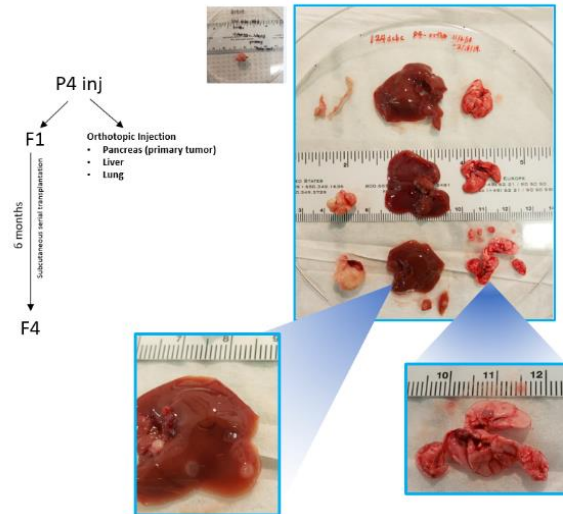
Figure 122. The metastasis lesions obtained from orthotopic injection of minimally expanded passaged barcoded cells. The left schematic represents the experimental design, while the right graphs shows the primary tumor (pancreas), liver, and lung (from left to right); bottom two shows the enlarged picture of liver and lung where the nodes of metastasis could be appreciated.

A subset of LTSR "cells" that could re-populate majority of the tumor mass also give rise to majority of the metastasis distal tumor

Analysis of clones (barcodes) in common and their relative tumor mass representation

The data shows (removing 4 reads as detection threshold) the initial clonal complexity at is ~200,000 barcodes ("P4 what is injected") in the parallel experimental arms. Between "F4 tumor" after 6 months and "P4 primary tumors", they share 468 common barcoded lineages. Looking at the tumor representation of those 468 common lineages, we see that it makes up ~96% of the "F4" tumor and ~82% of the "P4 primary tumor". This indicates that those initially barcoded 468 "CSCs" can sustain tumor long term (as seen from F4), as well as, have higher relative fitness (out of total ~2M barcodes) give rise to 82% of the tumor mass (P4 primary tumors). This suggests that, the majority of the tumors are comprised of cancer stem cell derived lineages.

By looking at the common barcodes that are overlapped between the "P4 liver metastasis", the "P4 lung metastasis" and the LTSR lineages from "F4", we see that there is a sub set of lineages (n=187) that are able to metastasize to both the liver and the lung.
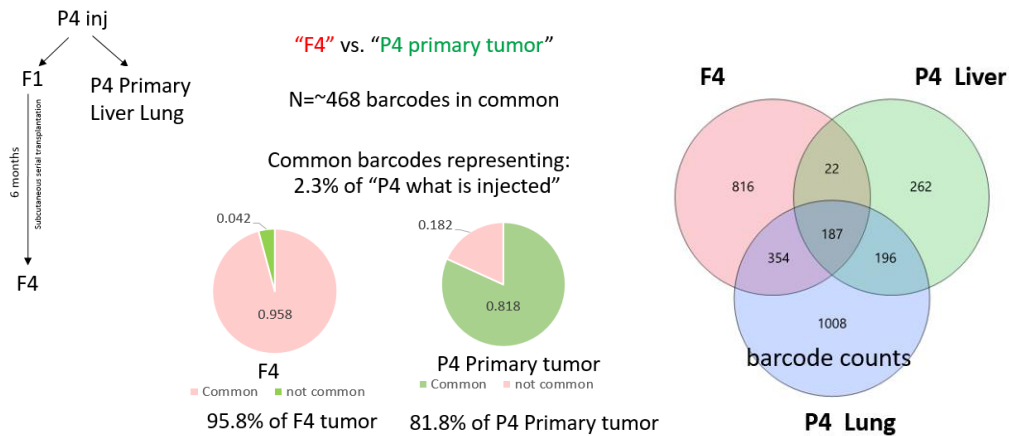
Figure 123. Common barcode lineages between F4 and P4 primary tumor, liver metastasis, lung metastasis. Comparing the common barcodes (n=468) between "F4" and "P4" tumor, the common barcodes makes up majority of both of the tumors at 95.8% and 81.8% respectively (left); Venn diagram of common barcodes shared between "F4" tumor and P4 liver and lung metastasis.

Consider that in the beginning, there were 200,000 individual barcoded lineages that were relatively equal in abundance. However, after six months, there were only 1,379 barcodes left. Out of those 1,379 barcodes, 187 of them has the ability to metastasize. From Figure 124 below, if we track those 187 barcodes that can metastasize and look at their relative representation in each populations, we see those 187 lineages only consists of 1.093% of the "P4 (what is injected into animals)", whereas they represent 93.9% of the "F4" tumor. Furthermore, those 187 clones make up 71.4% of the "P4 primary tumor" mass, as well as 98.3% of the liver metastasis, and 72.5% of the lung metastasis tumor mass. The fact that the clones that are able to metastasize and dominate the distal sites are the same clones that make up essentially the entire tumor mass after six months of serial transplantation indicates that cancer stem cells (LTSR cells) and their

lineages have more fitness, and that they are more inclined to disseminate and grow at distal organs.
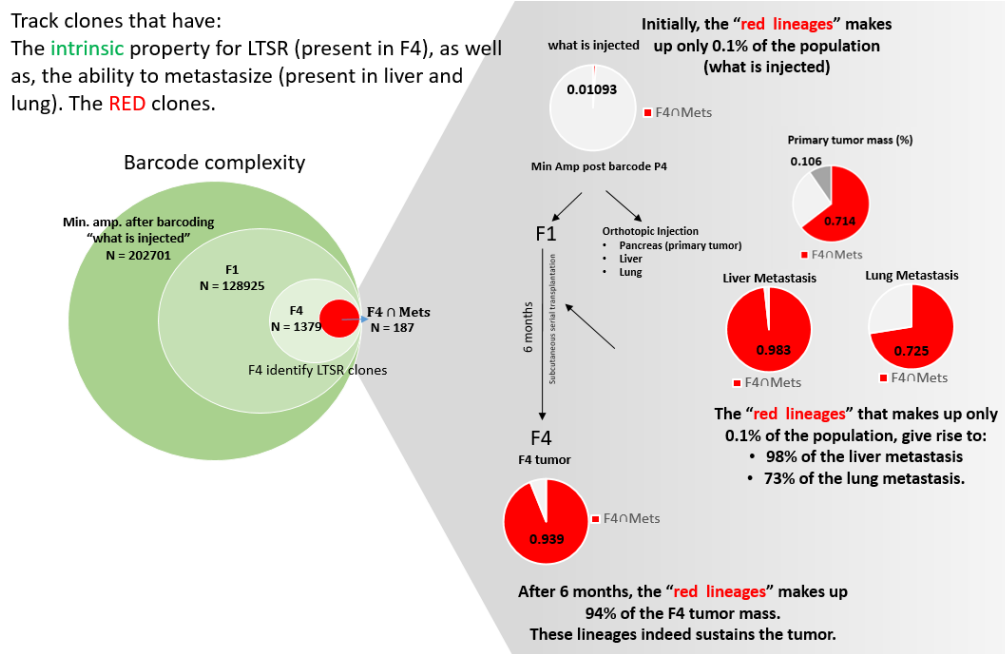
Figure 124. Diagram of barcode complexity throughout different samples.

In terms of barcode distribution for each sample, below is an analysis of the number of clones and relative clonal representation distribution in each sample. The Y-axis represents %, (in log scale) and the number next to each bar represents the count of barcodes that falls into the bin of a particular representation. If we look at the most left sample, "P4 what is injected", and compare it to "F4" and "P4 primary tumor", we see that the individual barcode distribution is relatively equal and does not have clones that dominate the population. As for "F4" or "P4 tumor", we see there are clones that are highly abundant, which represents up to 3-5% of the tumor mass each. If we look at the "P4 liver" and "P4 lung", we can see metastasis lesions are dominated by only a few clones, with each representing up to 15-20% of the metastasis tumor mass. In addition,

in the figure, we track the 187 clones that are able to disseminate to the liver and lung and shaded

their representation darker in each the bar graph. We can see that the most abundant clones in

the "F4" and "P4 primary tumor" are all able to metastasize. However, we also observe clones

that are lower in abundance with the ability to metastasize as well.

The darker color on histograms are barcodes that are overlapped between Liver and F4

Y-axis (log scale) percentage representation of sample(tumor)
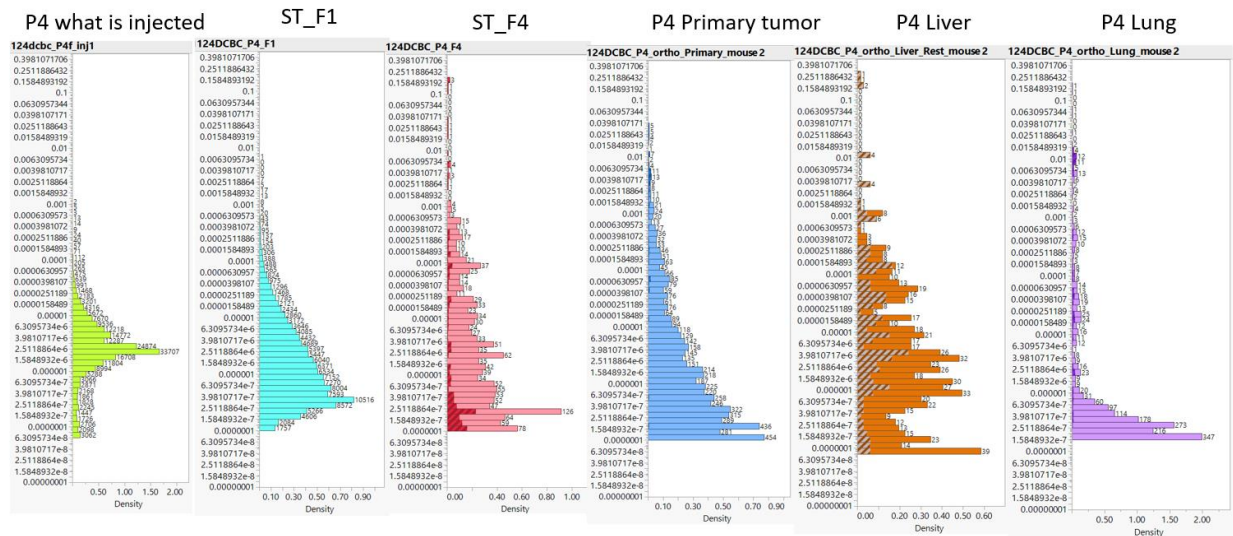
Figure 125. Bar graph showing the relative clonal abundance distribution in each sample. The Y axis shows the relative percentage representation in a sample (in log scale); X-axis shows the density. As we can see, passage four (the most left plot) all the clones are relatively equal in representation, whereas F4, there clearly are clones that are more abundant then the other.

To look at the distribution of the clones that are able to metastasize relative to the "what

is injected" entire barcode distribution, in the dotted plot below, we show the 187 metastasis

clones in red and the entire 200,000 barcodes in "P4 what is injected" and their relative

distribution. We see that the clones that are able to metastasize is evenly spread across the

sample, from high to low abundant clones (Figure below).

This result shows that the tumor cells that are able to gain relative strong fitness in the

primary tumor and able to metastasize in both liver and lung are derived from cells that have

LTSR capability. Therefore, this puts the capability of lineages that are able to metastasis and grow to the cell intrinsic LTSR property. In other words, metastasis capability originates from the cancer stem cell compartment.
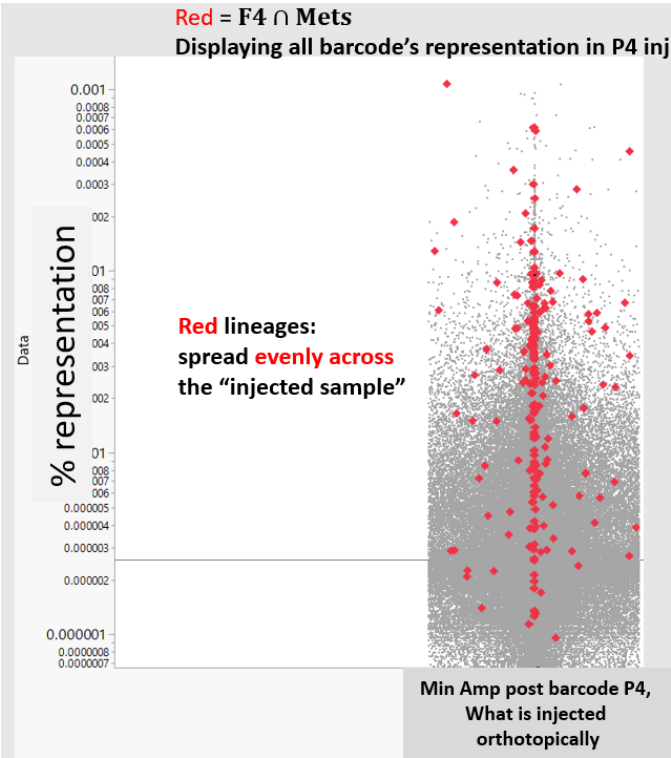
Figure 126. Dot plot for clones that are able to metastasize and their relative abundance (%) in "P4 what is injected". Red dots represent the 187 clones that are able to metastasize to the lung and liver; grey dots represent the clones in "P4 what is injected". The y-axis represents the relative representation in the sample displayed in log scale. The red dots are evenly spread across.
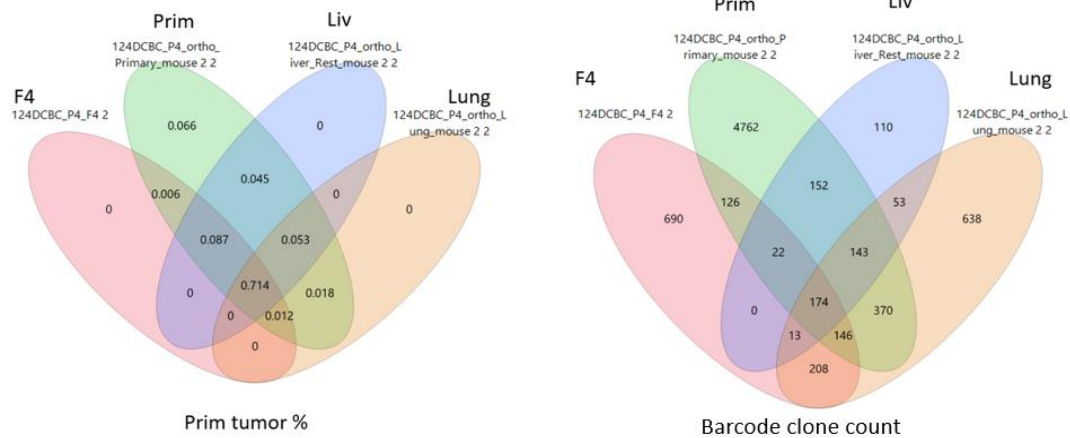
Figure 127. Figure showing the shared barcode number and their relative representation in "P4 primary tumor". For example, there are 174 barcodes shared between "F4", "P4 primary", "P4 liver" and "P4 lung", and their cumulative abundance makes up 71.4% of the primary tumor mass.

An assessment of spatial clonal heterogeneity in the orthotopic CRT tumor

As we know from studies involving autochthonous model and multiregional sequencing, clonal expansion occurs during tumor formation; and spatially, regions could be made up of different clonal lineages (with respect to mutational landscape), resulting in a "patchy" pattern in clonal distribution[39,122]. Clinically, these regional expansions of clones can provide challenges in terms of how well core biopsies represent the composition and characteristics of the entire tumor. Understanding how clones distribute and interact with stroma is an important aspect to understanding tumor evolution and behavior characteristic.

Transplantation model using cell lines derived from patient tumors is widely used amongst research communities. However, little is known about tumor lineage spatial distribution in such model. How closely can the injected tumor cells recapitulate clonal expansion and regional distribution of what has been observed in patient's remains unclear. Knowing this not only would

help us better understand the model we use as a researcher, but also help us understand the intrinsic behavior of clonal expansions.

The clonal replica tumor model presents a unique opportunity in charactering spatial clonal expansion of tumor. While injecting the barcoded lineages, the cells were mixed evenly into a homogenous population; therefore, the physical distribution of "barcoded cells" (lineages) should be evenly distributed in the injection mixture. With the injected lineages homogeneously mixed, will the tumor derived from this also display an "evenly distributed" phenotype resembling a "spotty" clonal distribution pattern? Or, will it resemble a "patchy" regional distribution pattern, where, although initially the lineages are homogeneously distributed, certain regions of the tumor will expand much faster than the rest? If it is the former observation, it would tell us that the cells intrinsic engraftment/proliferation capability are relatively comparable across all cells in a lineage, and that, in our model, environmental pressure has a minimal effect. If it is the latter observation, it would suggest that there might be a portion of the cells that have greater engraftment/proliferation capability then the other cells, as well as inferring that the cell extrinsic factor playing an important role in shaping the clonal expansion and the tumor's clonal composition spatially.
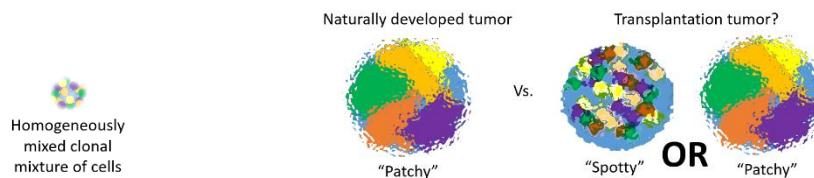
Figure 128. Illustration of clonal spatial distribution models. Evenly mixed cell lines containing heterogeneous clones (most left) transplanted into animal hosts and develop tumor over time; whether it would develop into a "spotty" pattern or a "patchy" regional distribution pattern remains unknown.

To sample clonal spatial distribution across the tumor, we first fix the tumor in FFPE blocks and sectioned it at 10-micron slices at the midsection of the tumor and stained with H&E. Using

laser capture microdissection (LCM), we sectioned 41 regions across the tumor as seen in Figure 129, where each square represents a LCM region. At the same time, on an adjacent FFPE slice, we stained for HLA (in red), ECAD (in green) and DAPI (in blue), which would provide functional context to the tumor.
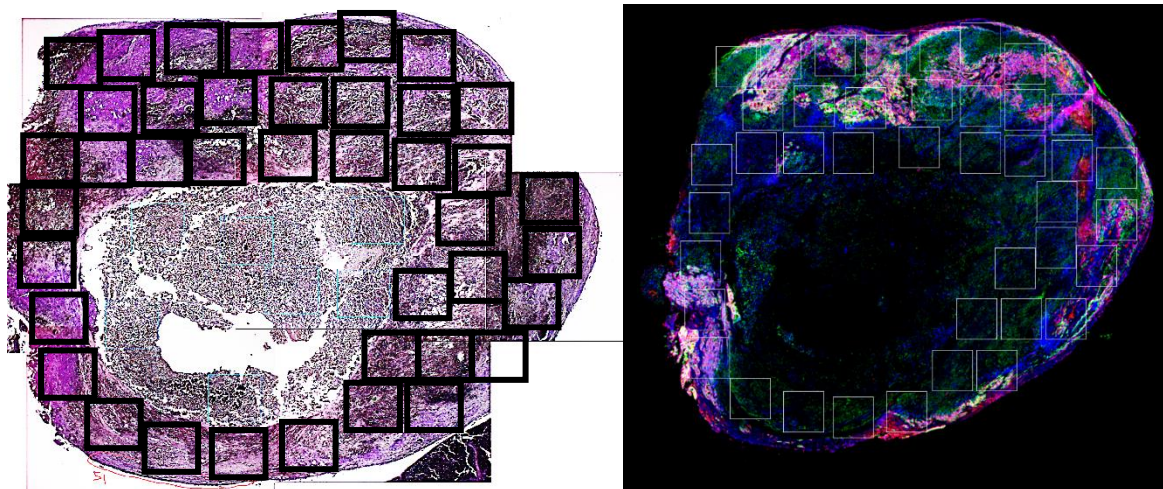
Figure 129. LCM sectioning regions of a transplanted pancreatic tumor. Left figure shows the H&E of cross sections of pancreas injected tumor; Right figure shows the IF image of the adjacent cross section of pancreas injected tumor, where green = ECAD, red = HLA staining. The square indicates independent LCM section's relative position. LCM and staining performed by Nick Yen.

Once each of the LCM regions were collected into individual tubes, we used a one-tube, FFPE-to-NGS library sample preparation workflow developed by me to obtain the barcode composition information in each LCM sections (proprietary) . In brief, each LCM region are fully lysed in an optimized condition so that the DNA is released into solution. To minimalize sample loss, in the same reaction tube, lysis buffer is inactivated and the barcode PCR reaction is performed in the same tube by addition of reagents. Each PCR primer for individual LCM sections has a unique index to allow for de-multiplexing after pooling all the PCR products together; as well as, containing NGS adaptor for sequencing. The multiplexed library is then purified and sequenced in one NGS run (NextSeq). The barcodes were analyzed via an in-house pipeline

determined by me and data filtered for noise and analyzed by Nick Yen. This section of the thesis is a collaboration between Nick Yen and I-Lin Ho.

Characterization of LCM sections' clonal composition

As an example of how barcoded clonal lineages distribute in each LCM section, we focus on the first four sections. Each LCM section's barcode composition and relative abundance is characterized and can be cross referenced with protein expression level via IF staining.

For example, LCM section #2 is mostly dominated by one barcode clonal lineage, LCM section #3 is co-dominated by two clones, and LCM section #4 is represented by multiple clones; furthermore, from the IF image, the level of expression of protein functional markers could be quantified for each section.
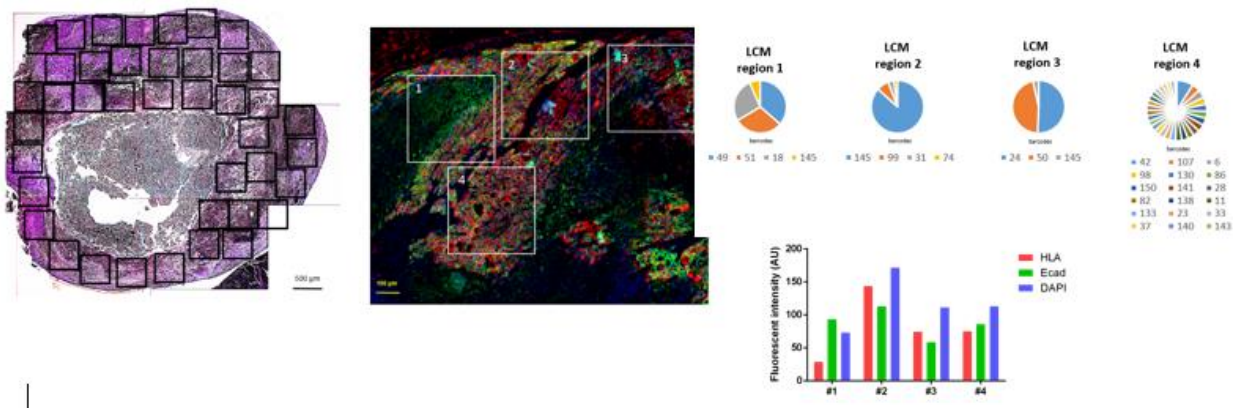
Figure 130. Barcode composition and functional marker for each LCM section. IHC of cross section of PATC124 primary tumor with LMC position in square (Left); IF staining and higher magnification of a section of primary tumor (green ECAD, red HLA, blue DAPI) with 4 LCM sections cut out for barcode analysis (2nd left); Quantification of staining intensity in each section, 1 to 4 (bottom right); Barcodes detected and their relative representation in each section (top right); Section 1 comprised mainly of 4 clones; Section 2 is mainly monoclonal; Section 3 is mainly bi-clonal; Section 4 has a mixture of many clones.

After compiling the barcode data for each of the 41 LCM sections and perform unsupervised clustering of the top barcode lineages found in each section. We can see the clonal regional distribution in its entirety. In terms of clonal distribution, there are roughly six different regions composed of different sets of clones. For example, on the top right of the heat-map, 29 sections share relatively high abundance of common 17 clones that are not found in the other sections. Also, two sections immediately adjacent to those are composed of essentially every clone. Furthermore, within each of the six regional clonal distribution patterns, the individual clone's relative abundance are highly correlated and consistent. This data suggests that tumors, despite the cell lineage initially during injection were evenly mixed, do indeed, expand in a regional pattern; stressing the possibility environmental pressure, or clonal clusters that cooperate /or compete during expansion.
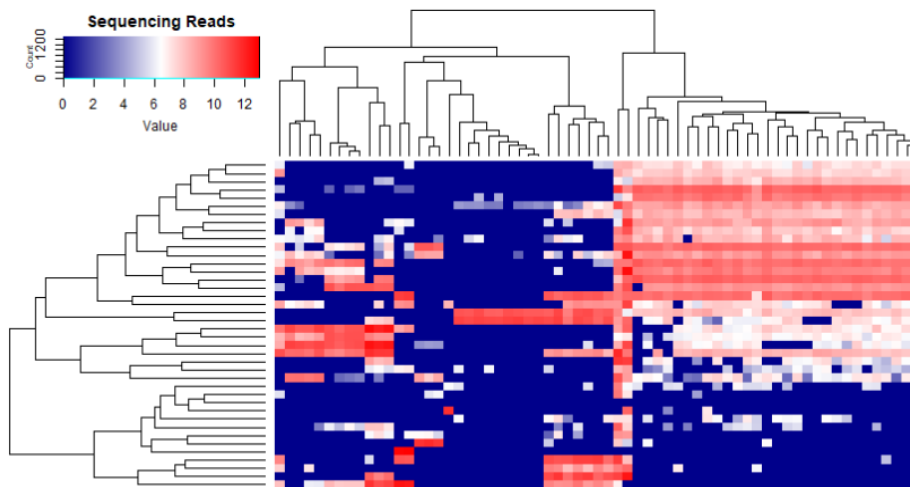
Figure 131. LCM regions vs. clonal abundance distribution. X-axis are composed of 41 LCM regions and Y-axis unique lineages (each line represents an individual barcode). The color of represents the relative abundance of each barcode (red=high, blue = low). Analysis done by Nick Yen.

Clonal lineage distribution pattern in tumor are both patchy and spotty

A patchy regional distribution pattern of how clones expand is only part of the picture. In our model, each clone does expand regionally, however, the clonal lineage can actually be found across the entire tumor. Thus, combining the two observed phenotypes, we concluded a model of "spotty" distribution of "patchy expansion" for individual clones. As an example, if we look at barcode #514, and #288 individually in terms of distribution of abundance across the entire tumor, we see exactly a spotty but patchy phenotype.

This purpose of this data is partly to gain an understanding of how our tumor model clonally distribute spatially and how closely do they mimic a naturally developed tumor *in vivo*. In addition, the data serves as a feasibility data for future hypothesis generation while linking clones with functional marker to study interactions between specific clones with stroma, other cell types, other clones, as well as, specific clonal intrinsic functional properties that link to phenotype.

This set of experimental data also ties into the observation of divergence in clonal abundance as tumor expands mentioned in the previous sections, where spatially different micro-environment differences could contribute to the unpredictability nature of cancer clonal expansion.
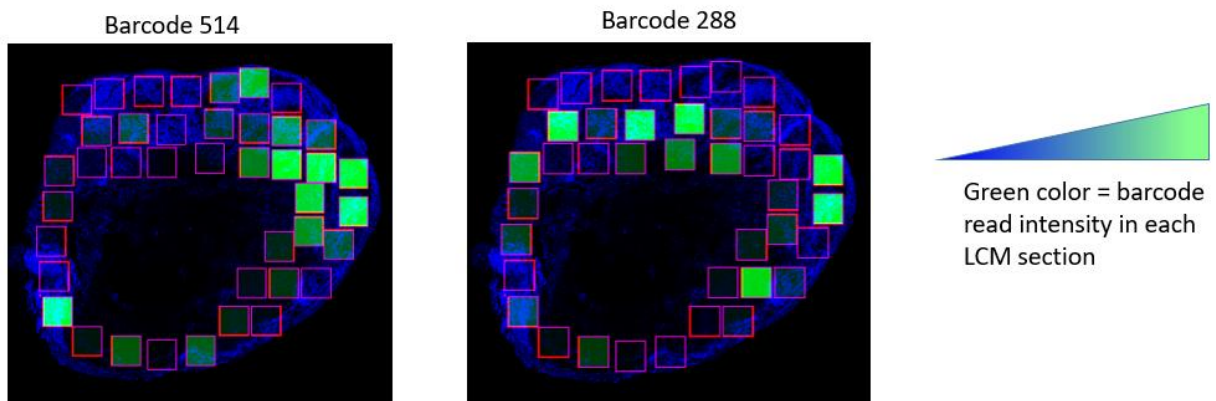
Figure 132. Geographical distribution of specific barcoded cell lineage (#514 and #288) in the primary tumor. By reconstructing cell lineage spatial distribution, we can observe grow dynamic and expansion pattern and possible clonal inter-dependency. Analysis by Nick Yen.

LCM sectioning of metastasis lesions in the liver reveals majority of metastasis lesions are composed of monoclonal lineages in tumor mass and oligoclonal in composition

From the same set of CRT tumor samples (part of the CRT longitudinal cohort), we also survey paired liver metastasis lesions with LCM and sequence their clonal composition. Since the tumors are from the same animal, we could trace individual metastasis lesion's spatial origin in the primary tumor section.

From the data, we can see that most of the liver metastasis lesions are composed of monoclonal barcoded lineages. For example, from the right graph below, looking at the liver metastasis column, metastasis node #1,3,6,7,8,11 are all composed by barcode #514 (middle column), and the barcode #514 could be found almost everywhere in the 41 LCM regions of the primary tumor FFPE slide. Another example is liver metastasis node #5, where is it composed of three clones (barcode #22, 1049 and 736), and those barcodes only comes from ~10 out of 40 sections, ¼ of the primary tumor in a spatial context.
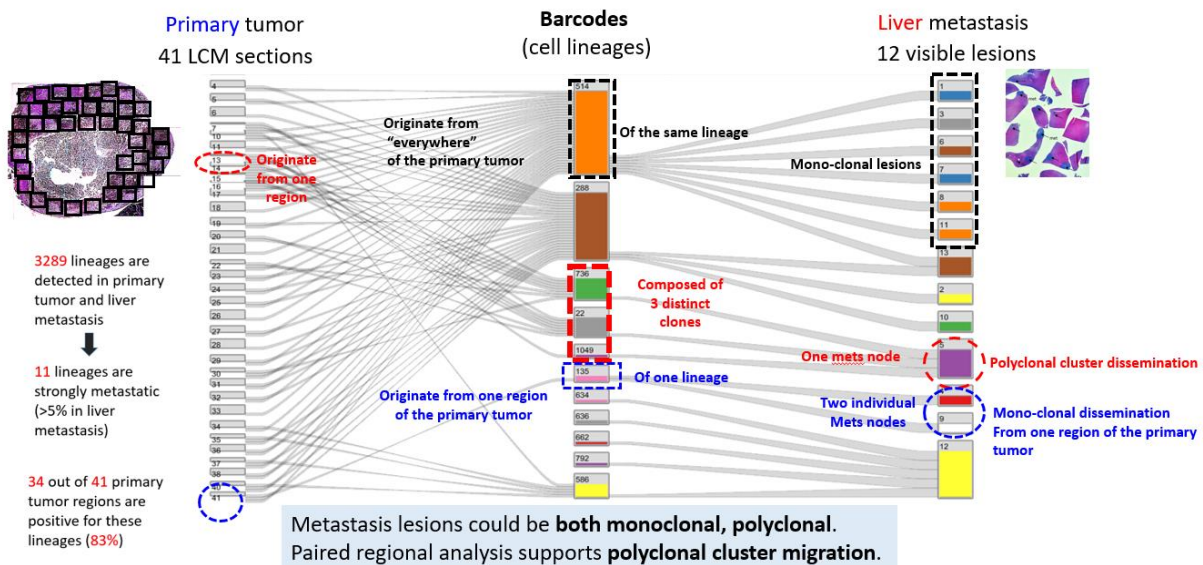
Figure 133. FFPE slices of paired liver and primary tumor from the same CRT mouse stained with H&E (top liver, bottom primary). The black boxes on the primary tumor indicate the LCM region

178

sampled. The right graph shows the connection and barcode composition between primary tumor and liver metastasis nodes. Analysis by Nick Yen.

Clonal metastasis mass does not necessarily correlate with clonal primary tumor representation

From the same LCM data above, we took the four most abundant liver metastasis lesions and calculated their relative tumor representation in the primary tumor. We found two of the big metastasis lesion barcodes are highly represented in the primary tumor (barcode ID 4611_12634 and 5304_9756), whereas the other two metastasis lesions are composed of clones that are not highly representative of primary tumor. This data aligns with what was seen from the previous sections, where in general, there is an abundance correlation with primary tumor and liver metastasis, however, clones that are lowly represented in the primary tumor could still give rise to relatively large liver metastasis nodes.
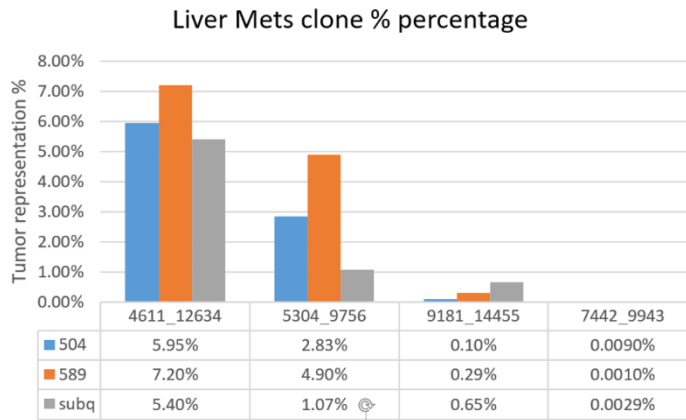
Figure 134. Four clonal ID identified through large liver metastasis nodes via LCM and their relative clonal representation in the paired primary tumor. The ID "504", "589" and "subq" indicates different mouse's tumor. "504" and "589" are two orthotopic CRTs, while "subq" refers to the subcutaneous tumor using the same barcoded cell line. The percentage represents how much percent representation of the specific lineage represents in those samples.

Large metastasis nodes are comprised of oligoclonal lineages, however, only one clone dominate the metastasis lesion mass

While the LCM data determined that majority of the metastasis lesions are monoclonal in nature, it is derived from a 2-D section of the liver lesions. To gain a more confident representation of metastasis lesions, we sequenced the entire individual metastasis lesions of three (3) of the CRT animals at day114.

Figure 135 below is a representative picture of the size and the relative scale of individual metastasis nodes and mouse liver (in which the tumors were excised from). From the barcode analysis, the big nodes in each individual animal are comprised of monoclonal barcodes. The barcodes that are able to grow into the big lesions also make up of the rest of the barcodes detected in the rest of the liver.
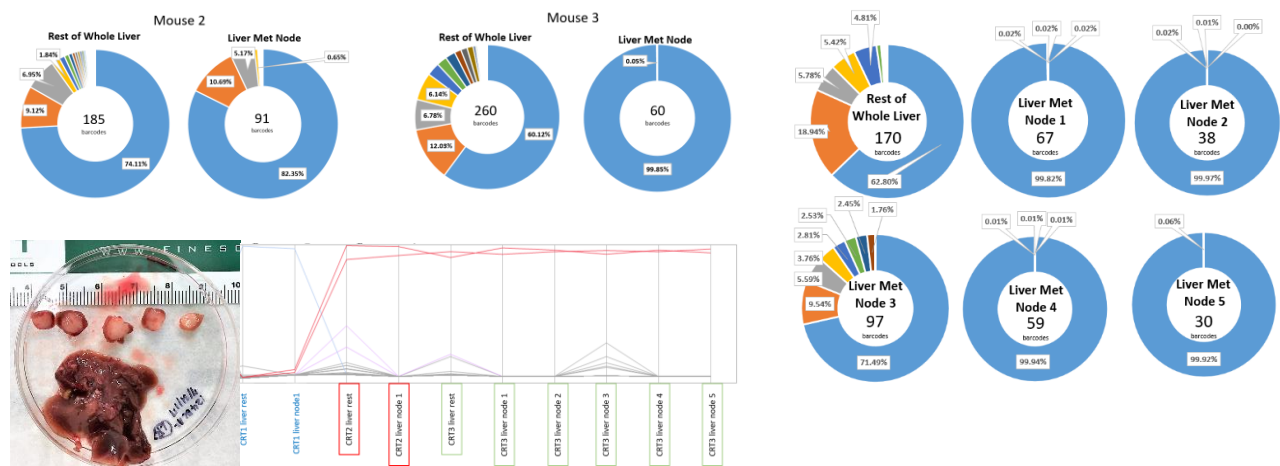
Figure 135. Metastasis liver clones are oligoclonal in composition but monoclonal dominance in nature. Bottom left, a representation of liver metastasis node size. Five nodes were excised out from the liver, all samples were sequenced (nodes and the rest of the liver); Parallel plots of the lineages shared across the three CRT mouse and liver metastasis nodes and "liver rest" (the rest of the liver). The X-axis represent relative abundance of the clones (represented by lines), the higher the line is, the more representation it has. In CRT 1, the liver population as well as the large metastasis node are all dominated by the "blue clone". The CRT2 and CRT3 mouse, in the same manner, are dominated by the red clone (it has been determined that the two barcodes belongs to the same clone, a dual-barcoded lineage). Majority of the large metastasis lesions are monoclonal dominance in mass, but oligo clonal in composition (shown in the pie charts).

Establishment of isogenic clones for functional characterization of metastasis vs. non-metastasis prone lineages

To functionally characterize clonal lineages with distinct phenotypes observed from the CRT longitudinal model. Through clonal isolation (described in previous chapters, which involved single cell sorting into tissue culture plates), we isolated six clones with differential tumor fitness and metastasis phenotype identified through the longitudinal tracing experiment. Three (3) of the clones display higher fitness in the primary tumor and metastasize, whereas the other three clones do not metastasize at all. The isolated clone's primary tumor abundance over time and their dissemination patterns in the blood, liver and lung from the CRTs longitudinal study are shown below. In addition, we performed subcutaneous injections of these six individual clones in nude and NSG mouse and found that these six clones all have the ability to engraft and form large tumors. In other words, all isogenic clones isolated are tumorigenic.

Focused clonal profile in primary tumor

		Day0 n=2	W4 avg n=2	W6 avg n=2	W10 avg n=4	W12 avg n=2
Mets 12SG	4611_12634	24991724	62476849	56656419	150637518	17853298
30FG	5304_9756	9640076	22004695	17795363	100706490	11030871
FR5	12763_9747	22545369	15352013	5935855	10402163	367201
Non-Mets 32FG	2833_15435	124211	292721	40516	617495	56519
FR13	1592_6046	65859	87061	27656	15295	14306
16SG	6023_6472	39889	2691	1365	459	1

■ 4611_12634   ■ 5304_9756   ■ 12763_9747   ■ 2833_15435   ■ 1592_6046   ■ 6023_6472

Figure 136. Six isogenic clones with tumorigenic potential and their abundance over time in the primary tumor from the CRT longitudinal tracing experiment.
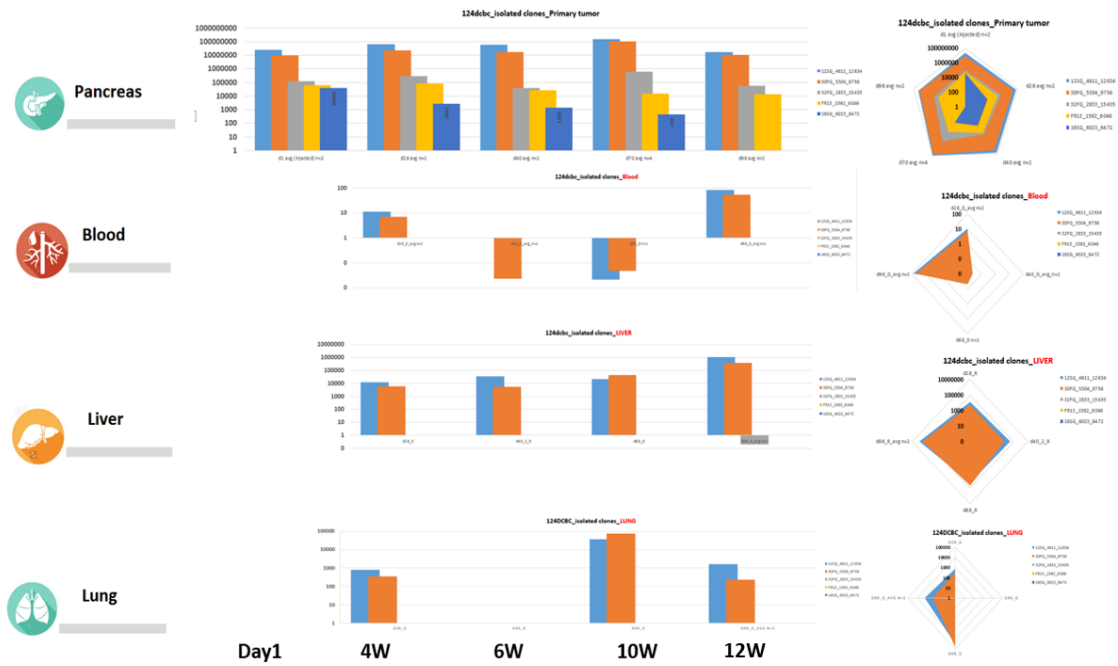
Figure 137. Six isogenic clones with tumorigenic potential and their relative abundance (%) over time in the primary tumor, blood, liver and lung from the CRT longitudinal tracing experiment. The scale is in log scale (left); Spider plot of the 6 clones and their relative abundance representation at 4w, 6w, 10w, and 12w data points are in the 12, 3, 6, and 9 o'clock position.

Isolated clones are heterogeneous in terms of mutational landscape and CNV profile

We performed whole exome sequencing for each of the six clones and analyzed their mutational landscape and copy number variance (CNV) profile. As see by the analysis on mutations in cancer hallmark genes, all clones share the same mutation of common drivers KRAS and TP53, DNA damage checkpoint ATM, ATR. As for the other genes, each clone has a more heterogeneous mutational profile. Furthermore, from CNV analysis, it shows these 6 clones are also heterogeneous in chromosome copy number states. There are not enough samples to conclude whether or not a specific genetic profile is correlated with metastasis phenotype.

182
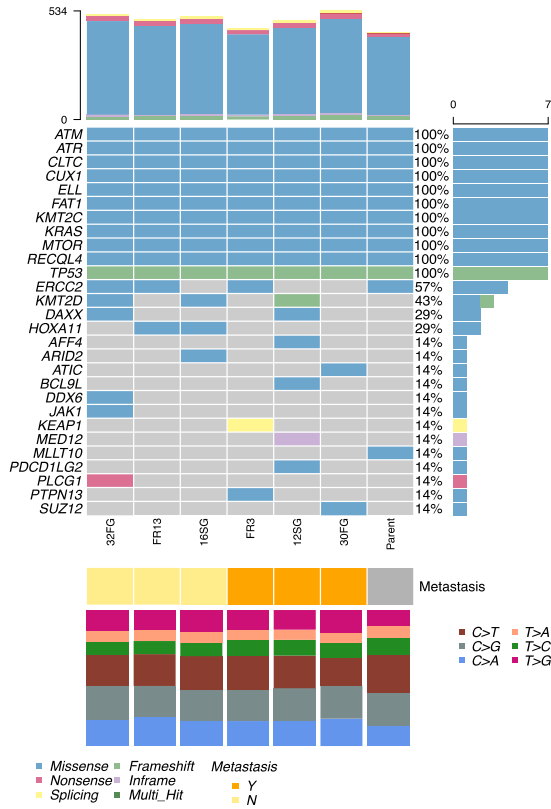
Figure 138. Cancer hallmark gene mutation in each of the 6 isogenic clones and their parental population. Isogenic clones with differential metastasis potential indicated by the orthotopic longitudinal CRT model are isolated from the parental cell population and submitted for whole exome sequencing and mutations analyzed. By Zhao Li.

Red: Copy gain
Blue: Copy loss
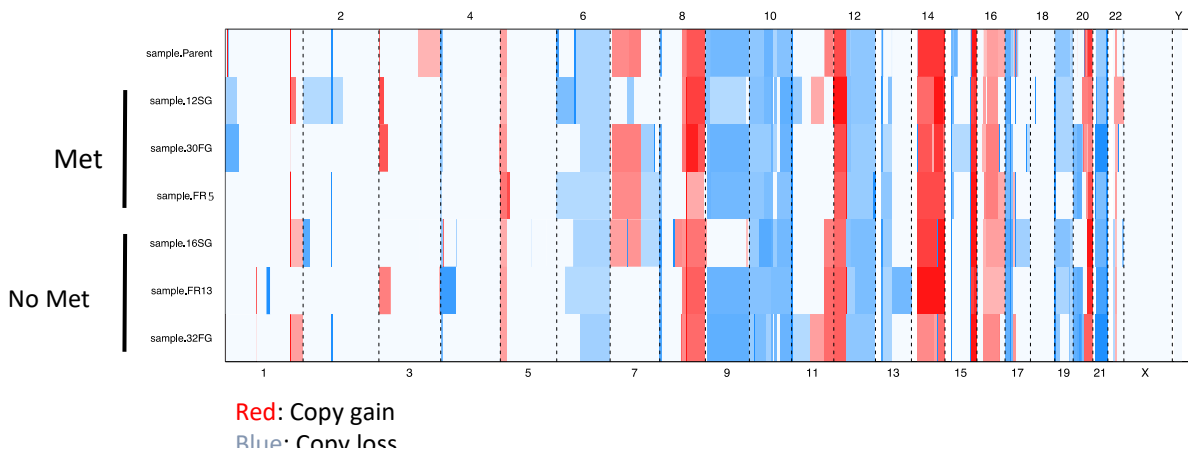
Figure 139. Copy number profiling on the 6 isogenic clones (copy number gain in red, copy number loss in blue, top graph). By Zhao Li.
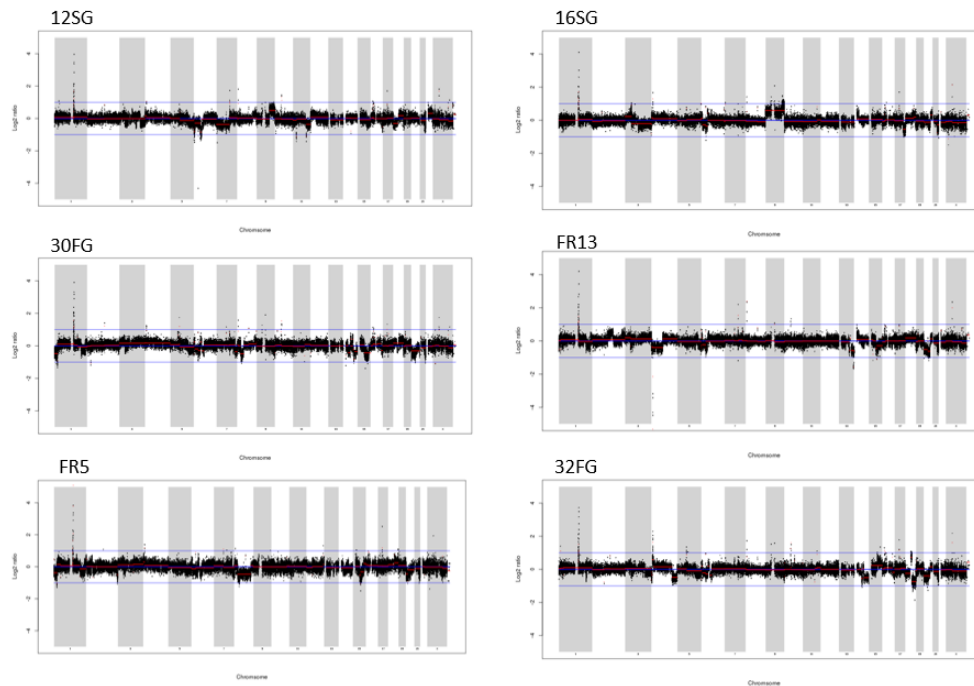
Figure 140. Copy number profile across all chromosomes of the six isogenic clones. Log2 ratio (with respect to a normal diploid genome) of the copy number variations (CNV) mapped to the corresponding chromosomal coordinates for all clones (black), and segmented CNV (red)By Zhao Li.

Clones with pro-fitness and metastasis potential are upregulated in NMDA pathway

Using supervised-clustering of differential expressed genes based on their metastasis potential, we identified a set of 200-gene signature that are up-regulated in either the "pro-met" or "non-pro-met" group. Suggesting there are indeed expression differences that could be linked to the phenotypes. Furthermore, through gene set enrichment analysis, we identified enrichment of pathways such as N-methyl-D-aspartate (NMDA) receptor, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor, and long-term potentiation pathways enriched in pro-metastasis clones[123].
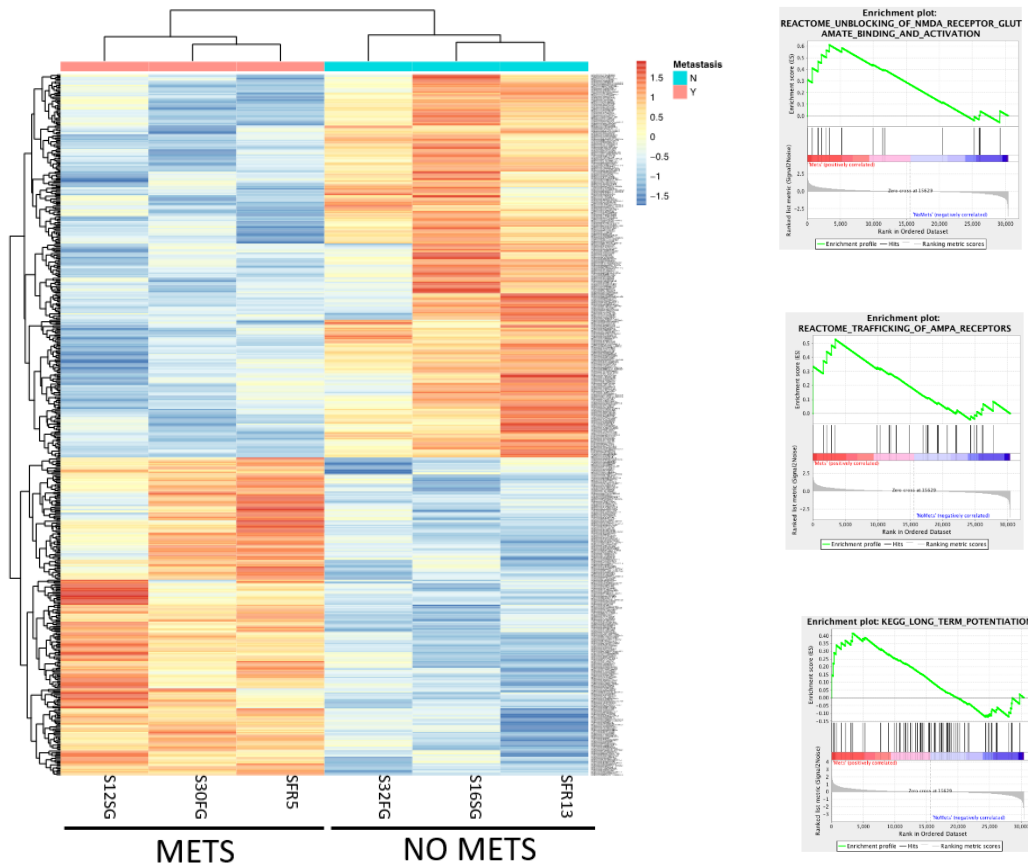
Figure 141. Differential gene expressed from "pro-mets" vs. "non-pro-mets" isogenic clones using supervised clustering (top 100 and bottom 100 gene).Left heat map; GSEA analysis for "pro-mets" vs. "non-pro-mets" clones, top right to top bottom showing NMDAR, AMPAR, Long-term potentiation pathways enriched in "pro-met clones". By Zhao Li.

NMDAR is an ionotropic type of glutamate receptor. It is found in the central nervous system where it is directly involved in synaptic signaling, where when active, causes calcium influx into the cytosol [124]. It has been reported that glial tumors has an excess level of glutamate. In addition, an increase in glutamate concentration is correlated with tumor aggressiveness by *in vitro* and xenograft models. In none CNV cells, glutamate level has also been demonstrated to be correlated to aggressiveness, invasion phenotype in breast cancer, prostate cancer, lung cancer, and pancreatic neuroendocrine tumors (PNET) as well as PDAC [125,126]. Specifically, NMDAR has been observed to have increased expression level at the tumor periphery in GEMM models by immunochemistry and upregulation of NMDAR causes a MAPK, CaMK dependent invasive phenotype [78-79, 127-128]. From Cbio portal, pancreatic cancer in specific, there exists about 1% of the patient which carries NMDAR mutation (GRIN2A subunit). Looking at the location of mutations, most mutations are spread out and does not seem to correlate with protein expression nor survival, leading to speculation that the overexpression could also be contributed by epigenetic modifications.

Figure 142. NMDA receptor activation schematics. Adapted from "Anti-N-methyl-D-aspartate Receptor Encephalitis – A Synopsis" [129].

NMDAR upregulation in "pro-metastasis" clones is validated *in vitro*

Taking the two most distinct "pro-metastasis" (12SG, 30FG) and two "non-pro-metastasis" (FR-13, 16SG) clones, we tested their mRNA expression level by qPCR. As seen from the figure below, GRIN2A (NMDAR, subunit 2A), GRIA2 (AMPA) and CaMK4 (calmodulin-dependent protein kinase type IV) all have significant upregulation in the "pro-metastasis" clones. The gene to normalization used is GAPDH for each sample.
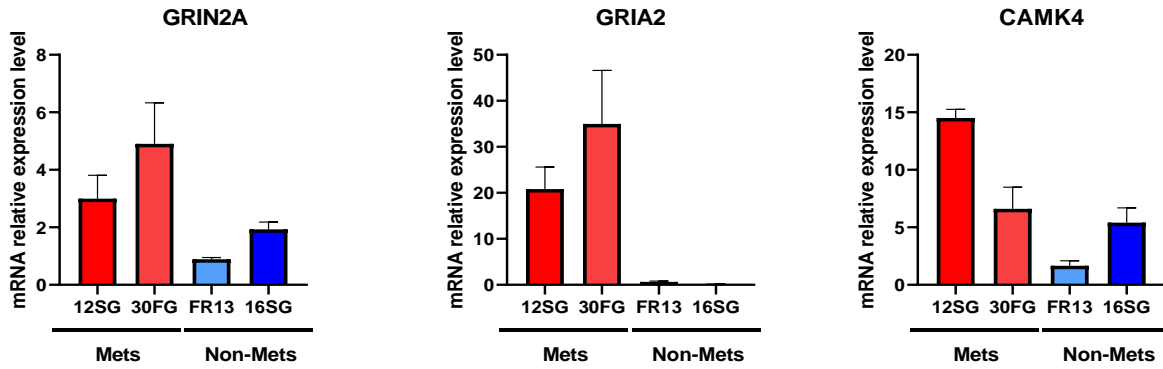
Figure 143. Relative mRNA expression of GRIN2A (NMDAR, subunit 2A), GRIA2 (AMPA) and CaMK4 (calmodulin-dependent protein kinase type IV) in four isogenic clones using GAPDH as normalization. Experiment by I-Lin Ho.

Taking the "pro-metastasis" (12SG, 30FG, 32FR) and "non-pro-metastasis" (FR-13, 16SG) clones, we tested their relative protein expression level by Western blot. As seen from the figure below, NMDAR2A has relatively more expression level in the "pro-metastasis" clones than in the "non-pro-metastasis" clones (using FR13 as normalization). Glial tumor cell lines were used as a control for the level of protein expression.
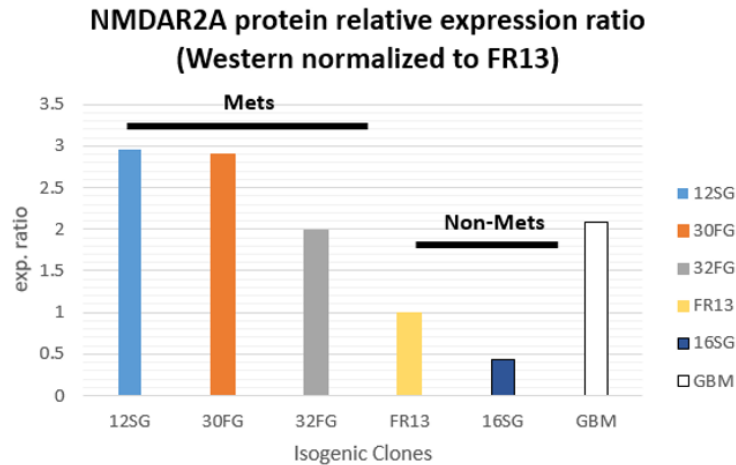
Figure 144. Relative protein expression of NMDAR by Western blot, subunit 2A between isogenic clones (normalized to FR13). Experiment by I-Lin Ho.
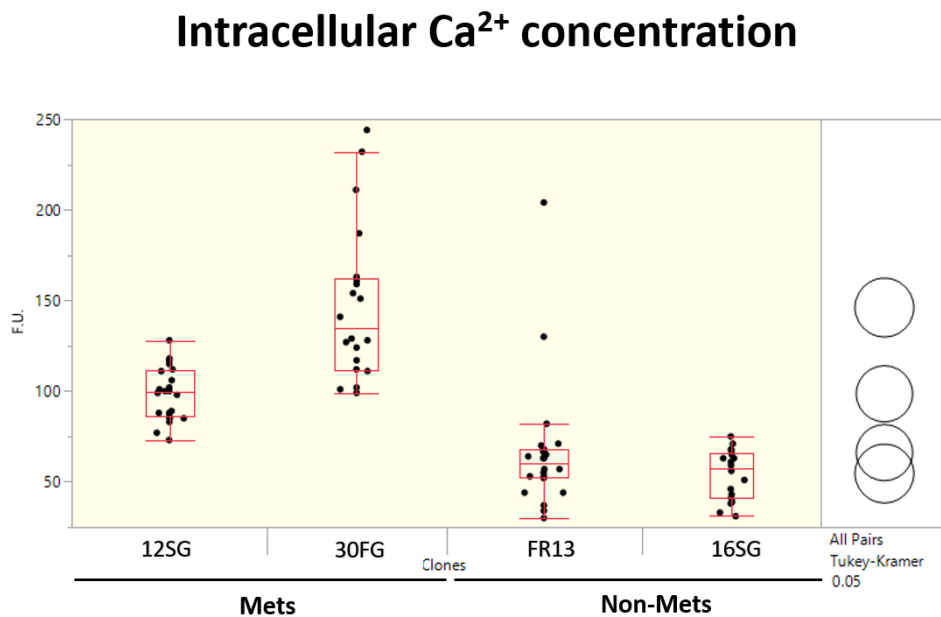
Figure 145. Intracellular basal level calcium concentration. Two of the isogenic clones isolated with higher fitness (12SG, 30FG) that express higher level of NMDAR also have higher level of intracellular calcium concentration compared to the non-metastatic clones. Experiment by I-Lin Ho.

Furthermore, we tested the clone's basal level of invasion ability. Using an invasion assay, testing the top and bottom two most pro-metastasis and non-metastasis clones, we demonstrated, at the baseline level, the pro-metastasis clones do invade more.
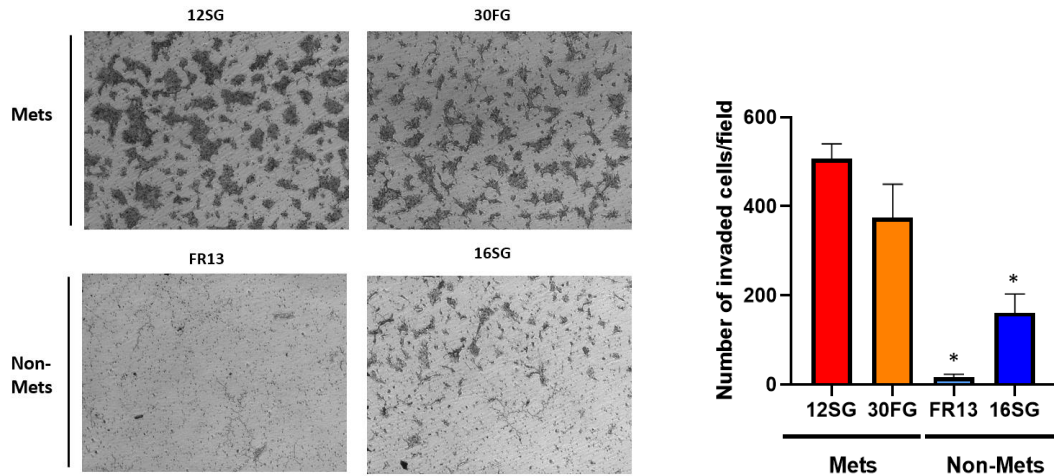
Figure 146. Clonal basal level measured by invasion assay.

Isogenic clones' invasion capability matches their CRT in vivo experiment phenotype. Left graph is the image of cells that could migrate, the more cells indicate a higher invasive capability. Right graph is a quantification of the image in bar graph format. By I-Lin Ho.

Memantine suppress primary tumor growth of *in vitro*

We identified a FDA approved drug named memantine, which is a non-competitive NMDA antagonist that could potentially have an effect on reducing the invasive phenotype of pro-metastatic clones. We first titrated a drug concentration (measured by MTT assay) that would not cause apoptosis. Using such concentration, we apply the drug on individual isolated clones. Through invasion assay comparing treated vs. non-treated conditions, we see that memantine effectively reduced the invasion phenotype of pro-metastasis clones.

Memantine
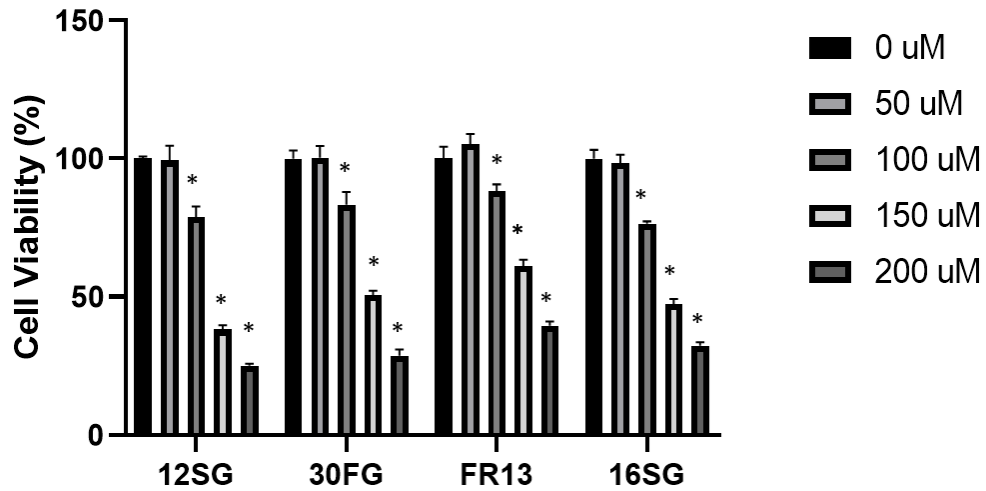
Figure 147. Cell viability assay (MTT assay) vs. memantine concentration. All four clones despite of their fitness or metastasis phenotype (in CRT experiment) all display sensitivity to memantine. At 50µM concentration memantine, the drug has minimum impact to cell viability. Using a dosage less than 50µM to test the drug's efficacy on clones' invasion ability is concluded from this experiment. Experiment by I-Lin Ho.
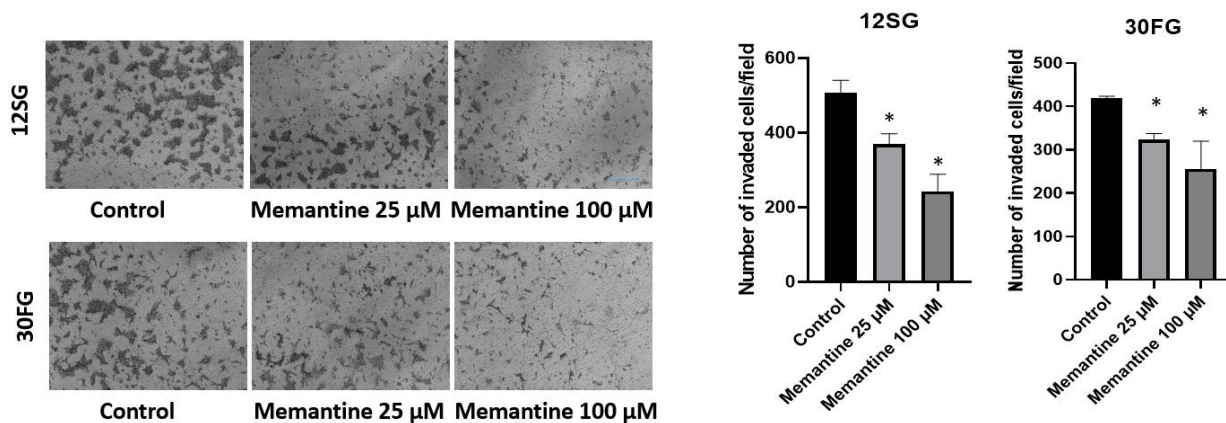
Figure 148. Invasion assay under the effect of memantine on pro-metastatic clones. Testing the most metastasis clones (12SG and 30FG) under the treatment of memantine. The left image are images from invasion assay, where cell amount correlates positively with cell's capability to invade. Right graph is the quantitated result of the images. Here, it demonstrates memantine can reduce the invasive ability of the most metastatic clones. Experiment by I-Lin Ho.

Memantine suppress primary tumor growth of *In vivo* reconstructed tumor

To test the effect of memantine on tumor growth in an *in vivo* setting. We orthotopically injected a total of 3 million cells composed of equal amounts (25% each) of isolated sub-clones "12SG", "30FG", "16SG", and "FR13", where "12SG" and "30FG" having a strong fitness and metastasis phenotype, and "16SG" and "FR13" having low fitness and low metastasis phenotype according to the CRT longitudinal data set. The experimental group were treated via intra-peritoneal injection of memantine at 11.54mM, ~200-250µL per mouse (10mg/Kg), q.d. for 5 days a week, 2 days off, while the control group receives 200uL PBS. The experiment lasted for 2 months.

We observed a significant decrease in tumor weight between the controls (n=4) vs. the treated cohort (n=7). Indicating memantine can effectively impact tumor growth.
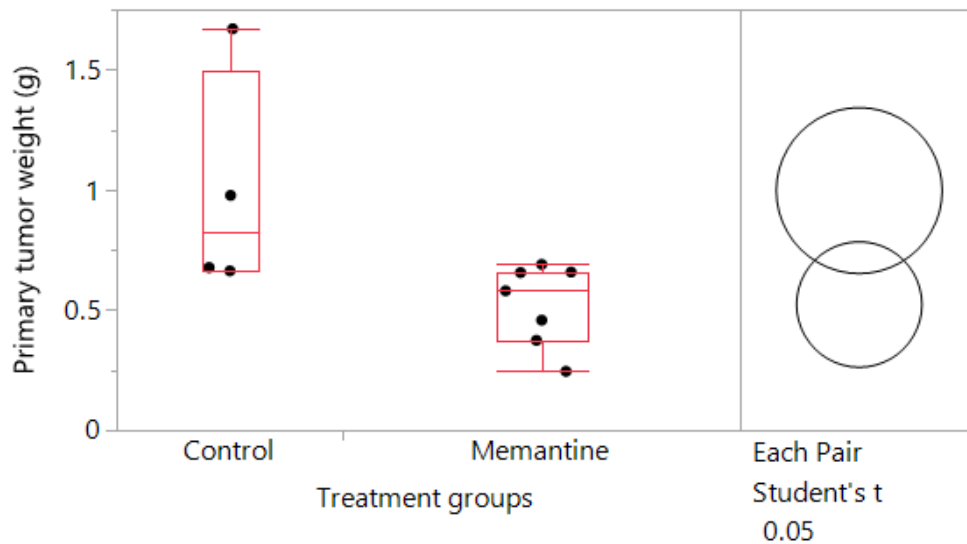
Abs(Dif)-LSD

	Control	Memantine
Control	-0.48811	0.04108
Memantine	0.04108	-0.36898

Positive value indicates significant difference.

Figure 149. Tumor weight *in vivo* of memantine vs. control group. Tumor weight were measured (g) after 2 months individually. Between the control group and the memantine treated group, there is a significant decrease in the tumor weight of the memantine treated tumor.

Memantine reduce metastasis of *in vivo* reconstructed tumor

With the same cohort described in the previous section. The livers were collected at the same time to survey the metastasis abundance for both the non-treated control group and memantine treated group. Visually, by eyes, there were obvious liver metastasis lesions in the liver of the non-treated group. On the other hand, the memantine treated cohort had zero to little visible (by eye) liver metastasis lesions. The observations is consistent across the cohort.
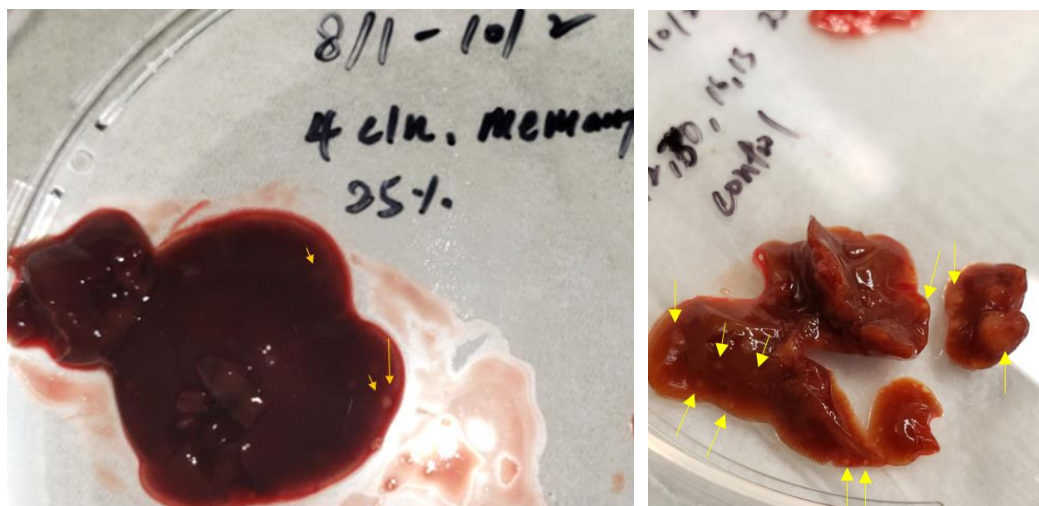
Figure 150. Representative picture of liver of *reconstitute tumor cohort with* memantine treated vs. control group. The left picture is a representative picture of the liver from orthotopically injected reconstituted tumor that received memantine treatment, where little visible liver metastasis was seen; right picture is a representative picture of the control group, where many liver metastasis is seen (indicated by yellow arrow). The observation is consistent across experimental and control cohort. Samples were collected on the same day.

Utilizing a more sophisticated quantitative method to assess the effects of memantine on tumor growth and metastasis, we processed the entire primary tumor, liver and the lung for DNA along with the spiked in scale (known cell number with unique barcodes). Then, through PCR amplification of barcodes and normalizing each barcode read to the spiked in scale's read

number, we could determine how many cells are present in the primary tumor, liver, as well as the lung. From the below figure, we could see the inhibitory effects of memantine on the reconstituted tumor's primary tumor expansion (~1/3rd smaller, from ~160M down to ~60M cells), as well as reduction of the amount of metastasis tumor in the liver (~1/7th, from ~11M down to ~sub-2M cells) and the lung (1/9th, from ~900K cells down to ~sub-100K cells). Therefore, we validated that *in vivo*, intrinsic upregulation of NMDAR pathway in clones that have more fitness and metastasis potential is targetable through memantine.
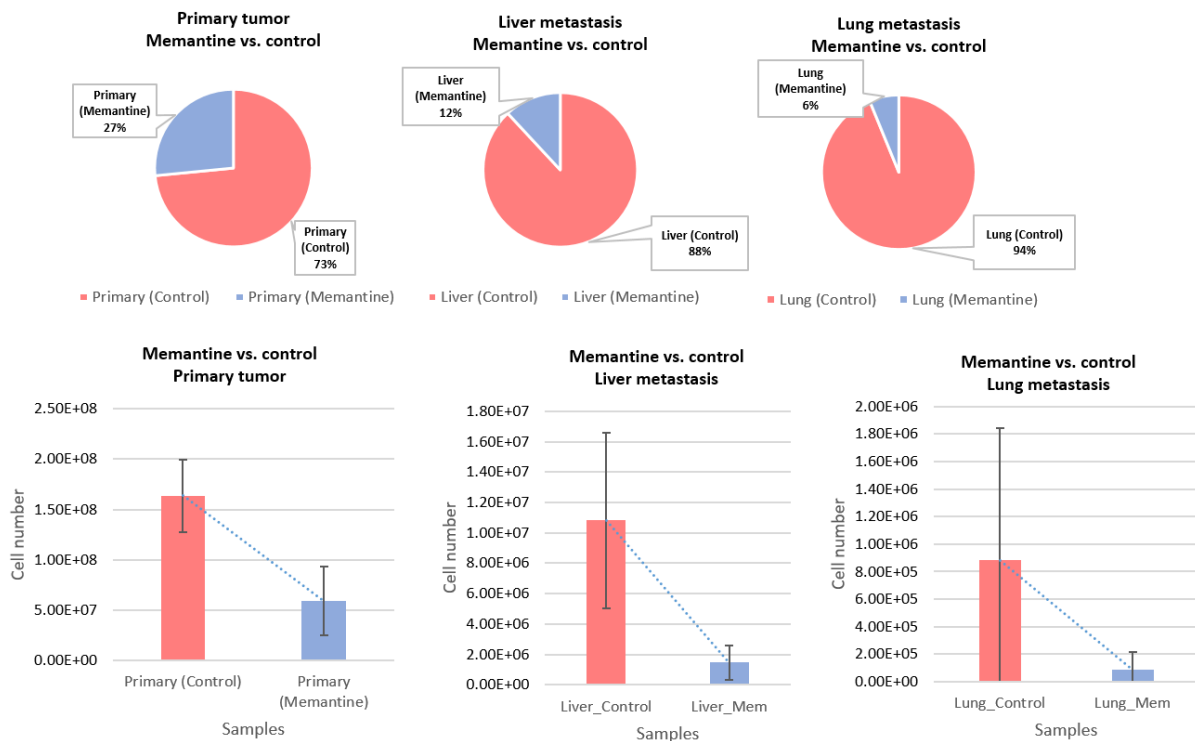
Figure 151. Quantitative assessment of tumor volume in the primary tumor, liver and the lung of reconstitute tumor cohort comparing memantine treated vs. control group. The top pie charts indicate the tumor volume's relative ratio detected in the primary tumor (pancreas), liver and lung (from left to right), in both the memantine treated (n=3) and the control group (n=3); The bottom bar graphs indicate the tumor volume in quantitative cell number (converted by normalizing scale during sample DNA processing).

Moreover, let us look at whether or not the reconstituted tumor lineages retain similar functional properties in terms of fitness and metastasization potential *in vivo.* As a reminder, we orthotopically injected a total of 3 million cells composed of equal amounts (25% each) of isolated sub-clones "12SG", "30FG", "16SG", and "FR13", where "12SG" and "30FG" having a strong fitness and metastasis phenotype, and "16SG" and "FR13" having low fitness and low metastasis phenotype from the longitudinal CRT data set. After two months post-injection, from the below figure, if we look at the primary tumors (Primary tumor control), we can see the two clones that had more fitness and metastasis potential (color in shades of red) dominating the reconstituted tumors' representation (>95%); on the other hand, the clones with lower fitness and less metastatic potential only represents a small portion of the primary tumor (color in shades of blue). In addition, the reconstituted tumor model recapitulates the metastasis phenotype from the longitudinal CRT data set), where the two clones that are more pro-metastasis indeed dominate the tumor at distal sites. Interestingly, although the memantine treatment has significant impact on tumor volume, it does not seem to have particular differential effects on particular clonal lineages (as seen from the table below, clonal relative ratio is similar between treated vs. control group). This could be due to the fact that the complexity of the barcodes in the reconstituted tumor is limited, hence, not able to fully evaluate the differential clonal response to memantine. Currently, we are treating the CRT tumors with higher barcode complexity with memantine to assess the differential clonal response in detail, as described in chapter one.

Figure 152. Qualitative assessment of lineage representation in the primary tumor, liver and the lung of reconstitute tumor cohort comparing memantine treated vs. control group. The top bar graph indicates the relative tumor representation (after two months) of the four clones (injected at equal amount). Barcodes "4611_12634", "5304_9756" are the clones with more fitness and metastatic potential based on the longitudinal CRT empirical data set, which are in shades of the color red; clones "1592_6046" and "6023_6472" are the clones with less fitness and lower metastatic potential, color shaded in blue; the bottom table indicates the precise number of tumor representation of the four clones in each site.

Pro-Metastasis signatures pre-exist in primary tumor and is predictive of patient survival

From the differential expression of genes obtained from the metastasis (high fitness) and the non-metastasis (low fitness) isogenic clones, we established a 200-gene set signature associated with "pro-metastasis". Matching the gene set signature to the TCGA pancreatic patient RNASeq data, we were able to separate out subgroups of patients having either enriched or decreased expression relative to the metastasis signature. Plotting the patient survival of these two groups, we find the "metastasis signature" can predict the prognosis of these two groups of

196

patients. Furthermore, by matching the metastasis gene signature to the single cell RNASeq data

of the primary tumor cell line (PATC124), we found there are 453/3,397 cells already enriched in
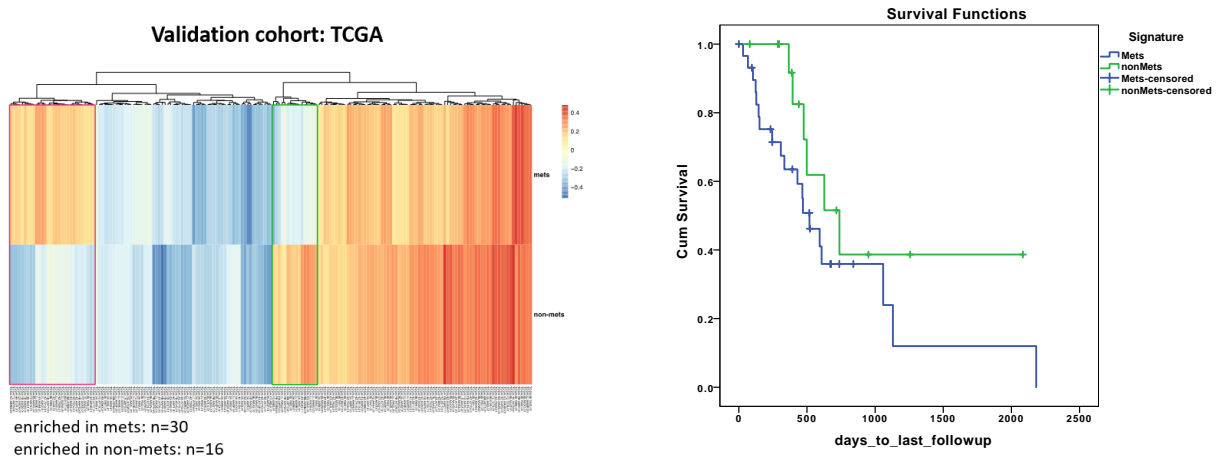
the signature.

Figure 153. Pro-metastasis signature matched to TCGA patient cohort can predict clinical prognosis. The left graph indicates the patients from the TCGA cohorts (column) and the top and bottom DE genes from the metastasis signature. 30 patients (red box) matched the metastasis signature, and 16 patients (green box) matched the non-metastasis signature. The right graph is the Kaplan-Meier survival curve for patients that match the metastasis and the non-metastasis signature.
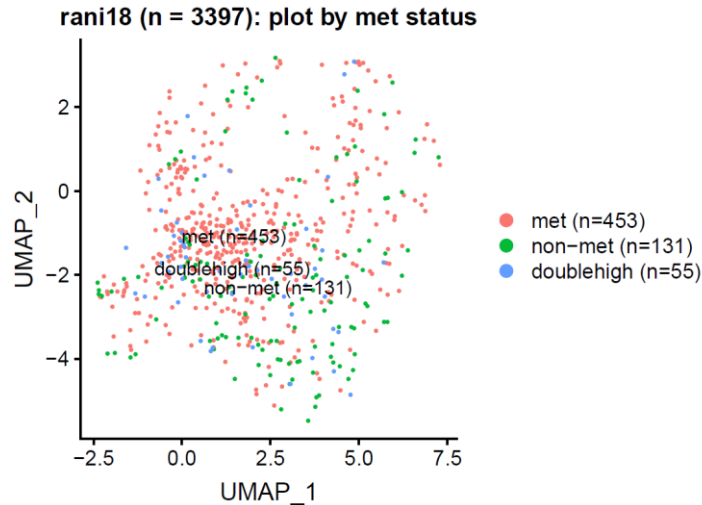
Figure 154. Pro-metastasis signature pre-exist in cells as measured by single cell RNA-Seq. Single cell RNAseq was performed on the parental PATC124 PDX-derived primary cell line. Out of 3,397 cells, 453 cells matches the pro-metastasis signature and 131 cells matches the non-metastasis signature. This indicates that the gene expression profile pre-exists in the parental population at a significant level. Judging from the distribution of the cells that match either signature, the cells seems to be dispersed throughout the clusters and not belonging to a specific compartment of the tumor cells.

Chapter Summary and Discussion

Clonal competition is one of the major functional mechanisms that shape how tumors evolve. As a tumor develops and expands, cell intrinsic properties that translate to functional fitness in terms of proliferation define the makeup and characteristics of such tumor. Therefore, a

tumor can be viewed as a complex ecosystem where sub-clonal lineages inter-compete and co-evolve, thus molding the tumor as a whole [32, 129,130, 134]. To understand the tumor's characteristics, it is logical for researchers to focus efforts on studying its building blocks – functional sub-clonal lineages – and how each of them responds to external pressure. Indeed, significant efforts in the field, with intentions to de-convolute how tumors progress and evolve over time, have been made to profile tumors at various stages of the disease. With genetic profiling of patient tumors, scientists have been able to determine main driver mutations that regulate progression in most types of cancers. For example, *KRAS, CDKN2A, TP53,* and *SMAD4* are associated with PDAC and *BRACA1, BRACA2, and TP53* with breast cancer [44, 132-135, 152-155]. Although big data analysis has provided us with a valuable grand view of cancer as a disease, tumors are, in reality, far more complex and heterogeneous. Cancer should be viewed as a "dynamic" disease that has the ability to adapt. Each patient's tumor sub-lineage may harbor unique mutations or modifications that may potentially influence a cell's response and adaptation to external stress [80, 81, 137, 138]. To be able to truly understand how a potential genetic or epigenetic change can influence a cell and its sub-lineage's functional state, it is essential for researchers to have the ability to longitudinally follow the same sub-lineage's progression over time.

Currently, there are no adequate *in vivo* models that allow for unperturbed longitudinal lineage tracing with good resolution and tumor representation [139-142]. As a common example, serial biopsy through fine needle aspiration in combination with whole exome sequencing at a single cell level provides the longitudinal aspect and single cell resolution for lineage tracing, but is severely limiting in the scope of representing the composition of the entire tumor. Moreover each serial sampling perturbs the tumor's ecosystem and the partial removal of the tumor may introduce artificial responses, which may prevent us from surveying the natural clonal dynamics of the tumor as it expands. Additionally, besides descriptive –omics profiling and an extrapolated lineage progression pattern, this method lacks the capability to functionally characterize clones of

interest and further investigate mechanisms related to specific clonal phenotypes. For reasons aforementioned, we aim to establish a high resolution, representative model that can quantitatively reflect clonal dynamics during tumor progression and dissemination without perturbing the tumor ecosystem.

Using human pancreatic ductal adenocarcinoma (PDAC) patient-derived xenograft (PDX)-derived cells, we generated orthotopic clonal replica tumors (CRTs), which are cohorts of animals bearing human tumors that are essentially identical in their clonal composition [109]. Additionally, we used a bootstrap simulation method, with a stringent criterion for similarity, and demonstrated that we could view each individual CRT in the same cohort as interchangeable, independent single variables over time [143]. Therefore, by surveying independent CRTs and their entire sub-clonal composition and abundance at multiple time points, we can empirically gain an unbiased representation of clonal dynamics during their natural expansion. In other words, similar to a cartoon flip book or a zoetrope, the model provides the capability to robustly trace lineage dynamics in parallel animals bearing clonally identical human tumors by creating a progressive time lapse video for each sub-clone based on additive snapshots.

By analyzing the clonal dynamics in the primary pancreas tumors, we immediately observed one single clone exhibiting massive clonal expansion within the first four weeks of *in vivo* transplantation. Surprisingly, while all the top abundant clones were relatively similar in percent abundance at the time of injection, only one clone displayed such an aggressive proliferation post transplantation. A possible explanation relating to external influence for this observation is that under *in vitro* culturing conditions, environmental pressure is relatively even (i.e. oxygen content, nutrition concentration) compared to that under *in vivo* conditions. After transplantation, pockets of distinct microenvironment niches favoring growth are unevenly distributed (i.e. hypoxia, blood vessel distribution), which thus leads to this difference in clonal expansion. The uneven clustering spatial distribution of clonal growth is further validated through

200

laser capture microscopy (LCM) sectioning across the entire primary tumor, where we observed, intra-clonally, a somewhat "patchy" spatial distribution that was not entirely regional. Interestingly, the same dominate sub-clonal expansion is seen across all CRTs, suggesting a strong lineage intrinsic linkage in relative clonal fitness.

Besides the clonal expansion dynamic, we also captured the "stochastic" divergent nature of tumor expansion by monitoring the sub-clonal abundance correlation between CRTs surveyed at the same time points. At early time points, each sub-clone of CRTs, progressed very similarly from week 4 to week 10 (correlation ~93.8%) despite becoming less and less concordant. It is not until the last time point, week 12, that we see a sudden increase in clonal divergence amongst CRTs (correlation ~73.4%). Keep in mind that we injected the same barcoded cells into all animals (correlation 99.9); however, over time, the clonal composition among tumors naturally becomes less and less similar. This suggests that any "minor" external factors amongst individual animals can shape lineage evolution, and that the difference can be appreciated and amplified with the progression of time.

Looking comprehensively at all the dynamics of sub-clonal lineages, we observed an astonishing and unexpected degree of inter-clonal oscillation during unperturbed tumor expansion (Fig. 78, 79, 89-94). Dynamics included 1) clones that displayed gradual loss or gain of fitness over time, 2) clones that displayed transient amplification followed by gradual loss of fitness, 3) clones that were mostly dormant followed by dramatic aggressive outgrowth, 4) clones that displayed a sigmoidal progression with gain-loss-gain of fitness, 5) clones that gained fitness for most of the time then followed by a sudden drop in fitness, and 6) clones that displayed a "U" shape fitness progression, where loss of fitness was followed by gradual gain of fitness. Frankly, some of these empirically obtained dynamic patterns might seem hard to believe at first, but all data and lineages were treated and processed with the same criteria. Therefore, if one can believe a clonal dynamic pattern of a "gradual gain of fitness over time", one should be open to believe

that the other observed clonal patterns, albeit unconventional, reflect true tumor biology. In addition, all of the above clonal progression dynamics were validated and observed in another independently barcoded CRT cohort, thus suggesting that these observed dynamics in the CRT model reflects sub-clonal tumor progression biology.

Clustering of clonal dynamics revealed a high degree of alternating clonal dominance (ACD) during unperturbed tumor expansion. Certain clonal lineages that were progressively dominant over time would experience a sudden drop in fitness at around week 10. Concurrent with this pattern's timeline, some clones that were low in abundance experienced sudden expansion and overtook the original dominating clones around week 10 (Fig.80). These ACD patterns were observed both with qualitative analysis (in clonal relative tumor % representation) and quantitative analysis (in absolute cell unit) of clonal abundance. These data again highlight the importance of viewing the tumor as an ever-changing entity, and that over time, different lineages would acquire different degrees of fitness in response to assumed variance in the microenvironment between individual mice as well as due to possible tumor self-induced environmental changes during progression. The potential cause of ACD in the tumor could be related to the tumor's constant propensity to expand. During expansion, clonal lineages are competing for resources, and the clones with the best fitness dominate the tumor's representation. However, by speculation, once a tumor reaches a certain size, necrosis, hypoxia, and nutrient deprivation start to factor into the clonal equilibrium equation, impacting clones with different intrinsic properties. Therefore, at this point, dominant clones may start to decrease in fitness, while other clones that are intrinsically better suited to adjust and adapt to the new environment will eventually become the more dominant population of the tumor. If this is true, one can expect the transcriptomic or epigenetic profile of the same clonal lineage in a tumor at an early stage to be drastically different than that at a late stage because the tumor itself could have gone through a radical shift in internal environmental properties and its clones must thus adapt and respond to

such changes. It also signifies that although tumor mass may stay relatively constant, the tumor sub-clonal composition may change drastically over time. This is in line with the concept of adaptive evolution as well as the concept of adaptive resistance in the context of treatment. Of note, adaptive evolution has been described by Rosenberg and colleagues, where under starvation, *E. coli* underwent loss-of-function mutations in proliferation associated genes and back mutated to allow for survival fitness [144-148, 224,225].

Moreover, looking at the less represented lineages in tumors at each time point surveyed, there were clones that were constantly proliferating, gaining in fitness persistently throughout and popping up – which thus suggests a rich "reserve repertoire" of clones waiting to thrive (Fig. 80). This also revealed the highly competitive nature amongst clones at the bottom of the pyramid. This naturally occurring ACD also has practical implications. For instance, it suggests that clinical decisions should be made on recently acquired biopsy profiling information, as dominant clones could alter over time. In addition, the notion that targeting the contemporary dominant clones of a tumor is not enough, as there are presumably many other lineages with distinct intrinsic properties waiting to take over [149-150]. Furthermore, the longitudinal CRT model that revealed alternating clonal dominance and the rich repertoire of low abundant sub-lineage's persistent dynamics may also serve as a vital tool to refine treatment strategy concepts such as adaptive therapy, where lineage-specific functional characteristics are taken into consideration during treatment to prevent tumor resistance.

In addition to studying the longitudinal lineage progression of the primary tumor, we also surveyed the all distal organs of CRT animals for the dissemination dynamics of tumor lineages over time. To this end, at each time point, we harvested the whole lung, liver, and blood from the same mouse in which CRTs' primary tumors were collected. Using the same clonal barcode detection methods, we obtained both longitudinal qualitative (relative %) and quantitative (cell unit) information on specific sub-lineages that displayed unique patterns of dissemination, seeding

at distal organs, metastasis dormancy and latency, and organotropism. We were also able to catalogue the clonal composition of metastasis lesions and the degree of the stochastic nature of metastasis amongst different organs and their outgrowth patterns. From a grand point of view, only a fraction of the sub-clones had the ability to metastasize (~1%-5%); further, barcodes were detected in all organs as early as week four, suggesting that tumor cells can indeed seed early on [156-159]. The number of tumor cells per lineage that sheds into the blood increases as the tumor progresses; however, for the liver and lung, we observed a sub-population of lineages that displayed a "transient metastasis" phenotype, of which the early stage (at week four) had more cells per lineage than late stage. In other words, some sub-clones may have the ability to disseminate early on and may even accumulate into a mass, but they can never be sustained and will eventually lose fitness over time. Looking at the clonal lineage progression dynamics in the distal organs (liver and lung), we see a more one dimensional dynamic than what has been observed in the primary tumor. In general, lineages that are responsible for metastasis in a distal organ displayed a long term "latency" of ~10 weeks. The fitness of each clone remained relatively stable even though the clone steadily increased in mass. After a long period of "latency", the clones present at distal organs exhibited an explosive expansion, where the metastasis proliferated at a large degree (Fig. 102). In terms of tumor mass, the metastasis and ascites' accumulative mass exceeded that of the primary tumor at late stage. By surveying the clonal composition of individual metastasis lesions (or nodes), we found that the metastasis lesions were oligoclonal but displayed monoclonal dominance in nature. Whether or not the oligoclonal lesions occur during seeding or after the establishment of metastasis nodes has yet to be characterized [42, 160]. However, one thing we are certain of is that the complexity of inter-clonal competition in the liver seems to be far less than what is occurring in the primary tumor, which corresponds with the observation of wider spatial distribution pattern of liver metastasis. Differences in selective environmental pressure between organs may also contribute to the latency of a clone to adapt before massive outbursts in growth.

Additionally, we observed an interesting correlation between the outburst of metastasis and the timing of ACD in the paired primary tumor. Most of the ACD in the primary tumor occurred around ~week 10, which was also when metastasis clones started to aggressively proliferate. While this was purely hypothetical at this time, I wondered if metastatic outgrowth might be linked to a change in the metabolic state of the primary tumor. In other words, as the primary tumor proliferated to a certain size, factors such as nutrient deprivation and hypoxia could increase, and/or the tumor could become more necrotic, resulting in a tumor growth self-induced internal microenvironment and tumor metabolic change. These changes in the primary tumor could influence distal metastasis lineages and stimulate growth through chemokines or exosomes. This hypothesis is in line with Notta and Co.'s paper in 2016, where through tracking CNV, the authors found that the majority of pancreatic tumors harbored complex genome rearrangement signatures related to mitotic error and that PDAC followed punctuated equilibrium as its main evolutionary trajectory [151]. Further, Marusyk and Co's paper in 2014 indicated that the sub-clonal population drove tumor expansion through enhancing the proliferative capability of other cells by overcoming extrinsic constraints—however, this could be outcompeted and lead to the collapsing of the tumor [101]. This hypothesis can be tested in the future to determine whether ACD-related tumor progression self-induced change in metabolic state is a major factor in triggering distal metastasis proliferation, and if the reverse of concomitant tumor resistance exists[168].

On the topic of organotropism, we observed that the sub-clones that metastasize in the lung only partially overlap to those of the liver. Multiple sub-clones thrive in an organ specific manner. By correlating two lungs or two livers collected at the same time point, we find that clonal lineage complexity overlaps ~60% in each distal organ (lung to lung, liver to liver). This suggests that metastasis progress is highly stochastic, although their paired primary tumor's correlation is >90%. The analysis of sub-clonal abundance of lung vs. lung and liver vs. liver at the same time point suggests that metastatic outgrowth in the liver is more consistent than that in the lungs

(Fig. 112, 113). This highlights a possible mechanical aspect of tissue dependent cell entrapment or intercellular adhesion's contributions to metastasis, as well as the existence of different organ specific niches[161-166].

By comparing sub-clonal abundance in the primary tumor versus the sub-clonal capability to metastasize, we determined that tumor lineage abundance had a high correlation with their ability to disseminate. However, although clonal abundance was a major factor, not all clones behaved in this way. In the CRT cohorts, we also observed multiple sub-clones that were lowly represented but could metastasize and proliferate, as well as multiple sub-clones that were highly represented in the primary tumor throughout the longitudinal study but did not metastasize. Therefore, the data suggest that the factors that contribute to metastasis sees to be both influenced by lineage fitness in the primary tumor, as well as by the intrinsic nature to disseminate in the sub-clone.

To dive more into the nature of metastatic clones, an experiment was designed to normalize each sub-clone's abundance (equal) in the primary tumor during injection and measure its ability to metastasize (orthotopic injection right after barcoding followed by minimal amplification). Additionally, in parallel, through *in vivo* serial transplantation and using the same population of cells right after barcoding, each lineage's ability to sustain long term tumor growth was determined (4 generations, 6 months, generation F1 to F4). The data showed that lineages that were able to metastasize contributed to only a tiny fraction (1.093%) of the injected population and that their relative abundance in the injected population was equally spread throughout the entire abundance distribution (Fig. 124). This suggested that the metastatic potential of a sub-clone was comparatively an intrinsic property and independent of the initial abundance. Furthermore, those same lineages that were able to metastasize contributed to 93.9% of the *in vivo* serial transplanted tumor mass in F4, indicating that part of the clonal lineages that could sustain tumor for long-term and contribute to the majority of the tumor mass had the intrinsic ability

to metastasize. In other words, the data strongly suggest that an intrinsic long-term self-renewal (LTSR) property is a crucial factor for metastasis. Therefore, the metastasis phenotype is governed by both intrinsic LTSR and fitness properties (abundance) of a lineage in the primary tumor, and those two factors are essentially intertwined [157-159, 169-171].

Since the lineage's intrinsic LTSR properties contribute to clonal fitness and potentially metastasis, we aimed to isolate isogenic clonal lineages that were metastatically naïve (yet to go through the dissemination process) and then perform functional characterization for mechanisms leading to the observed differential phenotypes in the CRT longitudinal study. Through a minimally perturbed high-throughput clonal isolation method, we generated multiple isogenic clones that displayed a wide array of clonal dynamics from the same CRT studies. We focused on six clones displaying the most extreme phenotype – three clones that had high fitness in the primary tumor and metastasized, and three clones that had low fitness in the primary tumor and did not metastasize. *In vitro* validation determined that the three pro-metastasis clones had a higher invasive phenotype than the non-metastasis clones, and that all six clones were able to engraft and establish tumors *in vivo*. Interestingly, the six clones all shared common PDAC driver mutations (*KRAS, TP53*), but all had unique diverse mutational landscapes as well as copy number profiles. Through differential expression of the RNAseq data, we identified N-methyl-D-aspartate receptor (*NMDAR, GRIN2A* gene) to be constitutively activated and relatively highly expressed in the pro-metastatic clones. Treating the clones with memantine, a non-competitive antagonist of NMDAR, effectively decreased the invasive ability of the metastasis clones both *in vitro*, as well as *in vivo* through reconstituted tumor. This suggests that NMDAR is one of the major pathways regulating metastasis, and that targeting this pathway can potentially reduce the degree of metastasis in patients. Furthermore, based on the top differentially expressed genes amongst clones displaying differential metastasis phenotypes, we generated a "pro-metastatic" gene set signature. Matching the "pro-metastatic" signature back to the single cell RNAseq data

from the CRT tumor cell lines, we found a subset of cells that were already enriched in the signature (~13%), suggesting that part of the tumor derived from patients already had the tendency to metastasize. Finally, through matching the "pro-metastatic" signature to TCGA patient data, we found that the signature could effectively predict patient prognosis in the TCGA PDAC cohort.

In conclusion, molecular technological advancements have revealed that tumors are much more complex than what we had initially thought, and that merely measuring tumor mass to determine or predict tumor behavior is vastly insufficient. The longitudinal CRT model provides the ability to not only measure the clonal heterogeneity of a tumor, but also the capability to probe into the longitudinal aspect of individual lineages of a tumor. This study has allowed us to appreciate the degree of functional diversity occurring naturally during tumor expansion, and demonstrates that tumors are a highly dynamic entity that is ever-evolving[187, 188].

Chapter 3 Abundance independent preservation of clone-to-clone ratio in cancer and the prospective concept of "tumor clonal fingerprint"

Introduction

Tumor clonal fingerprint – tumors could be defined by intrinsically maintained clonal-to-clonal lineages composition

In this chapter, we discuss a prospective concept of tumors maintaining their functional characteristic through preserving sub-lineage hierarchy. Through tracing of tumor lineages that has been cultured and stabilized in *in vivo* or *in vitro* conditions (will go into detail later), we observed an interesting result. The result shows that a population of lineages displaying higher fitness *in vitro* and lower fitness *in vivo*; at the same time, there is another population of lineages displaying the opposite phenotype with higher fitness *in vivo* and lower fitness *in vitro*. However, for either population of lineages, independent of the fitness, within that population the relative clonal-to-clonal ratio is preserved. This suggests the possibility of an inherited tumor inter-clonal hierarchy, where the ratio amongst certain lineages is preserved independent of environment at an equilibrium state. In other words, there exists an inter-clonal ranking profile that could potentially represent the functional make up of a tumor; furthermore, the inter-clonal ranking profile is intrinsically maintained in the tumor at low abundance under unfavorable environments. This observed inter-clonal ranking could hypothetically be served as a tumor's ID, a "**T**umor **C**lonal **F**inger**p**rint" (TCFP), to help us define tumor in the context of functional sub-units (sub-lineages) make up as well as apply it to guide further treatment effectiveness strategies.
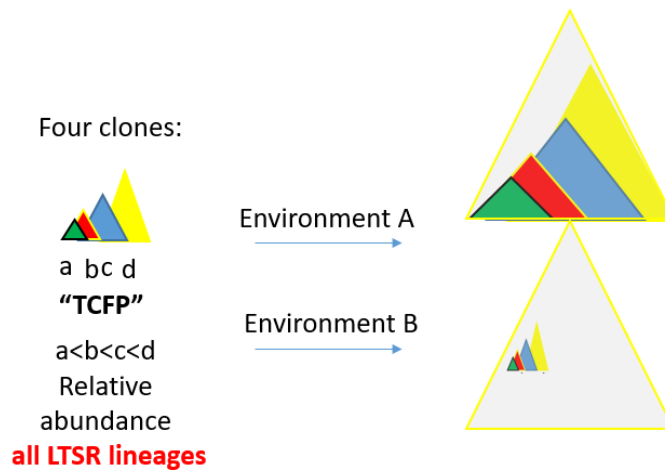
Figure 155. Perspective Tumor Clonal Fingerprint (TCFP) model concept schematics. Relative clonal ratio is maintained independent of environment. Under environment A, subset of clones would gain fitness and expand exponentially, as oppose to under environment B, the clones would lose fitness. However, the relative clonal ratio of each lineages is maintained; therefore, the tumor's clonal "profile" in the context of functional sub units (clones) is preserved despite of environment.

This concept of how a tumor's functional characteristic may be defined and preserved by clonal lineages and their hierarchy (in terms of relative ratio) is in line with the clinical concept of "adaptive therapy". Adaptive therapy refers to applying a controlled treatment regimen to tumor while avoiding the change in the tumor's overall characteristic (sensitivity to treatment), as opposed to hitting the tumor with intense treatment resulting in minor population of resistant cells giving rise to an entire tumor that is "different" in characteristic and resistant. Considering adaptive therapy under the context of sub-lineages, which is the functional sub units of tumor, adaptive therapy aims to maintain the relative ratio between individual lineages; therefore, the functional characteristic of resistance of the tumor is preserved. This contrasts with having intense treatment that disrupt the relative clonal equilibrium, leading to the change of relative clonal ratio, where more resistant clones dominate the relapsed tumor. This is in line with the observation of TCFP. If this TCFP exists and can represent the accumulative functional compartments of tumor, we

could potentially use TCFP as a matric to measure the effect and follow tumor progression under

treatment.
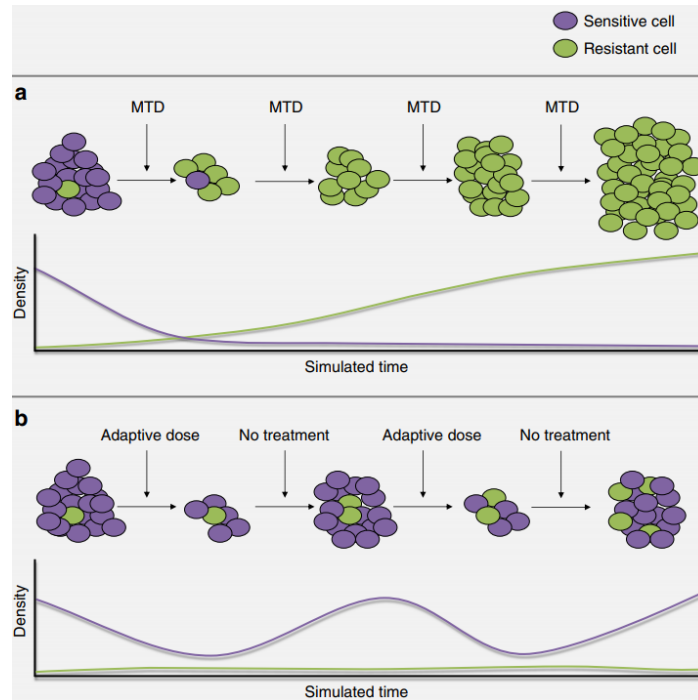
Figure 156. "Illustration of the designed evolutionary dynamics in adaptive therapy. a,b The purple cells are sensitive to the treatment and the green cells are resistant. The graphs represent the simulated density of each population over time during treatment. The top row represents standard therapy in which the maximum tolerated dose is given continuously after initiation. The cells sensitive to treatment are eliminated quickly. This intensely selects for cells that are resistant to the treatment, in this case T− cells, and eliminates the competition effects of the T+ population, resulting in competitive release with rapid treatment failure and tumor progression. The bottom row represents an evolution-based strategy in which therapy is halted before all of the sensitive cells are eliminated. In the absence of therapy, the sensitive cells out-compete the resistant cells due to their fitness advantage. This "steers" the tumor back to the pretreatment so that it remains sensitive to treatment. The resistant cells, or T− population, will increase slightly with each cycle so that this treatment eventually fails. However, mathematical models demonstrate control may be durably maintained for up to 20 cycles - significantly longer than continuous therapy". Figure and legend quoted from Zhang, J., Cunningham, J.J., Brown, J.S. et al. integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nat Commun 8, 1816 (2017). https://doi.org/10.1038/s41467-017-01968-5

Here, I will walk through the experiments in a stepwise manner and discuss the results

sequentially. Part of the data has already been discussed during previous chapters; however, to

make this an independent chapter for readers, I would reiterate the key experimental details. For

those who are familiar with the experimental details, you may skip directly to the analysis section: "Inverse correlation of clonal abundance from *in vitro* vs. *in vivo* stabilized experimental arms".

Results

Lineage tracing via barcoding of tumor cells

As discussed in the CRT section, by utilizing lentiviral barcoding technology on early passage of patient derived xenograft primary tumor cell line with a low multiplicity of infection and puromycin selection of barcoded cells (integrated vector contains puromycin resistance), we can create a culture of tumor cells each with a unique barcode integrated in the genome. Since the barcode sequence is integrated into the genome, once the cells proliferate and divide, the barcode sequence would be duplicated and passed onto daughter cells and their progenies. Due to this, we can use these barcodes to trace cell lineages over time.
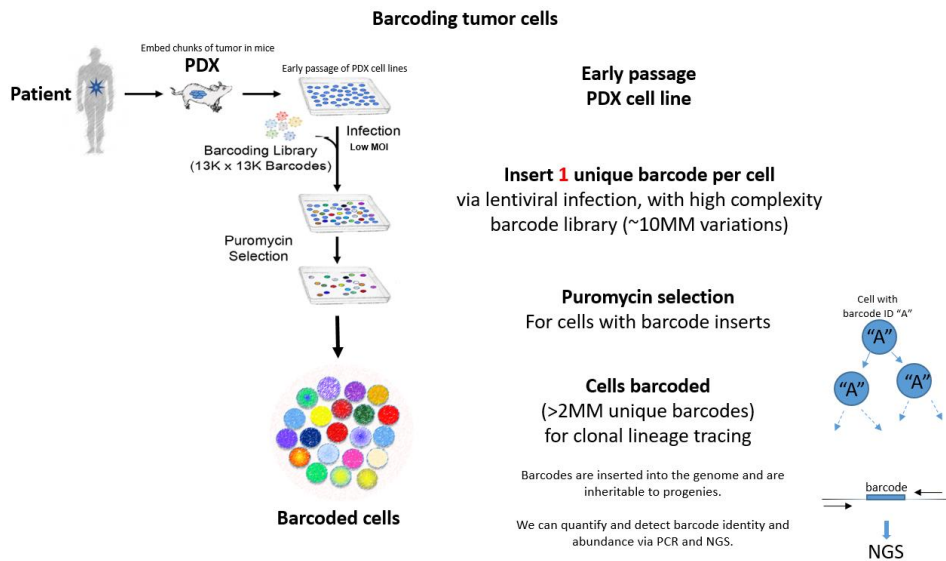
Figure 157. Schematics of barcoding of PDX derived tumor cell lines. Early passage of PDAC patient PDX derived cell lines are infected with lentiviral barcode library (~10M unique barcode types). Using a low MOI, most cells will get one unique barcode integrant. After puromycin selection of cells with barcode, we create a pool of cells that could be used for lineage tracing.

Parallel *in vivo* vs. *in vitro* tumor lineage tracing

To measure which cell lineages are endowed with long-term self-renewal properties, after puromycin selection, we expanded the barcode tumor cells ("passage 1") and then took half of the culture and performed serial transplantation *in vivo* for four passages over a period of 8 months (F1, F2, F3, and F4). *In vivo* serial transplantation is a functional assay to measure which cell lineages could sustain tumor growth and re-engraft over a long period of time. Therefore, the barcode lineages that are still present in the F4 tumors after 8 months, by functional definition, should be the cell lineages that have long-term self-renewal properties.

As for the other half of the "passage 1" barcoded cell culture (identical barcode population as the *in vivo* serial transplantation experiment), we continued to passage them *in vitro* (P2, P3, P4…P21) over a period of 8 weeks. At the end of the *in vitro* passaging, the cells were injected into NSG mouse to see which cell lineages, after *in vitro* passaging for a period of time, could engraft *in vivo* and form tumor.
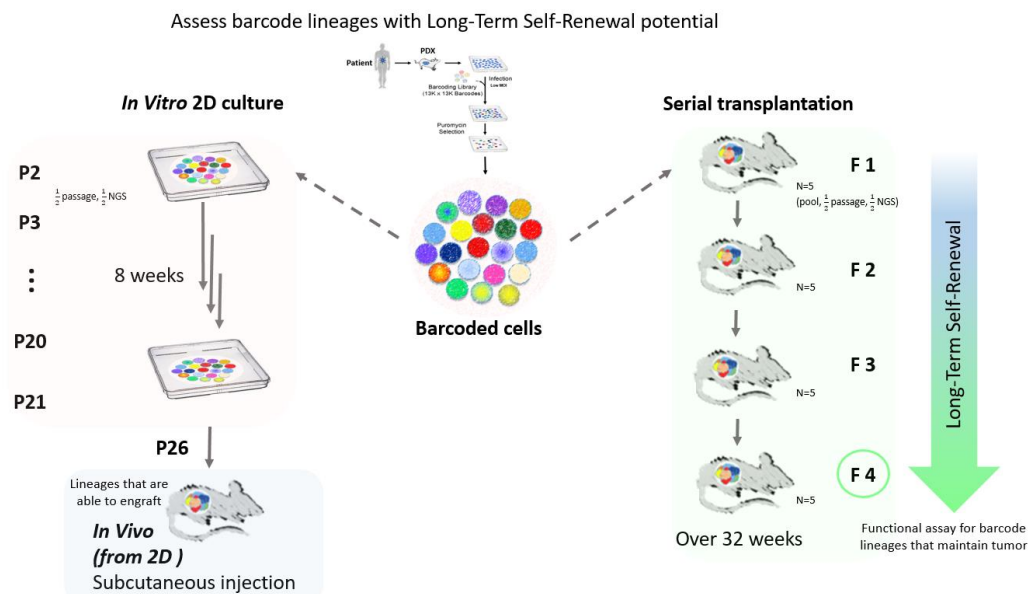
Figure 158. Parallel lineage tracing experiments *in vitro* and *in vivo* conditions. Barcoded PDAC cells are briefly expanded and subjected to two different long- term culturing conditions: *in vivo* or *in vitro*. The *in vivo* serial transplantation from F1 to F4 (each generation has 5 animals) would

inform us of lineages with long-term self-renewal properties (what is left in F4), whereas the *in vitro* passages (p2 to p21) would allow us to expand cells efficiently to create clonal replica tumors.

Tumor lineages established from *in vitro* stabilization is part of the long-term self-renewal compartment

By tracking the barcode complexity (barcode lineages) from the *in vivo* serial transplantation experiments (F1 to F4), as well as the *in vitro* passaging P2 to P21 and finally *in vivo* injection, we can see that although the total barcode complexity started the same with ~2,000,000 unique barcodes, as the passages increase, certain barcode lineages would disappear and exhaust (the blue lineages); certain lineages displayed transient amplification, where it would increase in abundance but eventually exhaust (green lineages); finally, there lineages that increase in tumor/culture representation and sustain most of the tumor volume and culture representation (in red, Figure 159). In terms of lineage behavior, both the *in vivo* and the *in vitro* experimental arms behave similarly, where >99% of the barcode lineages eventually exhaust and only a fraction of the lineages were able to sustain the tumor and culture over a long period of time. What is amazing is that, in the beginning, at "passage 1", both the experimental arms started off with ~2,000,000 barcoded clonal lineages, and over time, the majority of the lineages eventually exhaust and a common set of lineages were present in both the *in vivo* and the *in vitro* arm of the experiment in the end. This is denoted by the t-SNE plot, where the grey portion are the lineages that were exhausted and the red portion represent the clones that sustained the culture and tumor, and that these clones converged over time. This suggests that a minor fraction of common clonal lineages are able to maintain the *in vivo* tumor or *in vitro* culture, and that this is a lineage intrinsic behavior independent of culturing environment (Figure 159).

Furthermore, if we compare the "F4 tumor" clonal composition with the "*in vivo* tumor from *in vitro* stabilized cell culture", we find a common set of clonal lineages that make up the majority

of both tumors (70% and 97% respectively). Simply put, the clonal lineages (n= 1204) that make up basically the entire tumor derived from the *in vitro* stabilized culture falls in the long-term self-renewal compartment, as defined by functional *in vivo* serial transplantation experiment (those lineages present in F4 tumor).
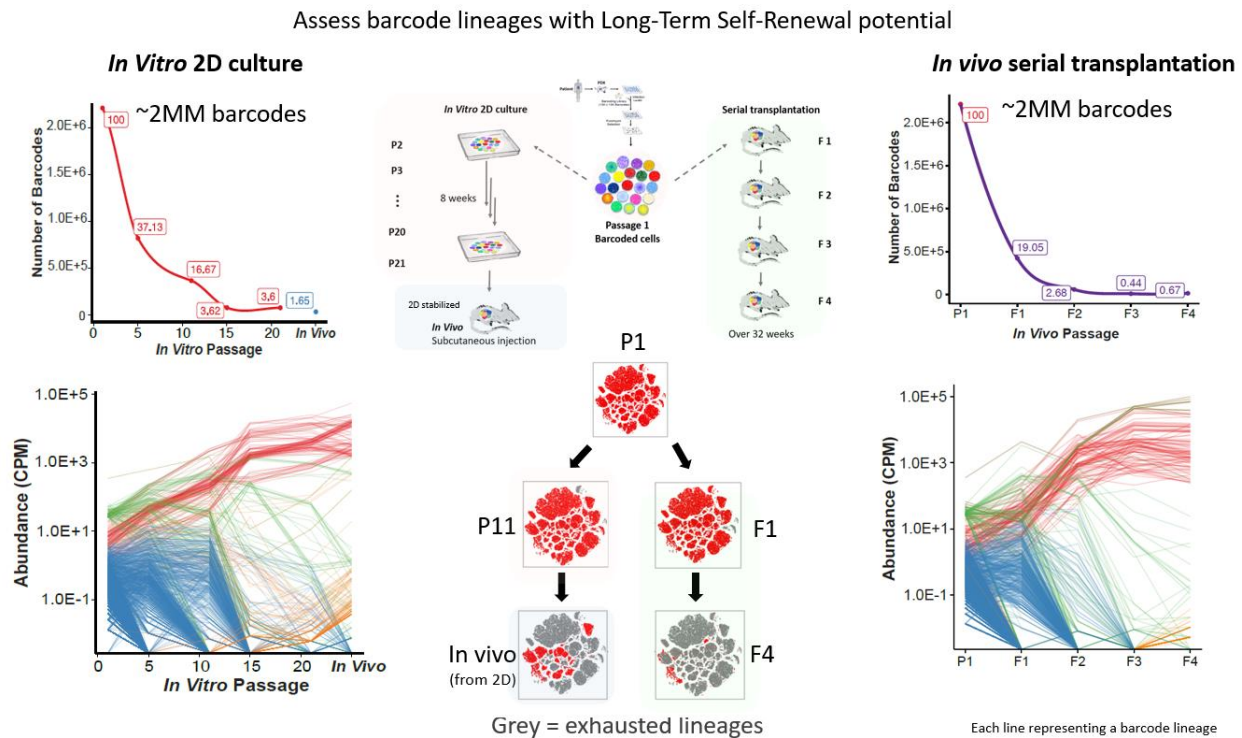
Figure 159. Barcode lineage tracing experiment *in vitro* and *in vivo*. Lineages reach equilibrium over time despite of culturing condition. Starting with the same population of barcoded cells, both *in vitro* and *in vivo* serial passaging display similar lineage dynamics over time, where most barcoded cells experience exhaustion (drop in barcode complexity) with only a portion of lineages sustain grow. The lineage complexity eventually becomes stabilized at 3.6% and 0.67% respectively (top left curves for *in vitro*, and top right curves for *in vivo*). Tracing the individual lineages abundance over time (bottom left and bottom right), we observe clonal exhaustion over time (in blue), as well as transient amplifying lineages (green) and the lineages sustaining tumor (red). T-SNE plot in the middle shows over time, regardless of *in vitro* or *in vivo* condition, the exhausted cells/lineages are the same (grey); also, the lineages that remain are also overlapping and converging (red).

Inverse correlation of clonal abundance from *in vitro* vs. *in vivo* stabilized experimental arms

Now that we have established that tumor lineages from the *in vitro* and *in vivo* experimental arms are LTSR clonal lineages, an interesting observation is made regarding the relative abundance of those individual LTSR clones in these two experimental arms (*in vitro* vs. *in vivo*). From graphing out the "*In vivo* tumor (cells stabilized in 2D)" vs. the "F4 tumor (serial transplantation)" side by side and looking at individual clonal abundance in these tumors, we observed an inversed relationship between the two samples. In other words, clones that are highly abundant in the "*In vivo* (2D)" tumor, is lowly represented in the "F4" tumor (Figure 160) and vice versa. To further confirm this observation of abundance inverse relationship, I plotted an X-Y scatter plot of relative clonal abundance by ranking of "*in vitro* passage P21" vs. "*In vivo* (2D stabilized)" tumor (Figure 160, each dot on the plot representing a lineage). From the plot, as expected, we can see that the correlation of the "*in vitro* passage P21" vs. "*In vivo* (2D stabilized)" is relatively high, since the "*In vivo* (2D stabilized)" tumor is derived from the *in vitro* passages in the same experimental arm. On the other hand, if I plot the same scatter plot for the "*In vivo* (2D stabilized)" vs. "F4", we see that the two tumors' clonal abundance, albeit being made up of the same LTSR clonal lineages, is evidently inversely correlated.

In Vitro 2D culture

Serial transplantation

Converged lineages
(each dot represents a unique barcode)

In vivo (from 2D) %

high

low

low   F4%   high

Axis – clone's relative % abundance representation in each tumor

- Two main populations
- Inverse abundance;
  (what's high is low in the other)
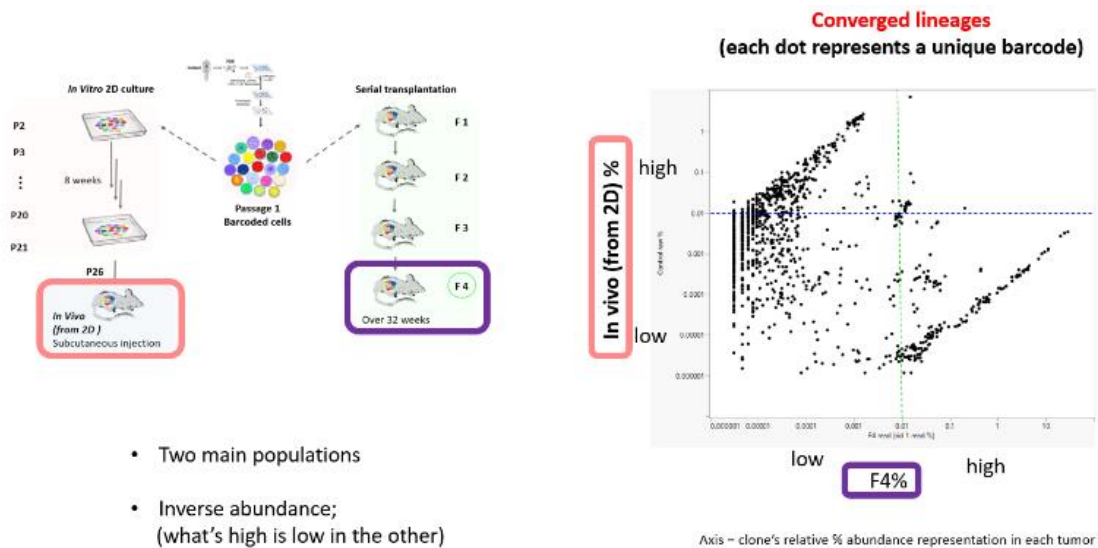
Figure 160. X-Y scatter plot of *In vitro* stabilized lineage and *in vivo* stabilized lineage abundance showing inverse correlation.

Most importantly, what is perplexing to me is the order of the clones aside from their inverse abundance. From the X-Y scatter plot of "*In vivo* (2D)" tumor vs. "F4" tumor, we can noticeably see two populations, yet the ranking of each clone is relatively maintained regardless if they are low in abundance in one tumor or highly present in the other, resulting in two well-correlated populations (Figure 160).

Since these barcoded clonal lineages were cultured and passaged under different environments (starting out the same and then split into the two experimental arms, one *in vivo*, and the other *in vitro* 2D culture and then transplant *in vivo*), it is logical to assume certain lineages would adapt and thrive under one environment, while some lineages would adapt better in the other, and thus resulting in the inverse abundance of tumor representation from "*in vivo* (2D)" tumor vs. "F4" tumor. However, I was not expecting to see clonal lineages retaining their relative ranking (clone-to-clone ratio) in such high correlation. I was expecting to see something less organized, where certain lineage population would grow better in the *in vivo* environment over

217

time, and vice versa in the *in vitro* environment, without the clonal lineages maintaining their relative order /ranking (Figure 161).
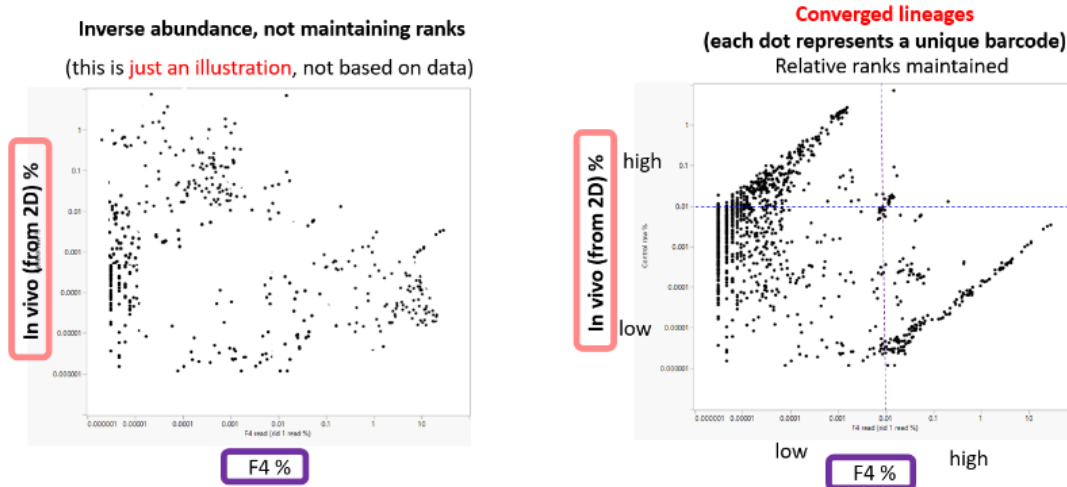
Figure 161. Inverse correlation with and without maintaining relative ranking. Hypothetical scatter plot of inverse correlation without maintain relative clonal ranking (left) vs. what is observed, where *in vitro* and *in vivo* stabilized clones are inverse in abundance and also maintaining relative ranking (clone-to-clone ratio).
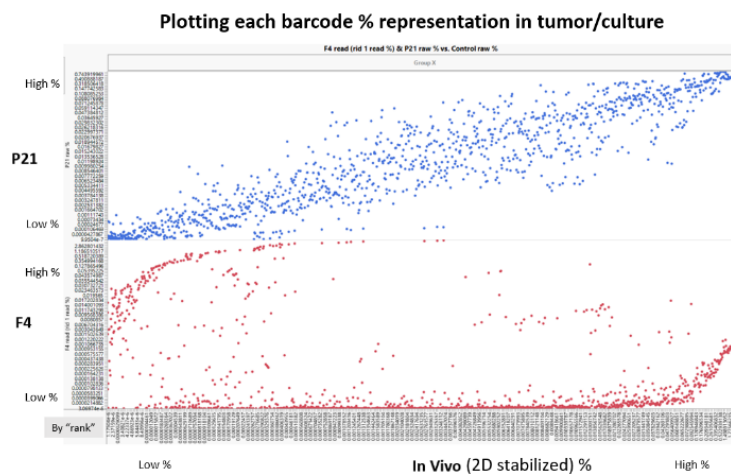
Figure 162. X-Y scatter of the lineage % representation of the last *in vitro* passage P21 vs. the injected tumor derived from P21. As an analysis control, a side by side comparison of *in vitro*

passage P21 vs. P21 derived "*in vivo* (2D stabilized) tumor" and *in vivo* F4 tumor vs. "*in vivo* (2D stabilized) tumor". We expect the P21 having positive and good correlation with "*in vivo* (2D stabilized) tumor" and it is what the data shows.

Longitudinal lineage tracing suggests clonal-to-clonal clonal ranking in tumors are preserved

To further validate the observation of preservation in clonal ranking, I first plotted the lineages representation in both linear and log scale, as well as based on % tumor representation and relative ranking to confirm that there are two major populations that are indeed positively correlated. For example in Figure 163, lineages representing above 0.1% of the "*in vivo* (2D)" tumor mass and below 0.01% of the "F4" tumor mass has a Pearson's correlation of 0.81 (p=0.00001); while the lineages representing above 0.1% of the "F4" tumor mass and below 0.01% of the "*in vivo* (2D)" tumor mass has a Pearson's correlation of 0.86 (p=0.00001).
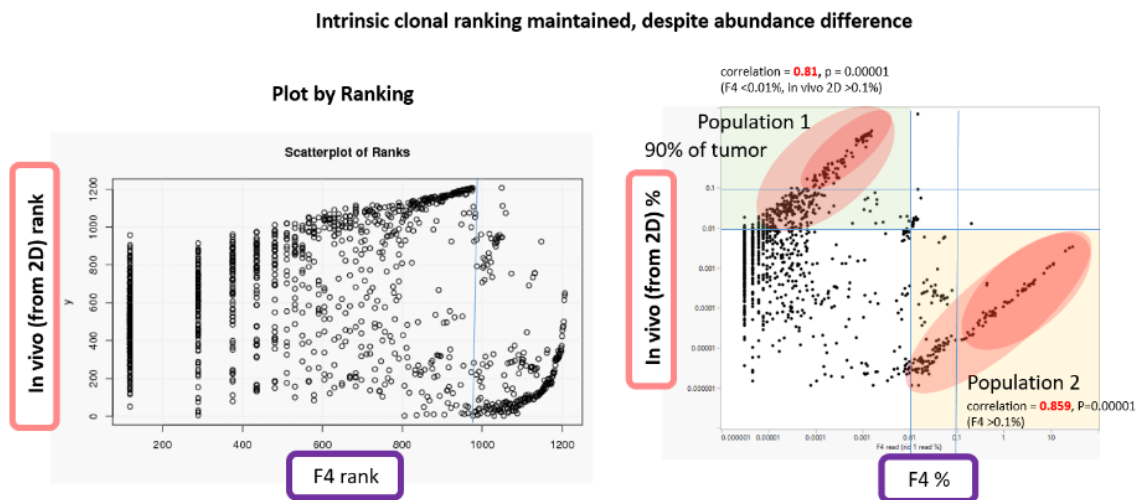
Figure 163. X-Y scatter plot of *in vitro* vs. *in vivo* stabilized clonal relative abundance by ranks. Left plot shows *in vitro* vs. *in vivo* LTSR clones plotted by rank. Right graph shows the plot in relative % tumor representation value. Both plot shows an inverse correlation in abundance, but

with high correlation in relative ranking (0.81% - 0.859% Pearson's correlation for clones that makes up 90% of the tumor mass).

One possibility of artificially creating such phenotype where common clonal lineages that are inversely correlated in abundance but maintain highly correlated in terms of relative lineages ranking is contamination. Imagine having two separate barcoded samples, and a tiny amount of "Sample No.1" got into the other sample "Sample No.2". In this example, the contamination would cause a false positive read out in "Sample No.2" at a lower abundance; in addition, the relative clonal ranking would have high correlation with what is highly abundant in "Sample No.1" due to the contamination (coming from "Sample No.1"). However, the probability of the observed phenotype is due to contamination is not likely. During the experiment, the barcoded tumor / cultures were processed separately in multiple batches to prevent cross contamination. Furthermore, the PCR primers has separate indexes for each sample and therefore, further preventing false positives due to cross contamination. Moreover, looking at the X-Y scatter plot of lineages from "*in vivo* (2D stabilized)" tumor vs. "F1", "F2", "F3" and "F4" tumor longitudinally, we can clearly see that the two positively correlated populations have an organic (gradual) separation starting from "F1" to "F4" in a span of 8 months (Figure 164). This suggest that the observation we are seeing here is a natural behavior and not likely due to artificial reasons such as cross contamination of barcodes. Lastly, some of these lowly represented clones do come back up during late passages, regaining a certain degree of fitness, suggesting the dormancy nature of some of these lineages and that, once again, these barcodes are not present due to contamination.
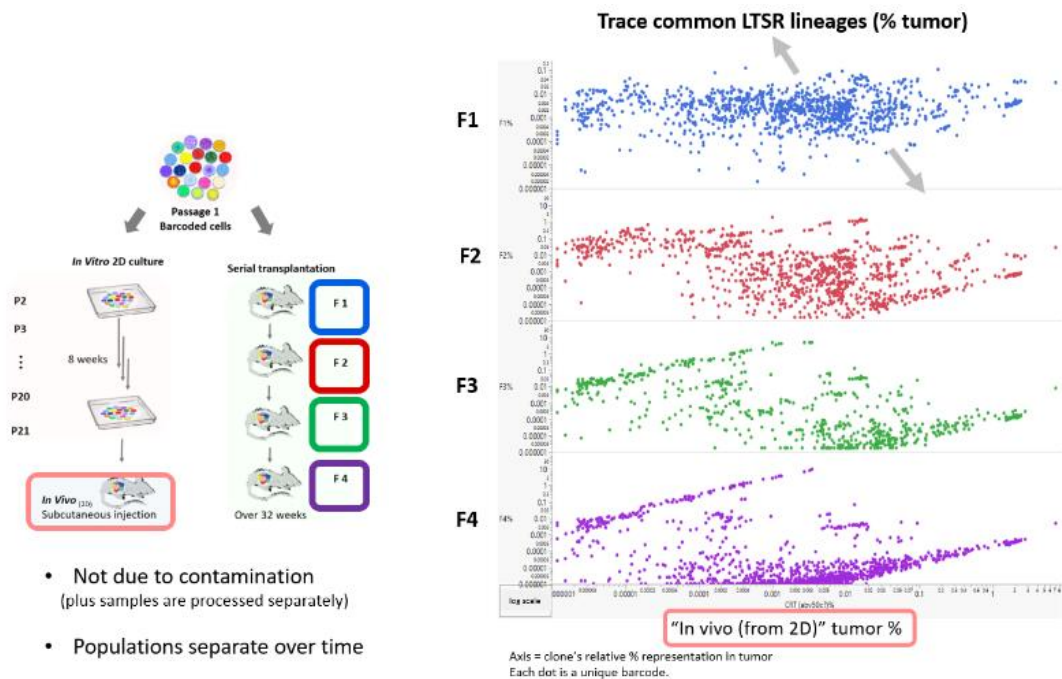
Figure 164. Longitudinal tracking of clonal abundance gradual divergence in *in vivo* serial transplanted tumor. Fixing the X axis as the "2D stabilized" tumor and tracing individual LTSR clone's relative abundance throughout the *in vivo* serial transplantation F1 to F4. The data shows a gradual shift over 8 months period of time in separation of the two inversely correlated clonal populations.

Now that we are confident that the observation is true, we can ask interesting questions regarding the cause of the inverse in population abundance for lineages "stabilized" under different environments and how the clonal ranking relationship is maintained intrinsically independent of clonal abundance.

Environmental influence on clonal equilibrium states in tumors

To determine the environmental influence on clonal equilibrium states in tumors, the experiment started with the same pool of barcoded cells - half of the cells were passaged *in vivo* over time ("F1", "F2", "F3", "F4"), while the other half were passaged *in vitro* over time and subsequently injected into mouse to form tumors ("*in vivo* 2D"). By looking at the dynamics (in

terms of abundance) of barcodes common in the tumors, we can cluster them in two main clusters using PCA analysis (Figure 165). For those two clusters of clones, we can see they are indeed separating over time as the serial transplantation passage increases (Figure 165, Figure 166). This indicates that the environmental factors, "*in vivo*" vs. "*in vitro*", are causing specific groups of clonal lineages to gain or lose relative fitness in the tumor. This is quite interesting because by functional definition, all these clonal lineages are considered to be long-term self-renewal clones that have the ability to maintain the tumor over a long period of time. However, due to environmental factors and the conditions they are cultured in (specifically "*in vitro*" vs. "*in vivo*"), the degree of abundance in tumor is different under different environments. Similar to chemo-resistant residual lineages, although low in abundance, are still important functionally and have the ability to give rise to relapse tumors and adapt to external perturbations.
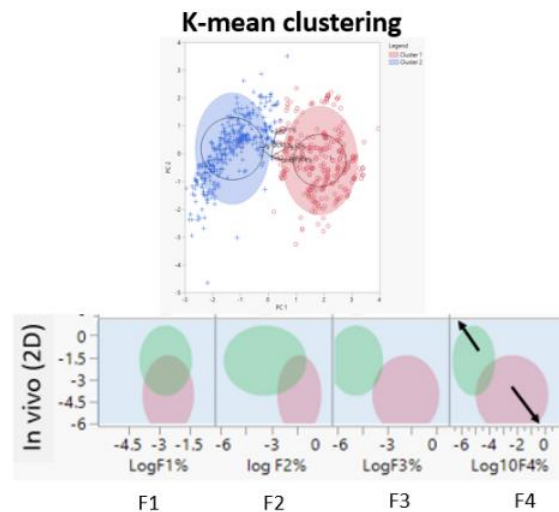
Figure 165. K-mean clustering of the LTSR lineages overtime. K-mean clustering shows that the two populations of LTSR lineages gradually separate over time during the 8 months of *in vivo* serial transplantation when plotted against the "2D stabilized" tumor.

Figure 166. Longitudinal tracking of clonal population divergence in *in vivo* serial transplanted tumor vs. In vitro stabilized tumor. Fixing the X axis as the "2D stabilized" tumor and tracing individual LTSR clone's relative abundance throughout the *in vivo* serial transplantation F1 to F4. The data highlights a subpopulation the lineages that are highly abundant in F4 tumor. Tracing back in time (in F1), we see the separation of the populations are gradual, where initially the particular cluster is a part of the larger clusters in F1 tumor, but separating out over time. The stabilization of the equilibrium is a gradual process and takes time.

Figure 167. Parallel plot of LTSR lineage dynamics overtime in *in vivo* serial transplantation vs. "2D stabilized" tumor.Y-axis representing the relative abundance in tumor representation and X –

axis representing the tumors "F1", "F2", "F3", "F4" and *in vivo* 2D" tumor. Each line represents a unique LTSR lineage. Following the two populations of lineages in terms of % abundance in each tumor stabilized under different conditions. Green indicating lineages with low representation in F4 tumor, while red represents lineage with high representation in F4 tumor and vice versa. We can see that although these lineages are all present and is part of the LTSR compartment, the abundance is inversely correlated.

Furthermore, by looking at individual lineage dynamics over time, we can appreciate that the tumors are naturally coming to an equilibrium state (Figure 167). From tracing the clonal abundance in these two populations, we can see from the parallel plot that initially from "F1" to "F2" (Figure 167), there are roughly half the lineages increasing in fitness (red lineages), while the others have a loss in fitness (green lineages); although this trend of fitness continues on, it starts to correct itself as observed from "F2" to "F3", where most of the lineages that have better fitness (red) continue to dominate the tumor representation, but a portion of those lineages starts to lose competitiveness. From "F3" to "F4", we can see that the tumor is "stabilizing", where the clonal change in fitness is less drastic from before and that the lineages that are being outcompeted (green lineages) start to maintain steadily at a low level of the tumor representation. On the other hand, the lineages that are dominating the tumor (red lineages) are also stabilizing, where majority of the lineages are not changing much in tumor relative abundanceFigure 167.

This experimental observation suggests that tumor acts like an "organism" where internal clonal dynamics adapt to environmental conditions, and through internal clonal competition, the overall clonal fitness may eventually reach a steady state. In addition, the data suggests that functional clonal lineage is a significant factor in determining and shaping the tumor under differential external pressure. Since we are passaging cells in a way that keeps as much of the heterogeneity as possible between passages (0.5 fraction of previous passage) and allowing the tumor to expand under a constant environment for a long time with minimum perturbation of the system, we could gain insights to the latency a tumor requires to establish an equilibrium state.

For example, in this case, the tumor cell line requires at least 20 doubling events to establish inter clonal equilibrium.

Bi-modal state of proliferation

Another interesting observation from this data set is that we do not see a population of lineages that is growing well in both conditions. A subset of lineages either grow better in the *in vivo* condition or grow better in the *in vitro* condition. This suggests there might be a factor that is regulating or determining the proliferation signals. This is purely a speculation at this state, but interesting to think about nonetheless.

Symmetric vs. asymmetric division could be a mechanism for maintaining clonal lineage rankings under differential environmental pressure

Based on the observation, there are differential preference of clonal lineages under "*in vitro*" vs. "*in vivo*" environment. If differential preference of environments is true and is reflected through relative lineage abundance, it is logical to think that there should be a population of lineages that is not growing well under both experimental conditions (since both *in vitro* and *in vivo* experimental setting are not the same as the original human tumor environment) – and this is exactly what we see. In the figure below, the highlighted population that is maintained at 0.01% of the tumor mass in both "F4" and "*in vivo* 2D stabilizes" tumor is not highly abundant, nor lowly represented in tumor (Figure 168). These lineages that are not thriving in either the "*in vivo*" and "*in vitro*" experimental conditions must actively go through cell division in order to maintained tumor representation around 0.01% of the tumor mass. If these lineages are not actively dividing, their tumor representation would drop and eventually be outcompeted. Therefore, if these "0.01% lineages" are actively proliferating, the mechanism for these clonal lineages to maintain their

relative tumor representation is likely related to cell proliferation rate. And hence, it is possible that the mechanism of maintaining clonal relative ranking could be controlled by lineage intrinsic division rate under certain environments.
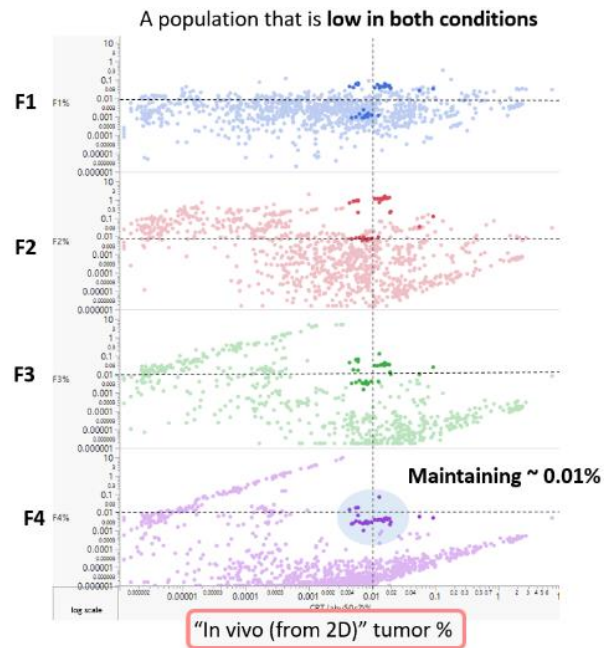
Figure 168. Longitudinal tracking of clonal abundance divergence in *in vivo* serial transplanted tumor. Fixing the X axis as the "2D stabilized" tumor, and tracing individual LTSR clone's relative abundance throughout the *in vivo* serial transplantation F1 to F4. The data highlights a subpopulation of LTSR lineages maintain at 0.01% relative tumor representation over time.

We have previously demonstrated that there are indeed hierarchy amongst tumor cells, where only a fraction of the barcoded cells could give rise to LTSR lineages. Another data supporting tumor hierarchy came from isolating single cells via FACS sorter and culturing the single cells to generate isogenic clones (see "single cell colony formation" section). In that experiment, we observed only certain single cells can regenerate the full culture, while other single cells either fail to proliferate or can only sustain growth for a short period of time and eventually exhaust. In addition, from this single cell proliferation potential experiments, we learnt

226

that these "stem like" cells that can regenerate the entire culture is independent of cell cycle state

and that different patient tumors consists of different percentages of these "stem like" cells (i.e.

10% vs. 5% from two PDAC tumors).
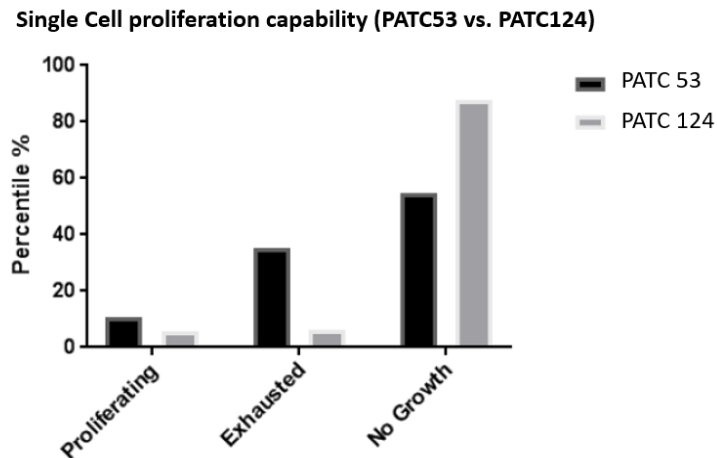
Figure 169. Bar graph comparing the percentages of cells with ability to re-establish culture between two patient tumor cell lines. For the two patient tumor cell lines (PATC53 and PATC124), the bar graph shows the % of cells after single cell FACs that could establish a new isogenic line, as well as % of cells that eventually exhaust or never divide. Albeit both cell lines follow a similar trend where most cells will exhaust, the percentage of cells with capability to re-establish colonies are different (PATC53 ~10%, PATC124~5%).

To answer the question whether this tumor cell hierarchy also exist within an isogenic

clonal lineage, we performed another round of single cell sorting on each isogenic cultures and

measured their single cell proliferation potential. We found that cell potentiation hierarchy also

exist within a single clonal lineage, where only a portion of single cells would have the ability to

regenerate the culture. Interestingly, similar to the finding that each tumor consists of different

fractions of "cancer stem like cells", we observed the percentages of "cancer stem like cells" also

vary from lineage to lineage. Here, the "stem like" cells are discussed in the context of their

functional ability to establish culture from a single cell. These data demonstrate that there are

indeed "stem-like" cells in the tumor, and their % varies from lineage to lineage.
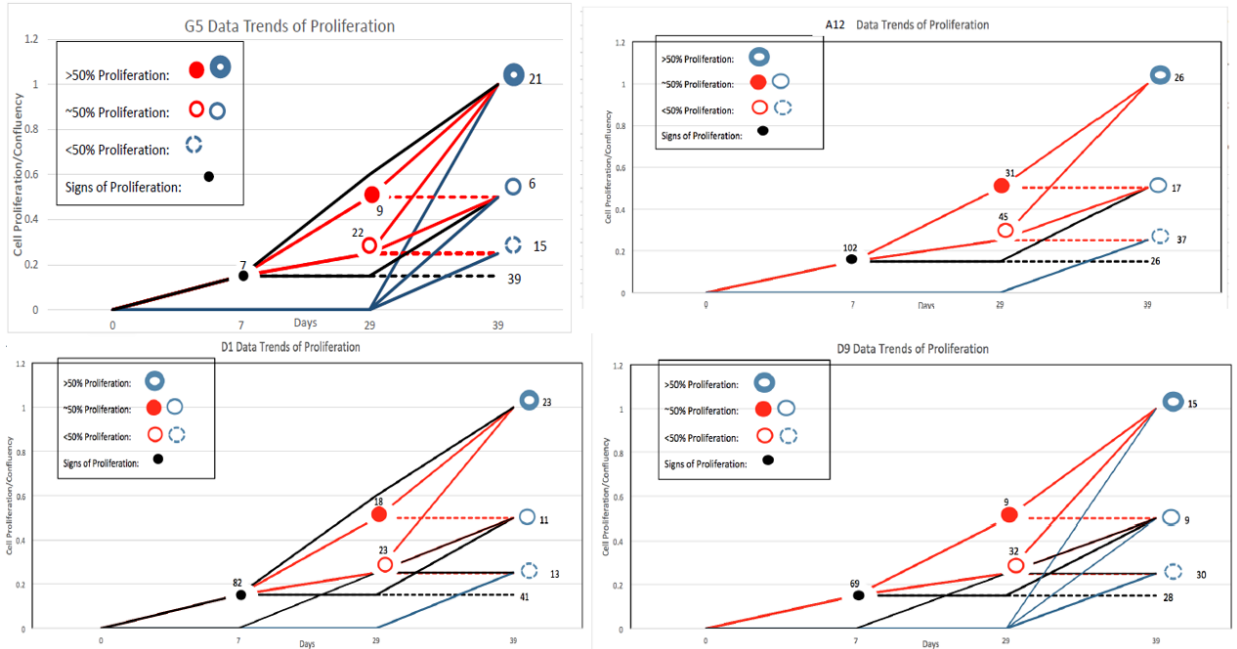
Figure 170. Clonal dependent intra-clonal cell proliferation and growth dynamics. Each panel represents a specific clone single cell sorted in 96-well tissue culture plate and observing individual cell's proliferation capacity. Each line represents a category of abundance (culture confluence) observed. The numbers indicating the count of cells displaying such abundance (culture confluence). X-axis indicates time of culturing. Y-axis indicates confluence of cell culture wells.

Clones*	Number proliferated**	Percent Proliferation	Number Exhausted***	Percent Exhausted	Number of no growth****	Percent of no growth
CL 09	21	10.7%	60	30.6%	115	58.7%
CL 11	23	11.7%	65	33.2%	110	56.1%
CL 01	26	13.3%	80	40.8%	90	45.9%
CL 12	15	7.7%	67	34.2%	110	56.1%
Parental	189	11.6%	184	11.3%	1266	77.6%

* 196 wells of single cells were sorted for each clone; 1632 wells were single sorted for parental culture (mixed clones); Cells were observed over 1 mo period for proliferation.

** Wells (cells) exhibit signs of proliferation and are able to establish confluence.

*** Wells (cells) exhibit signs of proliferation initially, however, stopped growing.

**** Wells (cells) stayed 1 cell throughout.

Figure 171. Intra-clonal cells' capability to re-establish a culture is heterogeneous. Isogenic clone's cell % that could re-establish an entire culture is represented in bar-graph format (top) and table with detailed % data (bottom). This suggests the "self-renewal" (ability to re-establish culture) hierarchy exists and the ratio of "stem" cells are actively re-equilibrated from clones established from a single cell.

It has been observed that a percentage of cells in a given tumor have the capability to remain quiescent for long periods of time. These dormant cells in tumors are potentially a cause for relapse after treatment with drugs that target cell division. These "stem-like" cells in cancer are more quiescent and proliferate at a less frequency and perform asymmetric division to give rise of its progenies. Asymmetric division is where the cell would only generate one daughter cell instead of generating two daughter cells. Furthermore, under certain conditions, the cells would alter between the states of performing symmetric and asymmetric division [199].

To verify that tumor cells in the experiment (PATC124) does indeed have cells dividing asymmetrically, we performed a cell division tracking experiment. PATC124 cells were dyed with a cell membrane dye PKH26 (red) prior to 3D culturing. Each subsequent cell division would dilute the membrane dye; therefore, allowing us to track differential cell division frequencies amongst cells through cell to cell dye intensity difference. By plating single cells onto the 3D culture plate over time, we observed that about 10% of the cells give rise to spheroids while the majority of the single cells fail to proliferate. Amongst those spheroids, most of them have no red dye remaining, meaning that all the cells are actively dividing. However, a fraction of the spheroids retained a strong PKH26 red dye signal (Figure 172). The data demonstrated that within the experimental cell line, there are differential degree of symmetric vs. asymmetric division. Thus, clonal lineage abundance could be potentially regulated through asymmetric vs. symmetric cell division under an environmental dependent manner, offering a potential explanation on how relative clonal rankings are maintained in tumor.
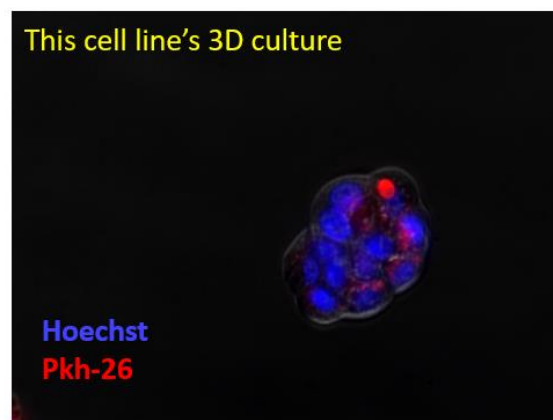
Figure 172. Asymmetric division measured by PKH-26. PATC124 cultured in 3D with PKH dye (red) tracking differential cell division rate over time. Intensity of red PKH-26 dye correlates inversely with cell division rate. Red=PKH-26, blue = Hoechst.

Modeling of intrinsic and extrinsic factors leading to clustering of clonal population in tumor

To model the impact of asymmetric vs. symmetric division on clonal representation over time, we performed a simple simulation. Assume that there are four clonal lineages, clone "A", "B", "C" and "D", where clone "A" and "B" undergo symmetric division (giving rise to two progenitors) while the clone "C" and "D" undergo asymmetric division (giving rise to one progenitor cell). As these lineages divide over time, we can see from Figure 173 below that clones "A" and "B" grow exponentially, while clones "C" and "D" hover at a lower representation. In addition, irrespective of whether the clones undergo symmetric or asymmetric division, they could maintain their relative clonal tumor representation ranking as shown in the figure below, where clone "A" > "B" and "C" > "D" is maintained throughout cell divisions.
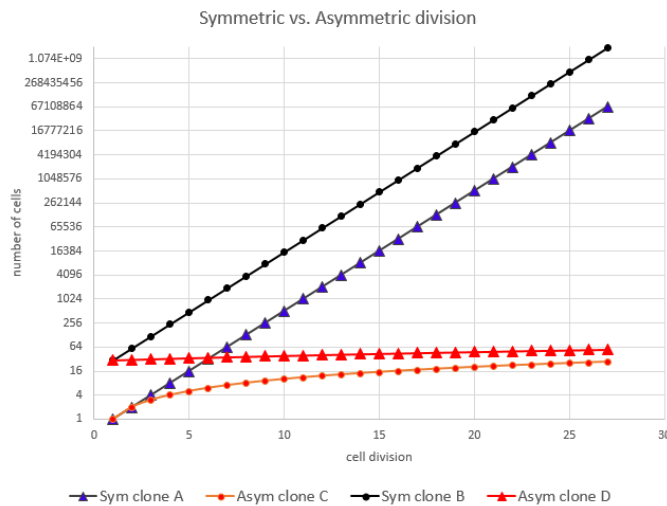
Figure 173. Simulation of linear vs. exponential amplification cell abundance. X-axis represents number of cell division and Y-axis represents the number of cells accumulated. Simulation demonstrates regardless of cells that go through symmetric or asymmetric division, relative clonal ranking can still be maintained (or clone-to-clone ratio).

Obviously, the actual cell proliferation scenario is much more complex than the assumptions made in the above simulation. The above simulation only assumed the extreme cases where differentiated cells will not divide and only progenitor cells will. In cancer, the line between differentiated and undifferentiated cells and their ability to proliferate long term is more ambiguous, which will be discussed later in the discussion section. For modeling, we would lump all the above factors and call them "cell intrinsic factors".

With the help of Nick Yen, to model lineage intrinsic and environmental impact on clonal lineage populations, we assume that the cell growth is based on three factors: 1. Intrinsic factor "I" – a normal distribution of proliferation potential (discussed above); 2. Extrinsic factor "E" – where environment stimulation causes increase or decrease in fitness of certain clonal lineages (clones either favor environment A or environment B in the simulation); 3. Random factor "R" – factors that can contribute to difference that are not considered. If the number of cells is represented by "N" and cell division represented by "n", the following equation could be used to describe the cell abundance vs. the above three factors:

$$N_{n+1} = N_n \times 2^{I \times E} + R$$

If we simulate the cell division guided by the above equation over 50 divisions (P1 to P50), where if a clonal lineage favors an environment, the value of "E" would be larger and vice versa. From Figure 174 below, we can see that the lineage population separates over the span of P2 to P50 gradually over time. The simulation model we observed here resembles the empirical observation from the barcode lineage tracing experiment, where the environmental factor is either "*in vivo*" or "*in vitro*". This simulation explains how lineage intrinsic bias towards a specific environment could cause the separation of lineages abundance into distinct populations.

- Assuming the cell growth between each passage is determined by 3 factors:
  1. Intrinsic- proliferation ability (normal distributed)
  2. Extrinsic- environmental factor (clones favor either environment A or environment B)
  3. Randomness- perturbation in each passage process or background signal of sequencing reads
- The simulation over 50 passages shows that environmental factor can cause a distinct separation of 2 populations
- $N_{n+1}=N_n*2^{IE}+R$
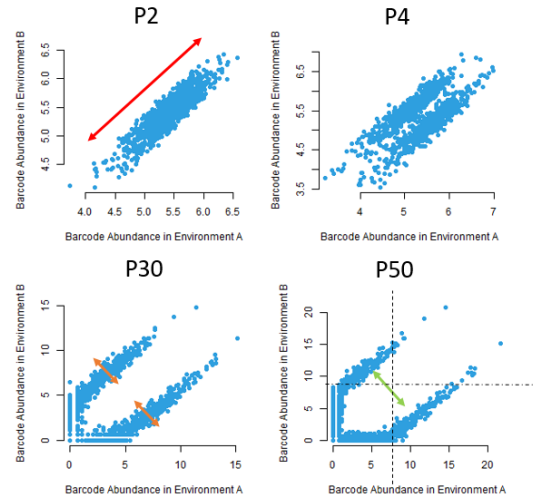
Figure 174. Modeling of lineage abundance under environmental dependent proliferation rate difference.

Clonal abundance in tumor under different environment is a dynamic process

From the simulation above, we can see that if the clones have environmental preferences, those extrinsic factors could influence intrinsic clonal proliferation rate, and finally, is reflected through their relative tumor representation over time.

To observe if this clonal shift in abundance is indeed a dynamic process, we performed another lineage tracing experiment. Using the same barcode strategy on the same tumor cell line (PATC124), we first stabilized the barcoded cells in 2D culture "*In vitro*" condition for 30 passages. Afterwards, we then injected those cells *in vivo* into NSG mouse and performed *in vivo* serial transplantation for three passages (F1, F2 and F3) that span across 24 weeks. After the tumors were collected, the barcode information was extracted via PCR and NGS by methods described in the previous chapters (Figure 175).

By comparing the common barcodes from "F3" and *in vitro* passage "P30", as well as a day 31 "*in vivo* tumor" created from *in vitro* passage 30, we can see that the clonal lineages start to separate out into distinct populations over time. This demonstrated that although the clones

are "stabilized" under one environment (*In vitro* from P1 to P30), once the environment shifts (from *in vitro* to *in vivo*), over time, these lineages would establish a new equilibrium. Moreover, the relative clonal ranking is also closely preserved.
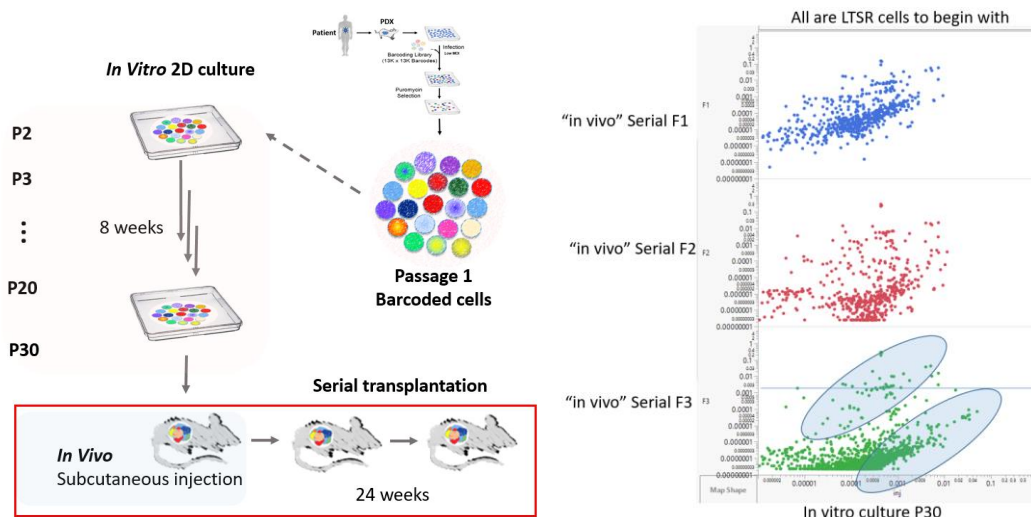
Figure 175. *In vitro* stabilized lineages shift in clonal equilibrium through serial *in vivo* transplantation. Switching external condition of the "*in vitro* stabilized" cell culture that has achieved clonal equilibrium by *in vivo* serial transplantation (left); observation of gradual clonal fitness separation post environmental switch from *in vitro* to *in vivo over* 24 weeks.

Chapter Summary and Discussion

To effectively trace clonal evolution longitudinally in an unbiased manner and study the effects of external culturing conditions, we took tumor cell lines derived from patient PDXs and barcoded each cell with a unique barcode. Then, we used the same set of barcoded tumor cells (shortly after barcoding) and subjected them into two different environments – *in vitro* passaging and *in vivo* serial transplantation. Throughout the passages under both conditions, the complexity of lineages dropped from ~2M lineages down to a few thousands over time. After a long period of time in culture (32 weeks *in vivo*, 8 weeks *in vitro*), which allows the cell population to naturally reach an equilibrium state, we found a common set of barcodes (lineages) maintained in both

environments. By functional definition, this set of common barcodes (~1,200 lineages) has long-term self-renewal properties, which contribute to the maintainence of the *in vivo* tumor as well as *in vitro* cell culture over a long period of time. Moreover, in terms of tumor or total cell culture abundance representation, this set of the common lineages makes up most of the population in both experimental conditions (70% - 97%). The data strongly suggest that there is a common pool of long-term self-renewal (LTSR) lineages that maintain the tumor's growth overtime, regardless of the environmental difference defined by the context of the experimental settings.

Interestingly, these common LTSR lineages in the *in vivo* environment have an inverse relationship in terms of abundance representation to those in the *in vitro* environments at the equilibrium state. For example, the part of the LTSR lineages that had lower representation in *in vivo* stabilized tumors had higher representation in the *in vitro* stabilized tumor (*in vitro* stabilized and transplanted in NSG mouse) and vice versa. This observation is not entirely surprising, as the influence of environmental pressure on the selection of sub-clonal lineages has also been observed under clinical setting under treatment [29]. However, further analysis of the correlation of the stabilized *in vitro* vs. *in vivo* abundance of these common LTSR lineages under equilibrium state showed that the clones were highly correlated in terms of relative clonal ranking (>0.8), despite whether or not they were the dominant or underrepresented populations (Figure 163). To summarize the observation in more simplistic terms:

1. There is a common set of LTSR lineages maintaining both the *in vitro* culture and the long-term *in vivo* tumor.

2. This common set of LTSR lineages can be separated into two main populations: one population dominants the sample (tumor/culture) and the other remains at low abundance.

3. At the equilibrium state, the dominant clonal population in one environment is the low-represented clonal population in the other environment, thus displaying an inversely correlated relationship.

4. When comparing an individual clone to the clone's counterpart within the sample population (i.e. lineages that are highly abundant in one environment and lowly represented in the other), we find that the correlation remains high amongst clone to clone abundances (or clone-to-clone relative ratio).

The above observations may suggest the possibility that, regardless of the overall fitness, there may be an intrinsic mechanism that maintains the relative abundance between individual clonal lineages (or clone to clone ratio). In other words, regardless of whether the environment is favored by a specific set of clonal lineages within a tumor, there may exist a mechanism that preserves the relative clone-to-clone ratio.

To understand whether or not this observation of "clone-to-clone ratio preservation" was gradually achieved over time, we followed these common lineages endowed with long-term self-renewal properties longitudinally (from day 1 of the experiment). To this end, we compared individual lineage abundance from the *in vivo* tumors to the *in vitro* stabilized tumors (*in vitro* passage 26, injected in NSG mouse) at different time points (month 2, 4, 6, and 8). We found that over time, due to environmental differences, these lineages started to separate into two major populations based on relative tumor fitness (one favored the *in vitro* environment, while the other favored the *in vivo* environment). These two populations, in nature, were either composed of lineages that gained fitness and gradually dominated the tumor, or lineages that lost fitness and decreased in tumor representation in either environment. Through a gradual shift, these two populations' dynamics started to become stable and eventually reached equilibrium (Figure 164). This observation of environmental differences' influencing the gradual separation of two main LTSR populations over time provided confidence in what we observed ("abundance independent

preservation of clone-to-clone relative ratio"), and was indeed reflective of biological phenomenon.

If the "abundance independent preservation of clone-to-clone relative ratio" is indeed a natural phenomenon and that the cumulative properties of individual lineages as a sum define the tumor's functional characteristics, then it is possible that tumor characteristics can be actively maintained through the preservation of each sub-clone's relative contribution (in other words, their relative ratio) even at a low level. As a metaphor, every human is born with a unique set of fingerprints, and that unique set of fingerprints remains throughout his or her life. Therefore, we could view this "abundance independent preservation of clone-to-clone relative ratio" as a potential "fingerprint" of a tumor – or, "tumor clonal fingerprint (TCFP)". The discovery of TCFP revealed by this set of data may potentially be used to profile and define a tumor's functional characteristics. A simplified illustration of TCFP is in the figure below.
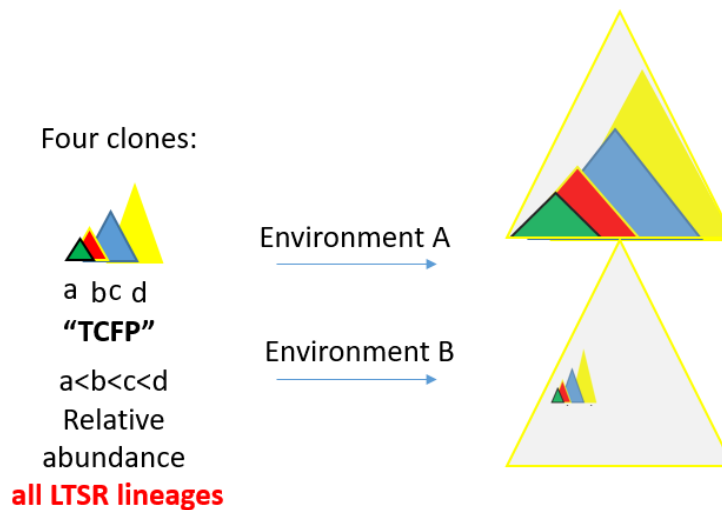
Figure 176. Model of "Tumor Clonal Fingerprint". Relative clonal rankings retention under different environmental-dependent fitness changes. Assumes four different clones in a tumor "a", "b", "c", and "d" in an abundance order of a< b < c < d. Under environment A, all clones have better fitness, and all proliferate while maintaining relative clonal ranks. In comparison, in environment B, all clones do not thrive and proliferate at lower rates but the relative clonal ranks are still retained.

To explore some of the possible mechanisms maintaining TCFP, we first investigated the intrinsic proliferation properties of a single lineage. Using a single cell proliferation assay, we determined that there were indeed a "stem-like" hierarchy in terms of the ability to re-establish the entire culture from a single cell (Figure 170, 171). The fraction of the cells that could re-establish an entire culture was different amongst lineages (ranging from 7.7% to 3.3%), and differed from tumor to tumor. For example, patient PDX derived cell line PATC124 had "stem-like" cells of ~5.6% while PATC53 had >10% (Figure 169). To further solidify a potential explanation as to how LTSR lineages were persistently maintained even at low abundance while maintaining TCFP, we surveyed the cell for evidence of quiescence and populations of cells with different division rates. Using 3D-cell culture combined with a cell division tracking dye, we identified that there were indeed cells that were more quiescent than the others and divided at a lower frequency (Figure 172). All these observations inferred the possibility that lineages in the long-term self-renewal compartment possessed differential division rates and used potential asymmetric vs. symmetric division as a mechanism (data using functional marker numb is not shown). Such differential division rates combined with a mechanism of asymmetric vs. symmetric division may be a part of the intrinsic factors that contribute to the regulation and maintenance of TCFP in tumors [189-198].
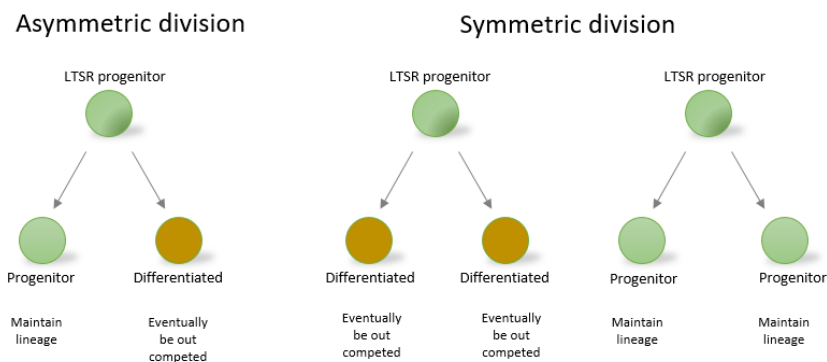
Figure 177. Model of asymmetric and symmetric divisions.
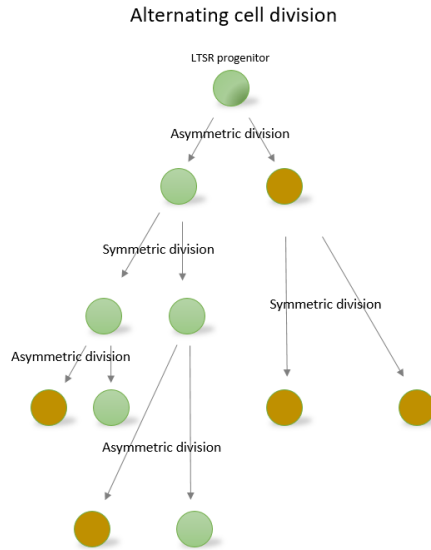
Figure 178. Model of alternating asymmetric and symmetric divisions within a lineage.

Obviously, the intrinsic factors and the key mechanisms that govern the maintenance of the "tumor clonal fingerprint" requires a lot more experimental proof. Asymmetric and symmetric division contributing to the TCFP hypothesis represent only one of the many possible mechanisms contributing to this biological phenomenon. Although we do not currently know the exact mechanisms, the possible concept of TCFP is nonetheless fascinating. The idea of extrinsic factors' influencing the ability of tumor cells to proliferate is not a new concept, as it has been described under the context of microenvironment niches for metastasis, mechanical stress, as well as DNA damage that alters symmetric vs. asymmetric division rates [199-202]. For instance, different culturing media conditions can result in different number of colonies forming, and/or we different tumor initiation rates may arise under different genetic mouse backgrounds [203-205]. It is clear that the environment influences clonal fitness (division rate), as shown here by the shifts of the two populations of lineages in abundance over time. Working with Nick Yen and combining this with the potential symmetric vs. asymmetric division rates (exponential vs. linear), we came up with a simple equation to describe how the environmental influences the cell division rate –

$$N_{n+1} = N_n \times 2^{I \times E} + R$$

Where "N" represents number of cells, "I" represents cell intrinsic division rate decision, "E" represents external factors influencing proliferation rate, and "R" represents random factors cannot be taken into account.

This concept of a "tumor clonal fingerprint", where long-term self-renewal lineages reach different equilibriums under different environments while retaining relative clonal ranking, can help us how define cancer and understand the disease more accurately. Since the experimental observation here is obtained through an unbiased measurement of clonal lineages via barcodes, we can appreciate that tumors can be structurally defined by clonal lineages and their relative abundances. Under certain conditions, there are lineages that will expand exponentially and take over most of the tumor mass. While there are lineages that do not favor the current environment and decrease in relative tumor abundance, they are still actively dividing at a much slower rate while maintaining a very similar relative lineage ranking (i.e. expand in a linear fashion). In other words, although these lineages are low in abundance, they still retain the tumor characteristics defined by the composition and relative ratio of clonal lineages - the tumor clonal fingerprint. Since the tumor clonal fingerprint is preserved when shifting underrepresented lineages to a favorable environment, these lineages are expected to increase in proliferation potential and increase tumor representation. Furthermore, as demonstrated by re-transplanting *in vitro* stabilized cell populations *in vivo* and continuous serial transplantation for 6 months (Figure 175), the level of clonal fitness vs. environmental dependent equilibrium state is a dynamic process.
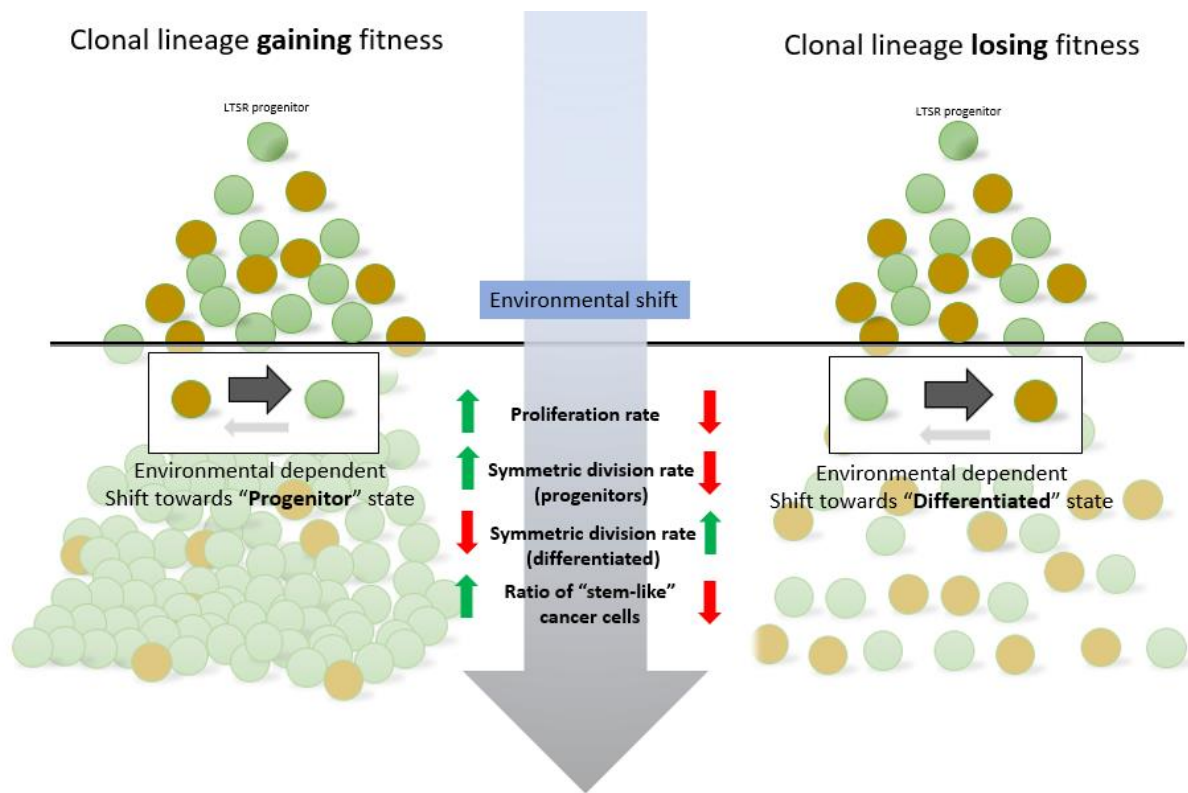
Figure 179. Model of clonal lineage fitness dynamics under environmental shift.

From a clinical point of view, the concept of TCFP hints that the complexity and the character of a tumor may be preserved at relatively lower abundances. If we could actively monitor tumors in the context of a clone-to-clone ratio, this may aide us in surveying treatment effectiveness. For example, TCFP may help in monitoring the treatment effectiveness of chemotherapy. While chemotherapy can shrink tumor mass, it is poor at perturbing intratumoral clonal complexity, and therefore, may have minimum impact of eliminating resistant clonal lineages that are responsible for relapse [109]. Finding suitable treatment regimens that can perturb the fitness of the "balanced ratio of collective lineages" in a tumor may be crucial in systematically breaking down the disease and preventing relapse. I believe this new concept of a "tumor fingerprint" and the empirical method to identify TCFP through CRT can aid in future research in targeted therapy and personalized treatment strategies.

Closing Summary and Future Directions

The tumor is a dynamic ecosystem comprised of multiple functionally diverse sub-lineages that co-exist and evolve. Currently, clonal progression models based on genomic analysis are challenged by the practical shortcomings of sample representation and detection sensitivity. Furthermore, besides end point –omics profiling, experimental models to functionally test observed phenotypes as well as to generate and validate new hypothesis are also limited. To bridge this gap and develop the first longitudinal model of cancer evolution, we have established a unique clonal tracing platform by exploiting the hierarchical nature of human tumors – the Clonal Replica Tumor (CRT) platform. The CRT platform can generate a large experimental cohort of animals bearing clonally identical tumors, which provides a high degree of biological reproducibility that enables the precise temporal tracking of functionally diverse sub-lineages in parallel in different animals. Most importantly, along with a high throughput clonal isolation technique, this approach allows us to isolate clones of interest based on their specific behavior from the *in vivo* experiments for further functional characterization and validation.

In ***chapter one***, we demonstrated the robustness and the tumor maintaining properties of CRT lineages. Using CRTs to study chemo-resistance in the context of differential clonal responses, we uncovered a wide array of pre-existing functional diverse clones that displayed a variety of sensitivity to multiple drugs. These data allowed us to appreciate the wide spectrum of harbored tumor chemo-resistance and indicated that experiments designed to assess treatment efficacy through measuring tumor size alone is inadequate. By characterizing isolated treatment-naïve clones with differential sensitivities to chemotherapy, we identified a novel transcriptomic signature that could detect pre-existing chemo-resistant populations of cells in treatment-naïve tumors by delineating relevant molecular pathways, such as DNA damage repair and OXPHOS.

This signature represents a new biomarker of response as it enables the prognostic stratification of patients with regards to their disease-free status and overall survival.

In ***chapter two***, we elevated the CRT concept and added a temporal element to create a novel longitudinal clonal tracking model. Along with robust statistical simulations and under an orthotopic transplantation setting, we demonstrated the capability to longitudinally track lineage dynamics during tumor progression and dissemination. This longitudinal clonal tracking model had the unique aspect of overcoming the practical limitation of serial sampling, and thus allowed us, for the first time, to obtain a comprehensive picture of natural tumor expansion dynamics with good representation and unprecedented sensitivity. Through empirically obtained data, we observed a plethora of unique clonal dynamics that were previously undescribed. In particular, we found an unexpectedly high degree of clonal abundance fluctuations during unperturbed tumor growth, including the phenomenon of alternating clonal dominance (ACD) that occurred naturally during tumor expansion. In addition, due to the high precision and sensitivity of this approach, we were able to probe into the dynamics of lowly abundant lineages, which have been usually neglected due to technical detection limitations. We unexpectedly uncovered a rich underworld of underrepresented clones that were constantly and persistently competing throughout the entire lifespan of the tumor for the chance to rise. These data captured and demonstrated the full extent of clone-to-clone competition, interaction, and the ever-evolving nature of tumors. At the same time, we also identified several interesting clonal dissemination patterns, including monoclonal dominance and the oligo clonal nature of metastatic lesions, as well as quantitative profiling of the entire longitudinal dissemination process. Furthermore, by tracking clonal colonization and expansion patterns in secondary organs, such as liver and lung, we found that metastatic growth, represented mostly by a long-term dormancy followed by explosive expansion, was more simplistic compared to the dynamics in the primary tumor. Moreover, by demonstrating that only cells endowed with long term self-renewal can spread and

sustain growth in secondary sites, we definitively demonstrated the intrinsic relationships

between stemness, clonal fitness, and metastasization.

In **chapter three**, we explored the effects of different environments in shaping tumor

clonality. Through serial passaging of lineages under different environment (*in vitro* and *in vivo*),

we gained a deeper understanding of how environmental pressures could alter the abundance

of tumor-maintaining lineages. Through this approach, we uncovered that tumors were always

maintained by a defined pool of cells with long-term self-renewal irrespective of the

environments that they are exposed to. Although the relative contribution of a single stem cell to

the tumor mass could change in different growth conditions, we demonstrated that distinct

environments exerted major effects on clonal relative abundance rather than on the overall

clonal complexity of tumors. In other words, lineages that were lowly represented in one

environment could gain fitness and expand when exposed to a new environment and vice

versa. Interestingly, despite low representation, lineages that lose fitness with the change of the

environment still constantly divide and persist over time as well as maintain an inter-clonal

ranking very similar to that they had when they were dominant. This observation once again

highlights the functional significance of underrepresented clones and their potential to contribute

to disease progression or relapse at any time upon extrinsic perturbations, such as therapeutics

or changes in the tumor microenvironment. This previously undescribed active preservation the

clone-to-clone relative ratio in a tumor – or "**tumor fingerprint**" as we have named it — can aid

in profiling tumor characteristics and be used to explore better treatment strategies in the

context of functional sub-clones.

Combining what we have learned from all three chapters, we are able to have a grasp

just how "plastic" and "dynamic" tumors really are, and can perhaps paint a picture of the

challenges ahead for therapeutic stratagems. At the same time, these data also present an

opportunity to realize the importance of conducting cancer research in the context of clonal

dynamics and interactions as well as with the understanding that tumors are complex ecosystems. Indeed, we plan to leverage the longitudinal CRT model's strengths and apply it to new studies on different tumor models. One of the many ongoing studies involves longitudinal clonal tracing under the context of combination therapy. Instead of treating the tumor with one drug at a time, we aim to monitor the clonal response and their dynamics during and after treatment as well as upon relapse under the context of multiple pharmacological perturbations. What we observe could directly contribute to the understanding of differential clonal interactions and competition, expansion dynamics during combination treatments, as well as specific clonal vulnerability or resistance to combinatory drugs. Furthermore, using our high throughput clonal isolation platform, we can conduct isolation, deep molecular characterization, and subsequent reconstitution of tumors with isogenic clones to perform further functional drug and genetic screening to verify and explore lineage specific vulnerabilities. This will enable the design of "smart" combinations to overcome or delay resistance.

Furthermore, the CRT platform is not restricted to the study of lineage dynamic tracing. Due to the robustness and biological reproducibility aspect of the model, we can apply it to the study of clonal interactions in a spatial context with high precision. As briefly demonstrated in chapter two, CRT models combined with histology and clonal geographical distribution analysis could allow us to study clonal-to-clonal spatial interactions and as well as clonal-stromal functional interactions. Questions such as clonal-specific associated stromal niches or the contribution of immune infiltration to treatment resistance can be answered by designing the proper experiments.

In addition, building upon our observations from the longitudinal metastasis study, the seemly synchronized events of the sudden changes in clonality (alternating clonal dominance) in the primary tumor and the explosive growth of clones in the metastatic sites are worth further perusing. The idea that metastatic tumor outgrowth can be induced by naturally occurring intra-

tumoral changes at the primary site during expansion is quite fascinating. To identify the factors changing in the primary tumor in concordance with expansion, we can serially sample entire tumors from parallel groups of animals in the same CRT cohort over synchronized time points. One group will be used to identify longitudinal clonal dynamics, while the other group will be processed for molecular profiling for both primary tumor and metastasis lesions (i.e. transcriptomic sequencing). Matching the transcriptomic changes with clonal dynamics may provide further hypotheses as to what is happening in the primary tumor at each time point of the longitudinal study. Furthermore, through histological analysis and immunostaining for functional molecular markers, we can validate and survey microenvironment changes such as perfusion, hypoxia, and nutrient deprivation during longitudinal tumor expansion at different time points. Moreover, performing multi-regional sectioning of tumors, coupled with barcode analysis as well as further cross-referencing with functional and histological data, can inform on clonal specific responses that relate to environmental changes in a high-resolution spatial context. The clinical implementation will be to identify the factors in the primary tumor that influence metastatic outgrowth in order to develop actionable prevention strategies.

In this thesis, I primarily focused our discussion on the technical aspects of establishing and demonstrating the capability of a novel longitudinal CRT platform. This platform can be a useful tool for studying cancer evolution in the context of tumor functional heterogeneity, as well as provide the unique opportunity to further validate and characterize isolated clonal lineages. Now that the CRT platform has been established, we can expand and apply it to study different cancers. For example, in a preliminary experiment, we barcoded a melanoma derived primary tumor cell line and found the clonal behavior to be vastly different than PDAC. This particular melanoma cell line was very aggressive and had a relatively higher fraction of long-term self-renewal cells, as well as displayed different inter-clonal dynamics during the stabilization and

serial passaging processes. Observations like this were not expected and could help us to form

specific testable hypotheses under the context of different cancer types.

All in all, we are just barely beginning to unveil the extent of the intricate tumor inter-

clonal behaviors and interactions. Now that we have a robust CRT platform that allows us to

systematically and empirically measure clonal dynamics with high resolution, we can use it to

discover new biological processes and generate new exciting hypothesis. After all, as scientists,

we are always learning from our discoveries.

Method detail

Generation of Clonal Replica Tumors (CRTs)

Barcoding and generation of Clonal Replica Tumors.

CellTracker 50M Lentiviral Double-Barcoded Library was purchased from Cellecta. After expansion isolated tumor cells were infected over night at low MOI (≤ 0.25) in presence of Polybrene (8µg/ml) in 245mm square plates (Corning) at 70% confluence. After 48 hours, cells were detached, and infection efficiency was evaluated by flow-cytometry (percentage of RFP) and cells plated in Puromycin at an optimal concentration able to achieve cell killing of control cells. When confluent, barcoded cells were trypsinized, washed and plated in two 245mm square plates, named as Passage 1. Medium was periodically changed, and cells were left to grow until confluency. When confluent, P1 plates were trypsinized, washed and pooled before being split in two equal parts: one stored for sequencing, the other plated in two 245mm square plates as Passage 2. When confluent, Passage 2 cells were detached, washed, pooled and split in two equal parts: one cryopreserved in BamBanker (Wako Chemicals), the other plated in two 245mm square plates as Passage 3. The procedure was repeated over and over again snap-freezing 'odd passages' for sequencing, cryopreserving 'even passages' for culture re-propagation and isolation of clones.

Quantitative Scale

To calibrate, normalize and quantify sequencing results a five-log cell-spike in ('*Conversion Scale*') was added to each sample before DNA extraction. Basically, 293T cells were independently infected at very low MOI (≤ 0.25) with lentiviral particles encoding five custom double-barcodes not present in the tracking library but sharing the same plasmid backbone (pRSI16). After evaluation of infection by flow-cytometry (percentage of RFP+), infected cells

were selected in Puromycin and expanded. Limiting dilutions of each unique barcoded cells were prepared and equal volumes of diluted cells were mixed together. Pooled unique barcoded cells were then aliquoted in a multitude of 1.5ml vials, all containing the same number of barcoded cells represented as follow: BC.3063.3418 5×10^0, BC.4451.4842 5×10^1, BC.4858.5013 5×10^2, BC.5022.5430 5×10^3, BC.5993.6943 5×10^4. One vial of scale was systematically added to each sample before DNA extraction. After sequencing, reads generated by unique barcodes of the scale can be used to normalize reads generated by library barcodes, enabling conversion of reads to more interpretable numbers of cells.

Name	Cells Number	Double Barcode	Sequence
P1	5	BC.3063.3418	CATGCACACAGTACACTGttcgGTGTTGTGTGCAACTGAC
P2	50	BC.4451.4842	CAACGTTGTGACTGTGACttcgGTGTGTCAGTACTGTGAC
P3	500	BC.4858.5013	GTGTCACAGTACTGACTGttcgTGTGTGGTCAACTGTGAC
P4	5000	BC.5022.5430	GTCATGGTCAACTGTGACttcgGTTGCAACTGTGTGACTG
P5	50000	BC.5993.6943	TGGTGTTGCATGTGACTGttcgCACATGACCAGTTGTGAC

Figure 180. Sequence of the NGS read-to-cell conversion scale.

In vivo Transplantation and Treatments

All studies have been performed in NOD.Cg-Prkdc[scid] Il2rgt[m1Wjl]/SzJ (NSG) mice purchased from Jackson Laboratory and under experimental protocols approved by the UTMDACC Institutional Animal Care and Use Committee. For *CRTs* cohort generation up to 50 mice were injected subcutaneously with cells derived from stabilized cultures, 5×10^6 cells/mouse suspended in 1:1 DMEM/Matrigel (BD). For drug challenging *CRTs* mice bearing overt tumors were randomized and treated with Gemcitabine (Syd Labs) 100mg/kg IP every three days, AZD6244 (AbMole BioScience) 75mg/kg OG or BEZ235 (AbMole BioScience) 40mg/kg OG daily, for three weeks (Pettazzoni et al., 2015). When control and relapsed tumors reached 0.8-1 cm in size, animals were euthanized and whole tumors collected, weighed and snap-frozen in liquid

nitrogen. Tumors were assessed weekly and volume calculated using the following formula, $V = L^2 \times W/2$ (Length; Width). For serial transplantation, Passage 2 cells (50% of cells from P2 pooled plates) were injected subcutaneously in 4 NSG mice. When tumors reached 1 cm in size, mice were sacrificed, tumors collected, pooled and digested to single cells. One half of the recovered cells was reinjected in 4 animals, one half used for sequencing. Tumors were passaged for four times *in vivo*.

CRT Sample Processing and Analysis

Genomic DNA extraction and PCR for NGS library production

Genomic DNA extraction and PCR for NGS library production was performed as previously described (Carugo et al., 2016). Briefly, frozen tumors were minced through mechanical procedure with sterile scalpels and suspended in Buffer P1 (QIAGEN, 1 mL Buffer/100 mg tumor) supplemented with 100 μg/mL RNase A (Promega). The dissociation step was performed in disposable gentleMACS M tubes (Miltenyi Biotech) with the gentleMACS homogenizer (Miltenyi Biotec). Cell pellets from *in vitro* samples were suspended in 1 mL Buffer P1/RNAse A. Before samples were transferred in a 15 mL polypropylene tube (Falcon) and lysed adding 1/20 volume of 10% SDS (Promega) and 1/20 volume of Proteinase K (QIAGEN), the scale was added. After mixing, the cell lysates were incubated at RT for 5 minutes and the tumors at 56°C for 20 minutes. Genomic DNA was sheared by passing the lysate 10-15 times through a 22-gauge syringe needle. Then, a first genomic DNA extraction step was executed adding 1 volume of Phenol:Chloroform:Isoamyl Alcohol (25:24:1 pH8.0, Sigma Aldrich). After centrifugation (12000 rpm, 12 minutes), the upper phase was transferred to a new tube and a second extraction step with Chloroform:Isoamyl Alcohol (24:1, Sigma Aldrich) was performed. Again, the upper phase was transferred to a new tube and added with 0.1 volumes of 3M NaOAc (Sigma Aldrich) and 0.8 volumes of Isopropanol (Fisher Scientific) to precipitate genomic DNA. Centrifugation of

tumor samples was performed at 14000 rpm for 20 minutes, the samples from *in vitro* cells were stored over-night at -20°C before centrifugation. DNA pellet was washed once in 70% Ethanol (Fisher Scientific) and centrifuged again for 5 minutes at 14000 rpm. The DNA pellet was finally air-dried and dissolved over-night in UltraPure distilled water (Thermo Fisher). The final DNA concentration was assessed by NanoDrop 2000 (Thermo Scientific) quantification.

For NGS libraries generation, the barcodes were amplified starting from the total amount of genomic DNA in 2 rounds of PCR using the Titanium Taq DNA polymerase (Clontech-Takara) and pooling together the total material from the first PCR before proceeding with the second round. The first PCR reactions were performed for 16 cycles with 13K_R2 (5'-AGTAGCGTGAAGAGCAGAGAA-3') and FHTS3 (5'-TCGGATTCAAGCAAAAGACGGCATA-3'). The second PCR reactions were performed for 12 cycles with P5_NR2 (5-AATGATACGGCGACCACCGAGACGAGCACCGACAACAACGCAGA-3') and Gx1_Bp (5'-TCAAGCAGAAGACGGCATACGAAGACA-3'). Primers for the second PCR reactions were optimized in order to introduce the required adapters for Illumina NGS technology. PCR amplification products were analyzed by agarose gel electrophoresis (2.5%, Lonza) for the expected 279bp size. Amplified PCR products from 2 replicates of the second PCR reactions were pooled together and extracted from agarose gel with the QIAquick gel purification kit (QIAGEN). The amount of purified PCR product was quantified using the High Sensitivity DNA Assay (Agilent Technologies) for the Agilent 2100 Bioanalyzer. Barcode representation was measured by Next Generation Sequencing on an Illumina HiSeq2000 with 13K_Seq (5'-AGAGGTTCAGAGTTCTACAGTCCGAA-3') as sequencing primer. PCR biases in barcode detection and quantification were excluded comparing NGS sequencing runs of high-complex barcoded samples (P1) amplified in standard conditions or starting from more DNA template or through more cycles of first PCR reaction.

Testing the sensitivity of our model - quantitative detection down to single digit cells with high confidence

Before setting up this orthotopic model, we know we wanted a method that is quantitative and sensitive enough to detect early events of metastasis, mainly seeding, micro-metastasis, or early circulating tumor cells, which are often minute in quantity compared to the mass of primary tumor. An example of early micro-metastasis is shown below using PANC1 (PDAC ATCC cell line) in the mouse liver, where the metastatic lesion is around 10 cells in total (the sample is stained for HLA in red, Ki67 in green, and DAPI in blue, the background of the liver is brightened to appreciate histology of the liver).

To address the quantitative aspect, we created a "cell conversion scale", which contains known amount of cells carrying unique "scale barcodes" that is spiked in during sample DNA extraction. This "cell conversion scale" acts as an internal reference during sample extraction and barcode amplification as it is processed alongside all the "lineage tracing barcodes". For example, if the sample only has an 80% yield on the DNA extraction, the spiked in "cell conversion scale" would also have an 80% yield since it is processed together. After barcode amplification and sequencing the barcodes using NGS, in the end, the relative barcode of interest and its ratio to the "cell conversion scale" would stay relatively constant. Since we know the exact number of "cell conversion scale" we spiked in (in units of cells), we could use this scale to convert the "barcode (lineage) of interest" to cell number.

In terms of how the "cell conversion scale" is created, we utilize five custom scales each with its own unique barcode ID that is not part of the lineage tracing barcode library that is used for experiment. With these five individuals "scale barcodes #1 – #5", we created five individual cell lines, each infected with one of the scale barcodes at one barcode per cell ratio (infected at low MOI and puromycin selected for positive barcode integrated cells). Then, by cell counter, I would create two sets of scales: A "small" scale, and a "large" scale. The "small" scale contains

smallest cell number starting from 10 cells of scale barcode#1, 100 cells of scale barcode #2, 1K cells of scale barcode# 3, 10K cells of scale barcode #4, and 100K cells of barcode #5. Whereas the "large" scale, the smallest cell number started from 50 cells. The "small" scale is used on samples to detect circulating tumor cells/ cell free DNA from the blood, as well as, early time points of lung and liver samples. The "large scale" is used with primary tumor and late stage metastasis samples.

Now that we have addressed the quantitative aspect of this model, we must carefully assess the S/N ratio to have an idea of how sensitive and confident we can trust our system to quantitatively assess biological phenotypes. This can be broken down into three parts: Sample preparation and amplification, Sequencing errors, and Data processing.

In terms of sample preparation and amplification, we compartmentalize reagents and primers used for PCR and UV crosslink and wash pipettes between each unique samples; we check for none – specific PCR amplicons using a no template control (water), as well as, whole liver and lung from mouse that is not used in the experiment (no barcoded cells present) for amplification and NGS library building. We see no visible bands of supposed amplicon side by side of positive controls. In addition, sample DNA were amplified with as low PCR cycle numbers as possible and utilizes a nested PCR approach to ensure amplification uniformity.

These no template controls (NTCs) were sequenced (by mimicking the same procedure as the positive control samples) and reveals that NTC and tissue background control only had very minute reads present. Which confirms the sample processing workflow is clean and the human error contributing to the process is at a minimum. However, although NGS is extremely sensitive, it is not perfect. The sequence in the end is an accumulation of all the "errors" introduced during PCR's polymerase, amplification of primer-dimers, as well as sequencing's chemistry and systematic errors. To have an insight into how to filter the NGS read data, I looked into the sequences that were miscalled, the distribution of barcodes with only 1 read and their

nature, as well as the linearity of the "cell conversion scale". In summary, most if not all of the barcodes with just 1 read have high level of miscalling and should be excluded (also, if a barcode is not called more than once, I do not have confidence in calling it). In addition, we can use a cutoff of cell number (after converting NGS reads to cells) for high confidence analysis. The linearity of the scales is good with R close to 1 (within 5 cells – 500K cells), however, due to the limitation of how many cells we can spike-in during sample DNA extraction, we cannot guarantee the linearity of anything above 500K cells (high representing lineages with high read number). In this case, the accuracy of cell quantification would be off for those highly abundant lineages. However, as a whole picture, it would not affect our conclusion, as the dominant clones would still be accurately represented (with slight degree of error) as they are in the biological sample representation.

All in all, utilizing the sensitivity of PCR and NGS alongside negative control experiments and data analysis noise trimming, we have high confidence in our model system in quantitatively calling biological phenotypes down to single digit cell number and retain overall the relative abundance of each cell lineage and their relative biological representations in each sample.
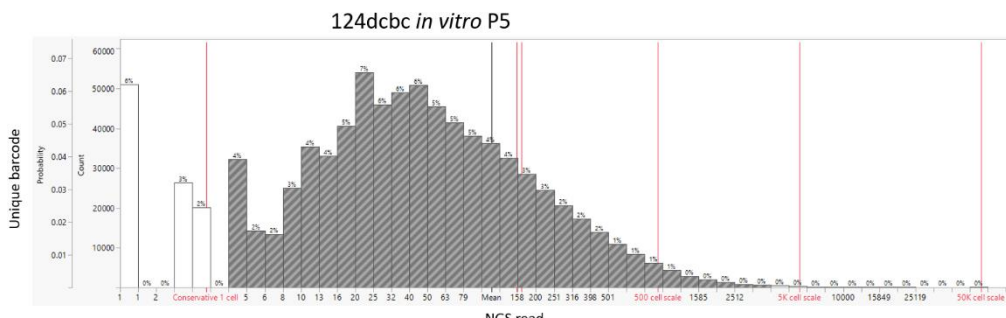
Figure 181. Barcode distribution in *in vitro* passage 5. X-axis = NGS read, red line highlighting the relative "cell" number converted by spiked in read-to-cell scale. Y-axis = count of barcodes.

An example of unfiltered barcode read distribution. By looking at how the barcode read distributes, we can have an idea of the amount of potential errors introduced by the sequencing workflow. In this histogram, taken from an *in vitro* passage, we see there are around 50,000 unique barcodes with only 1 read in total, which falls outside of a normal distribution pattern of how the other barcodes behave. After looking into what composes of these single read barcodes, many of which are barcodes similar by 1 base pair of the abundant barcodes in the sample. Alongside with that is the barcode is read only once and without a second read to verify within the samples, we will have more confidence in excluding any barcodes from analysis with only 1 read.
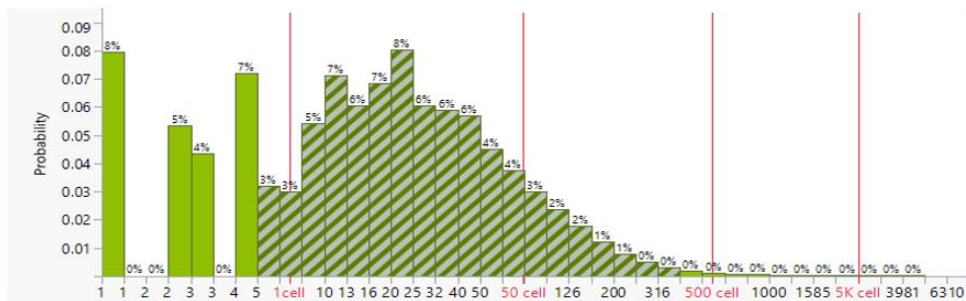
Figure 182. Barcode distribution in culture.X-axis = NGS read, red line highlighting the relative "cell" number converted by spiked in read-to-cell scale. Y-axis = count of barcodes.

Total barcode detected in "no template control (water)" and barcode null whole liver and lung background are few, indicating high specificity of the barcode detection workflow

To test the specificity of the barcode detection workflow from beginning to finish, we harvested the liver and the lung individually from mouse without barcoded cells injected. Taking those liver and lung (also water as a reagent control), we went through the entire process from gDNA extraction to the final NGS library. As seen from the figure below, the amount of barcode reads detected by NGS, compared to the samples with liver and lung metastasis (in the 10^7 read

number), the no barcode liver and lung and "water NTC" only has a very minor fraction of "noise" detected (in the 10 to 100 range). Since the barcode detected from no template control (water) and mouse liver and lung background is very minor, this indicates a high specificity and low noise for barcode detection in the barcode detection workflow.
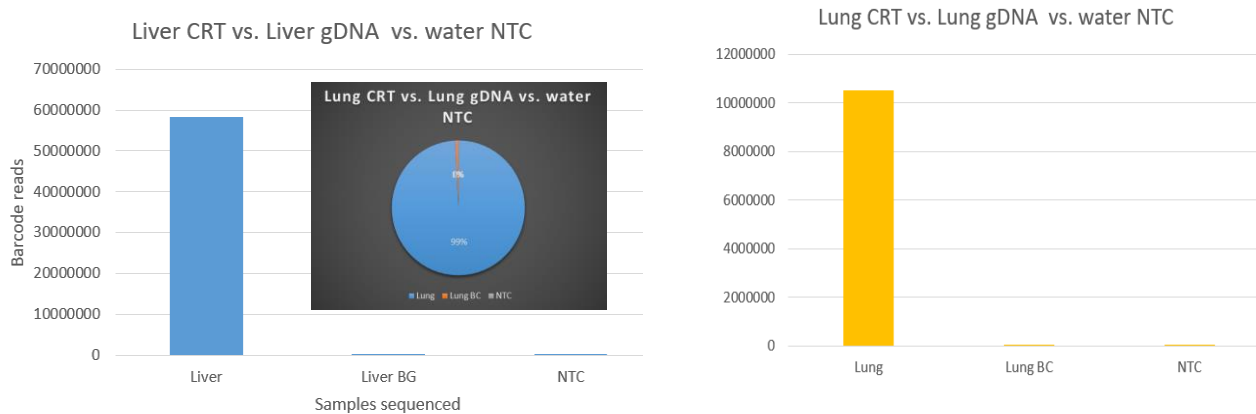
Figure 183. The total read number of barcode of liver/lung metastasis sample vs. the barcode-null liver/lung samples. The left graph shows the total number of barcode reads detected in the liver metastasis sample is close to 60,000,000 while the "no barcode liver" and "water NTC" sample only produce 10-100 barcode reads (background noise). The right graph shows the total number of barcode reads detected in the lung metastasis sample is close to 10,000,000 while the "no barcode lung" and "water NTC" sample only produce 10-100 barcode reads (background noise). These experiments are set up using the same reagents and processed in parallel.

The sensitivity of the barcode detection in the liver is calculated using the NGS to cell conversion scale

For each sample during DNA extraction, we spiked in a set of scales with known number of cells carrying unique barcodes that is not in the lineage tracing barcode library. The linearity of the scale is shown in the figure below with R value >0.98 for both primary tumor and liver. Converting the "noise" level of barcode reads from the previous section into cells using this scale, the sensitivity we claim for this model is ~ 20 cells.
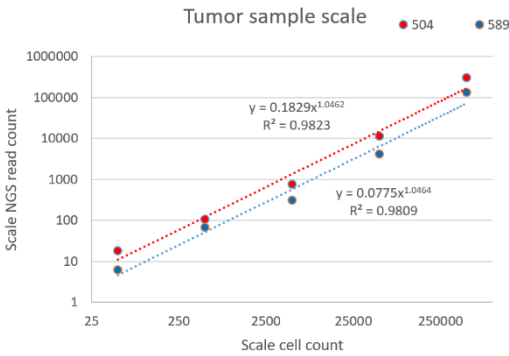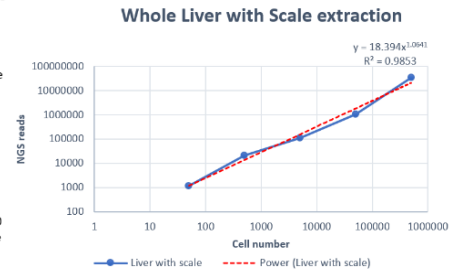
Figure 184. Cell conversion scale's linearity and sensitivity. Scales of 10 cells, 100 cells, 1K cells, 10K cells, and 100K cells are processed with primary tumor (#504 and #589, left graph) and scales of 50 cells, 500 cells, 5K cells, 50K cells, and 500K cells are processed with barcode-null, mouse liver. The scales showed good linearity (R^2 = 0.982, 0.981 and 0.985).

Barcode analysis pipeline for spatial barcode detection
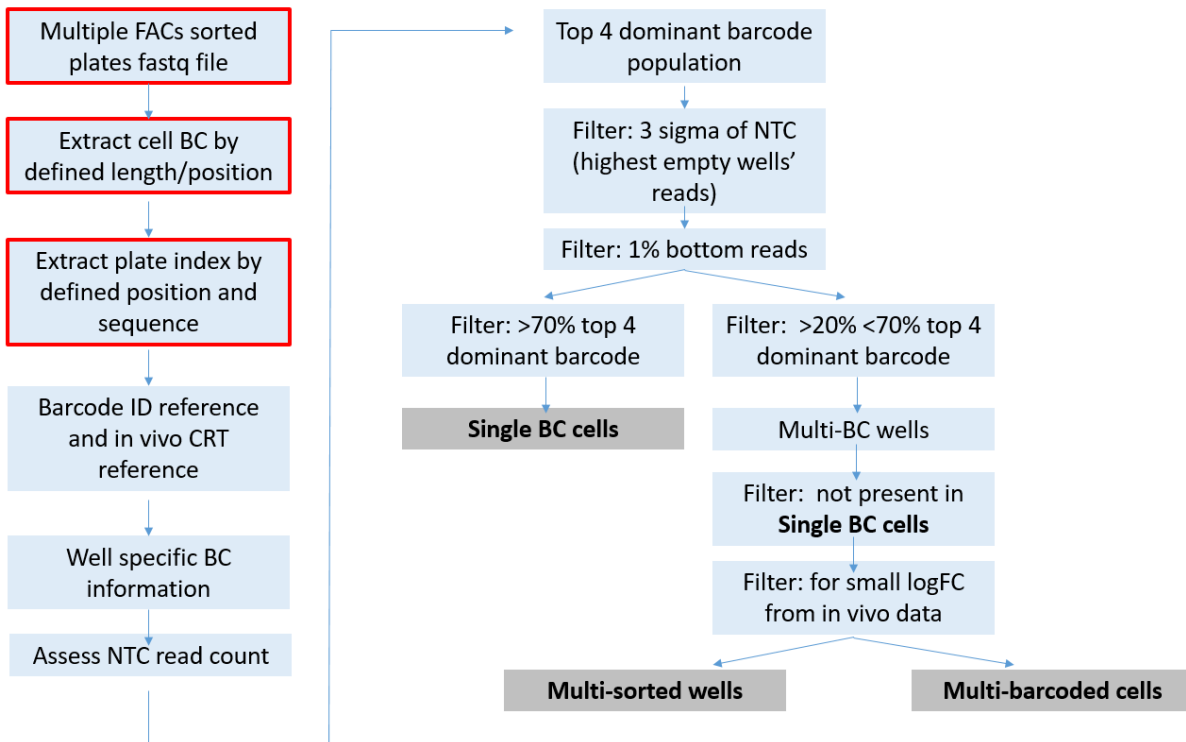
Figure 185. Schematics of novel analysis pipeline workflow to identify single barcoded vs. multi-barcoded cells.

High-throughput Clonal Isolation

Isolation of individual barcoded clones.

A major technical advancement of the *CRT* platform is that it enables the high-throughput, prospective isolation of nearly every *in vivo* computationally-identified clone by utilizing two rounds of NGS (Sup. Fig. 3a). Because barcoded cell cultures can be restarted at any time from cells used to generate *CRTs* stored in liquid nitrogen, we re-established cultures at lower clonal complexity by seeding barcoded cells in 96-well plates (5×10^3/well), a critical step toward the isolation of a pure population of treatment-naïve barcoded cells. When confluent, each seeded plate was replicated into two new plates (A and A'): one plate for sequencing and barcode detection (A), and the second plate for subsequent deconvolution and clone isolation (A'). The data from sequencing plate (A) provide a detailed map of the abundance (relative number of reads) of potential clones of interest across all wells, thus identifying "wells of interest" (i.e., wells with cultures characterized by relatively high contribution of a clone of interest). In *the second round* of NGS, wells of interest (from plate A') were further deconvoluted by establishing clonal cultures into new plates and barcode identified via high-throughput 'positional' sequencing. Briefly, cells from wells of interest were single sorted in multiple 96-well plates and left to grow until colonies are detectable using a high-content imaging system. Then, irrespective of confluency and growth kinetics, each plate is split into two plates to support clone identification and aliquot freezing for subsequent expansion. For clone identification, we established a rapid workflow that integrates the construction of a NGS library directly from trypsinized cells after PCR without the need for DNA extraction (Sup. Fig. 3b). We also incorporated two indexes (one for plate and one for well identification) in the NGS library adaptors to fast track multiple clones across a multitude of plates and wells in one single NGS run. This workflow enables cost effective, rapid identification of thousands of single clonal barcodes with extremely high sensitivity. As one might expect, only a fraction of single sorted cells grows as colonies (20-30%) (Sup. Fig 3c). This is an

intrinsic characteristic of tumors independent on the cell-cycle phase, where cells were transiting during sorting instead related to their ploidy, and indeed hyperploid cells were never able to restart a culture (Sup. Fig. 3c and Sup. Table 3). Barcode identity of individually isolated clones is then confirmed by Sanger sequencing and clones can be now expanded for storing and deep characterization.

A practical example of the unique opportunities provided by the CRT platform is shown in Supplementary Figure 3d-h, where we reported the isolation of 53 clones from de-complexed NGS run. Notably, the isolated clones capture the entire spectrum of drug response behaviors informed by *in vivo* CRT experiments (Sup. Fig. 3d and Sup. Tables 4-5) and include 19 clones of interest (7 gemcitabine resistant and 12 gemcitabine sensitive). After expansion and confirmation of their identity, as exemplified for two of them (Sup. Fig. 3f), clones can undergo multiple levels of characterization (Sup. Fig. 3g, h).

Deconvolution of barcode complexity- from cell culture to NGS.

Restarted barcoded cell cultures from previously frozen samples ('even passages', see Barcoding and generation of Clonal Replica Tumors) were left in incubator to allow recovery for 1-2 days. Cells were then detached using 0.25% trypsin-EDTA (Gibco), washed and resuspended in complete medium before being counted using Nexcelom Cellometer Mini cytometer. After counting, 5,000 cells were aliquoted in each well of 96-well flat-bottom TC plates (Corning/Falcon) and allowed growth till ~80% confluency in order for plated cells to divide at least once. Once the wells reached ~80% confluency, medium is removed, and wells washed with PBS. After adding to each well 30µL of 0.25% trypsin-EDTA (Gibco) to cover the entire bottom, cells were gently detached by pipetting. Using a multichannel pipette, 10µL of cells-trypsin were then transferred matching the well position into 96-well PCR 'sister' plates (Bio-Rad) for later NGS library processing. 20µL of DMSO-FBS (20:80) freezing solution were then added to each well of the remaining plates, now containing 20µL of trypsinized cells, before plates were sealed with

Parafilm and frozen at -80oC in insulated boxes to allow slow temperature cooldown. Frozen plates will be later used for flow-activated sorting to isolate single clones. In order to detect the barcode composition of each well of the frozen plates we took advantage of the replica 'sister' plates. Since the well barcode composition information acting the basis for downstream workflow (isolation of single clones from identified wells on interest), a high sensitive DNA extraction method is critical to guarantee no possible drop out of barcodes during detection. Conventional DNA extraction by column purification is not feasible nor ideal since the yield is < 100% and DNA material lost due to transfer between reaction tubes. Therefore, we optimized a direct cell (after trypsinization)-to-NGS library workflow to capture the entire DNA content (barcode) present in all the cells in all wells with high sensitivity and minimum hands on time. The rationale behind the workflow involved the use of a protease K based strong lysis solution to lyse cells and release DNA from the nuclei and histones. At the same time, proteinase K inactivated trypsin, which is a strong known PCR inhibitor. Then, after heat inactivation of the protease, we used primers with tailed NGS adaptors (providing unidirectional reads) to amplify the regions of interest via PCR. To evaluate the robustness of cell lysis, cells were stained with Hoechst-33342 (Thermo Fisher) and tested against "PK lysis buffer", "other commercial lysis buffer" and PBS and checked under the microscope. The optimized "PK lysis buffer" gives the most complete lysis of cells as seem by the lack of positive Hoechst stained cells (Sup. Fig. 3b). To test the reproducibility of the lysis conditions, eight replicates of 15,000 cells were lysed each with "PK lysis buffer", "other commercial lysis buffer" and "water", then used TaqMan RNAseP assay (Thermo Fisher) for PCR-compatible DNA read out. The "PK lysis buffer" is proven to be extremely efficient and reliable giving the lowest average Cq of 24.22 and the lowest standard deviation of 0.17 Cq (Sup. Fig. 3b). During the cell lysis step, another central function of protease K is to inactivate trypsin. To test the lysis condition's effect on PCR, we used human gDNA control input DNA (Thermo Fisher) of 25ng, 5ng, 1ng, 0.2ng and TaqMan RNAseP assay (Thermo Fisher) to test each condition's effect on PCR efficiency. In supplementary figure 3b, we show that PCR amplification efficiency

is impacted minimally after heat inactivation of protease K (efficiency of 97.78%, n=2). Also, trypsin, without the inactivation of protease K, inhibits PCR reaction completely (no Cq detected). And lastly, the entire "PK lysis buffer" workflow gives a high PCR efficiency of 95.49% (n=2).

Using the optimized direct cells to NGS workflow, to detect the barcode composition of each well of the frozen plates, we added a total 4.7uL of lysis buffer (10mM Tris-HCl pH7.5, 10mM NaCl, 10mM MgCl2, 0.19% NP40, 2µg Proteinase K) to each well of the 'sister' replica plates containing 10µL cells-trypsin, and incubated at 50oC for 40 min, and then at 95oC for 15min to inactivate Proteinase K.

Once the cells were lysed, the next step was to build NGS library directly in the same well. The rationale behind the tailed adaptor primer design is that, after PCR amplification, a uni-directional read NGS library will be produced, therefore, maximizing the read throughput and cell barcodes interrogated on the Ion Torrent NGS platform. We choose Ion Torrent as an NGS platform for barcode screening, due to its fast turnaround time and the ability to handle low complexity libraries, which is also a critical feature for the clonal isolation step (see below). Based on the above rationale, to amplify the barcoded region with tailed primers containing NGS library adaptors, we added directly to each well PCR-reaction mastermix according to the manual from Titanium Taq PCR Kit (Clonetech) with individual forward primer with unique index Ion-A-BC-GexBP1 (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGXXXXXXXXXXXGATTCAAGCAGAAGACGGCAT ACGAAGACAG-3', X bases refers to Ion Xpress 96 adaptor barcode sequence) and a common reverse primer GexSeqS-IonP1 (5'-CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATTGAGGTTCAGAGTTCTACAG TC-3') both at 600nM final concentration. Reactions were mixed (15-20µL) and PCRed with cycling profile of 95oC 3minutes, 6 cycles of 95oC-30 seconds, 56oC-10 seconds, 72oC-10 seconds, followed by 22 cycles of 95oC-30 seconds, 62oC-10 seconds, 72oC-10 seconds, and

then final extension at 72oC for 2 minutes and 4oC hold. Now each well contained amplified barcodes with NGS compatible adaptors with unique well indexes. Then, 10µL of PCR products from each well were pooled together and 500µL of the pooled product took and concentrated 10 fold with one time 1.8X AMPure XP purification (Beckman Coulter) by eluting with 50µL elution buffer. The 50µL of concentrated DNA were loaded on a 2% agarose gel for band purification. Expected bands (size ~185bp) were cut and DNA purified using QIAquick gel extraction kit (Qiagen). Following gel purification, DNA was purified with one time 1.8X AMPure XP purification (Beckman Coulter) and quantified, the library then was sequenced with Ion Torrent Proton NGS platform (Thermo Fisher).

Isolation of single barcoded cells, flow-activated cell sorting and positional sequencing.

After analysis of the NGS data from the de-complexed 'sister' plates and identification of the wells that have higher representation of clones of interest (applying read cutoff of 100 reads and considering % representation of the well), keeping plates on ice, thaw the identified wells from the 96-well TC plate previously frozen by adding 180µL of 37°C warm DMEM medium. Transfer the thawed cells to another 96-well TC plate and let the cell recover in the incubator. After two days, trypsinize, wash and suspend the cells in DMEM/FBS containing 1µg/mL Hoechst-33342 and SYTOX Green (Thermo Fisher) to sort live single cells. Alternatively, cells can be sorted also according to their DNA content, excluding polyploid cells in maximizing chances of isolating growing cells. Cells are sorted in 96-well TC plates containing DMEM medium supplemented with 20% FBS, and 1% Pen Strep. Single cell colonies are then left to grow periodically monitoring the plates by high-content imaging using Operetta (Becton Dickinson). When wells reach ~50% confluency, cells are detached and split using the approach described above for "Deconvolution of barcode complexity". As already described two/thirds of the trypsinized cells remain in the original 96-well TC plates where are slowly frozen in a final volume of 40µL for later re-culturing. The other third of the cells (10µL) is aliquoted into 96-well PCR

plates ('sister' plates) for barcode detection. Process multiple plates with clonal cell population at the same time following the lysis protocol describe previously. Since not all wells will have the same number of cells, the direct cell lysis method is used to ensure capturing the barcode information of wells with fewer cells. After cells are lysed, prepare the PCR reaction according to the manual described in Titanium Taq PCR Kit manual (Clonetech) using individual forward primer with unique index Ion-A-BC-GexBP1 (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAGXXXXXXXXXXXGATTCAAGCAGAAGACGGCAT ACGAAGACAG-3', X bases refers to Ion Xpress 96 adaptor barcode sequence Thermo Fisher Cat.4474517) and for *each plate a common unique indexed* reverse primer GexSeqS-bc-IonP1 (5' - CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATYYYYYYYYYCGAGGTTCAGAG TTCTACAGTC-3', Y refers to individual "plate index", Tab. 1) both at 600nM final concentration. Mix the reaction and PCR with cycling profile of 95°C 3 minutes, 7 cycles of 95°C-30 seconds, 56°C-10 seconds, 72°C-10 seconds, followed by 23 cycles of 95°C-30 seconds, 62°C-10 seconds, 72°C-10 seconds, and then final extension at 72°C for 2 minutes and 4°C hold. The cycles of the PCR can be adjusted according to the number of plates pooled. Each well now contains amplified barcodes with NGS compatible adaptors with unique well indexes, as well as unique plate indexes. Pool together all the PCR products from each well and each plate and load 50uL on a 2% agarose gel for band purification. Cut the expected band size ~190bp and purify DNA using QIAquick gel extraction kit (Qiagen). Following gel purification with one time 1.8X AMPure XP purification (Beckman Coulter) and sequence with Ion Torrent Proton NGS platform (ThermoFisher).

Expansion of isolated clonal cultures and Sanger barcode validation.

After analysis of NGS data we use plate and well indexes to trace back the correlated barcode with the well position from the frozen plates. Wells are thawed as previously described

and allowed to expand in TC dish for further Sanger validation of the barcode. Briefly, after expansion a fraction of clonal cells is directly lysed and genomic DNA used for PCR amplification (condition identical to the procedures described previously). For PCR reaction, use GEXBPi (5'-TCAAGCAGAAGACGGCATACGAAGACA-3') and SangerNR2 (5'-ACGAGCACCGACAACAACGCAGA -3') as forward and reverse primers at final concentration 600nM, and follow cycling condition 95°C 3 minutes, 38 cycles of 95°C-30 seconds, 60°C-20 seconds, 72°C-10 seconds, followed by final extension at 72°C for 2 minutes and 4°C hold. Purify PCR products using 2% agarose gel and cut expected band at 242bp. Purify DNA using QIAquick gel extraction kit (Qiagen) and followed by one time 1.8X AMPure XP purification (Beckman Coulter) for sanger sequencing. Use SangerNR2 as sequencing primer (5'-ACGAGCACCGACAACAACGCAGA -3'). In addition, a high throughput clonal screening can be integrated into this workflow. For example, during expansion of thawed plates for barcode validation, optionally, more replica plates can be generated for high throughput drug/compound screening purpose.

In vitro clonal competition assay.

Two isolated clones (5125_5793 and 13767_14001) were used to perform the *in vitro* reconstitution assay. One clone (5125_5793) was labeled with GFP, while the other clone (13767_14001) remained GFP-negative. The GFP-labeled clone was then mixed with the GFP-negative clone with equal number (75,000 cells each) and treated with 0, 100nM or 500 nM gemcitabine (Selleckchem) for 5 days followed by washout and allowed for 2 weeks recovery. The percentage of each clone in each group was analyzed by flow cytometry (Becton Dickinson Canto II Analyzer) (Sup. Fig. 3h).

Crystal Violet Cell Viability Assay.

The crystal violet cell viability assay was performed as follow. In brief, cells were seeded in 6-well plate (5×10^4 cells/well) and incubated at 37°C for 24 hours. For gemcitabine-only treatment, cells were treated with or without 100 nM gemcitabine (Selleckchem) for 2 days followed by replacement of fresh medium, and then kept at 37°C for 2 weeks. At the endpoint of the experiment, cells were washed twice with PBS and stained with 1 mL of 0.5% crystal violet in each well for 20 minutes with gentle shaking. After washing with PBS for three times, the cells were air-dry for 2 hours at room temperature and then scanned for images. For the optical density quantification, 1 mL of methanol was added to each well and incubated at room temperature for 20minutes with gentle shaking followed by detecting at 570 nm (OD570) with the POLARstar omega plate reader.

For combination treatment of gemcitabine and ATRi AZD-6738, cells were pretreated with 1 µM ATRi AZD-6738 (Selleckchem) for 3 hours and then co-incubated with 1 µM gemcitabine for another 24 hours; after changed with fresh medium, cells were kept at 37°C for 2 weeks. At the endpoint of each experiment, cells were washed twice with PBS and stained with 1 mL of 0.5% crystal violet in each well for 20 minutes with gentle shaking. After washing with PBS for three times, the cells were air-dry for 2 hours at room temperature and then scanned for images.

Analysis for basal DNA damage level by flowcytometry.

The detection of basal DNA damage level in each clone was performed by using the fluorescent anti-γH2AX antibody (Alexa Fluor® 647 Mouse anti-H2AX (pS139), BD Biosciences) according to the manufacturer's instructions. In brief, cells were seeded in 6 cm plates until they became 60-70% confluent. Cells were then detached, fixed and permeabilized followed by antibody staining for 20 minutes at room temperature. After washing for three times and DAPI staining, stained cells were analyzed with a Becton Dickinson Canto II Analyzer.

Immunofluorescence.

Cells from each clone were seeded on cover slips two days before treatment. Cells were then treated with or without gemcitabine 1 µM (Selleckchem) for 24 hours followed by 3 hours recovering. After fixation, permeabilization and blocking with 3% BSA, 0.1% Triton-X 100 and 5% normal goat serum, cells were stained with mouse anti-γH2AX (abcam) and/or rabbit anti-RAD51 (abcam) for overnight at 4°C. Secondary antibodies conjugated with rabbit Alexa-488 and mouse Alexa-594 (Molecular Probes) were used. After washing and DAPI staining, images were captured with a Hamamatsu C11440 digital camera, using a wide-field Nikon Eclipse-Ni microscope.

Hematoxylin and Eosin staining

Collected tissues were fixed overnight with buffered PFA followed by 70% ethanol and embedded in paraffin using Leica ASP300S processor. Tissue blocks were sectioned with microtome (Leica RM2235). Slides were then de-paraffined and stained with hematoxylin and eosin. Images were captured with a Nikon DS-Fi1 digital camera using a widefield Nikon Eclipse-Ci microscope.

Invasion assay

Invasion assay was performed with Corning® BioCoat™ Matrigel® Invasion Chambers according to the manufacturer's instructions. In brief, cells were starved with serum-free medium for 24 hours. $1\text{-}2\text{x}10^5$ cells in 0.5 ml serum-free medium were added to the inserts, and medium plus 10% FBS were added to the wells as chemoattractant with or without NMDAR agonist glutamate or antagonist memantine. After 27 hours, cells were fixed with 4% PFA followed by 0.5% crystal violet staining. Images were captured with a wide-field Nikon Eclipse-Ni microscope. The number of invaded cells was determined using Image J software.

Reverse Transcription and Quantitative Real-Time PCR

Total RNA from each clone was extracted with Qiagen RNeasy Mini Kit according to manufacturer's instructions. 1 µg RNA was used to generate cDNA using Thermo Fisher SuperScript™ VILO™ Master Mix. 10 ng of cDNA was used for quantitative PCR using Applied Biosystems PowerUp SYBR Green PCR Master Mix with gene-specific primers. Gene relative

expression level was determined by the comparative CT ($\Delta\Delta$CT) method and normalized to GAPDH.

Intracellular Ca^{2+} assay

Intracellular Ca^{2+} concentration was determined by Enzo FLUOFORTE® Calcium Assay Kit according to manufacturer's instructions. In brief, 8 x 10^4 cells per well were plated in 96-well plate for overnight. The cells were then stained with FLUOFORTE® Dye for 1 hour at room temperature followed by fluorescence detection at Ex=490 nm/Em=525 nm using PHERAStar HTS microplate reader (BMG Labtech).

Quantification and statistical analysis

Barcode data processing.

A custom analysis pipeline was used to quantify barcodes from FASTQ files. Reads were first filtered based on the presence of a 4bp spacer (TTCG) between the two barcodes, positions 19-22 bp. Reads which had the spacer, within a hamming distance of 1 bp, were used for downstream analysis.

The reads were then split into two separate Fastq files (1-18bp, 23-40bp), and both barcodes (BC1 and BC2) were aligned to the 13K library, preserving their pairing. Bowtie (v2.2.3) was used to perform the alignment allowing 1bp mismatch at either end of the barcode (Langmead and Salzberg, 2012). These parameters were optimized to enable maximal alignment, while minimizing alignment of reads to multiple barcodes. Furthermore, reads where both barcodes aligned to the library uniquely were preserved for counting. We used SAMtools to extract the reads and perform counting of paired reads (Li et al., 2009).

Barcode data analysis and normalization.

Counts obtained using methods described above were normalized for library size, by calculating counts per million for each barcode (i):

$$\mathrm{CPM_i} = \frac{\mathrm{counts_i}}{\sum\mathrm{counts}} 10^6$$

This enables comparisons of barcodes within and across samples. Furthermore, we used the quantitative scale to infer cell numbers represented by each barcode. For this, we used a log-linear model to regress centered CPMs over cell numbers (Sup. Fig. 1b, c), to infer barcode counts represented by 5 to 50,000 (Scale1-5). Barcodes for 5 cells were detected in 8/12 samples from PDX1 and 11/11 samples from PDX2. We see a strong linear trend comparing log CPM vs number of cells, which improves for higher cell numbers. We used barcodes, inferred to represent >50 cells for *vitro-vivo* (Fig. 1d and Sup. Fig. 1f) and *vivo*-serial transplantation (Fig. 1i) comparisons. For all pairwise analysis we used all barcodes common between the two samples, evaluated using Pearson correlation coefficient (Fig. 2c, f, i and Sup. Fig. 1i-k). For comparison of n>2 samples, for *in vitro* dilutions of PDX2 (Fig. 2a) we used barcodes shared by all samples and for *CRTs* of PDX1 and PDX2 (Fig. 2d, g) we used barcodes above 50 cells of abundance. Mean barcode counts across replicates was compared with the quantitative scale, to infer cell counts for the heat-maps (Fig. 2b, e, h).

Differential analysis of clonal lineage expansion

We leveraged computational frameworks previously developed for differential gene expressions analysis to perform differential barcode representation analysis (Robinson et al., 2010). Clonal lineages common between the three PDX2 *CRTs* (n=5683), were used for differential lineage expansion analysis. Different behaviors were interrogated for sensitivity and resistance to various pharmacological agents using the generalized linear modeling framework

provided by edgeR (Robinson et al., 2010) (Fig. 3b, c, d). Barcodes with FDR<0.2 were further manually curated before downstream validation and median centered for representation (Fig. 3e).

We used constrained correspondence analysis (https://www.jstor.org/stable/1938672?seq=1#page_scan_tab_contents) to dissect lineages into clusters based on response, as implemented in R package vegan (https://CRAN.R-project.org/package=vegan). Relative abundances were ordinated using CCA, constrained upon drug treatment (Fig. 3f). Log fold-change from edgeR was overlaid on the ordination to investigate the degree of overlap in terms of response to various treatments. We extracted lineages which were modulated by either treatment (FDR<0.25, FC>1 or FC<-1, n=2440), and compartmentalized them into eight quadrants in a 3-dimensional space (Fig. 3g and Sup. Fig. 2 f-h).

Simulations were performed to evaluate if the number of lineages observed in different categories of response can be expected by random sampling (Fig. 3h). We setup the in silico experiment to sample lineages determined significant in either direction (depleted or enriched), and any of the three treatments (GEM +/-, AZD +/-, BEZ +/-). Overlaps between sampled lineages were calculated and categorized according to response to the three drugs (+: enriched, -: depleted, 0: neutral), for 10,000 simulations (black boxplots, Fig. 3h). Estimates from simulations were then compared with observed values (red), and empirical p-values were calculated.

Non-linear dimension reduction (tSNE) and mathematical modeling.

Normalized counts (CPM) were used to filter barcodes representing >50 cells, by fitting observed reads counts to expected cell numbers (Sup. Fig. 1c). For each model, barcodes representing >50 cells, in more than one sample were retained for downstream analysis. Barcodes were then clustered using t-SNE

(http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf) and plotted using ggplot2 (https://doi.org/10.1198/jcgs.2009.07098). Normalized counts were overlaid on, while preserving the same clustering to tracing fitness dynamics in the two experimental arms, *vitro* and *vivo* (Fig. 1h).

To further understand the underlying mechanisms of fitness dynamics, we performed simulation experiments across passages, comparing the two experimental arms (*vitro* vs *vivo*). Specifically, we can argue that the barcode dynamics of a specific lineage, in subsequent passages is impacted by the state in previous passages, self-renewal ability and some stochastic behavior. Using this argument, we have described the barcode composition as follows

$$X_{ij} \sim X_{ij-1}(1 + s_i + z_i^x)$$

$$Y_{ij} \sim Y_{ij-1}(1 + s_i + z_i^y)$$

where, X_{ij} are counts *in vitro* for barcode i and passage j and similarly, Y_{ij} represents counts in *in vivo* passaging. Here, s_i represents fitness for barcode i and z_i^x, z_i^y are the fitnesses impacted by environment x and y, respectively. Furthermore, if we assume that the fitness between the two experimental arms (*vitro* vs *vivo* passaging) is not related, or $s_i = 0$, we can arrive at the following equations:

$$X_{ij} \sim X_{ij-1}(1 + z_i^x)$$

$$Y_{ij} \sim Y_{ij-1}(1 + z_i^y)$$

We simulated draws from a previous passage ($X_{ij-1} \text{ or } Y_{ij-1}$), assuming $s_i = 0$, and calculated the number of overlapping barcodes (over 1000 simulations). This null distribution is then compared with the observed values to estimate the empirical p-value (Sup. Fig. 1h).

Barcode analysis of IonTorrent data.

Fastq were de-multiplexed for row and column barcodes using Ion Torrent NGS software. A more stringent analysis strategy was used, considering differences in the sequencing platform and the nature of clonal sample. Subsequent analysis was performed using custom R scripts, developed using Bioconductor packages (Biostrings and ShortRead) (Morgan et al., 2009) (https://www.bioconductor.org/packages/release/bioc/html/Biostrings.html). Briefly, we extracted reads which follow the pattern "5'-TCAAGCAGAAGACGGCATACGAAGACAGTTCGNNNNNNNNNNNNNNNNNNNNNTTCGNNNNN NNNNNNNNNNNNNNTTCGGACTGTAGAACTCTGAACCTCRYYYYYYYY-3'", where Ns represent Barcode 2 and Barcode 1 respectively and Ys represent the plate barcode. Barcodes of each read were then compared with the library (CellTracker 50M Lentiviral Double-Barcoded Library, Cellecta), allowing for mismatches. Reads, where both barcodes aligned uniquely with the sequences in the library with minimum hamming distance, were used for downstream analysis. We then separated reads from different plates using the plate specific barcodes. A read number cutoff of 100 reads (n=1, no template negative control's top barcode read is 37) is applied to remove false positive barcode ID from each of the 96 samples (well index). We chose wells where clones of interest represented a dominant fraction of reads, to increase the chances of their isolation.

Further consideration was applied to Ion-Torrent data where each well was expected to have single barcodes (clonal isolation, step 2). In case same barcode is detected in multiple wells, wells with maximum percent representations that is higher than 75% was chosen for downstream expansion and isolation. At this point, all barcodes identified can be traced back to its frozen counterpart, knowing plate and well position. It is possible, although exceptional, that more than one unique barcode is associated with each sample due to negligible percentage of double barcoding and the accuracy of flow-activated single cell sorting.

271

Table 4 Plate Indexes

GExSeqS-bc-IonP1	5'
	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATYYYYYYYYT GAGGTTCAGAGTTCTACAGTC

GExSeqS-bc1-IonP1	5'
	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATTATATCCAC GAGGTTCAGAGTTCTACAGTC

GExSeqS-bc3-IonP1	5'
	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATTATCCATAC GAGGTTCAGAGTTCTACAGTC

GExSeqS-bc5-IonP1	5'
	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATTCCATATAC GAGGTTCAGAGTTCTACAGTC

GExSeqS-bc7-IonP1	5'
	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATTATAGGTA CGAGGTTCAGAGTTCTACAGTC

GExSeqS-bc9-IonP1	5'
	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATTAGGTATA CGAGGTTCAGAGTTCTACAGTC

Whole Exome Sequencing (WES) of isolated clones.

Exome capture was performed on genomic DNA per sample based on KAPA library prep (Kapa Biosystems) using the Agilent SureSelect Human All Exon V4 kit according to the manufacturer's instructions and paired-end multiplex sequencing of samples was performed on the Illumina HiSeq 2000 sequencing platform. Paired-end reads were generated by Illumina's bcl2fastq application and aligned to human genome reference (hg19) using Burrows-Wheeler Aligner (BWA) with seed length of 40 and maximum edit distance of 3 (allowing for distance <=2 in the seed) (Li and Durbin, 2009). BAM files produced by BWA were further processed according to GATK Best Practices, including removal of duplicate reads, realignment around indels and base recalibration (DePristo et al., 2011; Van der Auwera et al., 2013). Mutations were detected using Platypus (Rimmer et al., 2014) and MuTect (Cibulskis et al., 2013) by using an in-house normal reference created using 11 (normal) blood WES samples. Confident mutations were selected using an intersection of the two callers, and filtering out mutations with low depth (<20) and those annotations in dbSNP v129 and ESP-6500 [ref: http://evs.gs.washington.edu/EVS/]. All mutations detected in at-least one sample were kept, and mutation allele frequencies were re-derived by interrogating reads from BAM files, for all mutations across all samples. Apart from mutations, CNV was also derived from the WES data. An in-house R package (exomecn) was used to calculate read counts for each exon, and log2ratios were called comparing each sample with an assay matched (previously sequenced normal reference). Thus derived, log2ratios were normalized for total mapped reads in that region, and segmented using circular binary segmentation (CBS), as part of DNAcopy package of Bioconductor.

RNA Sequencing and transcriptomic analysis of isolated clones.

The RNeasy Mini Kit (Qiagen, Inc., Valencia, CA) was used to extract total RNA from all samples. Quantity and quality were measured using the RNA Nano kit on the Agilent Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA). RNA sequencing reads were derived from Illumina's intensity files as described in the WES section above. These were then aligned to hg19 transcriptomic reference using kallisto (Bray et al., 2016). Gene level TPM counts for protein coding genes (as defined by HGNC) and expressed in at least 3/11 (TPM>1) were used for all downstream analysis.

Gene level counts were resolved into pathway level enrichment scores using ssGSEA. These were then compared with response to different drugs (GEM, BEZ and AZD), using linear regression. Pathways associated with sensitivity and resistance were prioritized for downstream validation. In addition, we used the top resistant and sensitive clones and carried out differential gene expression analysis using edgeR. Top 100 most depleted and resistant genes were used to create a signature of response to gemcitabine.

TCGA gene expression data and survival data were downloaded from GDC and processed as previously described (Cancer Genome Atlas Research Network. Electronic address and Cancer Genome Atlas Research, 2017). We used nearest-template-prediction (Hoshida, 2010) to assign subtype scores to each sample, using the FDR threshold of 0.1. Survival curves for samples in the three categories (gemcitabine-resistant, gemcitabine-sensitive, un-categorized), were compared using Cox-proportional hazards model.

Phylogenetic analysis.

Phylogenetic trees were derived for the 13 samples using 3 different metrics namely, copy number variation (CNV), single nucleotide variation (SNV) and RNASeq gene level expression

(RNA). Data were derived as described in the previous sections. We used Euclidean distance and hierarchical clustering in R to derive trees and then edited them in Adobe Illustrator.

Estimating copy number variation in SNP6 array data.

DNA copy number for PDX2-derived cells was calculated using Affymetrix SNP6.0 arrays by the tool PICNIC (predicting absolute allele copy number variation with microarray cancer data) (Greenman et al., 2010) in a two-step process. First, the data were preprocessed by normalizing individual sample variation to the total DNA measured by determine the allelic responses of each probe based on normal samples. This information was used to transform all tumor sample measurements into a copy number intensity and genotype intensity at each probe locus. Next step involved segmentation and genotyping of the loci by fitting a bayesian HMM in which the possible states were the possible total and minor copy numbers.

Statistical testing and clustering

By Delia Wang and Ruitao Lin (discussion with Kim Do, Andrea Viale, and Chieh-Yuan Li)

Barcode cutoff procedure

For all barcodes, to remove potential noise caused by sequencing technically, we consider the below scenario for noise filtering:

1. For all primary, any reads below or equal to 3 reads, treated it as "0".

2. For all metastasis samples, any reads below or equal to 1 reads, treated it as "0"

For Primary tumor analysis:

3. Keep if barcode count is not "0" at any one of time points (N=46839)

a. Calculate the prevalence rates (N=46839): $p_x = x_i / \sum x_i$

b. Keep the barcode if percentage is not "0" at all the time points (N=1013)

Similarity test

We are using the similarity between "injected" (two of the aliquot tubes used to create CRTs) as a simulation criteria to gauge the similarity of clonal replica tumors at later time points. Using such criteria, we used bootstrapping (reference) method to simulate the data, then applying statistic T1 and T2, each informs on how similar the CRTs are at each time point. By statistical robust test, we have demonstrated that each CRTs indeed are similar at the same time point, as well as progressing similarly over time (until week10). Therefore, we can view CRTs in a cohort as an interchangeable single variable to satisfy the proposed longitudinal sample. This similarity test is demonstrate that the tumor growth patterns for different mice (CRTs) are similar in the beginning stage of the observation, while in the latter stage, the tumor growth diverges in different mice.

1) Let X=$(x_1,\ldots, x_n)$ denote the observed barcodes counts for mouse A, and Y=$(y_1,\ldots, y_n)$ denote the observed barcodes counts for mouse B, where N is the total number of barcodes observed.

2) In mathematics, the similarity between sample A and sample B can be quantified as $p_X \approx p_Y$, i.e., the prevalence rates for "important" barcodes between A and B are similar. Here p_X =$(p_{x1},\ldots, p_{x_{n_{A \cup B}}})$ (or p_Y =$(p_{y1},\ldots, p_{y_{n_{A \cup B}}})$) is the vector of prevalence rates.

3) We assume that X=$(x_1,\ldots,x_{n_{A \cup B}}) \sim$Multi$(N_X, p_x)$ and Y=$(y_1,\ldots,y_{n_{A \cup B}}) \sim$Multi$(N_Y, p_y)$, where $N_X = \sum x_i$ and $N_Y = \sum y_i$ are the total numbers of reads for mouse A and mouse B, respectively.

4) We define two statistics, T_1 and T_2 based on p_X and p_Y, to measure the similarity between two samples at each time point.

    a. Statistic 1: T_1 is the Pearson correlation coefficient between p_X and p_Y. If T_1 is close to 1, A and B is similar.

    b. Statistic 2: T_2 is the Euclidean distance between p_X and p_Y, i.e., $T_2 = ||p_A - p_B||_2$, $(d_{A,B} = \sqrt{\sum_{i=1}^{n=A \cup B}(p_A - p_B)^2}$. If T_2 is close to 0, A and B is similar.

5) To get T_1 and T_2, we use the bootstrap resampling approach to perform the similarity test, 10,000 times at each time point.

    a. Denote $\text{logit}(x) = \log(\frac{x}{1-x})$

    b. Generate $\tilde{p}_X$ and $\tilde{p}_Y$ via $\tilde{p} = \log[-\text{logit}(\hat{p})] + \epsilon$ where $\epsilon \sim r*N(0, \sigma_1^2) + (1-r)*N(0, \sigma_2^2)$ and $r=1$ if $\hat{p} > 10^{-6}$ and $r=0$ if $\hat{p} \leq 10^{-6}$

    c. Generate $\tilde{X} \sim \text{Multi}(N_X, \tilde{p}_X)$ and $\tilde{Y} \sim \text{Multi}(N_Y, \tilde{p}_Y)$

    d. Perform the selection procedures on $\tilde{X}$ and $\tilde{Y}$, and obtain $\tilde{p}_X$ and $\tilde{p}_Y$. Then calculate the test statistics $\tilde{t}_1$ and $\tilde{t}_2$.

    e. Repeat for 10,000 times and obtain the distributions of $\tilde{t}_1$ and $\tilde{t}_2$, respectively.

6) Perform the Similarity test.

    a. Null hypothesis: Two samples are similar

    b. Reject the null hypothesis if $\Pr(\tilde{t}_{simulated} < \tilde{t}_{observed}) < 0.05$ for T_1 or $\Pr(\tilde{t}_{simulated} > \tilde{t}_{observed}) < 0.05$ for T_2.

Calibrate parameters (sigma$_1$ and sigma$_2$) based on the similarity between two mice at injected (Week0)

We use the grid searching approach to determine parameters: sigma1 and sigma2. The rational of this calibration step is that samples are similar according to the experiment at injected (Week0).

1) Use observed prevalence rates, i.e., $\hat{p}_X$ and $\hat{p}_Y$, that is not 0 at all points (N=46839) to simulate at each time point.

    c. Based on $\hat{p}_X$ and $\hat{p}_Y$, we get $\tilde{X}\sim$Multi($N_X, \tilde{p}_X$) and $\tilde{Y}\sim$Multi($N_Y, \tilde{p}_Y$), where via $\tilde{p}=\log[-\text{logit}(\hat{p})]+\epsilon$ where $\epsilon\sim$r$*$N(0,$\sigma_1{}^2$)+(1$-$r)$*$N(0, $\sigma_2{}^2$) and r=1 if $\hat{p}>10^{-6}$and r=0 if $\hat{p} \leq 10^{-6}$.

2) Perform the selection procedures on $\tilde{X}$ and $\tilde{Y}$, , and obtain $\tilde{p}_X$ and $\tilde{p}_Y$.

3) Use the grid searching approach to test each combinations of 4 sigma1 (0.05 to 0.11, by 0.02) and 19 sigma2 (0.01 to 0.19, by 0.01).

    a. $diff_{observed}$ is the difference between the observed mean of mouse X and observed mean of mouse Y at injected (Week0), i.e., $diff_{observed}$ = mean($\hat{p}_{X_{injecteed}}$) $-$ mean($\hat{p}_{Y_{injecteed}}$)

    b. $diff_{simulated}$ is the difference between the simulated mean of mouse X and simulated mean of mouse Y at injected (Week0), i.e., $diff_{observed}$ = mean($\tilde{p}_{X_{injecteed}}$) $-$ mean($\tilde{p}_{Y_{injecteed}}$)

    c. $d_{calibrate} = diff_{observed}$ - $diff_{simulated}$

4) Select candidates for combinations of sigma1s and sigma2s that have closest observed difference and simulated difference at Injected (Week0), i.e., $d_{calibrate}$ is closed to 0.

5) Compare the distributions of simulated data and observed data at each time point and finally decide the combination of sigma's as sigma1=0.07 and sigma2=0.11.
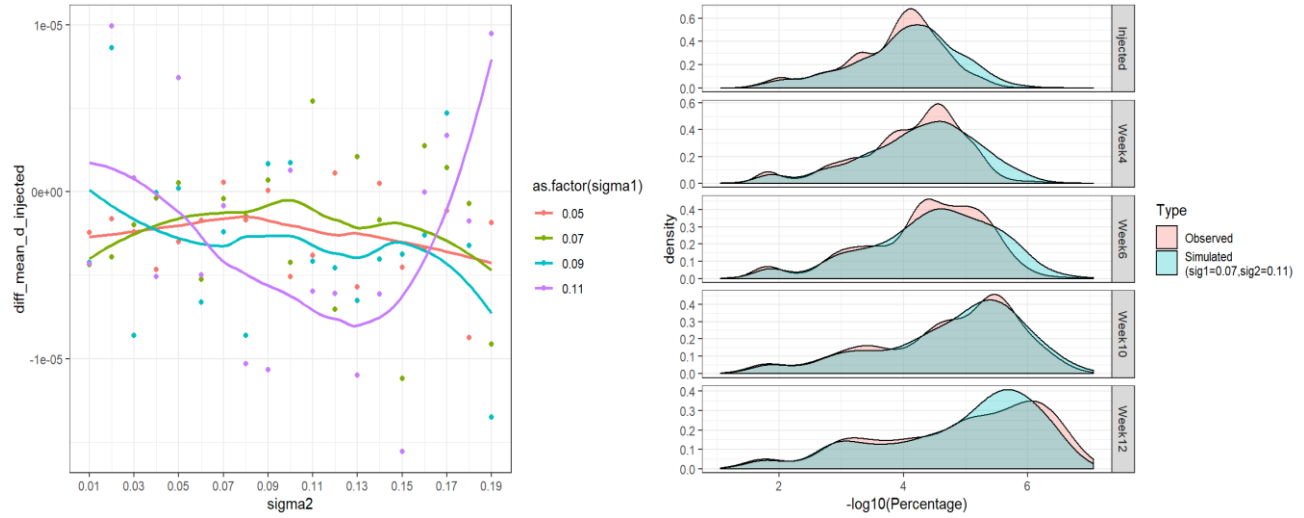
Figure 186. Sigma selection and simulated data. Left graph shows the different sigmas and their relative distribution across the difference of the means between the two injected vials of barcoded cells. The right graph shows the simulated data is well overlapped with the observed data based on selected sigma.

Perform bootstrap resampling approach 10,000 times at each time point to test Statistic1 and Statistic2 by using sigma1=0.07 and sigma2=0.11

Null hypothesis: Two samples are similar. Reject the null hypothesis if $\Pr(\tilde{t}_{simulated} < \tilde{t}_{observed})$<0.05 for T_1 or $\Pr(\tilde{t}_{simulated} > \tilde{t}_{observed})$<0.05 for T_2.

1) T_1: Pearson correlation coefficient between p_A and p_B.

    a. If T_1 is close to 1, then A and B is similar.

    b. If $\Pr(\tilde{t}_{simulated} < \tilde{t}_{observed})$<0.05, reject the null hypothesis. Two samples are not similar.

    c. If $\Pr(\tilde{t}_{simulated} < \tilde{t}_{observed})$>0.05, fail to reject the null hypothesis. We don't have enough evidence to say that the two samples are not similar.

2) T_2: Euclidean distance between p_A and p_B , i.e., $T_2 = ||p_A - p_B||_2$

a. $d_{A,B} = \sqrt{\sum_{i=1}^{n=A \cup B}(p_A - p_B)^2}$.

b. If T_2 is close to 0, A and B is similar.

c. If $Pr(\tilde{t}_{simulated} > \tilde{t}_{observed})$<0.05, reject the null hypothesis. Two samples are not similar.

d. If $Pr(\tilde{t}_{simulated} > \tilde{t}_{observed})$>0.05, fail to reject the null hypothesis. We don't have enough evidence to say that the two samples are not similar.
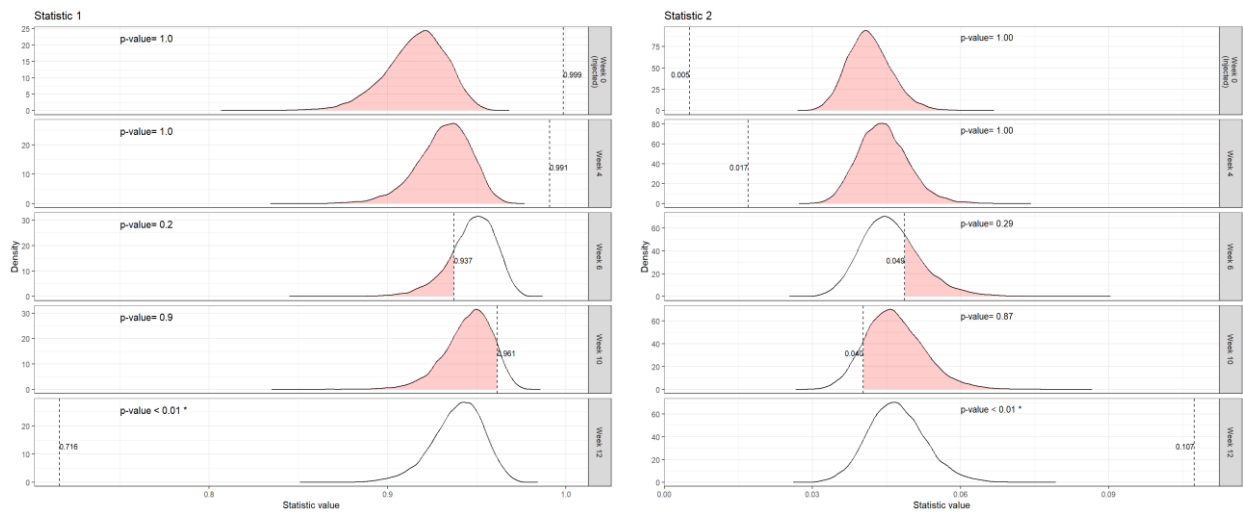
Figure 187. Simulated statistics of similarity. Left graph shows the observed value (dotted lines) and the "true" distribution of statistics 1 (Pearson's correlation); right graph shows the observed value (dotted lines) and the "true" distribution of statistics 2 (Euclidian distance).

Clustering

All primary tumor (reads>3, N=1013)

Difference of distance between prevalence rates for all primary tumor (reads>3, N=1013)

After selection procedures,

primary >3 at all time points, N=1013.

primary >3 + (liver/lung) >1 at all time points, N=287.

primary >3 + liver+ lung >1 at all time points, N=137.
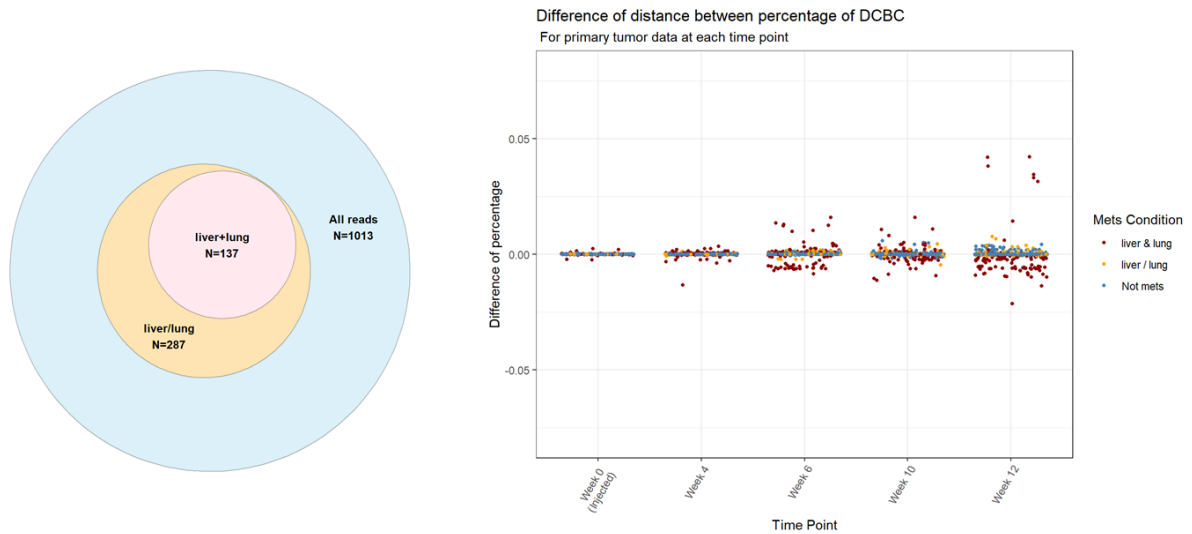
Figure 188. Distribution and counts of barcodes based on time and metastasis phenotype.

In all 1013 barcodes (reads>3), 287 barcodes will mets to either liver or lung at Week 12, count for 92.23% of all barcodes (N=2017709) and In all 287 barcodes (reads>3) that can mets, 137 barcodes will mets to both liver and lung at Week 12, count for 86.23% of all barcodes (N=2017709).

1) Blue dots (N=1013-287=726): primary tumor cannot mets

2) Yellow and red dots (N=287): primary tumor can mets to either liver or lung

3) Red dots (N=137): primary tumor can mets to both liver and lung

Classify the quartiles of clonal relative fitness based on percentage ranking at Week4 and at Week12 for all primary tumor (reads>3, N=1013)

Take average of percentage at each time point and quantify into four quartiles: [0-25%), [25-50%), [50-75%) and [75-100%]

Figure 189. Quartiles of the primary tumor

Change between Week4 and Week12 for all primary tumor (reads>3, N=1013)

Figure 190. Analysis of barcodes at week12 vs. week 4. The red lines indicate lineages that has an overall gain in fitness. The green lines indicate lineages that has an overall loss in fitness.

Clustering for all primary tumor (reads>3, N=1013)

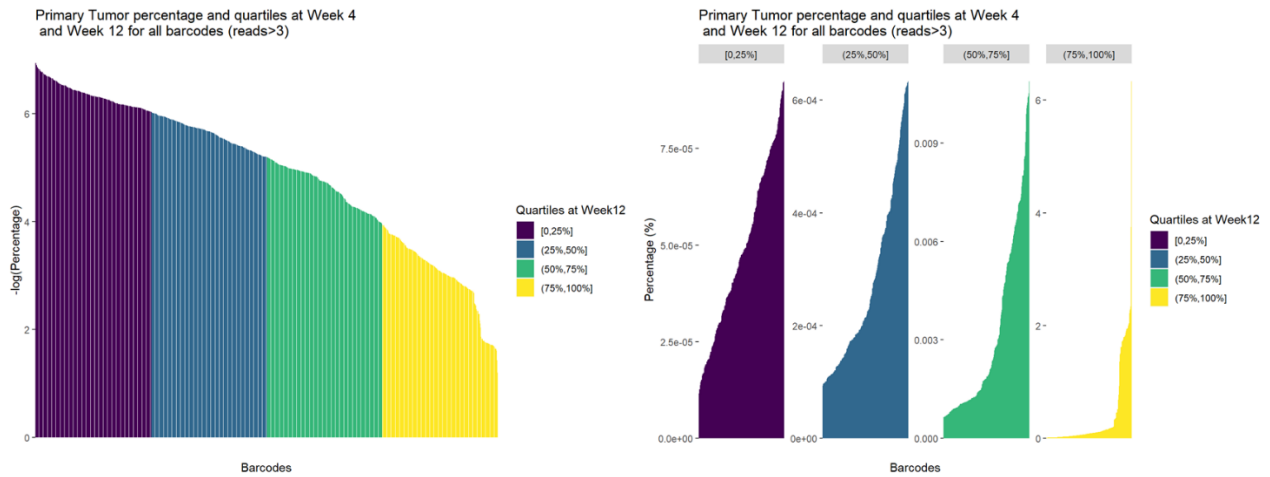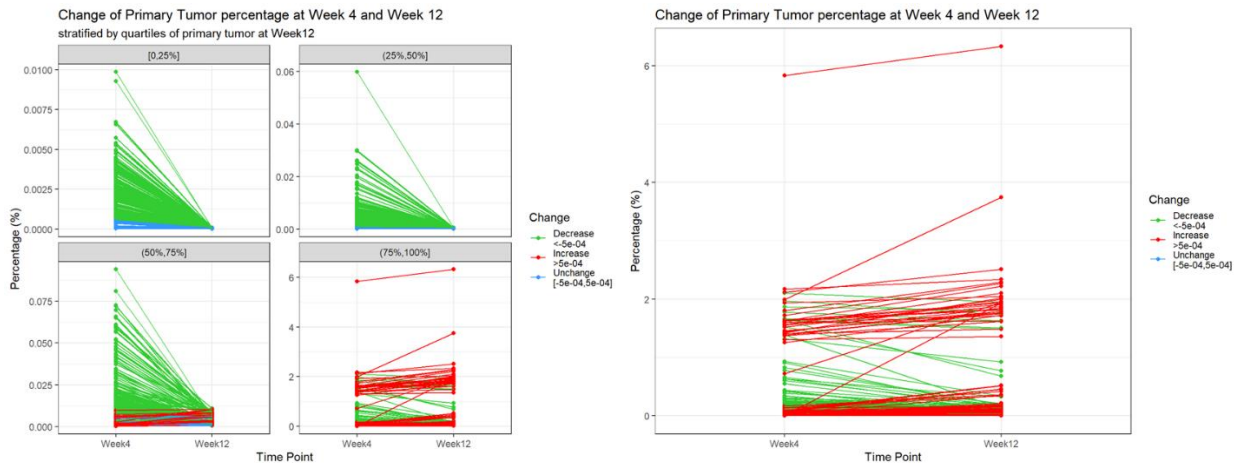1) CLARA (Clustering Large Applications, (Kaufman and Rousseeuw 1990)) [223] is an extension to k-medoids (PAM) methods to deal with data containing a large number of objects. This is achieved using the sampling approach.

   a. The k-medoids algorithm is a clustering approach related to k-means clustering for partitioning a data set into k groups or clusters. In k-medoids clustering, each cluster is represented by one of the data point in the cluster. These points are named cluster medoids.

   b. The term medoid refers to an object within a cluster for which average dissimilarity between it and all the other the members of the cluster is minimal. These objects (one per cluster) can be considered as a representative example of the members of that cluster so it is a robust alternative to k-means clustering. This means that, the algorithm is less sensitive to noise and outliers, compared to k-means, because it uses medoids as cluster centers instead of means (used in k-means).

   c. Instead of finding medoids for the entire data set, CLARA considers a small sample of the data with fixed size (sample size) and applies the PAM algorithm to generate an optimal set of medoids for the sample. The quality of resulting medoids is measured by the average dissimilarity between every object in the entire data set and the medoid of its cluster, defined as the cost function.

   d. CLARA repeats the sampling and clustering processes a pre-specified number of times in order to minimize the sampling bias. The final clustering results correspond to the set of medoids with the minimal cost.

2) In order to cluster based on the shape, rather than the average value of each barcode, we calculate the slopes between each time point first and then use CLARA to cluster "slopes" of barcodes. In addition, the use of slope indicates there is a progressive time dimension that is being analyzed.

3) We use total within sum of square (WSS) to get optimal number of clusters with good homogeneity.

    a. With-in-Sum-of-Squares (WSS): WSS is the total distance of data points from their respective cluster centroids.

Figure 191. Clustering of all lineages in the primary tumor. Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.

Cluster [0,25%] at Week 12 for all primary tumor (reads>3, N=254)
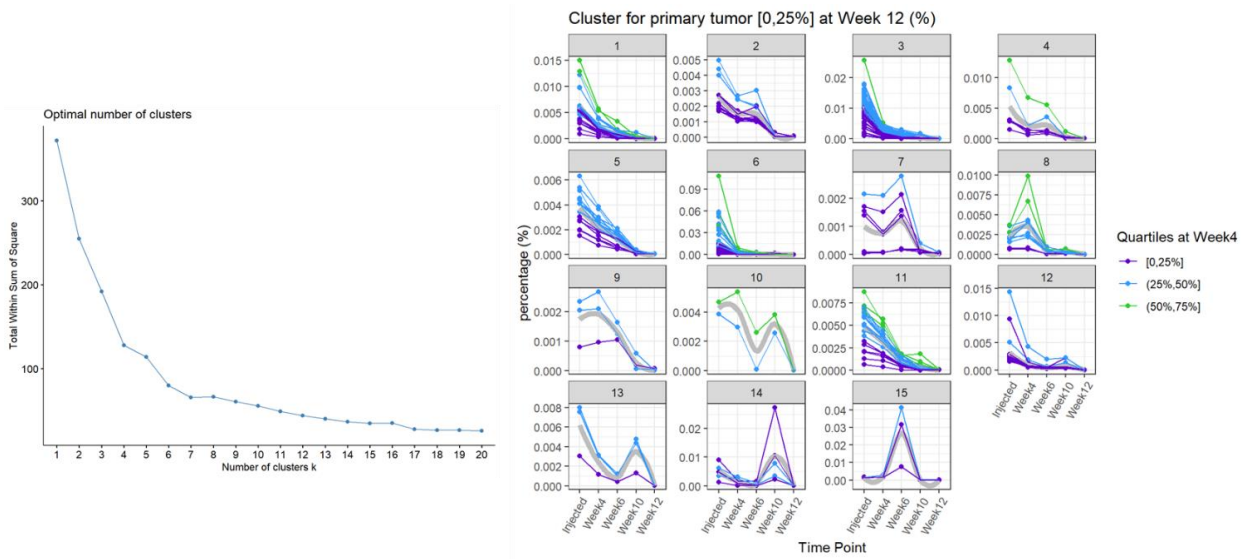
1) CLARA

Figure 192. Clustering of all lineages in the primary tumor by quartiles. Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.

Cluster [25,50%) at Week 12 for all primary tumor (reads>3, N=254)
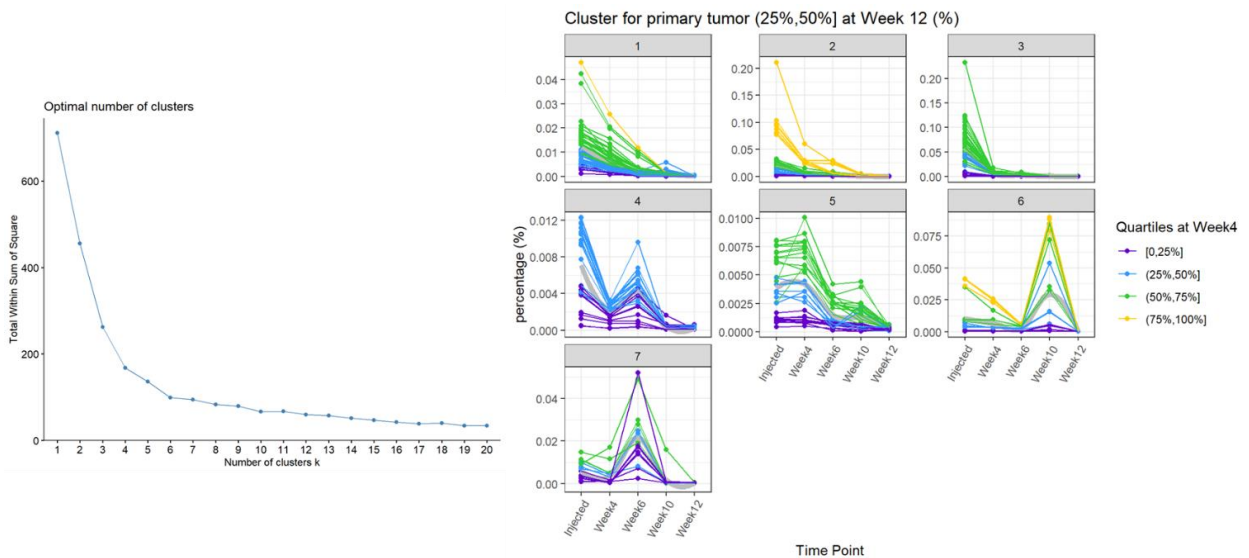
1) CLARA

Figure 193. Clustering of all lineages in the primary tumor by quartiles.Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.

Cluster [50,75%) at Week 12 for all primary tumor (reads>3, N=254)
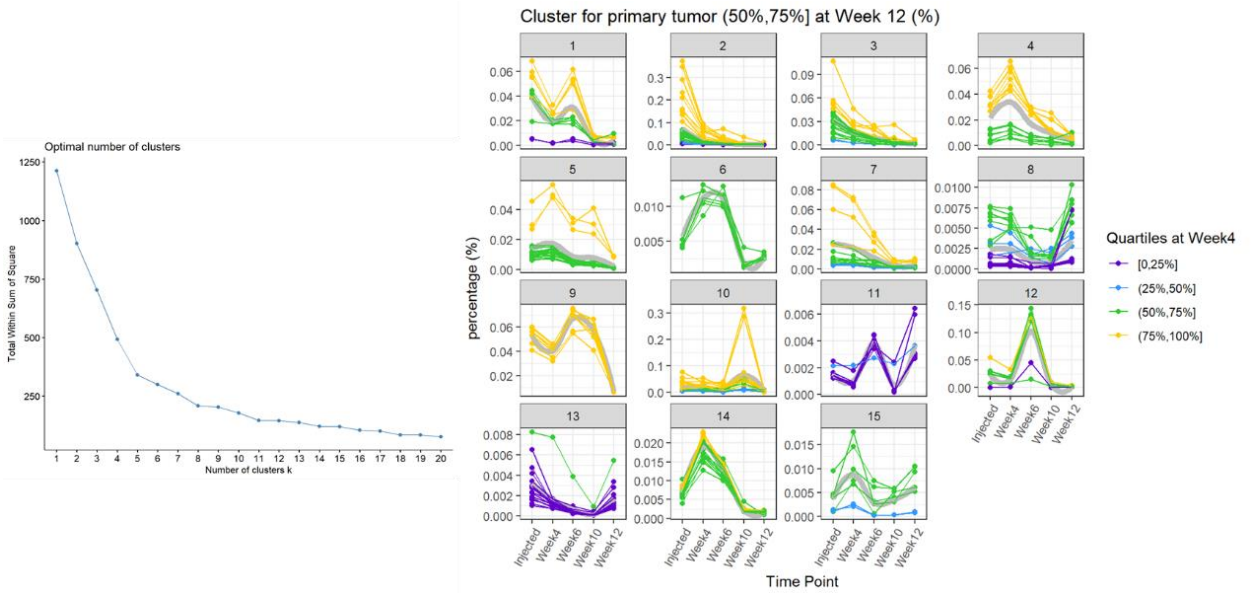
1) CLARA

Figure 194. Clustering of all lineages in the primary tumor by quartiles.Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.
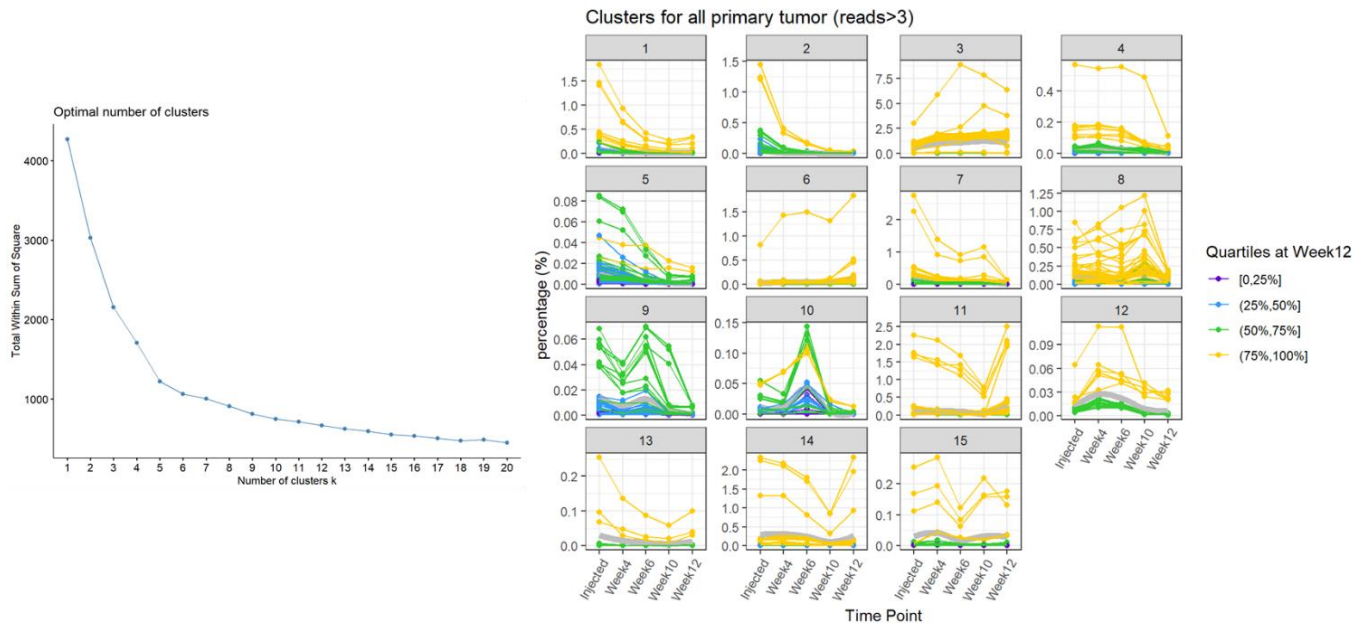
1) CLARA
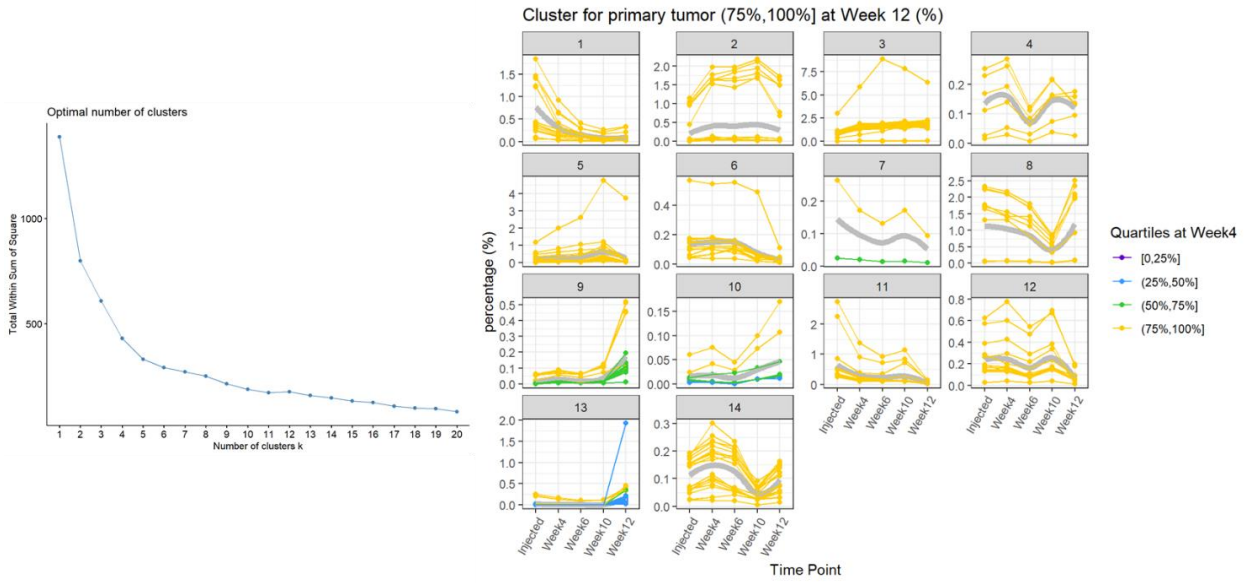
Figure 195. Clustering of all lineages in the primary tumor by quartiles. Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.

Primary tumor that can metastasize to either liver or lung (reads>3, N=287)

Classify the quartiles of clonal relative fitness based on percentage ranking at Week4 and at Week12 for primary that can mets to either liver or lung (reads>3, N=287)

Take average of percentage at each time point and quantify into four quartiles: [0-25%), [25-50%), [50-75%) and [75-100%]
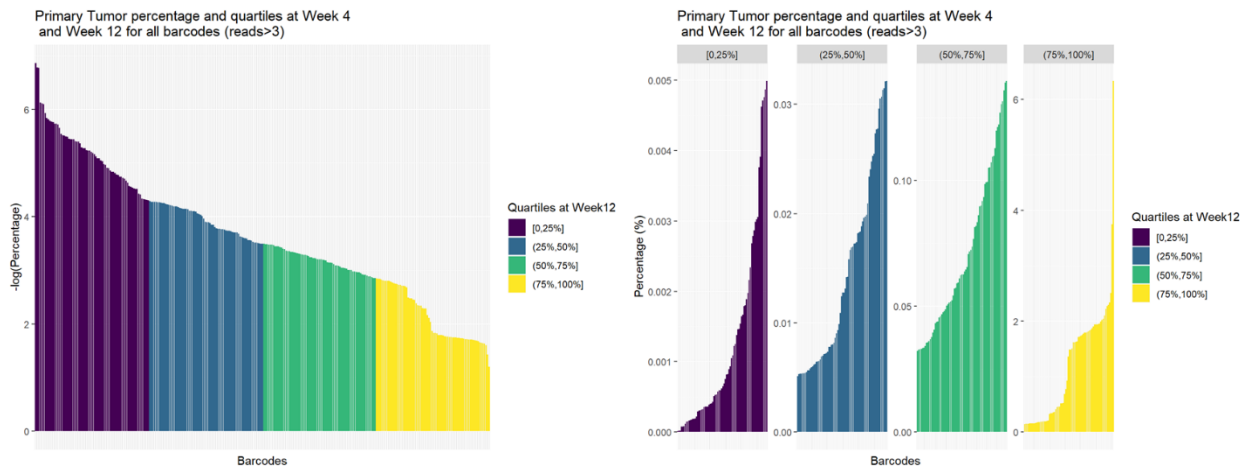
Figure 196. Quartile distribution of barcodes in the primary tumor that can metastasize to either liver or lung.

Change between Week4 and Week12 for primary that can mets to either liver or lung (reads>3, N=287)
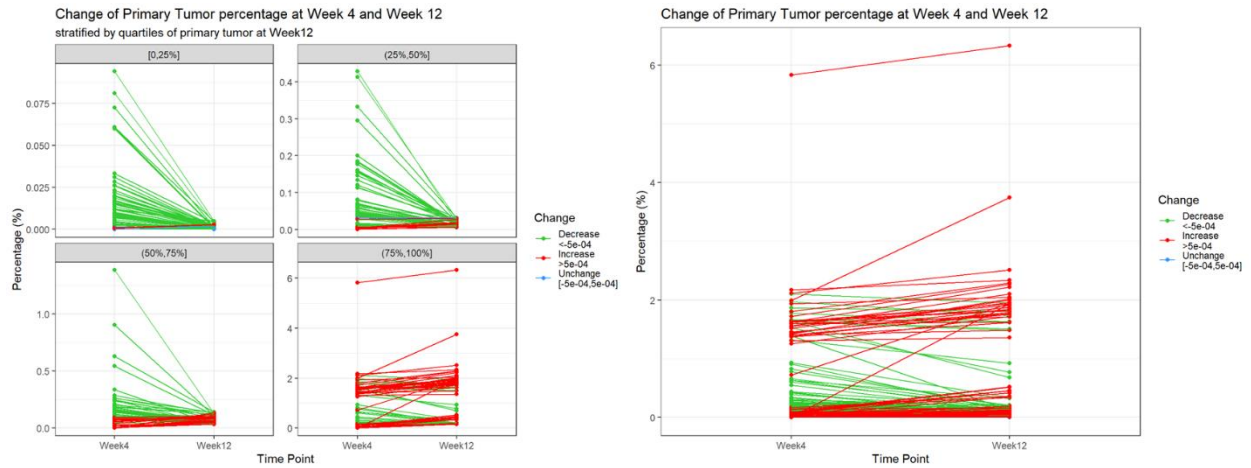
Figure 197. Analysis of barcodes that could metastasize to lung or liver at week12 vs. week 4. The red lines indicate lineages that has an overall gain in fitness. The green lines indicate lineages that has an overall loss in fitness.

Clustering for primary that can metastasize to either liver or lung (reads>3, N=287)

1) CLARA
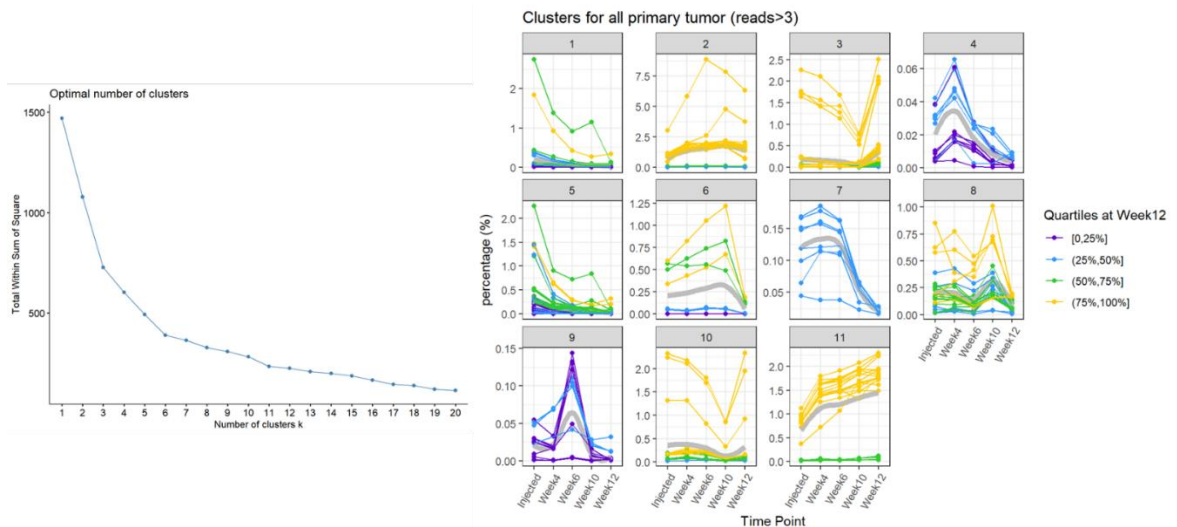
Figure 198. Clustering of all lineages in the primary tumor that could metastasize to lung or liver by quartiles. Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.
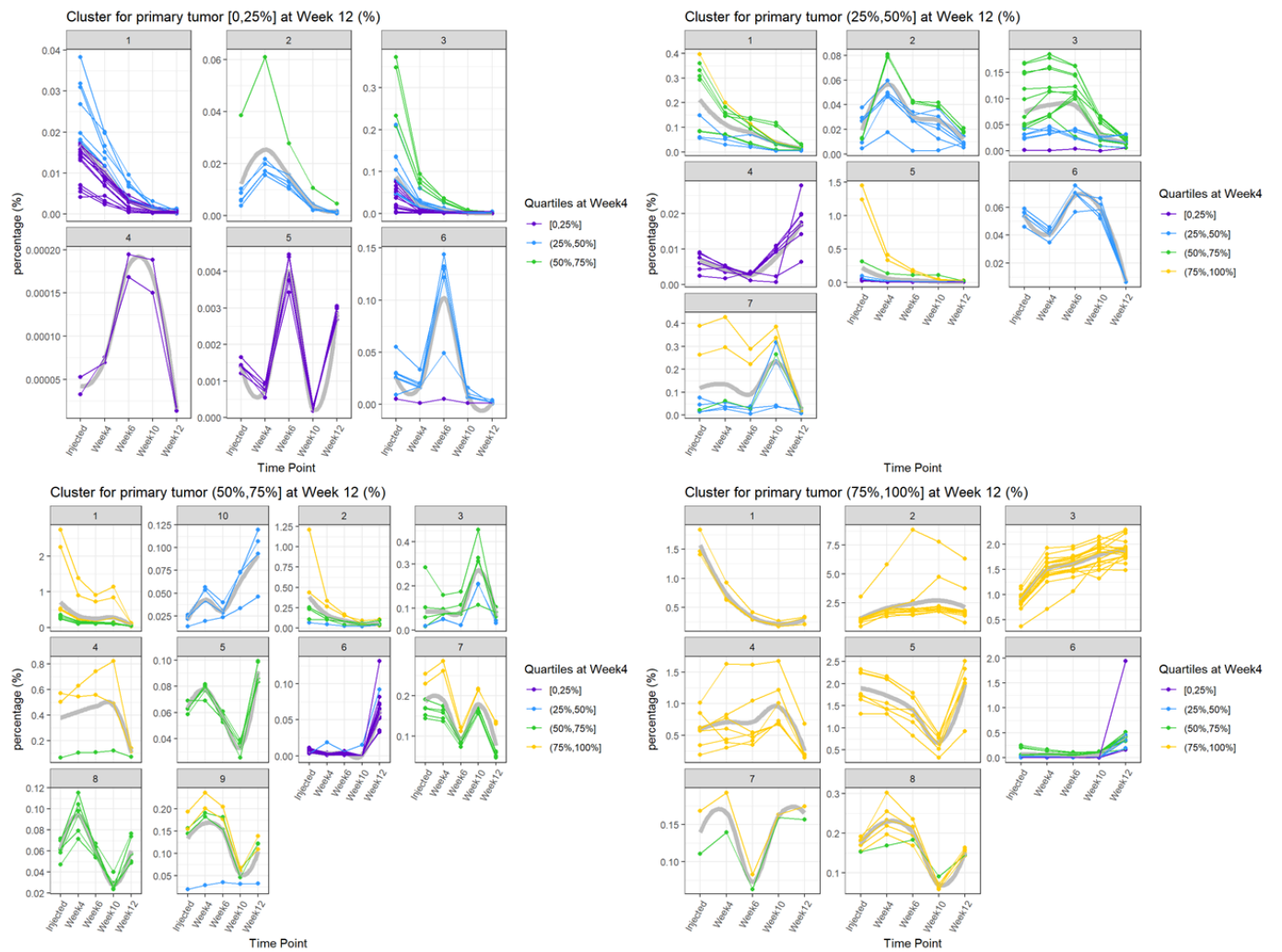
Figure 199. Clustering of all lineages in the primary tumor that could metastasize to lung or liver by quartiles.

Primary tumor that can mets to both liver and lung (reads>3, N=137)

Classify the quartiles of clonal relative fitness based on percentage ranking at Week4 and at Week12 for primary that can mets to both liver and lung (reads>3, N=137)

Take average of percentage at each time point and quantify into four quartiles: [0-25%), [25-50%), [50-75%) and [75-100%]
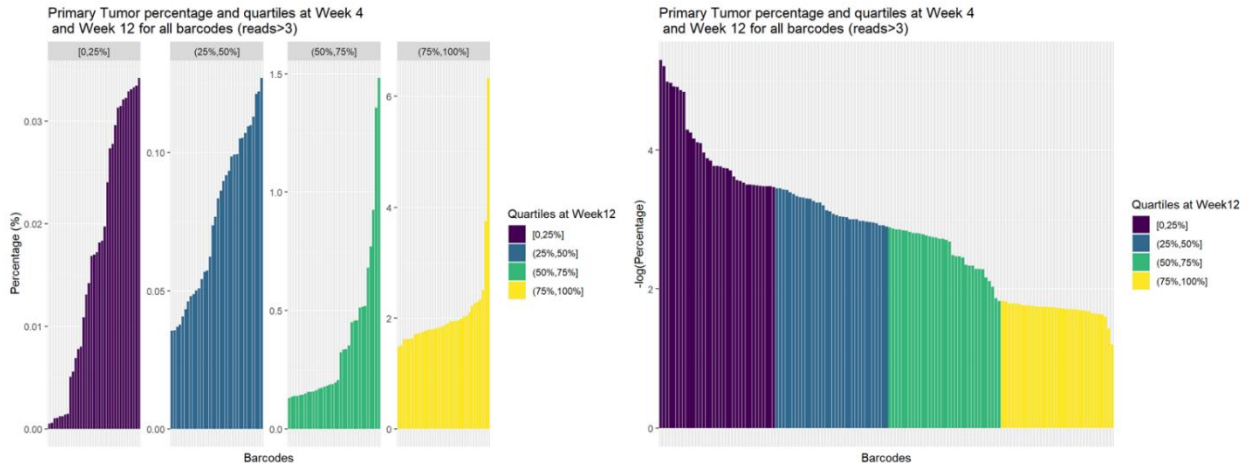
Figure 200. Quartile of barcodes in the primary tumor that could metastasize to lung and liver.

Change between Week4 and Week12 for primary that can mets to both liver and lung (reads>3, N=137)
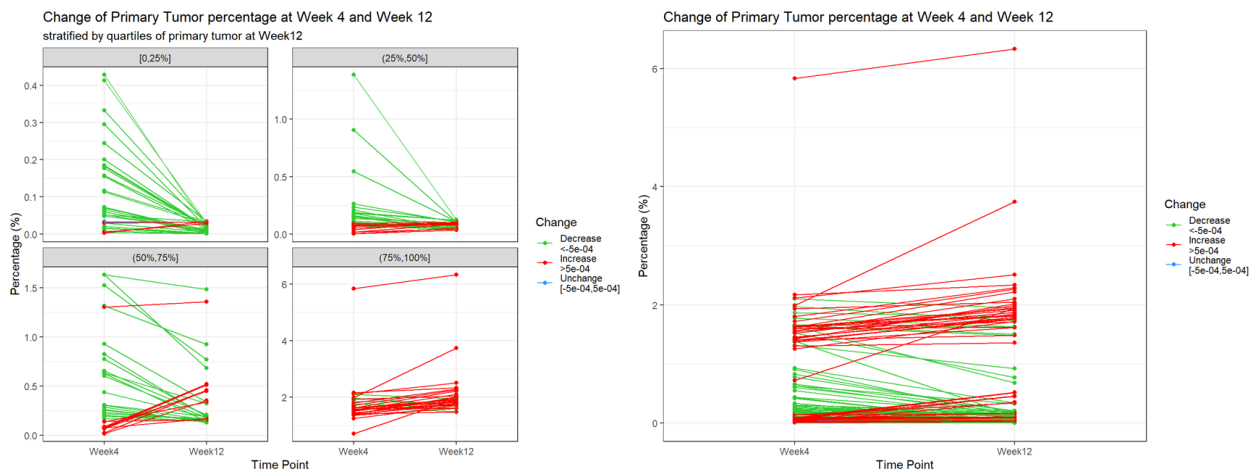
Figure 201. Analysis of barcodes that could metastasize to lung and liver at week12 vs. week 4.The red lines indicate lineages that has an overall gain in fitness. The green lines indicate lineages that has an overall loss in fitness.

<u>Clustering for primary that can mets to both liver and lung (reads>3, N=137)</u>

1) CLARA
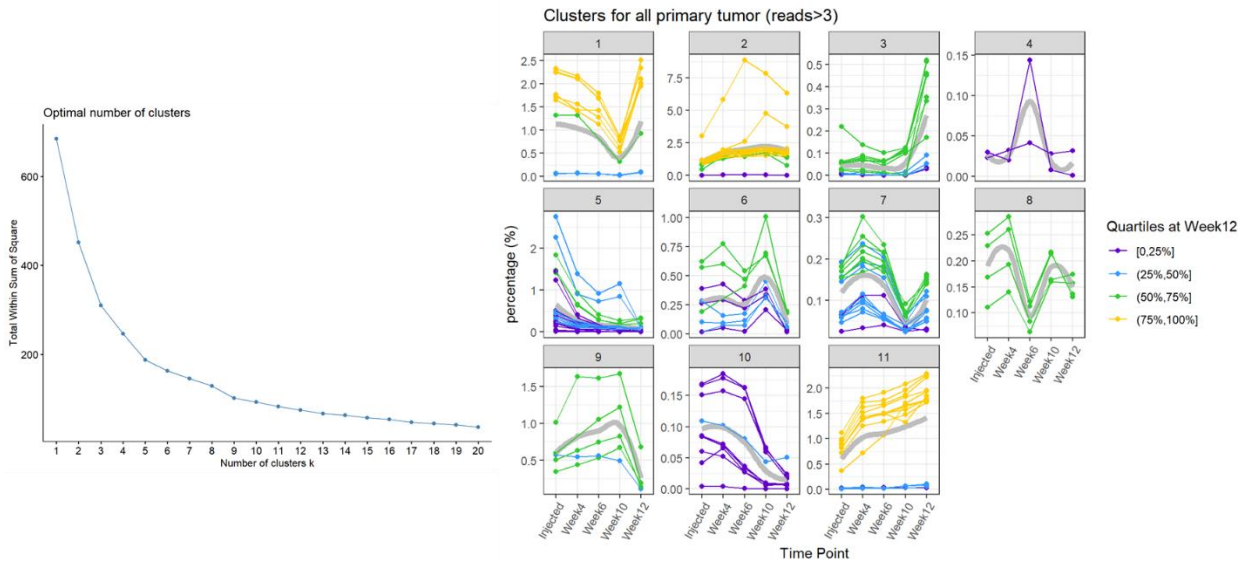
Figure 202. Clustering of all lineages in the primary tumor that could metastasize to lung and liver by quartiles. Left shows the optimal cluster number. Right shows the clusters of lineages and their dynamics.
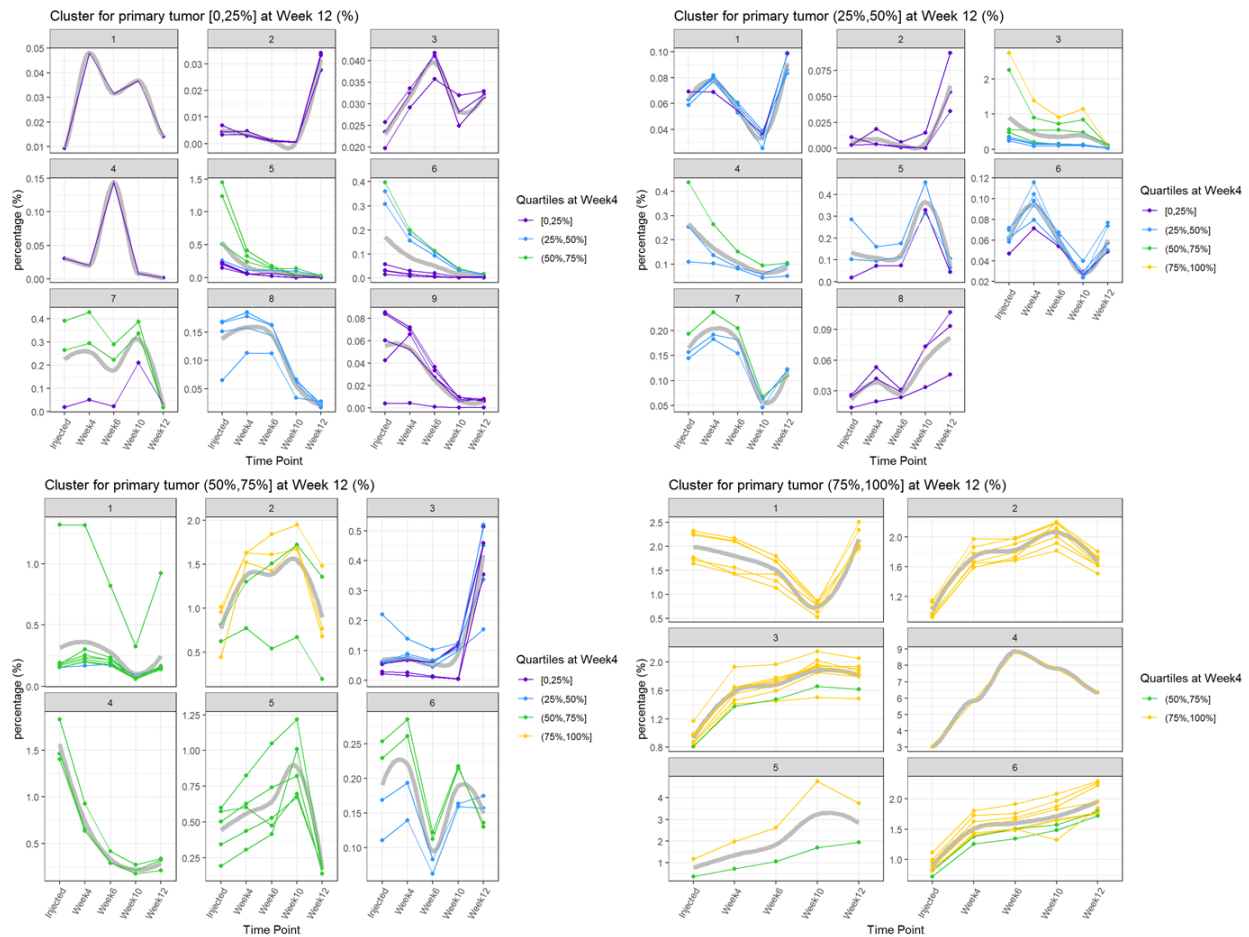
Figure 203. Clustering of all lineages in the primary tumor that could metastasize to lung and liver by quartiles.

All Metastasis lineages

1) In all 287 barcodes that will mets to either liver or lung, 236 barcodes will mets to liver at Week 12, count for 99.61% of all barcodes (N=2017709).

2) In all 287 barcodes that will mets to either liver or lung, 188 barcodes will mets to lung at Week 12, count for 94.07% of all barcodes (N=2017709).

3) In all 287 barcodes (reads>3) that can mets, 137 barcodes will mets to both liver and lung at Week 12, count for 86.23% of all barcodes (N=2017709).

Change between Week4 and Week12 for Mets liver (reads>1, N=236)
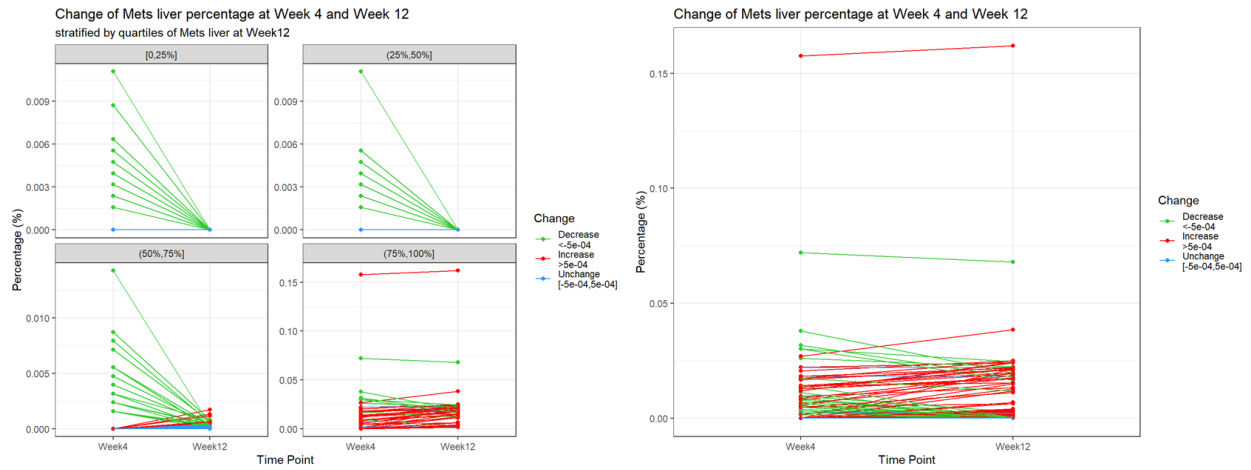
Figure 204. Analysis of barcodes that metastasize to liver between week 12 and week 4.

 Clustering for primary that can mets to both liver and lung (reads>1, N=236)

1) CLARA
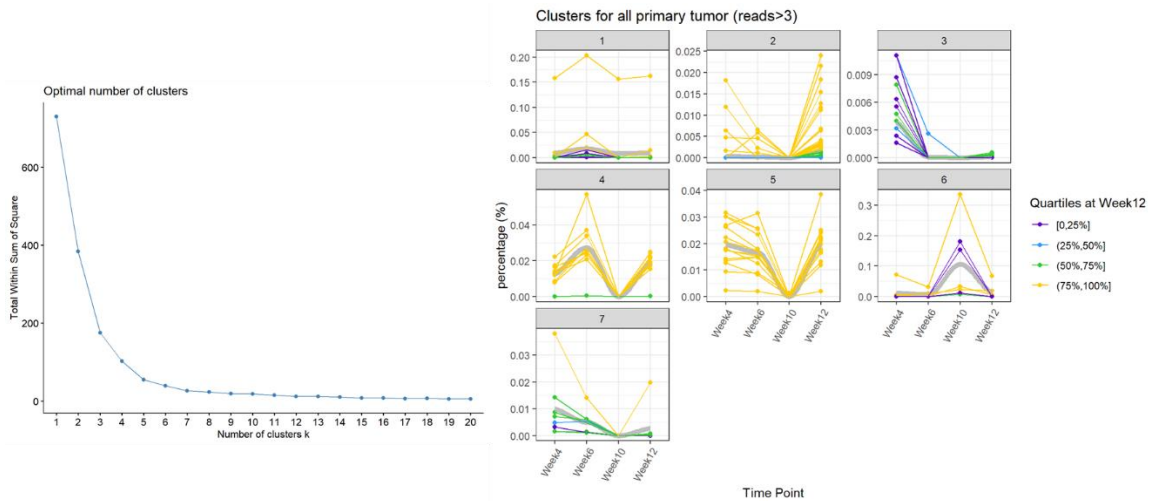
Figure 205. Clustering of clonal lineages in the liver. Left is the optimum of clusters; right is the clonal dynamics of metastatic clones in the liver over time.
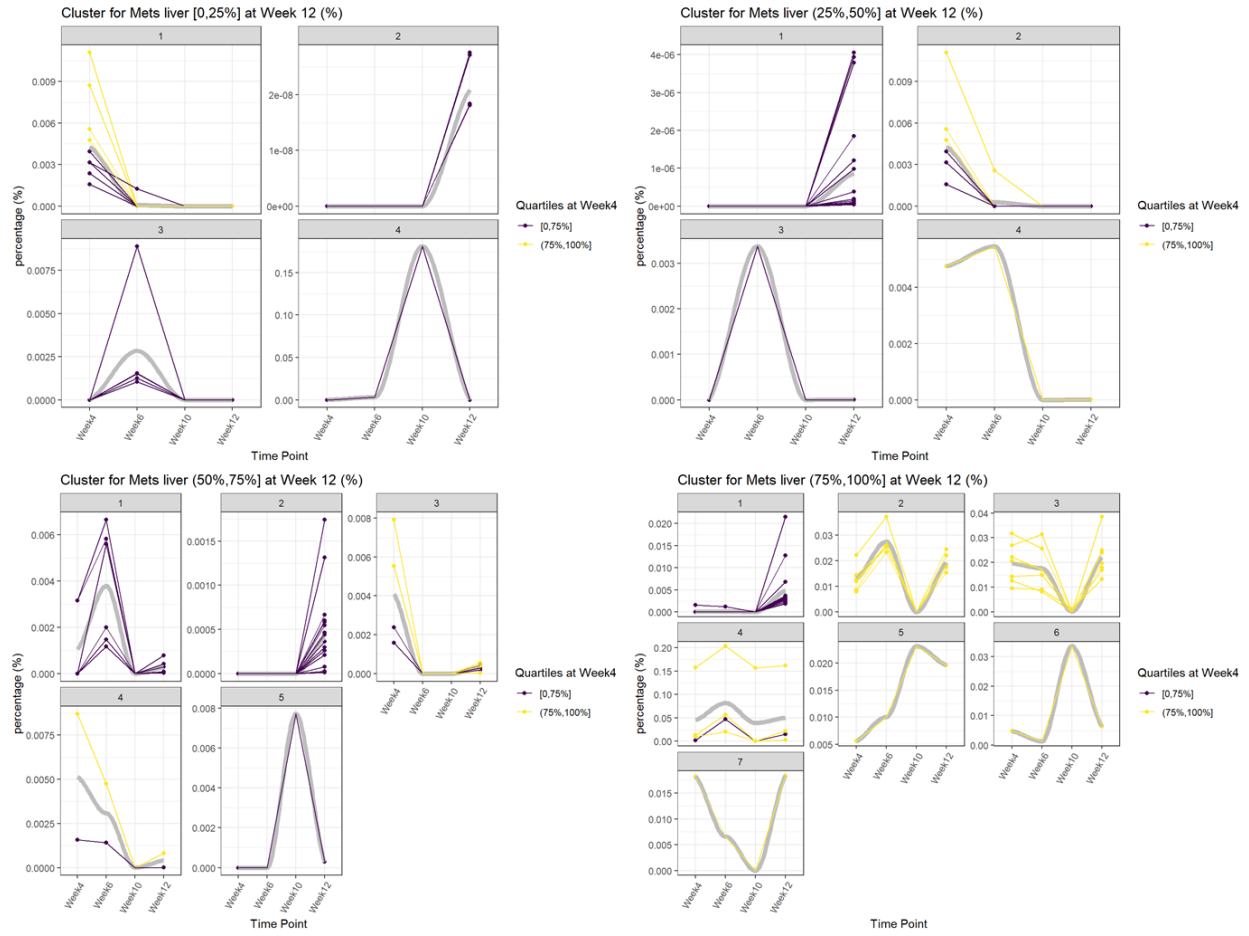
Figure 206. Clustering of clonal lineages in the liver. The clonal dynamics of metastatic clones in the liver over time.

Mets - lung (reads>1, N=188)

Change between Week4 and Week12 for Mets lung (reads>1, N=188)
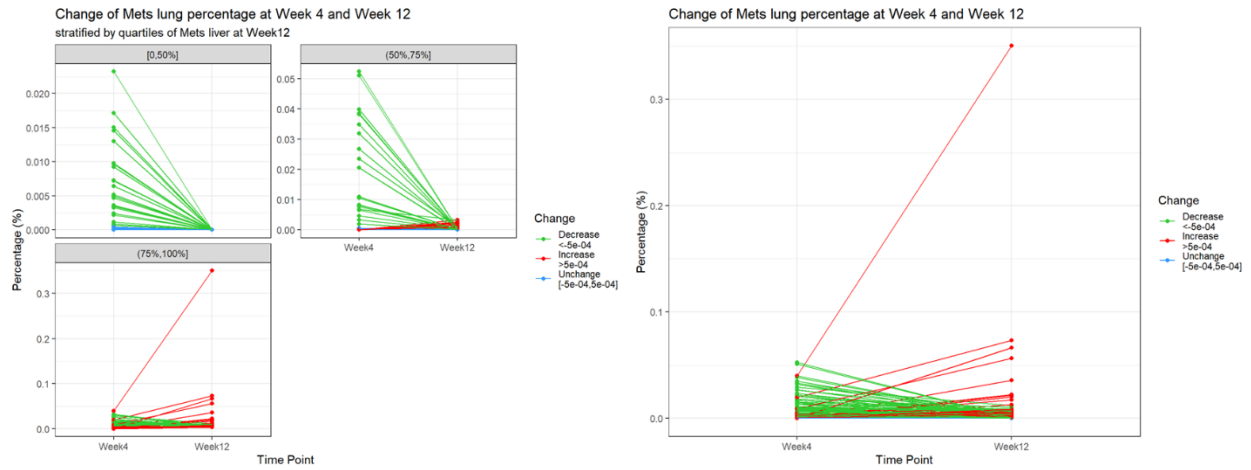
Figure 207. Analysis of barcodes that metastasize to lung between week 12 and week 4.

Clustering for all Mets lung (reads>1, N=188)

1) CLARA
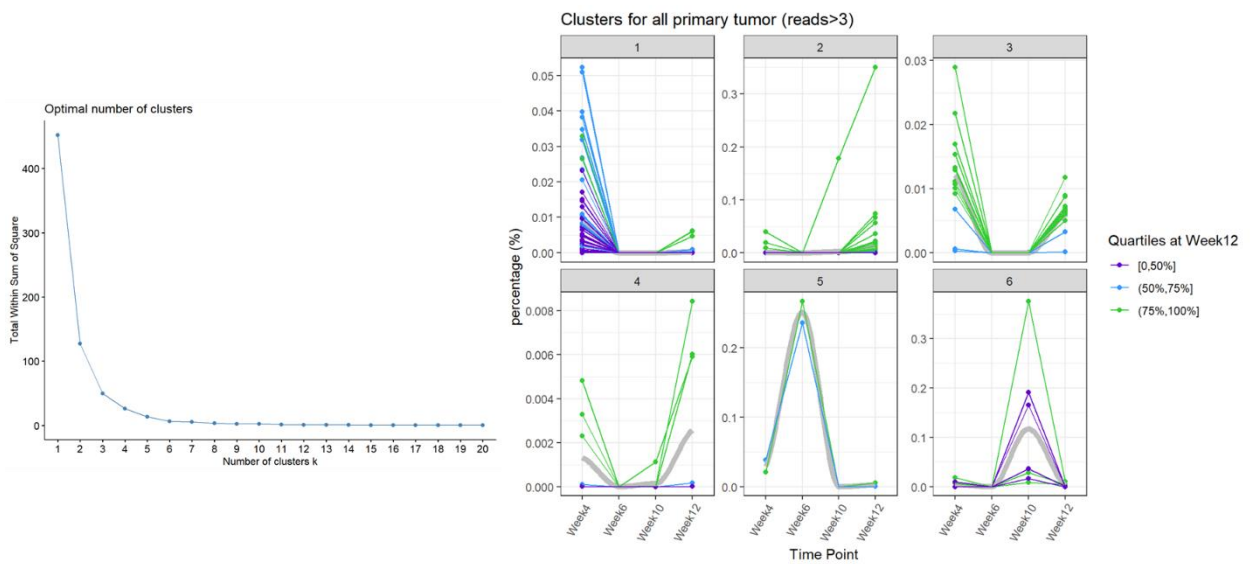
Figure 208. Clustering of clonal lineages in the lung.Left is the optimum of clusters; right is the clonal dynamics of metastatic clones in the lung over time.
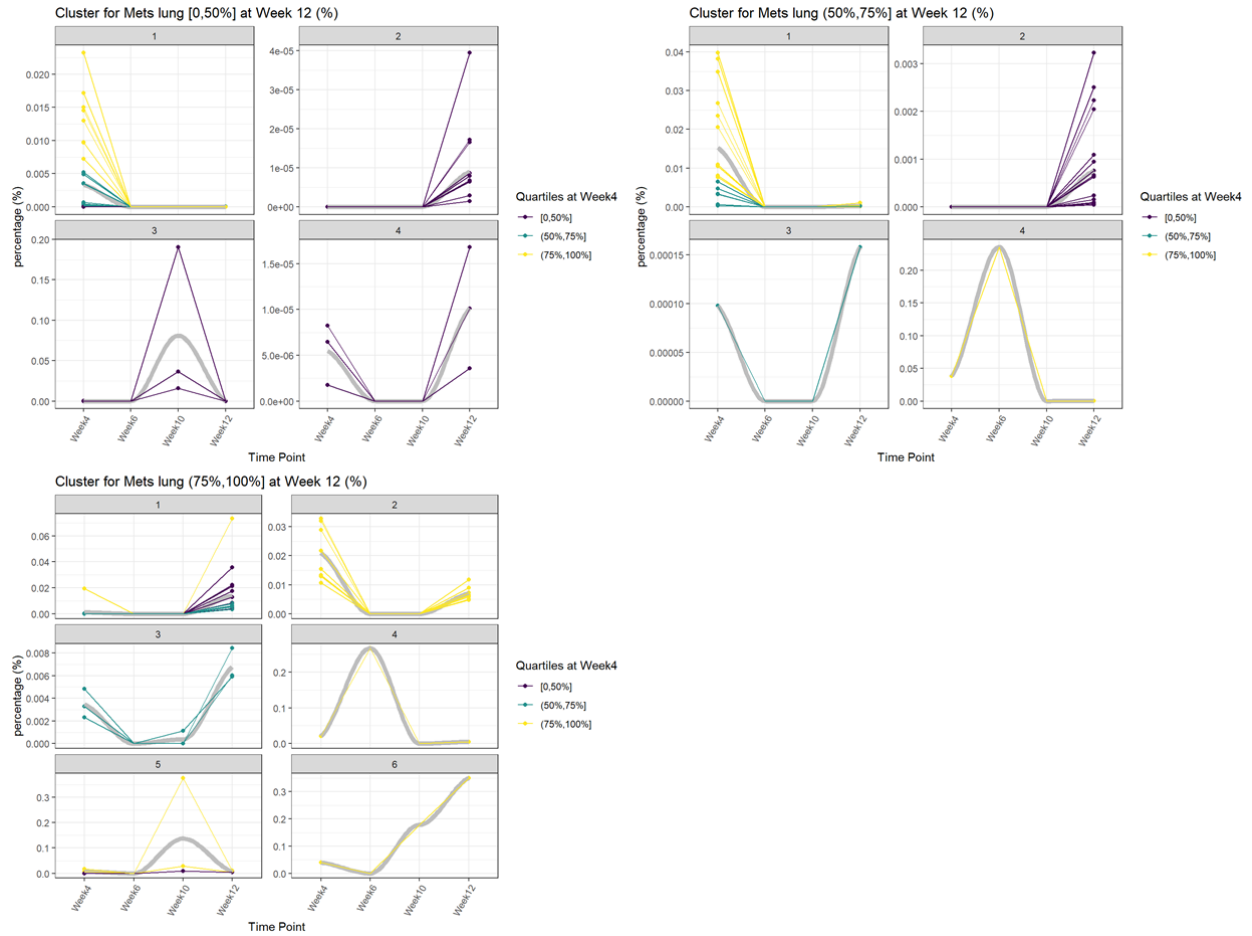
Figure 209. Clustering of clonal lineages in the lung. The clonal dynamics of metastatic clones in the lung over time. The clusters are stratified based on tumor representation at week 12.
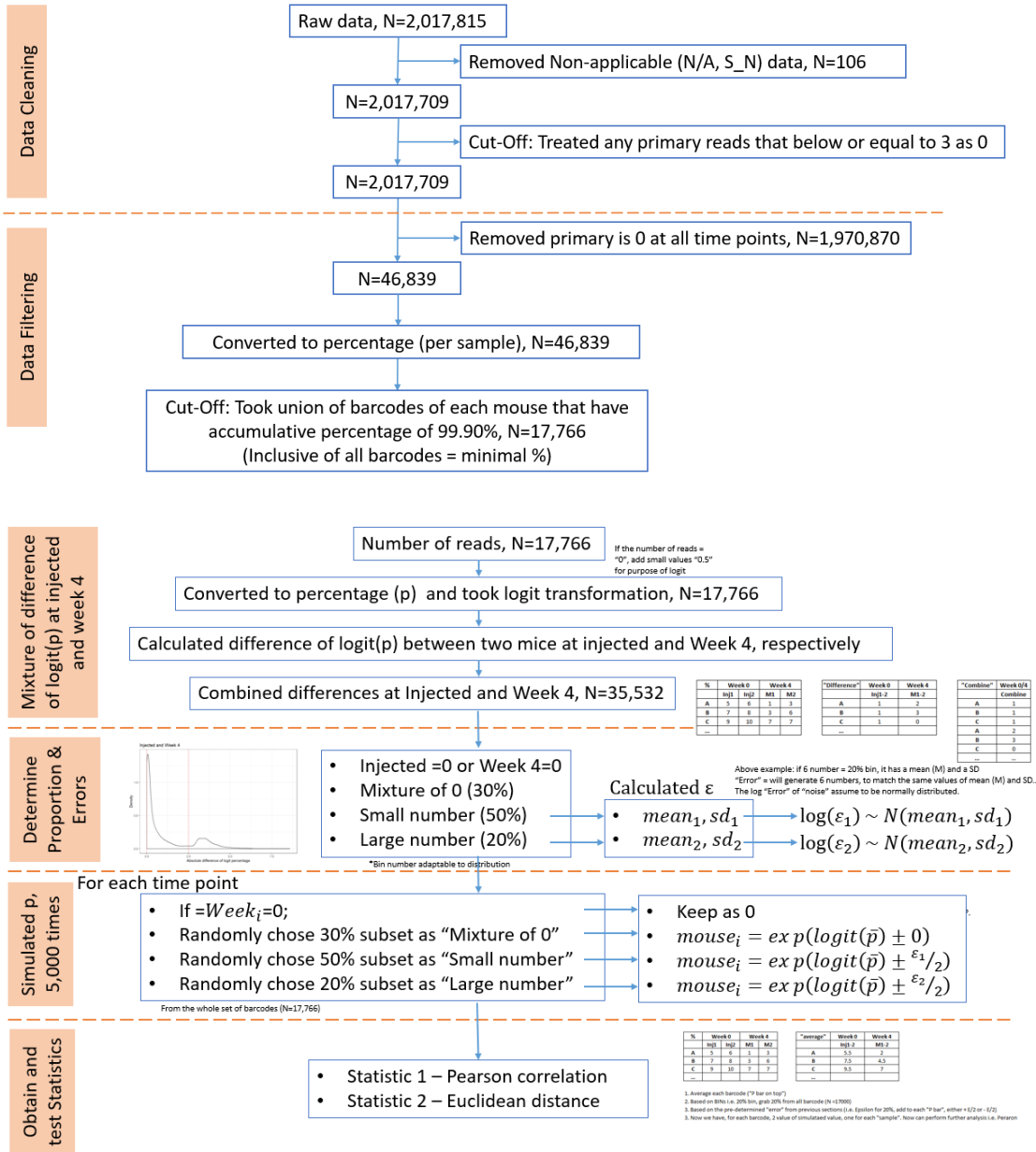
Figure 210. Flow-chart workflow of simulation. By Delia Wang.
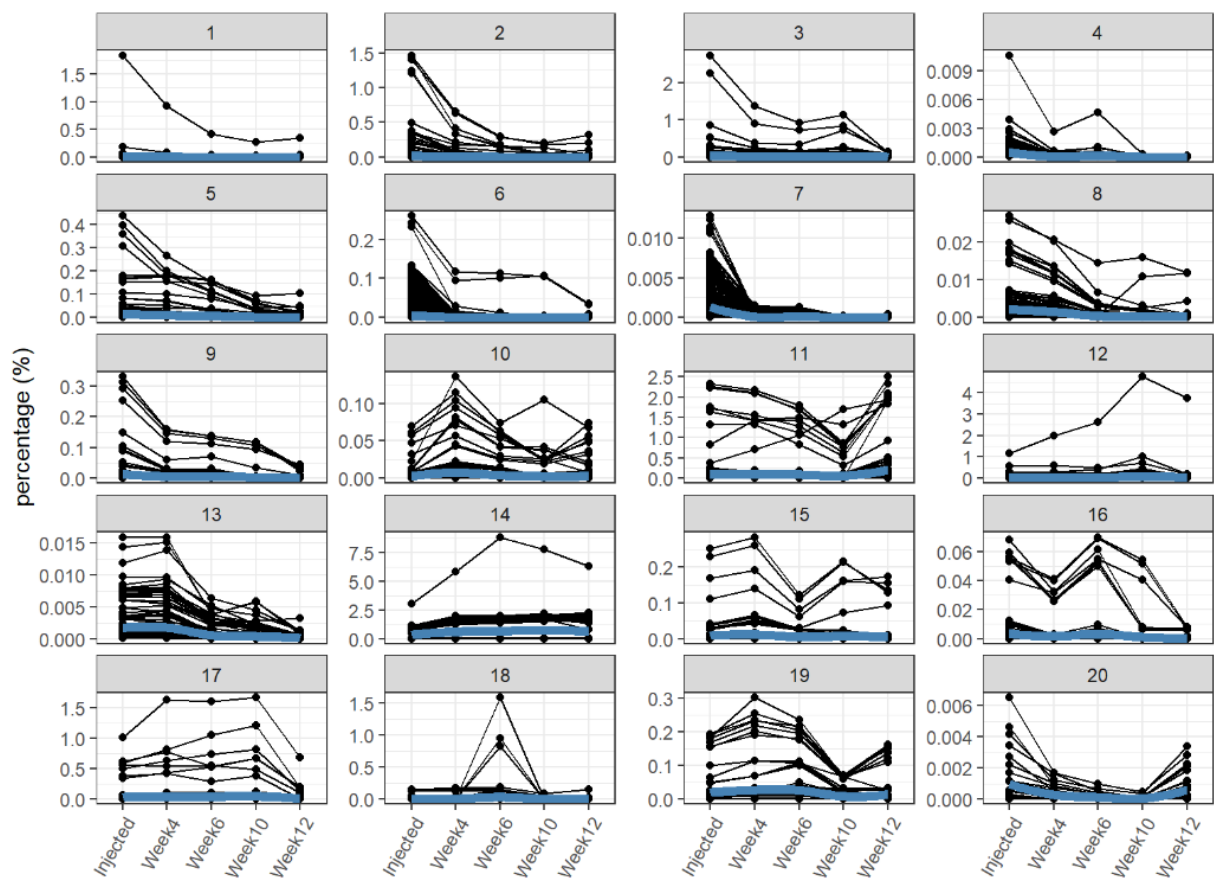
Figure 211. A comprehensive lineage dynamic in the primary tumor.
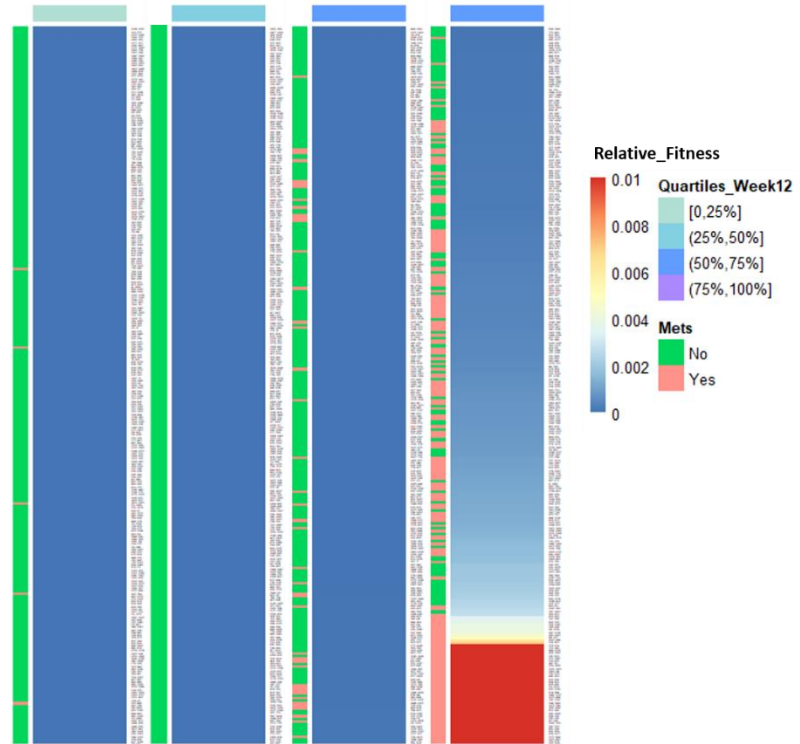
Figure 212. Heat-map for lineage primary tumor representation vs. metastasis outcome. As we can see from the heat-map, graphing the relationship of the metastasis outcome vs. lineage representation in the primary tumor (paired samples), abundance has a positive correlation with metastasis outcome; however, there are also lineage that are high represented that does not metastasize. This suggests that metastasis is not merely contribute by level of abundance.
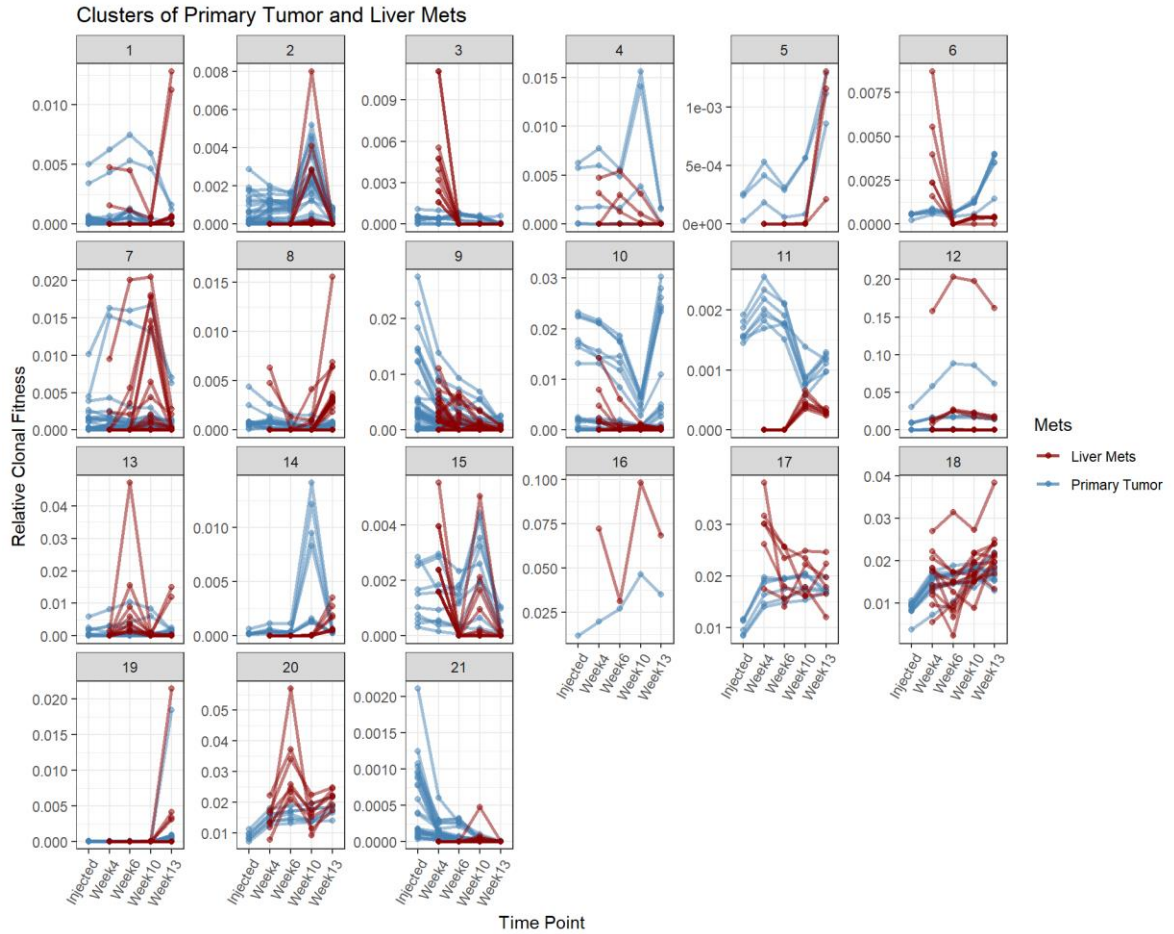
Figure 213. A comprehensive lineage dynamic in the primary tumor, as well as in the liver. Superimposed lineage dynamics in the liver vs. in the primary tumor. Y axis represents the % representation of the lineage in either liver or the primary tumor.
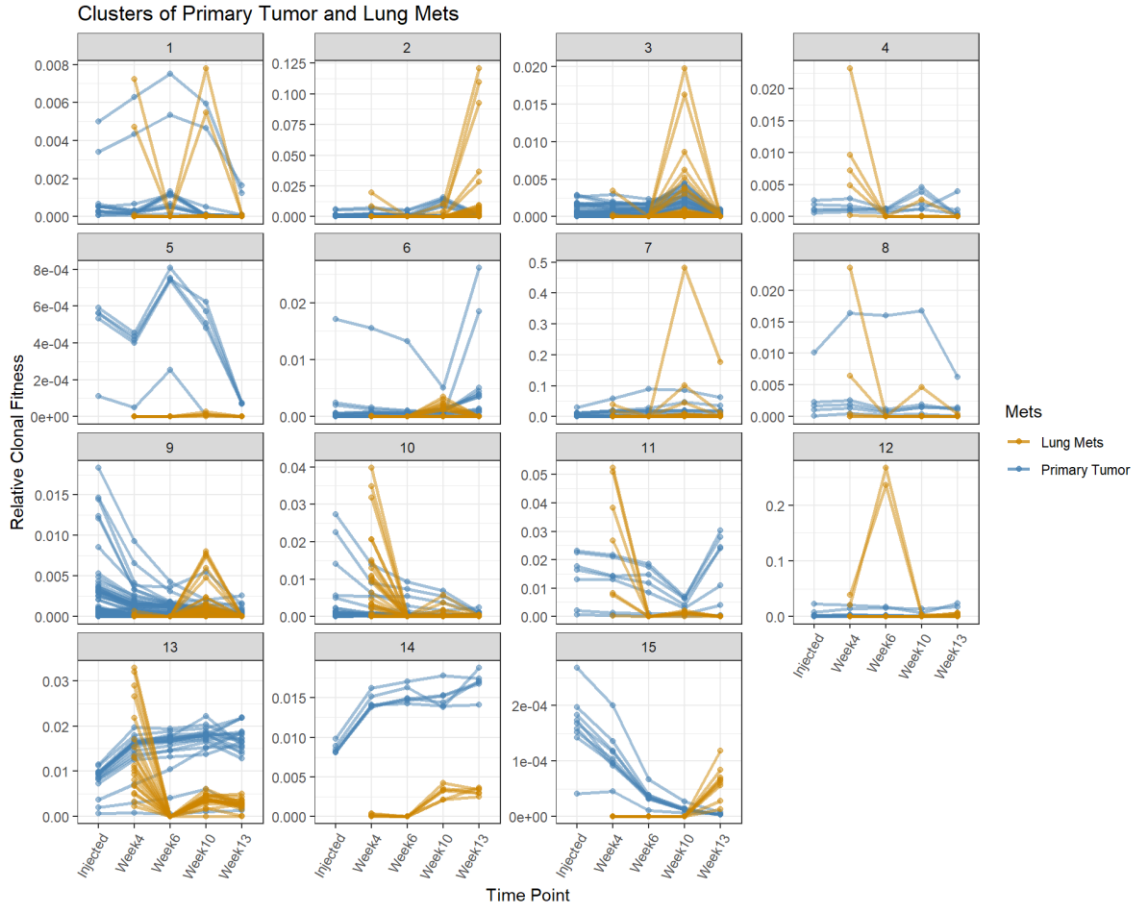
Figure 214. A comprehensive lineage dynamic in the primary tumor, as well as in the lung. Superimposed lineage dynamics in the lung vs. in the primary tumor. Y axis represents the % representation of the lineage in either lung or the primary tumor.

References

1. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

2. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Analytical Biochemistry, 242(1), 84–89. https://doi.org/10.1006/abio.1996.0432

3. Ahmadian, A., Gharizadeh, B., Gustafsson, A. C., Sterky, F., Nyrén, P., Uhlén, M., & Lundeberg, J. (2000). Single-nucleotide polymorphism analysis by pyrosequencing. Analytical Biochemistry, 280(1), 103–110. https://doi.org/10.1006/abio.2000.4493

4. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, et al. (2013) Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample. PLOS ONE 8(2): e55089. https://doi.org/10.1371/journal.pone.0055089

5. M.L. Metzker Sequencing technologies: the next generation. Nat. Rev. Genet., 11 (2010), pp. 31-46

6. Del Vecchio, F., Mastroiaco, V., Di Marco, A. et al. Next-generation sequencing: recent applications to the analysis of colorectal cancer. J Transl Med 15, 246 (2017) doi:10.1186/s12967-017-1353-y

7. Alekseyev, Y. O., Fazeli, R., Yang, S., Basran, R., Maher, T., Miller, N. S., & Remick, D. (2018). A Next-Generation Sequencing Primer—How Does It Work and What Can It Do? Academic Pathology. https://doi.org/10.1177/2374289518766521

8. Quail, M. A., Smith, M., Coupland, P., Otto, T. D. , Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics, 13(1), p. 341. doi:10.1186/1471-2164-13-341

9. Muers, M. Getting Moore from DNA sequencing. Nat Rev Genet 12, 586 (2011) doi:10.1038/nrg3059

10. Garrido-Cardenas JA, Garcia-Maroto F, Alvarez-Bermejo FJ, ManzanoAugugliaro F. DNA sequencing sensors: an overview. Sensors. (2017) 17:588. doi: 10.3390/s17030588 94.

11. Muir P., Li S., Lou S., Wang D., Spakowicz D., et.al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 2016; 17:53.

12. Kennedy, S. R., Schmitt, M. W., Fox, E. J., Kohrn, B. F., Salk, J. J., Ahn, E. H., Prindle, M. J., Kuong, K. J., Shen, J. C., Risques, R. A., & Loeb, L. A. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. Nature protocols, 9(11), 2586–2606.

13. Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner (2010) Target-enrichment strategies for next-generation sequencing. Nat. Meth. 7: 111–118.

14. Beadling, C., Neff, T. L., Heinrich, M. C., Rhodes, K., Thornton, M., Leamon, J., et al. (2013). Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping. J. Mol. Diagn. 15, 171–176. doi: 10.1016/j.jmoldx.2012.09.003

15. Meuzelaar, L. S., O. Lancaster, J. P. Pasche, G. Kopal, and A. J. Brookes (2007) MegaPlex PCR: a strategy for multiplex amplification. Nat. Meth. 4: 835–837.

16. Anderson, M., Ruff, D. (2017) Multiplex Amplification of Polynucleotides. US patent number 9822405.

17. Anderson, M., Ruff, D. (2018) Multiplex Amplification of Polynucleotides. US patent number 20180135104.

18. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J.l. Counting absolute numbers of molecules using unique molecular identifers. Nat. Methods 9, 72–74 (2011).

19. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplifcation noise with optimized single-molecule barcodes. Proc. Natl. Acad. Sci. USA 109, 1347–1352 (2012).

20. Bose, S., Wan, Z., Carr, A., Rizvi A., Vieira G., Pe'er D., Sims P. Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biol 16, 120 (2015).

21. Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–779. doi:10.1126/science.1247651

22. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161 (5):1202–1214. doi:10.1016/j.cell.2015.05.002

23. Navin N, Kendall J, Troge J, Rodgers L, Cook K, Stepansky A, Levy D, Lee Y, Esposito D, Muthuswamy L, Hicks J, Wigler M. Tumor Evolution Inferred by Single Cell Sequencing. Nature 7(472):90-4, 2011. PMID: 21399628.

24. Wang Y, Waters J, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, Multani A, Zhang H, Zhao R, Michor F, Meric-Bernstam F, Navin NE.. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature 512(7513):155-60, 2014. PMID: 25079324.

25. Baslan T, Kendall J, Troge J, Stepansky A, Wigler M, Navin N, Hicks J. Genome-wide Copy Number Analysis of Single Cells. Nature Protocols 7(6):1024-41, 2012. PMID: 22555242.

26. Leung ML, Wang Y, Kim C, Gao R, Jiang J, Sei E, Navin NE. Highly multiplexed targeted DNA sequencing from single nuclei. Nat Protoc 11(2):214-35, 2016. e-Pub 2016. PMID: 26741407.

27. Leung M, Wang Y, Waters J, Navin N. SNES: Single Nuclei Exome Sequencing. Genome Biol 25(16):55, 2015. PMID: 25853327.

28. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976; 194:23–28.

29. Gerlinger, M., McGranahan, N., Dewhurst, S.M., Burrell, R.A., Tomlinson, I., and Swanton, C. (2014). Cancer: evolution within a lifetime. Annu Rev Genet 48, 215-236.

30. Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., ... & Salm, M. (2017). Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine, 376(22), 2109-2121.

31. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer cell 27, 15-26.

32. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012; 481(7381):306–313.

33. Wagle N, Emery C, Berger MF, et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. J Clin Oncol. 2011; 29(22):3085–3096.

34. Fidler, I. J. Seed and soil revisited: contribution of the organ microenvironment to cancer metastasis. Surg. Oncol. Clin. N. Am. 10, 257–269 (2001).

35. Chambers, A. F., Groom, A. C. & MacDonald, I. C. Dissemination and growth of cancer cells in metastatic sites. Nat. Rev. Cancer 2, 563–572 (2002).

36. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147(2):275–292.

37. Massagué J, Obenauf AC. Metastatic colonization by circulating tumour cells. Nature. 2016;529(7586):298–306.

38. Leung ML, Davis A, Gao R, et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res. 2017;27(8):1287–1299.

39. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, Kamiyama M, Hruban RH, Eshleman JR, Nowak MA, et al. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467: 1114–1117.

40. Echeverria, G.V., Powell, E., Seth, S. et al. High-resolution clonal mapping of multi-organ metastasis in triple negative breast cancer. Nat Commun 9, 5079 (2018) doi:10.1038/s41467-018-07406-4

41. Hu, Z., Li, Z., Ma. Z., Curitis, C., Pan-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastasis. BioRxiv 825240; doi 10.1101/825249 (2019)

42. Haeger A, Wolf K, Zegers MM, Friedl P. Collective cell migration: guidance principles and hierarchies. Trends Cell Biol 25(9):556-66, 2015. e-Pub 2015. PMID: 26137890.

43. Venkatesan S, Swanton C. Tumor evolutionary principles: how intratumor heterogeneity influences cancer treatment and outcome. Am Soc Clin Oncol Educ Book. 2016;35:e141–9.

44. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. The New England journal of medicine. Mar 8.2012 366:883.

45. Zhang J, Fujimoto J, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014;346(6206):256–259. doi:10.1126/science.1256930

46. Hidalgo, M. Amant, F., Biankin, AV., et al. Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov 4, 998-1013 (2014).

47. Schmidt, M. Zickler P, Hoffmann G, et al. Polyclonal long-term repopulating stem cell clones in a primate model. Blood 100, 2737-2743 (2002).

48. Piacibello, W. Bruno S., Sanavio F, et al. Lentiviral gene transfer and ex vivo expansion of human primitive stem cells capable of primary, secondary, and tertiary multilineage

repopulation in NOD/SCID mice. Nonobese diabetic/severe combined immunodeficient. Blood 100, 4391-4400 (2002).

49. Hope, K.J., Jin, L. & Dick, J.E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. Nat Immunol 5, 738-743 (2004).

50. Mazurier, F., Gan, O.I., McKenzie, J.L., Doedens, M. & Dick, J.E. Lentivector-mediated clonal tracking reveals intrinsic heterogeneity in the human hematopoietic stem cell compartment and culture-induced stem cell impairment. Blood 103, 545-552 (2004).

51. McKenzie, J.L., Gan, O.I., Doedens, M., Wang, J.C. & Dick, J.E. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nat Immunol 7, 1225-1233 (2006).

52. Schmidt, M. Schwarzwaelder, K. Bartholomae, C. Zaoui, K. Ball C, Pilz I, Braun S, Glimm H, von Kalle C. High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nature methods 4, 1051-1057 (2007).

53. Stewart, M.H., Bendall, S.C., Levadoux-Martin, M. & Bhatia, M. Clonal tracking of hESCs reveals differential contribution to functional assays. Nature methods 7, 917-922 (2010).

54. Bardeesy, N., DePinho, R. Pancreatic cancer biology and genetics. Nat Rev Cancer 2, 897–909 (2002)

55. Jones S, Zhang X, Parsons DW, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008;321(5897):1801–1806.

56. Yachida, S., Iacobuzio-Donahue, C. Evolution and dynamics of pancreatic cancer progression. Oncogene 32, 5253–5260 (2013) doi:10.1038/onc.2013.29

57. Rahib L, Smith BD, Aizenberg R, et al. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pan-creas cancers in the United States.Cancer Res2014;74:2913–2921.

58. Baine M, Sahak F, Lin C, Chakraborty S, Lyden E, et al. (2011) Marital Status and Survival in Pancreatic Cancer Patients: A SEER Based Analysis. PLOS ONE 6(6): e21052. doi:10.1371/journal.pone.0021052

59. J. He, N. Ahuja, M.A. Makary, et al. 2564 resected periampullary adenocarcinomas at a single institution: trends over three decades HPB : the official journal of the International Hepato Pancreato Biliary Association, 16 (2014), pp. 83-90

60. Hata AN, Niederst MJ, Archibald HL, et al. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. Nat Med. 2016;22(3):262–269. doi:10.1038/nm.4040

61. Carugo, A. Genovese, G., Seth, S., Nezi, L., Rose, JL., et al. In Vivo Functional Platform Targeting Patient-Derived Xenografts Identifies WDR5-Myc Association as a Critical Determinant of Pancreatic Cancer. Cell Rep 16, 133-147 (2016).

62. Cellecta Inc. CloneTracker Barcode Cell Labeling Kits. v1c —2019/02/15 https://cdn.manula.com/user/8969/11778_12591_en_1513133850.pdf?v=20190215175937

63. Kim, M.P. Evans DB, Wang H, Abbruzzese JL, Fleming JB, Gallick GE. Generation of orthotopic and heterotopic human pancreatic cancer xenografts in immunodeficient mice. Nature protocols 4, 1670-1680 (2009).

64. Tentler, J.J. Tan AC, Weekes CD, Jimeno A, Leong S, Pitts TM, Arcaroli JJ, Messersmith WA, Eckhardt SG. Patient-derived tumour xenografts as models for oncology drug development. Nature reviews. Clinical oncology 9, 338-350 (2012).

65. Purton, L.E. & Scadden, D.T. Limiting factors in murine hematopoietic stem cell assays. Cell stem cell 1, 263-270 (2007).

66. Hope, K.J., Jin, L., and Dick, J.E. (2004). Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. Nat Immunol 5, 738-743.

67. Lan, X., Jörg, D., Cavalli, F. et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. Nature 549, 227-232 (2017).

68. McKenzie, J.L., Gan, O.I., Doedens, M., Wang, J.C., and Dick, J.E. (2006). Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nat Immunol 7, 1225-1233.

69. Nguyen, L., Cox, C., Eirew, P. et al. DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. Nat Commun 5, 5871 (2014).

70. Nguyen, L.V. Makarem M,. Carles, A. et.al. Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. Cell stem cell 14, 253-263 (2014).

71. Lu, R., Neff, N.F., Quake, S.R. & Weissman, I.L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature biotechnology 29, 928-933 (2011).

72. Gerrits, A. et al. Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood 115, 2610-2618 (2010).

73. Nolan-Stevaux O, Tedesco D, Ragan S, et al. Measurement of Cancer Cell Growth Heterogeneity through Lentiviral Barcoding Identifies Clonal Dominance as a Characteristic of In Vivo Tumor Engraftment. PLoS One. 2013;8(6):e67316. Published 2013 Jun 26.

74. Nguyen, L., Pellacani, D., Lefort, S. et al. Barcoding reveals complex clonal dynamics of de novo transformed human mammary cells. Nature 528, 267–271 (2015).

75. Burris, H.A., 3rd et al. Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. Journal of clinical oncology: official journal of the American Society of Clinical Oncology 15, 2403-2413 (1997).

76. Legendre, P. & Legendre, L. Numerical ecology, Edn. Third English edition. (Elsevier, Amsterdam; 2012).

77. Vendetti, F.P., Lau, A., Schamus, S., Conrads, T.P., O'Connor, M.J., and Bakkenist, C.J. (2015). The orally active and bioavailable ATR kinase inhibitor AZD6738 potentiates the anti-tumor effects of cisplatin to resolve ATM-deficient non-small cell lung cancer in vivo. Oncotarget 6, 44289-44305.

78. Li L, Hanahan D. Hijacking the neuronal NMDAR signaling circuit to promote tumor growth and invasion. Cell. 2013 Mar 28;153(1):86-100. doi: 10.1016/j.cell.2013.02.051.

79. Li, L., Zeng, Q., Bhutkar, A., Galván, J. A., Karamitopoulou, E., Noordermeer, D., Peng, M. W., Piersigilli, A., Perren, A., Zlobec, I., Robinson, H., Iruela-Arispe, M. L., & Hanahan, D. (2018). GKAP Acts as a Genetic Modulator of NMDAR Signaling to Govern Invasive Tumor Growth. Cancer cell, 33(4), 736–751.e5.

80. Schmitt, M.W., Loeb, L.A. & Salk, J.J. The influence of subclonal resistance mutations on targeted cancer therapy. Nature reviews. Clinical oncology 13, 335-347 (2016).

81. Shaw AT, Friboulet L, Leshchiner I, et al. Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F. The New England journal of medicine 374, 54-61 (2016).

82. Hutchinson, L. Tumour response, correlates of survival and clinical benefit. Nature reviews. Clinical oncology 12, 433 (2015).

83. Gupta, P.B. Fillmore CM, Jiang G, et al. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell 146, 633-644 (2011).

84. Viale, A. & Draetta, G.F. Metabolic Features of Cancer Treatment Resistance. Recent Results Cancer Res 207, 135-156 (2016).

85. Lan, X., Jörg, D., Cavalli, F. et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. Nature 549, 227-232 (2017).

86. Frei, E., 3rd, Karon M, Levin R., et al. The effectiveness of combinations of antileukemic agents in inducing and maintaining remission in children with acute leukemia. Blood 26, 642-656 (1965).

87. Y. Kuboki, C.G. Fischer, V. Beleva Guthrie, et al. Single-cell sequencing defines genetic heterogeneity in pancreatic cancer precursor lesions. J Pathol, 247 (2019), pp. 347-356

88. Li, F., Tiede, B., Massagué, J. et al. Beyond tumorigenesis: cancer stem cells in metastasis. Cell Res 17, 3–14 (2007) doi:10.1038/sj.cr.7310118

89. Kreso A, O'Brien CA, van Galen P, et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science 339, 543-548 (2013).

90. Shaw AT, Friboulet L, Leshchiner I, et al. Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F. The New England journal of medicine 374, 54-61 (2016).

91. Porter, S.N., Baker, L.C., Mittelman, D. & Porteus, M.H. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. Genome Biol 15, R75 (2014).

92. Cross WC, Graham TA, Wright NA. New paradigms in clonal evolution: punctuated equilibrium in cancer. J Pathol. 2016;240:126–36.

93. Wood HM, Conway C, Daly C, et al. The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. J Pathol. 2015a;237:296–306.

94. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated?. Biochim Biophys Acta Rev Cancer. 2017;1867(2):151–161. doi:10.1016/j.bbcan.2017.01.003

95. Perié L., et al. Determining lineage pathways from cellular barcoding experiments. Cell reports, ISSN: 2211-1247, Vol: 6, Issue: 4, Page: 617-24. 2014

96. Kretzschmar K, Watt FM. Lineage tracing. Cell. 2012;148:33–45.

97. Roy, E., Neufeld, Z., Livet, J. and Khosrotehrani, K. (2014), Concise Review: Understanding Clonal Dynamics in Homeostasis and Injury Through Multicolor Lineage Tracing. Stem Cells, 32: 3046-3054.

98. Surveillance, Epidemiology, and End Results Program. National Cancer Institute, 2007-2013. Web. 14 Apr. 2017. (https://seer.cancer.gov/statfacts/html/pancreas.html)

99. Mehlen P, Puisieux A. Metastasis: a question of life or death, Nature Reviews Cancer 6, 449-458 (June 2006). doi:10.1038/nrc1886

100. Kim S. New and emerging factors in tumorigenesis: an overview. Cancer Management and Research 7, 225 (2015).

101. Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature. 2014;514(7520):54–58. doi:10.1038/nature13556

102. Ramaswamy S., Ross KN, Lander E. et al. A molecular signature of metastasis in primary solid tumors. Nat Genet. 33(1):49–54. 2003.

103. Makohon-Moore, A., Zhang, M., Reiter, J. et al. Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. Nat Genet 49, 358–366 (2017). https://doi.org/10.1038/ng.3764

104. Fidler, I. J. & Kripke, M. L. Metastasis results from pre-existing variant cells within a malignant tumour. Science 197,893–895 (1977).

105. Nguyen DX, Massagué J. Genetic determinants of cancer metastasis. Nat Rev Genet. 2007;8:341–352.

106. Navin N., The first five years of single-cell cancer genomics and beyond, Genome Res. 2015 Oct; 25(10): 1499–1507.

107. Szczurek E., Krueger T., Klink B., Beerenwinkel N., A Stochastic Model of Metastatic Bottleneck Predicts Patient Outcome and Therapy Response. BioRxiv. 2018. doi10.1101/348086

108. Deer EL. et.al, Phenotype and genotype of pancreatic cancer cell lines. Pancreas. 2010 May;39(4):425-35. doi: 10.1097/MPA.0b013e3181c15963.

109. Seth, S., Li, C.Y., Ho, I.L., Corti, D., Loponte, S., Sapio, L., Del Poggetto, E., Yen, E.Y., Robinson, F.S., Peoples, M., et al. Pre-existing Functional Heterogeneity of Tumorigenic Compartment as the Origin of Chemoresistance in Pancreatic Tumors. Cell Rep. 2019, 26, 1518–1532

110. Reuter, J. (2011). Subcutaneous Injection of Tumor Cells. *Bio-101*: e166. DOI: 10.21769/BioProtoc.166.

111. Denny Liggitt, Suzanne M. Dintzis, in Comparative Anatomy and Histology (Second Edition), 2018

112. Van der Maaten, L. & Hinton, G. E. Visualizing data using t-SNE. J. Mach. Learn.Research 9, 2579–2605 (2008).

113. A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: A review", *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264-323, 1999.

114. Ishiwata, T., Matsuda, Y., Yoshimura, H. et al. Pathol. Oncol. Res. (2018) 24: 797. https://doi.org/10.1007/s12253-018-0420-x

115. Ebben JD, Treisman DM, Zorniak M, Kutty RG, Clark PA, Kuo JS. The cancer stem cell paradigm: a new understanding of tumor development and treatment. Expert Opin Ther Targets. 2010;14(6):621–632.

116. Kresp A., Dick J., Evolution of the Cnacer Stem Cell Model, Cell Stem Cell.(2014) http://dx.doi.org/10.1016/j.stem.2014.02.006

117. Capp J., Cancer Stem Cells: From Historical Roots to a New Perspective, Journal of Oncology. Cancer cell plasticity vol 2019.

118. Yang D. Teng, Lei Wang, Serdar Kabatas, Henning Ulrich, and Ross D. Zafonte. Stem Cells and Development.Nov 2018.1466-1478.http://doi.org/10.1089/scd.2018.0129

119. Rycaj K, Tang DG. Cell-of-origin of cancer versus cancer stem cells: assays and interpretations. Cancer Res 2015;75:4003–11

120. Gómez-Cuadrado L, Tracey N, Ma R, Qian B, Brunton VG. Mouse models of metastasis: progress and prospects. Dis Model Mech. 2017;10(9):1061–1074.

121. Massagué, J., Obenauf, A. Metastatic colonization by circulating tumour cells. Nature 529, 298–306 (2016)

122. Maddipati R, Stanger BZ. Pancreatic cancer metastases harbor evidence of polyclonality. Cancer Discovery (2015) 5(10):1086–1097

123. K. Chen, W. Qian, J. Li, Z. Jiang, L. Cheng, B. Yan, J. Cao, L. Sun, C. Zhou, M. Lei, et al.Loss of AMPK activation promotes the invasion and metastasis of pancreatic cancer through an hsf1-dependent pathway Mol. Oncol., 11 (2017), pp. 1475-1492

124. Paoletti, P., Bellone, C. & Zhou, Q. NMDA receptor subunit diversity: impact on receptor properties, synaptic plasticity and disease. Nat Rev Neurosci 14, 383–400 (2013)

125. Choi YK, Park KG. Targeting Glutamine Metabolism for Cancer Treatment. Biomol Ther (Seoul). 2018;26(1):19–28.

126. Hensley, C.T., Wasti, A.T. & DeBerardinis, R.J. Glutamine and cancer: cell biology, physiology, and clinical opportunities. J. Clin. Invest. 123, 3678–3684 (2013) An excellent review of glutamine metabolism and its potential as a therapeutic target.

127. North WG, Liu F, Lin LZ, Tian R, Akerman B. NMDA receptors are important regulators of pancreatic cancer and are potential targets for treatment. Clin Pharmacol. 2017;9:79–86. Published 2017 Jul 17.

128. Zeng Q, Michael IP, Zhang P, Saghafinia S, Knott G, Jiao W, McCabe BD, Galván JA, Robinson HPC, Zlobec I, Ciriello G, Hanahan D (2019) Synaptic proximity enables NMDAR signalling to promote brain metastasis. Nature 573(7775):526–531

129. Psych scene hub, Anti-N-methyl-D-aspartate (Anti-NMDA) Receptor Encephalitis – A
     Synopsis. 2019. (https://psychscenehub.com/psychinsights/anti-nmda-receptor-
     encephalitis-a-synopsis/)

130. Casás-Selves M, Degregori J. How cancer shapes evolution, and how evolution shapes
     cancer. Evolution (N Y). 2011;4(4):624–634.

131. Shlush L., Hershkovitz D., Clonal Evolution Models of Tumor Heterogeneity, American
     Society of Clinical Oncology Educational Book 35 (October 29, 2019) e662-e665.

132. Shen W, Clemente MJ, Hosono N, et al. Deep sequencing reveals stepwise mutation
     acquisition in paroxysmal nocturnal hemoglobinuria, J Clin Invest. 2014;124(10):4529-
     4538. https://doi.org/10.1172/JCI74747.

133. Zhao, ZM., Zhao, B., Bai, Y., Iamarino, A., et.al. Early and multiple origins of metastatic
     lineages within primary tumors. Proc. Natl. Acad. Sci. U. S. A., 113 (2016), pp. 2140-2145

134. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers [published
     correction appears in Cell. 2015 Aug 13;162(4):924]. Cell. 2012;149(5):994–1007.

135. Gundem, G., Van Loo, P., Kremeyer, B. et al. The evolutionary history of lethal metastatic
     prostate cancer. Nature 520, 353–357 (2015)

136. Kandoth, C., McLellan, M., Vandin, F. et al. Mutational landscape and significance across
     12 major cancer types. Nature 502, 333–339 (2013)

137. De Sousa E Melo F, Vermeulen L, Fessler E, Medema JP. Cancer heterogeneity--a
     multifaceted view. EMBO Rep. 2013;14(8):686–695.

138. Prasetyanti, P.R., Medema, J.P. Intra-tumor heterogeneity from a cancer stem cell
     perspective. Mol Cancer 16, 41 (2017)

139. Lareau CA, Ludwig LS, Sankaran VG. Longitudinal assessment of clonal mosaicism in
     human hematopoiesis via mitochondrial mutation tracking. Blood Adv. 2019;3(24):4161–
     4165.

140. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016;113(37):E5528–E5537.

141. Alves JM, Prieto T, Posada D. 2017. Multiregional tumor trees are not phylogenies. Trends Cancer 3:546–50

142. Loulier K., Barry R., Mahou P., Y et.al. Multiplex cell and lineage tracking with combinatorial labels. Neuron, 81 (2014), pp. 505-520

143. Singh K., Minge X.,Bootstrap: a statistical method. 2008 Rutger Statistics.

144. D. Metzgar, C. Wills Evidence for the adaptive evolution of mutation rates Cell, 101 (2000), pp. 581-584

145. Charlesworth D, Barton NH, Charlesworth B. The sources of adaptive variation. Proc. Biol. Sci. 2017;284(1855):20162864. doi:10.1098/rspb.2016.2864

146. S. Maisnier-Patin, J.R. Roth. The origin of mutants under selection: how natural selection mimics mutagenesis (adaptive mutation), Cold Spring Harb. Perspect. Biol., 7 (2015), Article a018176

147. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017;551(7678):45–50. doi:10.1038/nature24287

148. Holohan, C.; van Schaeybroeck, S.; Longley, D.B.; Johnston, P.G. Cancer drug resistance: An evolving paradigm. Nat. Rev. 2013, 13, 714–726.

149. Brady SW, McQuerry JA, Qiao Y, et al. Combating subclonal evolution of resistant cancer phenotypes [published correction appears in Nat Commun. 2018 Feb 5;9(1):572]. Nat Commun. 2017;8(1):1231. Published 2017 Nov 1.

150. Chowell D, Napier J, Gupta R, Anderson KS, Maley CC, Sayres MAW. Modeling the Subclonal Evolution of Cancer Cell Populations. Cancer Res. 78, 830–839 (2018)

151. Notta, F., Chan-Seng-Yue, M., Lemire, M. et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. Nature 538, 378–382 (2016)

152. Moskaluk, C. A., Hruban, R. H. & Kern, S. E. p16 and K-RAS gene mutations in the intraductal precursors of human pancreatic adenocarcinoma. Cancer Res. 57, 2140–2143 (1997)

153. Wilentz, R. E. et al. Inactivation of the p16 (INK4A) tumor-suppressor gene in pancreatic duct lesions: loss of intranuclear expression. Cancer Res. 58, 4740–4744 (1998)

154. Wilentz, R. E. et al. Loss of expression of Dpc4 in pancreatic intraepithelial neoplasia: evidence that DPC4 inactivation occurs late in neoplastic progression. Cancer Res. 60, 2002–2006 (2000)

155. Lüttges, J. Falehdari, H., Brocker, V., Schwarte-Waldhoff, I., et al. Allelic loss is often the first hit in the biallelic inactivation of the p53 and DPC4 genes during pancreatic carcinogenesis. Am. J. Pathol. 158, 1677–1683 (2001)

156. Rhim AD, Mirek ET, Aiello NM, et al. EMT and dissemination precede pancreatic tumor formation. Cell. 2012;148(1-2):349–361. doi:10.1016/j.cell.2011.11.025

157. Bernards, R. & Weinberg, R. A. A progression puzzle. Nature 418, 823–824 (2002).

158. Van 't Veer, L. J. et al. Gene-expression profiling predicts clinical outcome of breast cancer. Nature415, 530–536(2002).

159. Rørth P., Collective Cell Migration. Annual Review of Cell and Developmental Biology 2009 25:1, 407-429

160. Wirtz D, Konstantopoulos K, Searson PC. The physics of cancer: the role of physical interactions and mechanical forces in metastasis. Nat Rev Cancer. 2011;11(7):512–522. Published 2011 Jun 24. doi:10.1038/nrc3080

161. Mierke C. Physical break-down of the classical view on cancer cell invasion and metastasis. Eur. J. Cell Biol., 92 (2013), pp. 89-104

162. Glinskii OV, Huxley VH, Glinsky GV, Pienta KJ, Raz A, Glinsky VV. Mechanical entrapment is insufficient and intercellular adhesion is essential for metastatic cell arrest in distant organs. Neoplasia. 2005;7(5):522–527. doi:10.1593/neo.04646

163. Chen, W., Hoffmann, A.D., Liu, H. et al. Organotropism: new insights into molecular mechanisms of breast cancer metastasis. npj Precision Onc 2, 4 (2018) doi:10.1038/s41698-018-0047-0

164. Heath, A.S.; Yibin, K. Determinants of organotropic metastasis. Annu. Rev. Cancer Biol. 2017, 1, 403–423.

165. Lambert, A.; Pattabiraman, D.; Weinberg, R. Emerging biological principles of metastasis. Cell 2017, 168, 670–691.

166. Aguirre-Ghiso, J. Models, mechanisms and clinical evidence for cancer dormancy. Nat Rev Cancer 7, 834–846 (2007) doi:10.1038/nrc2256

167. Hamilton AM, Parkins KM, Murrell DH, Ronald JA, Foster PJ. Investigating the Impact of a Primary Tumor on Metastasis and Dormancy Using MRI: New Insights into the Mechanism of Concomitant Tumor Resistance. Tomography. 2016;2(2):79–84. doi:10.18383/j.tom.2016.00151

168. Shiozawa Y, Nie B, Pienta KJ, Morgan TM, Taichman RS. Cancer stem cells and their role in metastasis. Pharmacol Ther. 2013;138(2):285–293. doi:10.1016/j.pharmthera.2013.01.014

169. Li S, Li Q. Cancer stem cells and tumor metastasis (Review). Int J Oncol. 2014;44(6):1806–1812. doi:10.3892/ijo.2014.2362

170. J. C. Chang, "Cancer stem cells: role in tumor growth, recurrence, metastasis, and treatment resistance," Medicine, vol. 95, Supplement 1, pp. S20–S25, 2016.

171. Ganguly KK, Pal S, Moulik S, Chatterjee A. Integrins and metastasis. Cell Adh Migr. 2013;7(3):251–261. doi:10.4161/cam.23840

172. Hamidi, H., Ivaska, J. Every step of the way: integrins in cancer progression and metastasis. Nat Rev Cancer 18, 533–548 (2018) doi:10.1038/s41568-018-0038-z

173. Sarvaiya PJ, Guo D, Ulasov I, Gabikian P, Lesniak MS. Chemokines in tumor progression and metastasis. Oncotarget. 2013;4(12):2171–2185. doi:10.18632/oncotarget.1426

174. Kakinuma, T. and Hwang, S.T. (2006), Chemokines, chemokine receptors, and cancer metastasis. Journal of Leukocyte Biology, 79: 639-651. doi:10.1189/jlb.1105633

175. Fu, Q., Zhang, Q., Lou, Y. et al. Primary tumor-derived exosomes facilitate metastasis by regulating adhesion of circulating tumor cells via SMAD3 in liver cancer. Oncogene 37, 6105–6118 (2018) doi:10.1038/s41388-018-0391-0

176. Costa-Silva B, Aiello NM, Ocean AJ, Singh S, Zhang H, Thakur BK, et al. Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. Nat Cell Biol. 2015;17:816–26

177. Liu Y, Gu Y, Han Y, Zhang Q, Jiang Z, Zhang X, et al. Tumor exosomal RNAs promote lung pre-metastatic niche formation by activating alveolar epithelial TLR3 to recruit neutrophils. Cancer Cell. 2016;30:243–56.

178. Azmi AS, Bao B, Sarkar FH. Exosomes in cancer development, metastasis, and drug resistance: a comprehensive review. Cancer Metastasis Rev. 2013;32:623–42.

179. Bobrie A, Krumeich S, Reyal F, Recchi C, Moita LF, Seabra MC, Ostrowski M, Thery C. Rab27a supports exosome-dependent and -independent mechanisms that modify the tumor microenvironment and can promote tumor progression. Cancer research. 2012;72:4920–4930.

180. King HW, Michael MZ, Gleadle JM. Hypoxic enhancement of exosome release by breast cancer cells. BMC cancer. 2012;12:421. doi: 10.1186/1471-2407-12-421.

181. Jung T, Castellana D, Klingbeil P, Cuesta Hernandez I, Vitacolonna M, Orlicky DJ, Roffler SR, Brodt P, Zoller M. CD44v6 dependence of premetastatic niche preparation by exosomes. Neoplasia. 2009;11:1093–1105.

182. Grange C, Tapparo M, Collino F, Vitillo L, Damasco C, Deregibus MC, Tetta C, Bussolati B, Camussi G. Microvesicles released from human renal cancer stem cells stimulate angiogenesis and formation of lung premetastatic niche. Cancer research. 2011;71:5346–5356.

183. Hoshino A, Costa-Silva B, Shen TL, et al. Tumour exosome integrins determine organotropic metastasis. Nature. 2015;527(7578):329–335.

184. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. Nature reviews Cancer. 2009;9:239–252. doi: 10.1038/nrc2618.

185. Sosa MS, Bragado P, Aguirre-Ghiso JA. Mechanisms of disseminated cancer cell dormancy: an awakening field. Nat Rev Cancer. 2014;14(9):611–622.

186. Klein, C. Parallel progression of primary tumours and metastases. Nat Rev Cancer 9, 302–312 (2009)

187. Hayashi, A., Fan, J., Chen, R. et al. A unifying paradigm for transcriptional heterogeneity and squamous features in pancreatic ductal adenocarcinoma. Nat Cancer 1, 59–74 (2020)

188. A.P. Kusumbe, S.A. Bapat. Cancer stem cells and aneuploid populations within developing tumors are the major determinants of tumor dormancy. Cancer Res, 69 (2009), pp. 9245-9253

189. J.L. Dembinski, S. Krauss., Characterization and functional analysis of a slow cycling stem cell-like subpopulation in pancreas adenocarcinoma. Clin Exp Metastasis, 26 (2009), pp. 611-623

190. S. Pece, D. Tosoni, S. Confalonieri, G. Mazzarol, M. Vecchi, S. Ronzoni, et al. Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. Cell, 140 (2010), pp. 62-73

191. K. Sampieri, R. Fodde., Cancer stem cells and metastasis., Semin. Cancer Biol., 22 (3) (2012), pp. 187-193

192. Morrison S.J., Kimble J. Asymmetric and symmetric stem-cell divisions in development and cancer. Nature. 2006;441:1068–1074. doi: 10.1038/nature04956.

193. He S., Nakada D., Morrison S.J., Mechanisms of Stem Cell Self-Renewal. Annual Review of Cell and Developmental Biology 2009 25:1, 377-406

194. Brescia P, Richichi C, Pelicci G (2012) Current strategies for identification of glioma stem cells: adequate or unsatisfactory? J Oncol 2012: 376894.

195. Shahriyari L., Komarova N.L. Symmetric vs. Asymmetric Stem Cell Divisions: An Adaptation against Cancer? PLoS ONE. 2013;8:e76195. doi: 10.1371/journal.pone.0076195

196. Bajaj J., Zimdahl B., Reya T. Fearful symmetry: Subversion of asymmetric division in cancer development and progression. Cancer Res. 2015;75:792–797. doi: 10.1158/0008-5472.CAN-14-2750

197. Mukherjee S., Kong J., Brat D.J. Cancer Stem Cell Division: When the Rules of Asymmetry Are Broken. Stem Cells Dev. 2015;24:405–416. doi: 10.1089/scd.2014.0442.

198. Tse HTK, Weaver WM, Di Carlo D (2012) Increased Asymmetric and Multi-Daughter Cell Division in Mechanically Confined Microenvironments. PLoS ONE 7(6): e38986. https://doi.org/10.1371/journal.pone.0038986

199. Insinga A, Cicalese A, Faretta M, et al. DNA damage in stem cells activates p21, inhibits p53, and induces symmetric self-renewing divisions. Proc Natl Acad Sci U S A. 2013;110(10):3931–3936. doi:10.1073/pnas.1213394110

200. S.R. Pine, B.M. Ryan, L. Varticovski, A.I. Robles, C.C. Harris. Microenvironmental modulation of asymmetric cell division in human lung cancer cells., Proc. Natl. Acad. Sci. USA, 107 (2010), pp. 2195-2200

201. Q. Lin, Z. Yun., Impact of the hypoxic tumor microenvironment on the regulation of cancer stem cell characteristics. Cancer Biol Ther, 9 (2010), pp. 949-956

202. Plaks V, Kong N, Werb Z. The cancer stem cell niche: how essential is the niche in regulating stemness of tumor cells?. Cell Stem Cell. 2015;16(3):225–238. doi:10.1016/j.stem.2015.02.015

203. Scott JG, Hjelmeland AB, Chinnaiyan P, Anderson ARA, Basanta D (2014) Microenvironmental Variables Must Influence Intrinsic Phenotypic Parameters of Cancer Stem Cells to Affect Tumourigenicity. PLoS Comput Biol 10(1): e1003433. https://doi.org/10.1371/journal.pcbi.1003433

204. Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, et al. (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell 146: 633–44.

205. B. Langmead, S.L. SalzbergFast gapped-read alignment with Bowtie 2. Nat. Methods, 9 (2012), pp. 357-359

206. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing SubgroupThe Sequence Alignment/Map format and SAMtools. Bioinformatics, 25 (2009), pp. 2078-2079

207. M.D. Robinson, D.J. McCarthy, G.K. SmythedgeR: a Bioconductor package for differential expression analysis of digital gene expression data., Bioinformatics, 26 (2010), pp. 139-140

208. C.J.F. TerbraakCanonical correspondence-analysis—a new eigenvector technique for multivariate direct gradient analysis., Ecology, 67 (1986), pp. 1167-1179

209. Oksanen, J., Blanchett, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., et al. (2018). vegan: Community Ecology Package.

210. H. WickhamA layered grammar of graphics. J. Comput. Graph. Stat., 19 (2010), pp. 3-28

211. M. Morgan, S. Anders, M. Lawrence, P. Aboyoun, H. Pagès, R. GentlemanShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data., Bioinformatics, 25 (2009), pp. 2607-2608

212. H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoyBiostrings: efficient manipulation of biological strings. https://rdrr.io/bioc/Biostrings/ (2018)

213. H. Li, R. DurbinFast and accurate short read alignment with Burrows-Wheeler transform., Bioinformatics, 25 (2009), pp. 1754-1760

214. M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis , G. del Angel, M.A. Rivas, M. Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet., 43 (2011), pp. 491-498

215. G.A. Van der Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, et al.From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics, 43 (2013), pp. 11.10.1-11.10.33

216. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S.R.F. Twigg, A.O.M. Wilkie, G. McVean, G. Lunt er, G. Lunter, WGS500 Consortium Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications., Nat. Genet., 46 (2014), pp. 912-918

217. K. Cibulskis, M.S. Lawrence, S.L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E.S. Lander, G. GetzSensitive detection of somatic point mutations in impure and heterogeneous cancer samples., Nat. Biotechnol., 31 (2013), pp. 213-219

218. N.L. Bray, H. Pimentel, P. Melsted, L. PachterNear-optimal probabilistic RNA-seq quantification., Nat. Biotechnol., 34 (2016), pp. 525-527

219. Rapheal B., Aguirre, A., Hruban, R. Cancer Genome Atlas Research Network., Integrated genomic characterization of pancreatic ductal adenocarcinoma., Cancer Cell, 32 (2017), pp. 185-203.e13

220. Y. Hoshida. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment., PLoS ONE, 5 (2010), p. e15543

221. C.D. Greenman, G. Bignell, A. Butler, S. Edkins, J. Hinton, D. Beare, S. Swamy, T. Santarius, L. Chen, S. Widaa, et al.PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data., Biostatistics, 11 (2010), pp. 164-175

222. P.J. Rousseeuw, L. KaufmanFinding Groups in Data: An Introduction to Cluster Analysis. John, John Wiley and Sons (1990)

223. Galhardo RS, Hastings PJ, Rosenberg SM. Mutation as a stress response and the regulation of evolvability. Crit Rev Biochem Mol Biol. 2007;42(5):399–435. doi:10.1080/10409230701648502

224. Fitzgerald DM, Hastings PJ, Rosenberg SM. Stress-Induced Mutagenesis: Implications in Cancer and Drug Resistance. Annu Rev Cancer Biol. 2017;1:119–140. doi:10.1146/annurev-cancerbio-050216-121919

225. Campbell PJ, Yachida S, Mudie LJ, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010;467(7319):1109–1113. doi:10.1038/nature09460

Vitae

Chieh-Yuan Li (Alex) was born in Taiwan on January 25, 1984. The son of Chung-Meng Li and Jean Yen, and proud brother of Jill Li. He graduated from University of California, Berkeley with a Bachelor of Science in Chemical Biology and has worked in several biotechnology companies including Life Technologies by Thermo Fisher Scientific and Roche Molecular Diagnostics. As an R&D scientist, he was involved in developing genomic assays and technologies such as single-cell RNA-Seq, targeted-resequencing and IVD use assay. Alex is also the author of several patents in pioneering high complexity targeted DNA amplification and NGS sample preparation technologies. In summer of 2015, Alex entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Permanent address:

2121 Hepburn Street, Apt 815, Houston, Texas, 77054